PATHFINDER REPORT

451 Research® | Advisory

Datacenters of the future

A shifting landscape from the core to the edge

MAY 2017

COMMISSIONED BY

Raritan®

A brand of legrand®

About this paper

A Pathfinder paper navigates decision-makers through the issues surrounding a specific technology or business case, explores the business value of adoption, and recommends the range of considerations and concrete next steps in the decision-making process.

About 451 Research

451 Research is a preeminent information technology research and advisory company. With a core focus on technology innovation and market disruption, we provide essential insight for leaders of the digital economy. More than 100 analysts and consultants deliver that insight via syndicated research, advisory services and live events to over 1,000 client organizations in North America, Europe and around the world. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

451 Research shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.

NEW YORK
1411 Broadway
New York NY 10018
+1 212 505 3030

SAN FRANCISCO
140 Geary Street
San Francisco, CA 94108
+1 415 989 1555

LONDON
Paxton House
(Ground floor)
30, Artillery Lane
London, E1 7LS, UK
P +44 (0) 207 426 1050

BOSTON
75-101 Federal Street
5th Floor
Boston, MA 02110
Phone: +1 617.598.7200
Fax: +1 617.357.7495

Executive Summary

Most new datacenters operate at optimal availability and with infrastructural energy efficiency close to theoretical design targets. As such, it might be argued that the two biggest challenges of datacenter technology in the past 30 years have been addressed.

But despite this progress, the pace of change in the datacenter industry will continue and is likely to accelerate over the next decade and beyond. This will be spurred by increasing demand for digital services, as well as the need to embrace new technologies and innovation while mitigating future disruption. At the same time, there will also be a requirement to meet increasingly stringent business parameters and service levels.

This combination of business and technology drivers is likely to result in the emergence of new classes of datacenter facilities. These emergent datacenter types will share some traits with existing facilities but will also be tailored to new use cases. For example, it is expected that new edge datacenter capacity will be required to aggregate, process, store and analyze data from Internet of Things (IoT) infrastructure.

This paper examines these forces of change and disruption over the next decade, makes predictions about new datacenter types and specific use cases, and finally suggests ways to future-proof existing datacenters against disruption and to capitalize on innovation.

