McKinsey & Company

Technology, Media & Telecommunications Practice

AI power: Expanding data center capacity to meet growing demand

Soaring demand for AI-ready data centers offers many opportunities for companies and investors across the value chain. How quickly they grasp them could determine the pace at which AI is deployed.

This article is a collaborative effort by Bhargs Srivathsan, Marc Sorel, and Pankaj Sachdeva, with Arjita Bhan, Haripreet Batra, Raman Sharma, Rishi Gupta, and Surbhi Choudhary, representing views from McKinsey's Technology, Media & Telecommunications Practice.

October 2024

The race is on to build sufficient data center capacity to support a massive acceleration in the use of AI. Data center demand[1] has already soared in response to the role data plays in modern lives. But with the emergence of generative AI (gen AI), demand is set to rise even higher. And that is likely to presage a supply deficit.

As challenging as this could be, companies and investors along the entire data center value chain have an opportunity to help address the looming capacity crunch—if they understand the requirements of data centers designed for the AI age. A big chunk of growing demand—about 70 percent at the midpoint of McKinsey's range of possible scenarios—is for data centers equipped to host advanced-AI workloads. And the nature of those workloads is rapidly transforming where and how data centers are being designed and operated.

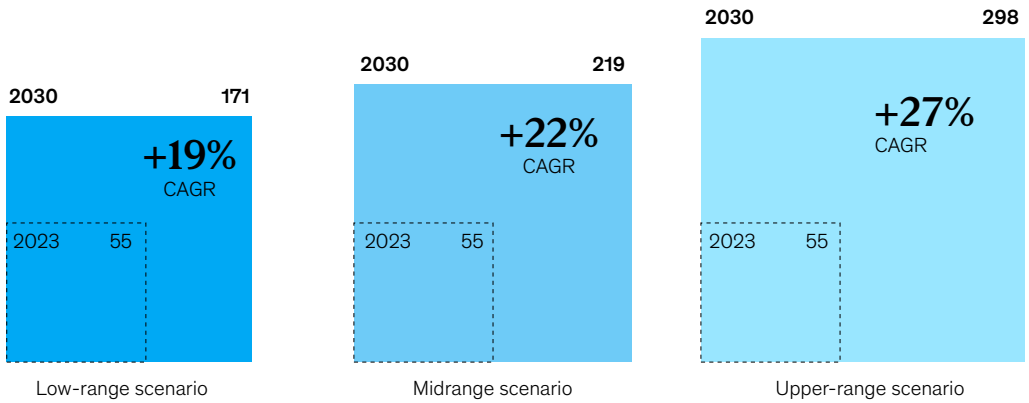
Exploding demand and lagging supply

Future demand for data center capacity will depend on factors that are still hard to accurately determine. The pace of adoption of advanced-AI use cases will certainly count, but so too will the mix of different types of chips deployed and their associated power consumption, as well as the balance between cloud and edge computing for AI workloads and the typical compute, storage, and network needs of AI workloads. This explains McKinsey's range of estimates. Our analysis of current trends suggests that global demand for data center capacity could rise at an annual rate of between 19 and 22 percent from 2023 to 2030 to reach an annual demand of 171 to 219 gigawatts (GW). A less likely yet still possible scenario sees demand rising by 27 percent to reach 298 GW (Exhibit 1).[2] This contrasts with the current demand

Exhibit 1

Global demand for data center capacity could more than triple by 2030.

Demand for data center capacity,[1] gigawatts



[1]Three scenarios showing the upper-, low-, and midrange estimates of demand, based on analysis of AI adoption trends; growth in shipments of different types of chips (application-specific integrated circuits, graphics processing units, etc) and associated power consumption; and the typical compute, storage, and network needs of AI workloads. Demand is measured by power consumption to reflect the number of servers a facility can house.
Source: McKinsey Data Center Demand model

McKinsey & Company

[1] Demand is measured by power consumption to reflect the number of servers a facility can house.
[2] Estimates are based on an analysis of AI adoption trends; the likely mix of application-specific integrated circuits (ASICs), graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and nonaccelerated central processing units (CPUs) used to run workloads; the mix between training and inference workloads; the emergence of inference optimized chips; efficiencies in model training; and the extent to which higher processing power requires higher power consumption.

