UII KEYNOTE REPORT 121M • JANUARY 2024
FINANCE AND STRATEGY

Five data center predictions for 2024

The digital infrastructure sector needs to invest and innovate to meet demand

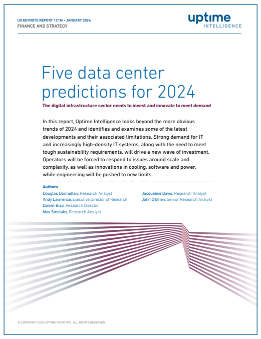
In this report, Uptime Intelligence looks beyond the more obvious trends of 2024 and identifies and examines some of the latest developments and their associated limitations. Strong demand for IT and increasingly high-density IT systems, along with the need to meet tough sustainability requirements, will drive a new wave of investment. Operators will be forced to respond to issues around scale and complexity, as well as innovations in cooling, software and power, while engineering will be pushed to new limits.

Authors

Douglas Donnellan, Research Analyst
Andy Lawrence, Executive Director of Research
Daniel Bizo, Research Director
Max Smolaks, Research Analyst
Jacqueline Davis, Research Analyst
John O'Brien, Senior Research Analyst

Synopsis

The critical digital infrastructure sector continues to enjoy robust growth. Rapidly evolving technologies will further drive and sustain this trend in 2024 and beyond — but will also create new challenges for operators. This report highlights some of these challenges and their implications. These include greater scrutiny over sustainability commitments; the adoption of power-hungry AI; the need for (and limitations of) direct liquid cooling; the evolution of data center management software; and the emergence of data center campuses that redefine the meaning of hyperscale.



UII KEYNOTE REPORT 121M • JANUARY 2024
FINANCE AND STRATEGY

Five data center predictions for 2024

The digital infrastructure sector needs to invest and innovate to meet demand

In this report, Uptime Intelligence looks beyond the more obvious trends of 2024 and identifies and examines some of the latest developments and their associated limitations. Strong demand for IT and increasingly high-density IT systems, along with the need to meet tough sustainability requirements, will drive a new wave of investment. Operators will be forced to respond to issues around scale and complexity, as well as innovations in cooling, software and power, while engineering will be pushed to new limits.

Authors
Douglas Donnellan, Research Analyst
Andy Lawrence, Executive Director of Research
Daniel Bizo, Research Director
Max Smolaks, Research Analyst
Jacqueline Davis, Research Analyst
John O'Brien, Senior Research Analyst

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED

Contents

Figures

Uptime 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 Intelligence, visit uptimeinstitute.com/ui-intelligence or contact research@uptimeinstitute.com.

Introduction

At the beginning of each calendar year, Uptime Intelligence compiles a short list of trends or predictions that will be relevant to the digital infrastructure sector for the year (and years) ahead. These lists aim to highlight crucial, yet often overlooked industry topics, which encourage a closer examination. Our five predictions for 2023 have proved to be largely accurate and will remain relevant for 2024. These are summarized in **Table 1**.

The predictions for 2024 highlight the opportunities arising from the sustained surge in demand for IT and the progress in new IT and facility technologies. However, we also address the accompanying challenges: how to cool high-density racks at scale, how to meet escalating IT demand, and the difficulties around expanding capacity while also meeting ever-stricter sustainability commitments.

Many of the challenges facing the digital infrastructure sector result from the ongoing success of the IT sector

Despite the notable growth of the digital infrastructure sector over the past five years, external global events (including the COVID-19 pandemic, Russia's invasion of Ukraine and extreme weather events) have disrupted supply chains and energy prices, and raised the cost of capital projects. Many of these challenges, however, result from the ongoing success of the IT sector itself, with the development of new software (including artificial intelligence, AI) and processors, and the economical delivery of IT services.

Investment by owners and operators of data centers, including cloud and hyperscale operators, is set to increase. Uptime Institute survey data reveals that more than half (61%) of enterprise operators and almost three-quarters of colocation operators (71%) expect their data center spending budgets to increase in 2024, and this is primarily driven by the need for capacity growth.

Laws passed in 2023 — and others slated for the years ahead — are creating additional hurdles for operators. These regulations primarily focus on reporting climate risks, improving energy performance and lowering carbon emissions. While not all organizations will be affected by these regulation, data center operating costs are still likely to increase.

In addition to the increasing legislation around climate change, 2023 also exposed the industry's lack of preparation in responding to extreme weather events as heat waves overwhelmed data center cooling systems in some regions. Recent record-breaking temperatures are expected to climb even higher in 2024 and will force many operators to reassess their resiliency strategies. Many are already increasing investment in this area.

To help make operational decisions relating to resiliency and energy efficiency, many organizations may plan to leverage AI technologies in 2024. However, according to Uptime survey data, trust in AI to make operational decisions has decreased over the past year. This is likely due to some of the unpredictable and inaccurate results of large language models. Even so, innovation in other forms of AI and machine learning is beginning to make an impact in the data center sector.

Securing and accommodating the necessary infrastructure for AI training models will be expensive and will require power-hungry IT and facility equipment. Given the costs and supply chain constraints, these deployments may be limited to only a few large-scale operators in the near term.