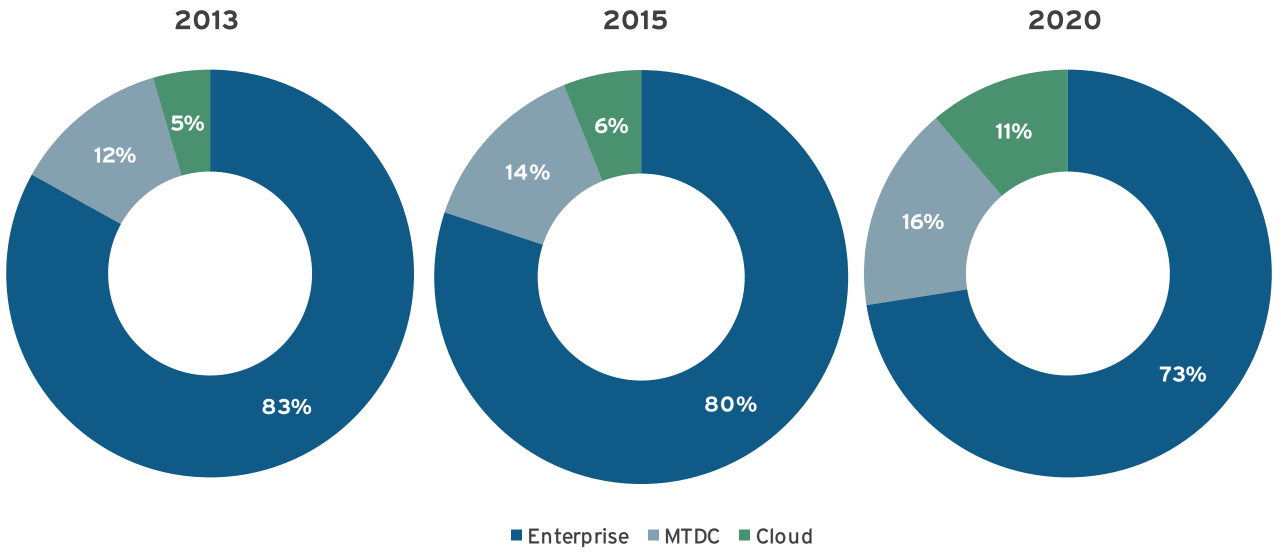
KEY FINDINGS

- New datacenter form factors (both large and small scale) and architectural requirements are emerging, driven by a combination of underlying forces of change and specific technology trends.
- The relative number and distribution of existing datacenter types are already undergoing change due to the impact of cloud providers and the move toward hybrid IT. Growth in edge demand will also be a significant contributor to new capacity and form factors in the longer term.
- Specialist datacenters (to deliver specific applications and services) will coexist with, but eventually replace, some generic facilities – mainly enterprise sites. In addition, we anticipate greater standardization and industrialization of datacenter design and construction, enabling a more modular buildout of capacity.
- Wild-card disruptive technologies – such as post-silicon technologies and quantum computing – could also result in major changes in design and operation, and could make some existing datacenter designs, business models and services obsolete in the longer term.
- Progressive operators will seek to future-proof new and existing physical datacenter infrastructure in order to minimize disruption and maximize capital expenditure. Increased agility should allow them to take advantage of innovative and emerging technologies.

Technology discussion

There are currently more than four million datacenters worldwide, according to 451 Research's Datacenter Monitor. These range from small enterprise-owned server closets and rooms (the vast majority of sites today) to commercial multi-tenant datacenters (MTDCs) and large hyperscale sites. Some of the dynamic forces of change in the industry are already having an impact on the distribution and relative numbers of these different facility types. Approximately 80% of all datacenter space (by square footage) globally was owned by enterprises in 2015. However, that percentage is expected to drop below 75% by 2020. The main reason for this shift is the migration of certain workloads from enterprise-owned sites into typically more cost-efficient cloud and colocation facilities.

Figure 1: Global Datacenter Space (Square Footage) Distribution



Source: 451 Research, Datacenter Monitor Q1 2017

The expectation is that the shift toward hybrid IT models will continue, causing enterprises to consolidate sites and move into larger, yet fewer, facilities. Cloud and colocation space will be increasingly viewed by enterprises as an extension (or replacement) of their own on-premises capacity. Public cloud and other cloud service providers are also driving significant demand for wholesale colocation space. It is expected that this migration from generalist enterprise sites to highly efficient colocation and cloud facilities will continue, and accelerate. However, there will still be a requirement for dedicated, premium enterprise sites among some organizations. New edge micro-datacenters will also to some extent replace 'legacy edge' enterprise server closets and rooms, as well as supporting new use cases.

UNDERLYING FORCES AND TECHNOLOGY TRENDS

Changes in datacenter design and operation are to some extent shaped by a number of underlying forces. These forces are in turn driving specific emerging technology trends, some of which are already beginning to have an impact on the construction and operation of new facilities. So-called 'wild cards' – theoretically very disruptive and hard-to-predict technologies such as post-silicon technologies and quantum computing – could also result in major changes in design and operation. Acting in conjunction, these forces and specific technologies are expected to result in new classes of facilities over the next decade that exploit innovative technologies but also increasingly meet specific business requirements.

FORCES OF CHANGE

- **Demand** – Even after taking into consideration incremental innovation in IT, there are significant concerns that datacenter space, power and bandwidth may struggle to keep up with the explosive global demand for IT services expected over the next two decades. Large amounts of new capacity will be required, but existing sites may also be required to take on some of the burden. If enterprise datacenters were fully utilized, this would be a major challenge, but most current datacenters run at very low IT utilization, with overprovisioned power and cooling. Thus, there is a significant opportunity to improve utilization, migrate workloads (off-premises) and optimize efficiency in existing sites, in addition to tapping new datacenter capacity.
- **Cost transparency** – Colocation and service providers are leading the way in terms of investing in technology to provide more transparency (e.g., power usage and IT capacity), showback and real-time costing. Using these tools, decisions about best execution venue, levels of availability required, latency, proximity, performance and service will be more informed given the greater visibility into the true costs (which may change dynamically). This could lead to more real-time and dynamic decision-making than would otherwise be possible.