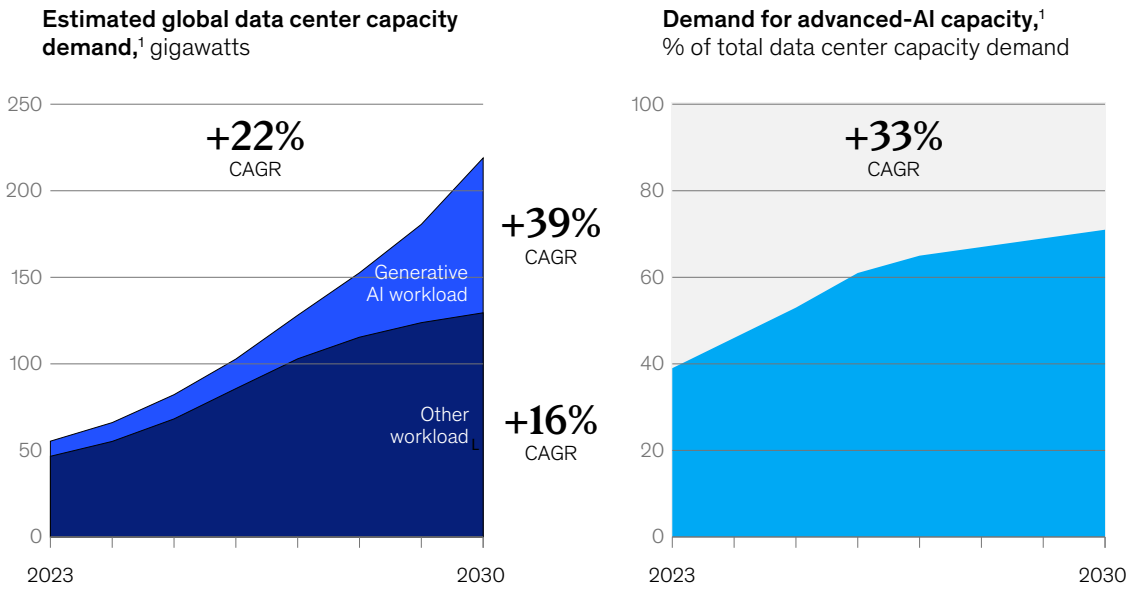
of 60 GW, raising the potential for a significant supply deficit. To avoid a deficit, at least twice the data center capacity built since 2000 would have to be built in less than a quarter of the time.

However, estimating the precise size of that deficit is hard because of uncertainties surrounding the pace of rising demand, the extent to which innovations might improve power efficiency, and limited knowledge concerning the longer-term expansion plans of data center owners and operators. But even if all currently known plans are delivered on time, there could still be a data center supply deficit of more than 15 GW in the United States alone by 2030.

Demand for AI-ready capacity is the main driver of this potential deficit—as it must provide the high computational power and power density required by AI workloads. Our analysis suggests that demand for AI-ready data center capacity will rise at an average rate of 33 percent a year between 2023 and 2030 in a midrange scenario. This means that around 70 percent of total demand for data center capacity will be for data centers equipped to host advanced-AI workloads by 2030. Gen AI, currently the fastest-growing advanced-AI use case, will account for around 40 percent of the total (Exhibit 2).

Exhibit 2

AI is the key driver of growth in demand for data center capacity.



[1]Midrange scenario is based on analysis of AI adoption trends; growth in shipments of different types of chips (application-specific integrated circuits, graphics processing units, etc) and associated power consumption; and the typical compute, storage, and network needs of AI workloads. Demand is measured by power consumption to reflect the number of servers a facility can house.
Source: McKinsey Data Center Demand model

McKinsey & Company

Hyperscalers dominate capacity demand and supply

Cloud service providers (CSPs) such as Amazon Web Services, Google Cloud, Microsoft Azure, and Baidu are the companies fueling most of today's incremental demand for AI-ready data centers. That's because of the capacity these hyperscalers require to run large foundational models developed in-house, such as Google's Gemini, or to host models developed by AI companies, such as OpenAI's ChatGPT.