The year ahead will task operators with balancing new technology integration against costly infrastructure updates and sustainability pressures — while also managing greater complexity and minimizing operational risk.

Table 1

Predictions for 2023 remain valid

2023 prediction	Summary
Geopolitics deepens supply chain worries	Data center supply chains are still reeling from long delays, and current geopolitical dynamics present new threats. Uptime sees particularly high risks around the advanced semiconductors that are vital to IT hardware and data center equipment, and around subsea cable systems.
Too hot to handle? Operators to struggle with new chips	Operators face various trade-offs for handling new-generation IT technologies. Driven primarily by silicon technology, data center capacity planning will need to accommodate a potentially fast-shifting balance between power, cooling and space.
Cloud migrations to face closer scrutiny	The cost of migration and the threat of spiraling cloud costs deter some mission-critical migrations in the years ahead. Regulators and executives try to understand and limit the risks associated with a high dependency on the public cloud, causing some organizations to proceed more cautiously than before.
Energy efficiency focus to shift to IT — at last	More stringent sustainability regulation and reporting requirements will force IT to deliver improved performance in energy efficiency. Fewer higher-performance, highly utilized servers could deliver major energy gains. Refusing to deploy these improvements will be increasingly difficult to justify.
Data center costs set to rise and rise	The costs of critical digital infrastructure are set to rise: a trend likely to persist in the medium term. Supply chain issues and higher cost of capital, energy and labor, have all contributed to rising prices.

UPTIME INSTITUTE 2024

PREDICTION 1

Operators — prepare for a sustainability reckoning

Key trends

- The data center sector will continue to use more power, and emit more carbon, as its footprint rapidly grows.
- Publicly stated net-zero goals and other commitments will become harder and more expensive to maintain. Climbdowns will become more common — while many will choose to only disclose the requisite information.
- Tighter reporting and accountability regulations will be a catalyst for greater investment and more action, especially in overall energy efficiency.

The data center industry has been living with the threat of greater sustainability legislation, or other forms of mandatory control, for more than a decade. The EU first introduced the voluntary Code of Conduct for data centers in 2008, warning that legislation would follow if carbon and energy footprints were not brought under control. In the UK, a carbon reduction commitment for data centers was instigated, but later withdrawn. Other locations, including California, Amsterdam and Singapore, have introduced tighter planning restrictions for data centers and even moratoriums on new developments — although some of these have been either watered down or suspended.

This “green honeymoon” period of half-hearted legislation is coming to an end — from 2024, new regulations and requirements will enforce much stricter carbon and energy reporting in many countries. Beyond 2024, the impact of tighter reporting and controls will spread and many organizations will be forced to backtrack on publicly stated net-zero goals.

A difficult period ahead

Uptime Intelligence is predicting a challenging period for the sector from 2024 to 2030 as organizations struggle to meet sustainability goals and reporting requirements, battle with regulators (and even some partners), and strive to align their corporate business goals with wider sustainability objectives.

There are already signs that monitoring bodies are becoming stricter in assessing corporate sustainability. In August 2023, for example, the UN-backed Science Based Targets initiative (SBTi) removed Amazon’s operations (including Amazon Web Services) from its list of committed companies because Amazon had failed to validate its net-zero emissions target. The CDP, previously known as the Carbon Disclosure Project and the most comprehensive global registry of corporate carbon emission commitments, reported that of the 19,000 companies with registered plans on its platform, only 81 plans were credible.

Meeting tougher public goals will not be easy for those operating critical infrastructure. The greater use of more power-hungry software and processors, the lack of renewable energy availability in the power grid, and the growing resiliency requirements in the face of climate change, for example, will all make it tougher to reduce carbon emissions.

The sector may be at an inflection point. The pressures associated with compliance may also encourage the widespread adoption of more aggressive and thoughtful sustainability strategies, as well as encourage progressive and effective investment.

A clear disconnect

Larger and listed companies in most major economies will soon have to report their carbon emissions and climate-related risks through directives such as the EU's Corporate Sustainability Reporting Directive (CRSD) and California's Climate Corporate Data Accountability Act (passed in September 2023). The US Securities and Exchange Commission will also soon require some emissions and risk disclosure from listed companies, as will the UK, through its forthcoming Sustainability Disclosure Standards law.

Most concerning for the digital infrastructure sector, however, is the EU's Energy Efficiency Directive (EED), published in October 2023. This has detailed reporting requirements for data centers that include IT and network equipment use; and Germany's Energy Savings Act also sets down PUE levels and requirements to reuse heat (with some exceptions).