- **Convergence** – Progress has been slow, but there are indications that the typically separate IT and facilities (equipment and staff) organizations are becoming more integrated. This should enable more holistic approaches to datacenter design and facility management that is integrated with IT workflow optimization. For example, the possibility of delivering 'virtual power' by consolidating and moving workloads using software and intelligent power equipment is becoming more of a reality. Datacenter infrastructure management (DCIM) software tools can also help to provide a holistic view of IT and facilities operations. Prefabricated modular (PFM) datacenter designs (including micro-datacenters) will also enable IT and facilities infrastructure to be more tightly coupled.
- **Industrialization (and standardization)** – The datacenter industry is very large and becoming more industrialized. For example, entire PFM datacenters can be built in factories. In the years ahead, it is likely that the massive scale of the global datacenter market will enable suppliers to produce or preconfigure an extensive catalog of equipment and designs, each optimized for customer-specific requirements and applications.
- **Research and development** – Research with a scientific imperative, but sometimes without a clear business case, will also shape future datacenter design and operation. A number of new and emerging technologies are likely to produce significant changes in the economics of running datacenters. These include quantum computing, silicon photonics, memristors and high-speed, high-bandwidth networks including 5G.

SPECIFIC TECHNOLOGY TRENDS

The broader forces of change in the datacenter industry are closely coupled with a number of specific technologies and trends that are reshaping the design and operation of new datacenter capacity.

- **Hyperscale cloud (driving efficiency and innovation)** – Datacenter operators at enterprise and colocation facilities are under increasing pressure to match the efficiency and cost-optimization of hyperscale operators such as Amazon, Facebook, Google and Microsoft. Cloud operators are able to exploit economies of scale and advances in IT infrastructure (compute, storage, networking) to drive up virtualization and utilization, and to test and apply innovations in datacenter architectures. Some rival operators will have to seriously consider whether it makes sense to invest in order to compete or whether it may be more prudent to explore partnership strategies in the long term.
- **PFM designs** – Prefabricated modular datacenters are assembled using one or more structural building blocks that are assembled and tested in a factory-like environment and shipped for final on-site integration. Forward-thinking operators that have already adopted the methodology will enjoy the competitive gains that PFM datacenter construction affords: standardization, compressed timelines, tighter budget controls, lowered risk, and better alignment with business goals by adding capacity in a more modular fashion driven by demand.
- **Software- and data-driven** – A growing proportion of datacenters are coming under software control in order to improve utilization, availability, resiliency and agility. Despite early concerns surrounding datacenter infrastructure management software implementation and its return on investment, DCIM is beginning to be recognized as an integral component of software-defined infrastructure. The recent development of cloud-based datacenter management as a service (DMaaS) promises to increase the value of DCIM data as it is aggregated and analyzed at scale. This could eventually allow for the data-driven, real-time, autonomic management of datacenters (potentially with few or no on-site staff) based on using large data sets.
- **Smart and transactive energy** – Datacenters have made considerable advances in improving energy efficiency and power usage effectiveness (PUE) ratios. However, the next stage in this process is to link energy use to demand, and take more control over energy supply. This could involve smart buying and selling of energy – including greater use of demand response – and managing IT power consumption by greater use of power management and power capping.
- **Connectivity** – The rise of public cloud is putting increased pressure on enterprise and MTDC providers, but it is also opening up new opportunities to connect and integrate public cloud with private cloud and non-cloud services. As a result, interconnectivity (enterprises connecting directly to cloud providers, partners, carrier networks, etc.) is becoming an increasingly critical service. Application and data resiliency – including disaster recovery – will in many cases be achieved at the network and software level, by replicating processes and data across a network of facilities spread within and between regions

