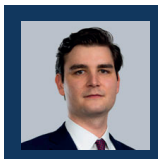
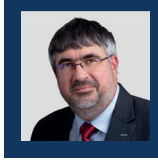
AI is transforming the data center



SEBASTIAN DOOLEY, CFA
Senior Fund Manager



PAUL LEWIS
Director, European Data Centres

Artificial intelligence (AI), formerly relegated to the pages of scientific journals and science fiction narratives, is now permeating the fabric of our everyday lives—and transforming the data center in the process. Because AI workloads are quite different from traditional workloads, supporting them demands a new approach to data center design and operation. As AI applications progress—for example, from training models to decision-making—its impact on the data center industry will continue to evolve.

Adding to the pressure on data center operators, demand for data centers to support AI is additive to demand for data centers to support traditional workloads. Data center operators will have to support these very different demand profiles—sometimes even within the same facility. This article explores how.

AI has been around since the 50s. So what's new?

For decades, artificial intelligence has been—at least according to popular media—poised to take over the world. And yet it's only recently that AI has begun to infuse daily life. (It's still not taking over the world. But it is affecting how data centers are designed and operated.)

As far back as the 1950s, early AI pioneers laid the groundwork for AI as we know it with the development of rudimentary neural networks, inspired by the human brain. Today, with generative AI we're witnessing the rise of AI that can not only analyze and process information but also create original content.

The increasingly rapid evolution of AI is driven by:

- **The data deluge** — The recent exponential growth in data generation, primarily driven by big data analytics and the Internet of Things (IoT), has provided AI algorithms with the extensive datasets necessary for effective learning and development.
- **Advancements in computing architecture and chip performance** — Processors like GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units) are specifically designed for parallel processing, significantly boosting performance compared to traditional CPUs.
- **AI's increasingly broad applicability and accessibility** — Powerful AI capabilities are becoming accessible to developers of all levels through user-friendly interfaces and cloud-based services. As a result, AI is more accessible and applicable to individuals and enterprises across various sectors.

Supporting AI workloads demands a new approach to data center design and operation

The differences between AI workloads and traditional workloads drive new requirements for the data center. The fact that AI workloads are generally more power intensive but with wide variability drives new power infrastructure and management requirements. The fact that AI workloads are typically deployed in higher density configurations drives the need for new approaches to cooling. And the fact that some AI workloads are latency sensitive drives facility location and size decisions.

Supporting AI workloads demands new power infrastructure and approaches to power management

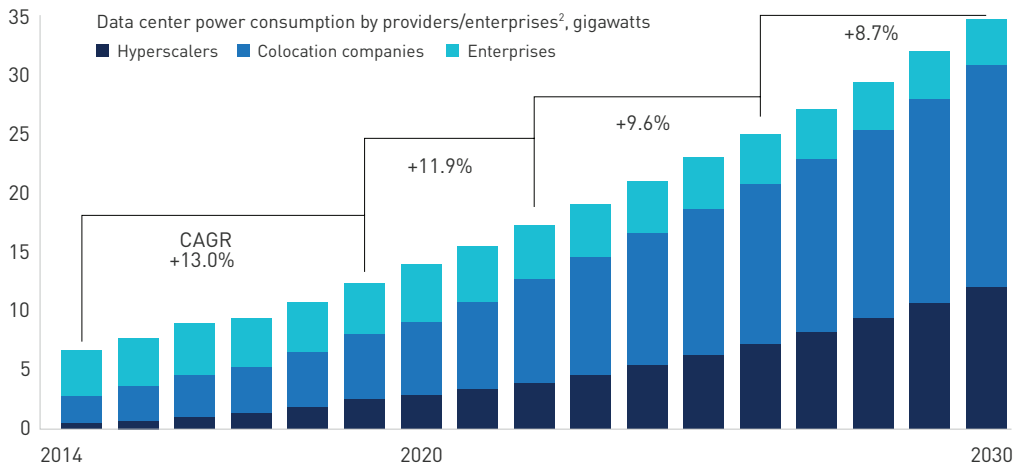
The scale of data centers is growing, with many data centers being built today over 50 megawatts. Data center campuses can be in the 100s of megawatts, with some over a gigawatt.

These massive power requirements are making new data center capacity increasingly hard to come by in many major markets, where utility power is increasingly constrained. As CBRE reports, "Sourcing enough power is a top priority of data center operators. Certain secondary markets with robust power supplies stand to attract more data center operators."[1]

Where there is sufficient utility grid capacity, an existing data center can be retrofit to increase power capacity. This involves a systematic upgrade of the power infrastructure—adding transformers, expanding electrical backup systems (uninterruptible power supplies and standby generators), and installing additional switchgear and power distribution units (PDUs).

Retrofitting a data center to enhance its power capacity in the absence of sufficient grid capacity presents a more complex challenge. The data center operator might consider on-site power generation or the development of a microgrid, which can operate independently or in conjunction with the main power grid. Ultimately, the operator may have to work with the local

Rising data center power consumption



This chart shows the continued rise in data center power consumption and the need for possible upgrades or new construction.
Source: McKinsey & Company, Investing in the rising data center economy, 17 January 2023.

utility to enhance the grid's capacity through distribution infrastructure upgrades or the construction of new power generation facilities.

In addition to demanding more power than traditional workloads, AI workloads are also more highly variable. Accommodating power usage spikes may require new power infrastructure as well as new approaches to data center power management.