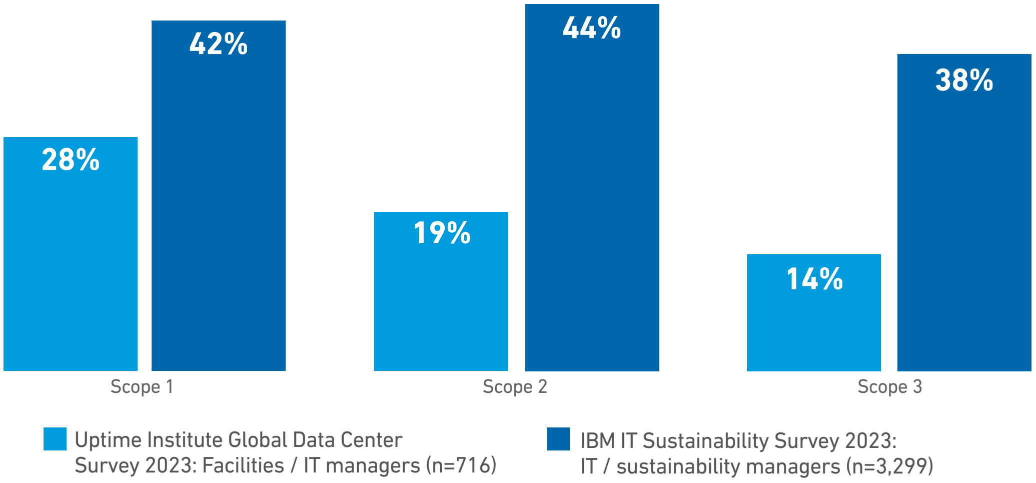
There is a move towards greater precision and accountability at a non-governmental level, too. International institutions, such as the World Resources Institute and the World Business Council for Sustainable Development, agree on the principles of carbon emission measurement and reporting, and these are then used by bodies such as the SBTi and the CDP. All are seeking to be more effective.

Despite all these developments, there is a still a startling disconnect between many of the public commitments by companies at an executive level, and what most digital infrastructure organizations are currently doing at a practical level. According to a 2023 survey by Uptime Institute, far less than half of managers polled in IT (digital infrastructure) organizations say they are currently tracking carbon emissions data. In contrast, a separate survey of a mix of executives and sustainability managers by IT supplier IBM in 2023 shows much higher tracking of this data (**Figure 1**).

Figure 1

Digital infrastructure's tracking of carbon emissions

Which Scopes of carbon emissions does your organization collect?



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2023; IBM IT SUSTAINABILITY SURVEY 2023

The data from these surveys underlines the corporate disconnect that Uptime Intelligence has frequently identified: most of those concerned with reducing energy consumption or collecting sustainability data have limited contact with sustainability or executive teams — even though they have the tools and knowledge to collect this data.

Further challenges

Accurate, timely reporting of carbon emissions and other data will be difficult for many digital infrastructure operators, especially when it extends to Scope 3 (embedded, third-party and supply chain emissions). Reducing emissions, either in absolute terms or relative to the overall business workload, will, however, be an even greater challenge, given three strong trends across the sector:

IT power consumption

Moore's law-type improvements in processor energy efficiency have slowed, while the use of power-hungry multi-core processors and graphics processing units (GPUs) have increased. These processors may do more work, but they also require more power. Improved cooling (e.g., direct liquid cooling), and better utilization of IT and management of workloads will be required to prevent runaway power consumption and over-heating, but may not reduce overall energy use.

Energy grid

Carbon reporting standards will increasingly require the use of in-region carbon-free energy (CFE) — or renewable energy. As more data center operators buy CFE to meet net-zero goals, the price of renewable energy will rise — if it is available at all. In many grid regions, it will take years (and possibly decades) to reach high levels of CFE.

Growth in workload

Energy use by data centers is currently estimated to be between 150 terawatt-hours (TWh) and 400 TWh a year (a wide range reflecting uncertainty or competing data models). Even without the extra energy needed to drive generative artificial intelligence (AI), the data center footprint and power use is expected to increase significantly, with some predicting energy use to double or more beyond 2030. This will strain power grids and supply chains, render carbon emissions targets yet more difficult to meet — and bring digital infrastructure operators into the crosshairs of regulators, environmental monitoring groups and campaigners.

PREDICTION 2

Demand for AI will have a limited impact on most operators

Key trends

- Advances in generative AI models have created runaway expectations of demand for data center capacity and high-density racks.
- A limited supply of chips for training large models will cap the significant pace of capacity growth for AI.
- AI's impact will be widely felt throughout the industry, but indirectly through straining equipment chains, pushing server chip power levels and making operators rethink their facility resiliency posture.

The recent intrigue and fascination with artificial intelligence (AI) marks a new high in the evolution of the technology. Now the data center industry is bracing itself for a significant increase in demand for net new capacity and the technically challenging requirements of providing sufficient power and cooling.

The intense activity in developing AI-powered services results largely from two developments: one in neural network architecture, called the "transformer"; the other in compute hardware, spearheaded by Nvidia's continued development of AI acceleration in silicon.

The launch of OpenAI's ChatGPT chatbot in November 2022, demonstrating previously unseen machine-learning sophistication, triggered an "arms race" as organizations competed to develop sophisticated AI-powered applications.

Demand for AI-accelerators is outstripping supply, and industry stakeholders are predicting a significant increase in new data center capacity to accommodate the higher rack densities required to drive generative AI workloads. This could also place strenuous demands on power and cooling infrastructures, as well as drive up the size and weight of extreme density cabinets.

While there is little doubt that demand for compute power and density will increase, Uptime Intelligence considers some of these expectations to be overblown. For many operators, the impact on data centers will be more indirect through AI's effect on supply chains, chip designs and attitudes towards facility resiliency.