- **Open architectures (Open Compute Project/Open 19)** – The Open Compute Project (OCP) hasn't made a noticeable impact yet (outside of hyperscale) but it could be significant in the long term as open ecosystems continue to evolve and grow. Open-sourced hardware and software promise to bring hyperscale-inspired designs and efficiency to the enterprise and colo markets, and are disrupting traditional facility architectures including distributed UPS, alternative rack designs, distributed connectivity and DC power distribution. The recently introduced Open19 specification, which has a lower bar to entry than OCP in some respects, also introduces new rack form factors
- **Edge datacenters** – The term 'edge computing' covers a range of workload types and use cases, including some that are established and some that are emerging. Demand at the edge is expected to be a significant driver for new datacenter types and form factors including micro-datacenters (small form-factor sites including prefabricated micro-modular sites) but also new centralized facilities (This is explored in more detail below).

Future datacenters and specific attributes

Accurately predicting the evolution of physical datacenter design and operation over the long term is obviously challenging. The number and types of datacenter form factors have changed significantly since the early days of the mainframe. We expect that pace of change to continue into the foreseeable future, with a move away from generalist, inefficient (usually) enterprise-owned sites.

It is possible to identify a number of future datacenter types (some of which exist today) that are likely to dominate in the next decade and beyond. Future datacenter types could include, but not be limited to, the following:

- Hyperscale (cloud operators but also some service providers)
- Cloud (non-hyperscale) and service provider
- Colocation (MTDC) and service provider
- Enterprise (dedicated, premium sites and fewer closets/rooms)
- Edge (micro-datacenters as well as core sites)
- HPC and specialized

These various datacenter types will be defined by some of the following criteria and attributes:

Business Models

Business models will vary between datacenter types based on ownership. For example, colocation and service-provider sites will be required to offer high availability and, very often, low-latency services – frequently achieved through proximity and connectivity. Similarly, some enterprises will have specific workloads, data requirements or governance issues that dictate that they must continue to design and operate their own premium datacenters. Given this, these organizations then have the opportunity to innovate and customize in ways that may not be open to commercial (colo, hosting) operators.

Scale

There is every indication that the efficiencies of scale that existing hyperscale sites enjoy mean that despite advances in compute capacity and more workloads moving to the edge, there will be a continued and sustained need for hyperscales in the future. At the other end of the spectrum, small-scale micro-sites are expected to become more widespread to support IoT and other applications. However, there will continue to be a consolidation of server rooms and closets (legacy edge) into colocation, cloud and, in some cases, micro-datacenters.

Resiliency

Resiliency requirement will be even more closely related to business case and function. For example, hyperscales may have lighter physical infrastructure (reduced UPS, lower-tier designs) because of lower service levels and ability to manage availability by load balancing. Some MTDCs may also build to different resiliency levels within the same facility, depending on customer requirements. There will also be greater adoption of software-based, distributed resiliency with less reliance on physical infrastructure (generators, UPS).

Efficiency

Efficiency will continue to be a requirement across the board. Some facilities, such as hyperscale sites, will heavily prioritize efficiency – in some cases, above most other criteria. A percentage of hyperscale sites will also continue to focus on sustainability and carbon reduction by utilizing more renewable energy (through power purchase agreements, renewable tariffs or, in some cases, on-site generation).

IT Density

Average rack power density, currently less than 5kW, is likely to continue increasing over time, driven by applications such as AI/machine learning, high-performance computing and big data. However, some wild-card technologies (e.g., quantum computing) have the potential to increase compute capacity while significantly reducing power requirements. Density will increasingly be tied to business function and workload. High-density zones may be created to enable more efficient cooling and power distribution. HPC and other specialist sites are likely to have high-density IT equipment (>25kW per rack) and consume more energy per unit of space/rack, for example. This means the cooling is also likely to be close-coupled: i.e., targeted to the requirements of a relatively small number of high-density racks.