To effectively support AI workloads, data centers must be capable of handling more variable loads with higher peaks. Accomplishing that might involve upgrading transformers, PDUs, and wiring to handle the increased power draw without compromising safety or efficiency. Intelligent power distribution systems can monitor and dynamically adjust the power supply to different racks or zones based on real-time demand, ensuring that AI workloads receive sufficient power during peak processing times.

Incorporating energy storage solutions, like battery banks, can help in smoothing out the power spikes. During periods of lower demand, these batteries can store energy, which can then be utilized during periods of high demand, thereby reducing the strain on the primary power infrastructure.

Beyond new power infrastructure, new approaches to data center power management are essential to supporting highly variable AI workloads. Power monitoring and management software can predict power usage patterns, monitor real-time consumption, and provide alerts for potential issues. Such predictive analysis helps in optimizing power usage and planning for future capacity needs.

Supporting AI workloads demands new cooling infrastructure and strategies

Because AI workloads are typically deployed in higher density configurations, the cooling requirements for AI workloads are substantially higher than for traditional workloads—driving the need for new cooling infrastructure and strategies.

Traditional forced air cooling, which operates by circulating cool air and expelling warm air, can encounter several challenges at higher densities. The sheer volume of heat generated in a high-density rack can overwhelm the air cooling capacity. Air has a lower specific heat capacity compared to liquid coolant, meaning air can carry less heat per unit volume. This limitation necessitates larger volumes of air to be circulated at a higher velocity.

Moving the heat exchange closer to the servers can increase cooling efficiency and provide supplemental cooling to localized loads, augmenting a more traditional "flooded room" cooling strategy. A more effective solution for high- density racks, liquid cooling systems work by circulating a coolant directly through heat exchangers or cold plates positioned close to heat-generating components. This approach allows for more direct and efficient heat transfer.

Retrofitting a data center equipped with forced air cooling to accommodate liquid cooling at the rack level is feasible. For facilities that already have a chilled water loop, the transition from traditional forced air cooling to liquid cooling is easier, as the existing chilled water loop can be extended directly to the racks. This extension typically involves the installation of specialized piping and heat exchangers within or adjacent to the racks to facilitate the direct transfer of heat from the high-density computing equipment to the water.

One of Principal's assets—a data center commissioned in 2009—provides a great example of the feasibility of retrofitting an existing facility to support more power (and the requisite additional cooling) with relatively modest infrastructure additions. A data hall originally designed to support 2.7 MW of capacity is currently being retrofit to support 12.5 MW, in the same footprint and with the original mechanical equipment.

Supporting AI workloads demands different facility location and size decisions

The fact that some AI workloads are latency sensitive drives facility location and size decisions. In some cases, that will mean smaller data centers in edge locations.

The impact of latency on AI workloads varies depending on the specific application and its requirements. Data center providers catering to AI applications must therefore optimize their infrastructure accordingly, whether it's through the use of edge computing to reduce latency for real-time applications or optimizing network and storage systems for high-throughput, latency-tolerant tasks.

Because AI model training, for example, may not require low latency, such workloads can be located in very large data centers where power is plentiful and inexpensive. But because inference requires low latency, once an AI model is trained it may need to be moved to a data center closer to where the data is produced and consumed. In many cases these data centers will be at the edge—whether that's the city where autonomous vehicles are operating or the availability zone where the enterprise keeps its data. They may also be smaller than most of today's deployments (e.g. 3 MW instead of 50 MW).

AI doesn't make existing data centers obsolete

Many AI deployments will be in facilities dedicated to AI. Deploying AI in a data center outfitted for traditional cloud or enterprise use would typically likely require at least some retrofitting to support the unique network architecture and deployment densities of AI workloads—for example, re-racking the space to widen cold rows or install new network backbone, or the kind of power and cooling infrastructure upgrades mentioned previously.

When a data center has been designed to be future-proof, retrofitting it to support new AI workloads is feasible and may be the most cost-effective option. One of Principal's assets—originally a bank-owned data center commissioned nearly 15 years ago—provides a great example. To support the workloads of the time, the facility was designed for high redundancy and low density. But the data center had the ability to deliver significant amounts of power and cooling so retrofitting it to support modern workloads was feasible. In fact, the retrofit data center was leased to a specialty cloud service provider to deploy an AI/ML strategy that at full deployment will be one of the largest supercomputers in the world.

While running multiple types of workloads in a single facility can make efficiency more difficult to achieve, a data center may be able to efficiently support both traditional workloads and AI workloads. For example, one data hall could have forced air cooling supporting traditional workloads at relatively low density while another data hall in the same building could have liquid immersion cooling to support an extremely high density deployment running AI workloads.

Bottom line

AI has ushered in a new era in data center design and operation. The unique demands of AI workloads, from increased power consumption and cooling requirements to the need for continual learning and updating, require a reimagining of traditional data center architectures. This evolution, while challenging, presents an opportunity for innovation and growth in the data center industry.

As AI continues to advance, it will be imperative for businesses and technology leaders to stay abreast of these changes and adapt their strategies accordingly. The future of data centers lies in their ability to effectively support the dynamic, power-intensive, and ever-evolving workloads.

For more information,
visit: principalam.com/datacentres

Principal Asset Management℠

Investing involves risk, including possible loss of principal. Past Performance does not guarantee future return. This document is intended for sophisticated institutional and professional investors.

[1] CBRE, Global Data Center Trends 2023, 14 July 2023.

[2] Demand is measured by power consumption to reflect the number of servers a data center can house. Demand includes megawatts for storage, servers, and networks.