New capacity demand and densification will be muted

There are two key reasons for new capacity demand and muted densification. First, the speed and scale at which AI training infrastructure grows is capped by chipmaking capacity. Even though Nvidia has been able to scale up production of its chips at its manufacturing partner TSMC, it is managing output carefully to avoid a potential supply glut following pent-up demand. Currently, Nvidia is forecast to ship up to 600,000 of the H100 / H200 chips by the end of 2023, and another 1.5 million to 2 million in 2024. These chips will form the basis of the large majority of AI training infrastructure for large generative models.

Assuming that most of these chips will be running AI workloads, Uptime Intelligence considers 2,000 megawatts (MW) to 2,500 MW of additional IT load capacity (uplift compared with baseline data center demand) to be a plausible scenario between the start of 2023 and the first quarter of 2025. Even if the whole sector adds hundreds of megawatts of data center capacity in the next 18 months (both net new and through planned expansions) to accommodate for a surge in AI infrastructure demand, this would be a significant — but not dramatic — undertaking. Moreover, relatively few operators and sites will see the bulk of this uptake. Currently, only a handful of customers account for the majority of AI chip sales, with major cloud infrastructure and web services companies (such as Microsoft, Google and Amazon Web Services, as well as their counterparts in China) topping the rankings.

Most recently built facilities with modern power distribution will be able to handle all but the largest AI training clusters

Second, while AI-optimized hardware can be much denser than typical IT equipment, it does not need to take extreme form. Nvidia's reference designs call for up to 50 kilowatts (kW) of power in its densest implementation, but hardware can also be spread out to meet power delivery or airflow limitations. Although much denser systems exist (some well above 100 kW per cabinet), they tend to be offered to a few large, multi-megawatt supercomputer installations to limit the footprint and to mitigate the cable run length limitations of high-speed interconnects.

In reality, most recently built facilities with modern power distribution (3-phase, higher voltages) will be able to handle all but the largest AI training clusters. This will occur through a combination of spreading out systems, upgrading breakers where necessary and adding more power circuits to reach the desired level of capacity per rack (e.g., 20 kW or 40 kW).

Generative AI's broader impact: slow but lasting

Besides adding demand for capacity and the adoption of high-density IT systems, Uptime Intelligence expects generative AI (and other forms of AI) to affect data center operators indirectly. In chronological order of the expected time of impact:

- **Prolonged demand-supply imbalance**. Lead times of data center equipment remain long. This is particularly evident when it comes to large systems, such as engine generators (which often have two-year waiting lists), switchgears, transformers, uninterruptible power supplies — but also some mechanical equipment and smaller components. The additional capacity needs of AI training clusters will contribute to a wider demand-supply imbalance by tying up even more of the equipment supply in the hands of a relatively few large and hyperscale data center operators.
- **AI arms race will push chip power envelopes further**. Server silicon power ratings have escalated markedly since around 2017. The scale of integration has outpaced transistor energy gains and many larger IT customers prefer more performant systems, even at the cost of higher power consumption. There is a renewed bout of fierce competition between chipmakers for performance supremacy in AI and other

technical computing tasks. Cloud infrastructure will also inevitably push up silicon design power further and faster. In a few years, mainstream servers with up to 1 kW of realized power use will be common.

This trend affects the design of all current and future server processors and accelerators, and shapes model line-up and pricing decisions. New generations of servers may only bring tangible benefits in application performance or economics for those with specific and suitable IT workloads.

Equally, keeping up with best-in-class server chips is costly: the same class of products (relative in the porfolio) can cost 50% to 100% more today than they did five or six years ago, despite growing competition. Chipmakers will offer only the largest buyers sufficient discount to offset this silicon inflation.

- **Promoting mixed-tier facilities**. Most data centers are designed and built to a single level of resiliency, typically aiming for very high levels of service availability despite its high cost (low resiliency facilities tend to be dedicated to supercomputing). The real possibility of AI-heavy services becoming a meaningful application category in mission-critical enterprise and colocation facilities has led to a reconsideration of mixed-tier facilities.

  Arguably, the criticality of IT racks running AI workloads (training in particular) does not warrant the costly overheads from delivering conditioned power backed up by batteries and engine generators, let alone all the component redundancy required to achieve concurrent maintainability. The same applies to many other applications serving low criticality functions, yet receiving gold-plated facility services. A wider spread of enterprise AI training systems might change the resiliency posture of some data center operators from uniform resiliency standards to a multi-tier service approach.

PREDICTION 3

Data center software gets smarter, leverages data — at last

Key trends

- The design and capabilities of data center equipment have changed considerably over the past 10 years and traditional data center management tools have not kept up.
- The increasing scale and complexity of modern data centers, along with industry-wide staffing shortages, create more demand for infrastructure automation.
- Hype around artificial intelligence is pushing more operators to experiment with machine learning and is changing their understanding of the value of facilities equipment data.

Despite their role as enablers of technological progress, data center operators have been slow to take advantage of developments in software, connectivity and sensor technologies that can help optimize and automate the running of critical infrastructure.

Most data center owners and operators currently use building management systems (BMS) and/or data center infrastructure management (DCIM) software as their primary interfaces for facility operations. These tools have important roles to play but have limited analytics and automation capabilities, and they often do little to improve facility efficiency.