Geography and Distribution

A number of large cloud operators have built out in specific locations (e.g., Europe) or have leased from MTDC providers, partly in order to comply with data regulations. This trend is likely to continue for future datacenter types. Hyperscale sites will also continue to be built in areas with low energy costs, tax incentives and climates that allow for free-air cooling. Edge capacity will be added in centralized datacenters as well as in metro sites outside of core datacenter hubs (see Figure 3).

Figure 2: Design and Operational Criteria of Future Datacenters

	DESIGN AND OPERATIONAL CRITERIA
BUSINESS	• Some highly cost-sensitive (hyperscale/cloud) • Others offering a mix of services, low latency.
SCALE	• From >10MW (hyperscale) to <100kW (edge micro-sites).
RESILIENCY	• Tier II or III on average, but some very low redundancy (HPC) • Some MTDCs adopting mixed-tier designs • Application-dependent resiliency becomes more widespread.
EFFICIENCY	• Highly prioritized by hyperscale • Lower priority for premium enterprise sites.
IT DENSITY	• Between 5kW and 10kW on average, but increasing • Specialized sites such as HPC will be 25kW to 80kW. • Some MTDCS will have mixed density.
GEOGRAPHIC	• Hyperscales value access to cheap, plentiful energy, free cooling • Commercially sensitivity and data governance/regulatory will guide new (non-hyperscale) cloud, MTDCs • Latency, bandwidth critical for some edge deployments.

EDGE DATACENTERS

Of all the trends shaping future datacenter development, the demand for edge computing is expected to be one of the more significant and, as such, deserves specific attention. Edge computing can be described as the distribution of compute and storage capabilities to the very edge of the network near the point of data generation and data use. This could be an enterprise factory floor or a carrier point of presence, a cell tower or a smart building.

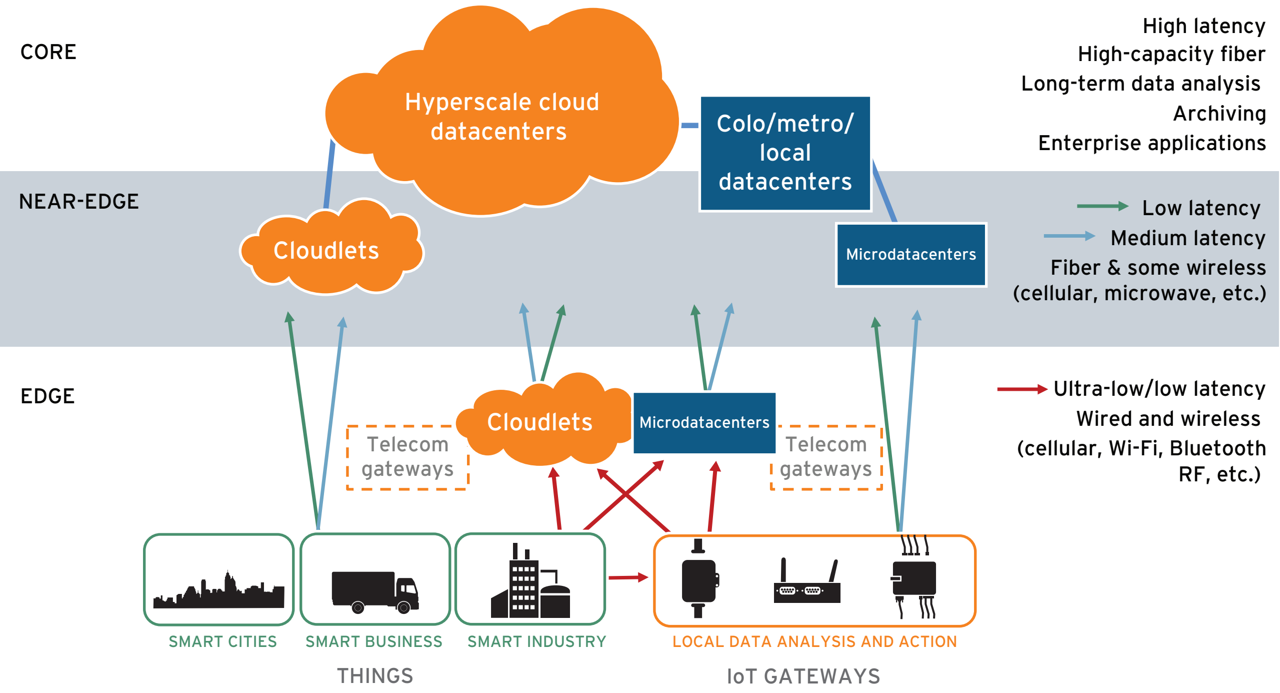
While use cases such as content distribution networks (CDNs) and local processing and storage are expected to drive edge compute demand in the short term, the Internet of Things (IoT) is expected to be one of the long-term drivers for new edge capacity. IoT's uses cases are vast but even within similar use cases, data paths and datacenter types will vary. In our view, it does seem likely that a number of IoT deployments will end up having data residing in a combination of public cloud and non-public cloud facilities, with the need for both distributed micro-datacenters and very large centralized sites.