Uptime Intelligence has long argued that data center management tools need to evolve toward greater autonomy. We first proposed the data center maturity model (see **Table 2**) in 2019 and applied specifically to DCIM. Four years later, there is the beginning of a shift toward Level 4 and Level 5 autonomy, albeit with a caveat — DCIM software alone will likely never evolve these capabilities. Instead, it needs to be combined with a new generation of data-centric tools.

Table 2

Data center management maturity model

Level	Description	Operating efficiency
Level 1	No integration of infrastructure data. Basic monitoring is provided by the equipment vendor software and the BMS.	Low
Level 2	Software installed to monitor environmental and equipment power use. Able to adjust basic controls (e.g., cooling) in line with demand.	Low
Level 3	Software can track physical data center equipment characteristics, location and operational status. Energy and environmental data are used to reduce risks and waste.	Medium
Level 4	Machine learning models are used for prediction, service management and multiple views, optimizing the data center in near real time. AI is applied to DCIM-based data lakes for advanced analytics.	Medium
Level 5	AI-driven integrated management software adjusts data center behavior and makes the best use of resources according to goals, rules and service requirements throughout its life cycle.	High

A new school of thought on data center management software is emerging, proposed by data scientists and statisticians. From their point of view, a data center is not a collection of physical components but a complex system of data patterns.

Many of the newer companies working in this field have close ties to the research community or, in some cases, to large-scale operators (such as Google). Vendors tracked by Uptime Institute have sprung out of academic institutions in California, the UK, Spain and Holland. This development has helped bring a cutting-edge, data-oriented approach to data center management software that contrasts with some of the more proven approaches.

Not all machine learning models require extensive compute resources, rich datasets and long training times

Some of the benefits offered by data-centric management software include:

- **Improved facility efficiency**. Through the automated configuration of power and cooling equipment, as well as the identification of inefficient or faulty hardware.
- **Better maintenance**. By enabling predictive or condition-based maintenance strategies that consider the state of individual hardware components.
- **Discovery of stranded capacity**. As a result of the thorough analysis of all data center metrics, not just high-level indicators.
- **Elimination of human error**. Through either a higher degree of automation or automatically generated recommendations for employees.
- **Improvements in skill management**. By analyzing the skills of the most experienced staff and codifying them in software.

All about the data

Data centers are full of sensors, which can serve as a source of valuable operational insight — yet the data they produce is rarely analyzed. In most cases, applications of this information are limited to real-time monitoring and basic forecasting.

The same data can be used to train artificial intelligence (AI) models with a view to automating an increasing number of data center tasks. When combining sensor data with an emerging category of data center optimization tools — many of which rely on machine learning — data center operators can improve their infrastructure efficiency, achieve higher degrees of automation and lower the risk of human error.

The past few years have also spawned new platforms that simplify data manipulation and analysis. These enable larger organizations to develop their own applications that leverage equipment data — including their own machine learning models.

Not all machine learning models require extensive compute resources, rich datasets and long training times. In fact, many of the models used in data centers today are small and relatively simple. Both training and inference can run on general-purpose servers, and it is not always necessary to aggregate data from multiple sites — a model trained locally on a single facility's data will often be sufficient to deliver the expected results.

New tools bring new challenges

The adoption of data-centric tools for infrastructure management will require owners and operators to recognize the importance of data quality. They will not be able to trust the output of machine learning models if they cannot trust their data — and that means additional work on standardizing and cleaning their operational data stores. In some cases, data center operators will have to hire analysts and data scientists to work alongside the facilities and IT teams.

Data harvesting at scale will invariably require more networking inside the data center — some of it wireless — and this presents a potentially wider attack surface for cybercriminals. As such, cybersecurity will be an important consideration for any operational AI deployment and a key risk that will need to be continuously managed.

Evolution is inevitable

At present, AI and analytics innovation in data center management appears to be driven by startups rather than established software vendors. Few BMS and DCIM developers have integrated machine learning into their core products, and while some companies have features in development, these will take time to reach the market — if they ever leave the lab.

Uptime Intelligence is tracking six early-stage or private companies that use facilities data to create machine learning models and already have products or services on the market. It is likely more of these companies will emerge in the coming months and years.

These businesses are creating a new category of software and services that will require new types of interactions with all the moving parts inside the data center, as well as new commercial strategies and new methods of measuring the return on investment. Not all of them will be successful.

The speed of mainstream adoption will depend on how easy these tools will be to implement. Eventually, the industry will arrive at a specific set of processes and policies that focus on benefitting from equipment data.

PREDICTION 4

Direct liquid cooling will not resolve efficiency challenges

Key trends

- Expectations around scale adoption of direct liquid cooling (DLC) in data centers are running high on the back of escalating server power levels, rack densification and sustainability pressures.
- While the use of DLC will inevitably become more common to handle high concentrations of heat, most installations will be mixed with air-cooled IT infrastructure. Also, many DLC users will opt for lower temperatures to prioritize cooling capacity, performance and resiliency over cooling energy efficiency.
- Due to the above factors and a gradual, measured uptake, any impact on power usage effectiveness or sustainability performance from DLC adoption will remain imperceptible for some time.

A growing number of data center operators and equipment vendors are anticipating the proliferation of direct liquid cooling systems (DLC) over the next few years. Recently, IT and facility equipment vendors, together with some of the larger data center operators, have started commercializing DLC systems for much broader adoption, largely in response to runaway cooling requirements of server chips.

They cite that a main benefit of DLC is improved energy efficiency. Specifically, the superior thermal performance of liquids (compared with air) will reduce the consumption of electricity and water in heat rejection systems and will also increase opportunities for year-round free cooling in some climates. In turn, the data center's operational sustainability credentials would improve significantly.

However, many of the operators deploying DLC systems in the next few years will likely prioritize speed and ease of installation into existing environments, with a focus on maintaining infrastructure resiliency — rather than aiming for maximum DLC efficiency.

Another major factor is time: the pace of adoption. The use of DLC in mission-critical facilities, let alone a large-scale change, represents a shift in cooling design and infrastructure operations, with industry best practices yet to catch up. Many data center operators will also deem the current DLC systems limited or uneconomical for their applications, slowing rollout across the industry.

Cooling in mixed company

Operators will need to manage a potentially long period when liquid-cooled and air-cooled IT systems and infrastructure coexist in the same data center. In many cases, this will mean a cooling infrastructure (for heat transport and rejection) shared between air and liquid systems. Adoption of DLC will not be limited to data centers with chillers, but will occur in many sites without facility water.

In these hybrid environments, DLC's energy efficiency will be constrained by the supply temperature requirements of air-cooling equipment, which puts a lid on operating at higher temperatures — and compromises the cooling energy and capital efficiency benefits of DLC.

Even though DLC eliminates many, if not all, server fans and reduces airflow requirements for major gains in total infrastructure energy efficiency, these will be difficult to quantify for real-world reporting because IT fan power is not a commonly tracked metric — it is hidden in the IT load.

It will take years for DLC installations to reach the scale where a dedicated cooling infrastructure can be justified as a standard approach in data centers.

A significant share of DLC adoption will likely represent an investment in cooling performance and IT capacity, rather than facility efficiency gains

Hidden trade-offs in temperature

The same favorable thermal properties that make it possible to reach higher cooling energy efficiency can also be used for cooling performance advantage instead. Even when presented with the choice, some operators will choose lower supply temperatures for their DLC systems' water supply because of a range of benefits — despite the costs involved.

Chiefly, a low facility water temperature reduces the flow rate needed for the same cooling capacity, which eases pressure on pipes and pumping, including the DLC systems' coolant distribution units. Low facility water temperatures also make data center planning and design decisions simpler by guaranteeing the ability to meet the future needs of server technology. Coolant temperature requirements will only become stricter due to the evolution of server silicon.

But IT, too, benefits from lower temperatures. Processors, for example, exhibit lower static power losses at lower temperatures, which is energy that can be either saved or redirected towards other aspects of IT. This is particularly valuable when an operator wants to maximize compute performance against cost, which is often the case in high-performance computing and, more recently, artificial intelligence (AI) training. Running servers at a cooler temperature will also reduce component failure rates in general.

These facility and IT benefits from low temperature will often outweigh the attraction of a leaner heat rejection infrastructure. In effect, a significant share of DLC adoption will likely represent an investment in cooling performance and IT capacity, rather than facility efficiency gains.

DLC changes more than the coolant

For all its potential benefits, a switch to DLC raises some challenges to resiliency design, maintenance and operation. Implementing concurrent maintainability or fault tolerance with some DLC equipment may not be practical. In addition, in the event of a failure in the DLC system, cold plates tend to offer less than a minute of ride-through time because of their small coolant volume. Operating at high temperatures can mean it only takes seconds before servers overheat if cooling fails. As a result, a conversion to DLC can demand that organizations maintain their infrastructure resiliency standard in a different way from air cooling.

Organizational procedures for procurement, commissioning, maintenance and operations will need to be re-examined because DLC disrupts the current division of facilities and IT infrastructure functions. For air-cooling equipment, the division of equipment between facilities and IT teams is clearly defined. No such consensus exists for liquid cooling. A resetting of staff responsibilities will require much closer cooperation between facilities and IT infrastructure teams. All this will take considerable time and effort.

In the long term (i.e., 10 years or more), DLC is likely to handle a large share of IT workloads, including a broad set of business applications. This will happen as standardization efforts, real-world experience with DLC systems in production environments and mature guidance take shape in new, more robust products and best practices for the industry. For growth in mission-critical facility infrastructure, DLC systems will have to meet additional technical and economic objectives.

In the near term, the business case for DLC will likely prioritize IT performance and ease of retrofitting with a shared cooling infrastructure. Importantly, choosing lower water supply temperatures and utilizing chillers appears to be an attractive proposition for added resiliency. As many operators deem performance needs and mixed environments to be more pressing business concerns — free cooling aspirations, along with their benefits in sustainability, will have to wait for much of the industry.

PREDICTION 5

Hyperscale campuses begin to redraw the data center map

Key trends

- Hyperscale colocation campuses will continue to be seen as a solution to the rocketing demand for compute and storage.
- These campuses will be built on huge areas of land, capable of supporting multiple tenants, including both cloud providers and enterprises expanding their digital footprint.
- North America will see the largest of these campuses. However, other regions are also expanding rapidly, including Asia-Pacific, where the largest number of new developments are taking place.