Key data, including data needed by other applications and people, will in some cases be made available at the 'near edge' in large datacenters where colocation and other metro datacenters are sited close to where the data is generated. Cloud heavyweights are rapidly building hyperscale datacenters with direct fiber links to leased colocation sites. This brings hyperscale cloud capacity closer to the edge – effectively functioning as 'near edge' datacenter capacity. Cloud providers will also utilize 'cloudlets' – distributed edge capacity for data caching or low-latency compute.

Once consumed or integrated, data will then typically be moved or streamed into large or hyperscale remote datacenters to be aggregated, analyzed (including through integration with other data and applications) and archived. These large facilities represent the 'core layer.'

The broad schema in Figure 3 illustrates the main layers of IoT and edge datacenters, with some of the different types of datacenters and data paths that IoT applications might require.

Figure 3: Datacenters for the Internet of Things and Edge Computing

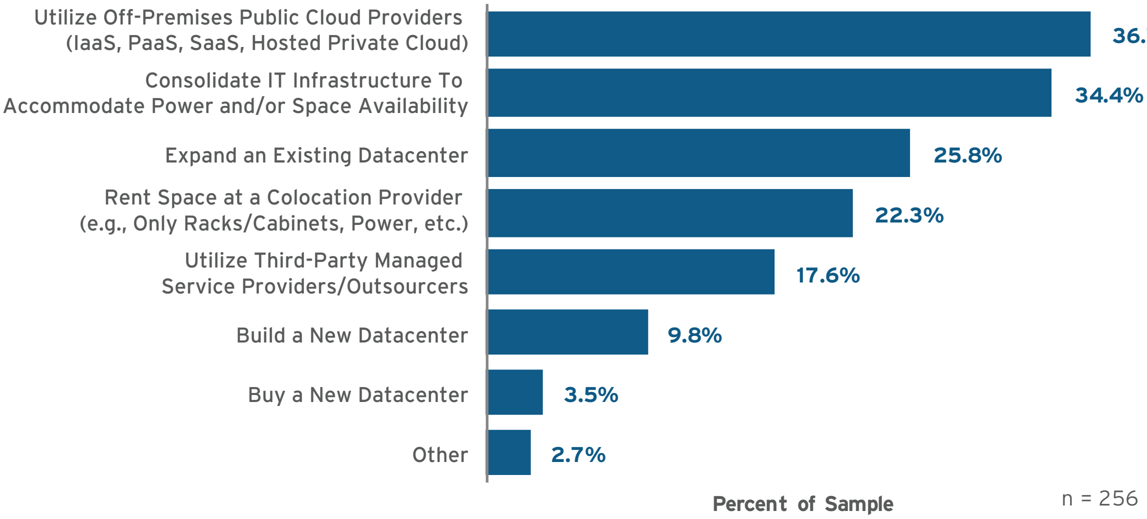


Source: 451 Research, Datacenters and Critical Infrastructure, 2017

Future-proofing today's infrastructure

It is sometimes said that the most efficient datacenter is the one that doesn't have to be built. Despite the benefits that new form factors will bring, operators are incentivized to ensure that existing sites are available, productive, highly utilized and efficient for as long as possible (See Figure 4). Datacenters built today with 10- to 15-year lifespans will also be the datacenters of the 2020s and into the 2030s. Operating hybrid IT environments provides increased flexibility in terms of capacity management, but consolidation and/or upgrades to on-premises IT infrastructure and implementation of virtualization can improve utilization and productivity of existing sites.