Massive new colocation sites are being proposed by some builders, investors and operators as a solution to the rocketing demand for compute and storage. These hyperscale colocation campuses are a new addition to cloud provider-owned and operated hyperscale campuses. The largest facilities are targeting gigawatt scale, but most will be under that threshold. Uptime considers hyperscale size to be capacity levels that are upwards of 100 MW.

Analysis of 35 of the most recent hyperscale colocation campus projects across the world reveals a mean average planned capacity of more than 400 MW. This differs considerably by region. For example, Asia-Pacific has the largest number of proposed projects, with a mean average capacity under 200 MW per campus. North America, however, has the largest MW campus projects and a mean average campus size of over 600 MW. These figures exclude cloud provider-owned hyperscale campus developments.

Table 3

Hyperscale colocation campus projects in progress

Campus location*	Provisioned power in megawatts (MW)	Number of projects	Average megawatts (MW)	Total spend ($ million)
North America	6,210 MW	10	621 MW	$45,000m
Asia-Pacific (excluding China)	3,832 MW	21	182 MW	$11,628m
Europe, the Middle East and Africa	750 MW	2	375 MW	$6,400m
China	500 MW	1	500 MW	$4,890m
Latin America	450 MW	1	450 MW	$400m
Total (worldwide)	**11,742 MW**	**35**	**426 MW**	**$68,318m**
Annual terawatt-hours (TWh) at 50% of provisioned power	51 TWh			

Provisioned power is likely capacity once everything in place; projects announced or identified since 2021; single location campuses (i.e., not multisite investments); public cloud vendor hyperscale projects excluded.
**Campuses 100 MW and above.*

UPTIME INSTITUTE 2024 uptime INTELLIGENCE

If all these projects were built and ran at half of their projected capacity, they would account for around 51 terawatt-hours (TWh) of energy use each year. Globally, current estimates for the annual energy use by data centers range from 200 TWh to 400 TWh.

The hyperscale colocation campus

Based on published plans so far, a hyperscale colocation campus will be made up of hyperscaler-size facilities that are physically colocated on the same site, typically covering millions of square meters. There could be multiple colocation providers in the same campus, depending on its size and scale.

Expensive infrastructure — such as high-bandwidth fiber, subsea landing stations, electricity substations and renewable energy generation and storage — can then be shared by multiple tenants.

Hyperscalers have a key role to play: their involvement both drives and guarantees demand, as well as providing credibility, financing, connectivity and expertise

Investment on this scale favors a consortium model, which may include stakeholders from across the data center ecosystem. And here, hyperscalers have a key role to play: their involvement both drives and guarantees demand, as well as providing credibility, financing, connectivity and expertise.

In most cases, there will be one or two lead operators, wholesale colocation companies or hyperscalers, with smaller operators also expected to take capacity. This will lead to the emergence of new data center clusters that, in part, resemble the existing clusters in Tier 1 markets today.

Hyperscale colocation campuses will be supported by redundant and new high-bandwidth fiber to major centers. They will also be typically located on brownfield sites with minimal environmental impact.

Designed to enable operators to meet low-carbon or carbon-free energy objectives, these projects may involve power purchase agreements, colocation with power plants and use of on-site renewable energy sources. This may include solar, wind, geothermal and even nuclear energy, as well as the use of microgrids to manage on-site / off-site resources.

For data center operators, these new campuses will give them the opportunity to expand across multiple data halls. Some of the characteristics include:

- Optimize energy consumption, with a power usage effectiveness (PUE) of better than 1.4, preferably closer to 1.2 in many locations.
- Apply a modular design of internal and external spaces to enable rapid reconfiguration.
- Use of artificial intelligence (AI)-based management systems to optimize monitoring, performance and availability of data center assets.
- Presence in other hyperscale colocation campuses, enabling repeatability of designs.
- Support for high densities, high-performance servers and graphics processing units (GPUs) to run AI models, which are likely to be drivers of demand. This will likely involve liquid cooling options.

Impact on the data center sector

Uptime Intelligence has identified four areas areas where these hyperscale colocation sites will have an impact on the sector:

- **The data center map.** As new clusters of data centers are built, the gravity of the industry will shift. New builds will be attracted as much to the new (and second-tier) clusters as to the current first-tier hotspots. This will likely bring down the cost of colocation, cloud and connectivity.
- **Supply chain.** Hyperscale colocation campuses will not only push up demand for all equipment, which is already high, but they will also enable operators to build at scale, encouraging automation and investment.
- **Sustainability.** The sites will aim to take advantage of a low-carbon energy source. But the huge and reliable demand from IT will also make large-scale innovation and investment possible, with the sites becoming centers for microgrids, long-duration batteries and on-site renewable energy generation.
- **Resiliency.** The new hyperscale colocation campuses (and associated fiber) outside of these centers, along with the development of new hubs in smaller cities, will likely provide more resiliency and diversity across national and international digital infrastructure.

Appendix A

Summary of the data center predictions for 2024

1. Operators — prepare for a sustainability reckoning

New reporting laws and toughening requirements will enforce stricter data center carbon reporting in many countries. These will challenge organizations' publicly announced sustainability goals and force operators to prove their targets are both realistic and evidence based. For many, this will be painful and expensive.

2. Demand for AI will have limited impact on most operators

The fervor around AI has the data center industry bracing itself for a significant increase in demand, and a need for more power and cooling. While the overall impact on data centers may ultimately be profound, the most demanding services will be delivered only by a few. For most operators, the impact will be indirect: the immediate challenge being how best to deliver a richer mix of densities and resiliency tiers from the same facility.

3. Data center software gets smarter, leverages data — at last

Operators have been slow to take advantage of developments in software, connectivity and sensor technologies that can help optimize and automate the running of critical infrastructure. This is beginning to change, with more operators embracing new tools and the intelligent use of data (including machine learning). But the market is still evolving, and there will be risks from complexity, poor implementation and tool selection.

4. Direct liquid cooling will not solve efficiency challenges

Operators have great expectations of direct liquid cooling in terms of improving efficiency and sustainability. However, these benefits will be out of reach for many organizations. A slow rollout of the technology, characterized by mixed environments, constrained optimization, and the continuing requirement for existing systems to run in parallel will limit its contribution to infrastructure efficiency — even if it is necessary.

5. Hyperscale campuses begin to redraw the data center map

The build out of new hyperscale colocation campuses, connected by wide-bandwidth fiber, will relieve pressure on traditional data center hotspots — and, in the long term, lower colocation prices. As a solution to rocketing demand for compute and storage, the hyperscale campus will emerge slowly — with the availability of fiber and power being critical factors.

Appendix B

Recap of the data center predictions for 2023*

1. Geopolitics deepens supply chain worries

Data center supply chains are still reeling from long delays, and current geopolitical dynamics present new threats. Uptime sees particularly high risks around the advanced semiconductors vital to IT hardware and data center equipment. Subsea cable systems are particularly vulnerable to hostile forces or agents.
(*Continued relevance: medium*)

2. Too hot to handle? Operators struggle with new chips

Operators will be faced with various choices in handling new-generation IT technologies. Driven primarily by silicon technology, data center capacity planning will need to accommodate a potentially fast-shifting balance between power, cooling and space.
(*Continued relevance: high*)

3. Cloud migrations to face closer scrutiny

The cost of migration and the threat of spiraling cloud costs will deter some mission-critical migrations in the years ahead. Governments, regulators and executives are trying to understand and limit the risks associated with high dependency on the public cloud. This, too, will cause some organizations to proceed more cautiously than in the past.
(*Continued relevance: high*)

4. Energy efficiency focus to shift to IT — at last

More stringent sustainability regulation and reporting requirements will force IT to deliver improved performance in energy efficiency. Fewer, yet higher-performance, highly utilized servers could deliver major energy gains. Refusing to deploy these improvements will be increasingly difficult to justify.
(*Continued relevance: high*)

5. Data center costs set to rise and rise

The costs of critical digital infrastructure are set to rise: a trend likely to persist in the medium term. Supply chain problems, record inflation and higher prices for capital, energy and labor have all contributed to rising prices.
(*Continued relevance: high*)

(**2023 predictions made in December 2022.*)

About the authors

Douglas Donnellan

Douglas Donnellan is a Research Analyst at Uptime Intelligence covering sustainability in data centers. His background includes environmental research and communications, with a strong focus on education.

ddonnellan@uptimeinstitute.com

Andy Lawrence

Andy Lawrence is Uptime Intelligence's Executive Director of Research. He is Uptime Institute's Executive Director of Research and has spent three decades analyzing developments in IT, emerging technologies, data centers and infrastructure. He also advises companies on their technical and business strategy.

alawrence@uptimeinstitute.com

Daniel Bizo

Daniel Bizo is Uptime Intelligence's Research Director. He has been covering the business and technology of enterprise IT and infrastructure in various roles, including more than a decade as an industry analyst and advisor.

dbizo@uptimeinstitute.com

Max Smolaks

Max Smolaks is a Research Analyst at Uptime Intelligence. His expertise spans digital infrastructure management software, power and cooling equipment, and regulations and standards. He has 10 years' experience as a technology journalist, reporting on innovation in IT and data center infrastructure.

msmolaks@uptimeinstitute.com

Jacqueline Davis

Jacqueline Davis is a Research Analyst at Uptime Intelligence covering global trends and technologies that underpin critical digital infrastructure. Her background includes environmental monitoring and data interpretation in the environmental compliance and health and safety fields.

jdavis@uptimeinstitute.com

John O’Brien

John O’Brien is Uptime Intelligence’s Senior Research Analyst for Cloud and Software Automation. As a technology industry analyst for over two decades, John has been analyzing the impact of cloud migration, modernization and optimization for the past decade. John covers hybrid and multi-cloud infrastructure, automation, and emerging AIOps, DataOps and FinOps practices.

jobrien@uptimeinstitute.com

All general queries

Uptime Institute
405 Lexington Avenue
9th Floor
New York, NY 10174, USA
+1 212 505 3030

info@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry’s most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.

Uptime Institute is headquartered in New York, NY, with offices in Seattle, London, Sao Paulo, Dubai, Singapore and Taipei.

For more information, please visit **www.uptimeinstitute.com**