Figure 4: Solutions for Addressing Lack of Floor Space or Power Capacity



Source: 451 Research, Voice of the Enterprise, Datacenter Transformation 2016

IT hardware in a datacenter is typically refreshed in a three- to four-year time-frame, but most facilities infrastructure is considerably harder, and more costly, to update or retrofit. It is standard practice to overprovision cooling and power overhead to allow for future IT capacity requirements (balanced against efficiency). However, this could be described as future-proofing for changes in load rather than for new physical technologies that might render aspects of the design obsolete. There are, however emerging and established strategies and technologies (some of the same that are shaping future datacenter types) that can help maximize efficiency and capacity management and also reduce the risk of obsolescence. These include:

- **Modular designs** – Adding new space using PFM (including containers or micro-datacenters) provides benefits in terms of short-term capacity requirements but also for long-term innovation. Architects and engineers can take advantage of prefabrication to explore novel (in some cases, radical) ways to improve the structural, mechanical and electrical features of the design. Factory integration and modular installation of next-generation PFM infrastructure will deliver such new designs to existing sites with less cost and complexity.
- **Flexible power distribution and storage** – The use of specific technologies such as upgradable and intelligent (three-phase) PDUs, medium voltage distribution, and busbar power distribution provides scalability and adaptability to changing load requirements for existing sites. Building in sufficient compute headroom in intelligent devices such as smart PDU controllers also helps to ensure their longevity. Operators will benefit from testing and adopting these technologies in existing sites, since they are likely to become increasingly standard in future datacenter builds. Emerging energy management, (on-site) generation and storage technologies, including micro-grids, could also eventually help future-proof sites for power availability and resilience.
- **Efficient and high-density cooling** – As with power distribution, flexibility in cooling design is key to effective capacity management and future-proofing. Operators that currently rely on mechanical cooling should investigate chiller-free designs, since free-air cooling is likely to become the default option in most future sites (depending on geography). New capacity should also be built out with the capability to support close-coupled cooling, including rear-door heat exchanges and/or direct liquid cooling (requiring water distribution to the rack or row).
- **Management software** – Tools such as DCIM can reduce risk while also enabling new efficiencies, better capacity forecasting and improved business agility. In addition, the data-driven insights from using DCIM, and DMaaS, to more closely monitor and manage existing sites can significantly improve facility efficiency and capacity planning.

Conclusions and outlook

Datacenter operators are facing a time of unprecedented change. In 10 years, it is likely that the datacenter landscape will look very different as it responds to the macro forces and technology trends previously described:

- Today's distributed datacenter landscape will likely develop into a more intelligent, highly interconnected network made up of new datacenter form factors. Compact, self-managing datacenter nodes and hubs (of data or connectivity, or both) will be embedded and pervasive, supported by and feeding into large datacenter campuses.
- New edge capacity will be made up of clearly distinct and different datacenter types. They will include hyperscale cloud and large colocation facilities that are sited close to the point of data generation to support many applications; new micro-datacenters at the edge; and smaller clusters of capacity that may not be large or critical enough to be technically described as datacenters.
- Large datacenters could effectively become their own power utilities and act as energy hubs. Being able to control the quantity, quality and security of their power source via, for example, private micro-grids will improve energy resiliency.
- Access to connectivity, reliable power and energy costs will continue to dictate the siting of new facilities. However, the importance of data governance and privacy will increase as datacenter operators respond to stricter regulations and public pressure. Growth in edge demand driven by IoT and other applications/use cases will also drive demand for new and specific datacenter form factors in largely urban locations.

During the next decade of unprecedented change and new disruptive technologies, it is clear that the datacenter – in all its new forms, shapes and roles – will continue to enable innovation and drive technological and business transformation.