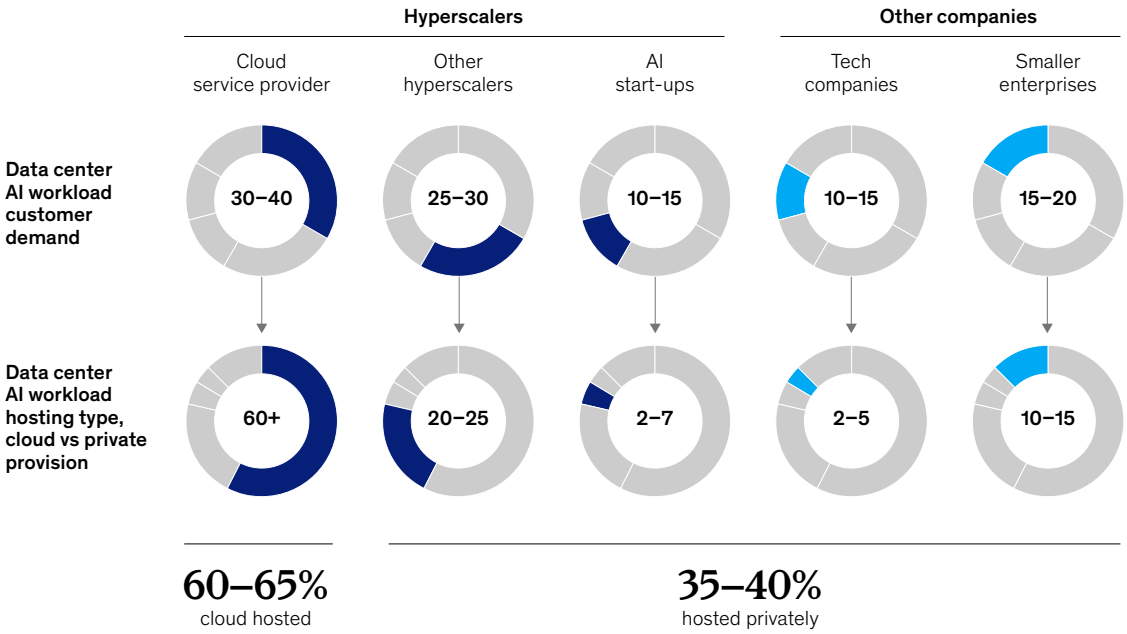
Most other companies are using (and sometimes refining) off-the-shelf models that are largely hosted on a public cloud. As the technology matures, more enterprises are likely to build and train their own models on their internal data, which could lead to demand for private hosting. Our estimate, however, is that by 2030, some 60 to 65 percent of AI workloads in Europe and the United States will be hosted on CSP infrastructures and other hyperscaler infrastructures (Exhibit 3).

To address the increasing demand, CSPs, which currently own more than half of the world's AI-ready data center capacity, according to McKinsey estimates, are rapidly constructing state-of-the-art facilities. However, because of impending supply constraints, they are also partnering with colocation providers (known as "colos") that are similarly expanding their infrastructures.

Exhibit 3

By the end of the decade, hyperscalers will host the lion's share of data center AI workloads.

Data center AI workload customer demand and hosting type in 2030 in Europe and the US, %



Source: McKinsey Data Center Demand model

McKinsey & Company

A small group of graphics processing unit (GPU) cloud providers is also emerging to meet the demand for AI-ready data center capacity. As the name suggests, these providers offer high-performance GPUs as a service to train AI models, then often work with colocation providers to build and operate the data center facility. GPU cloud provider CoreWeave, for example, had a fleet of approximately 45,000 GPUs by July 2024 and aims to operate in 28 locations globally by the end of the year.[3]

This new capacity is not likely to keep pace with demand, however. Tight supply is already apparent in the market. Prices charged by colocation providers for available data center capacity in the United States fell steadily from 2014 to 2020 in most primary markets but then rose by an average of 35 percent between 2020 and 2023. Additionally, new capacity due to come online in the next two to three years has already been leased out.[4] In Northern Virginia, dubbed the data capital of the world because of the high number of data centers concentrated there, the vacancy rate was less than 1 percent in 2024.[5] Making matters worse is the high demand for new data centers, which is causing supply constraints for power, key pieces of electrical infrastructure, and labor, thereby delaying completion of new facilities.

New location, design, and operational requirements

Data centers have seen steady changes and improvements over the past decade, gradually getting bigger, housing more power-hungry, high-density servers, and operating more efficiently and sustainably.

AI has forced the pace of progress, however. Most notably, data centers have exploded in size in terms of power consumption. Ten years ago, a 30-megawatt (MW) center was considered large. Today, a 200-MW facility is considered normal. The driving force for this is the computing power required for AI workloads, which, in turn, bumps up energy consumption.

All data centers consume significant amounts of energy, but AI-ready ones are especially demanding because of their high average power densities—the energy consumption of servers in the racks. Average power densities have more than doubled in just two years, to 17 kilowatts (kW) per rack, from eight kW, and are expected to rise to as high as 30 kW by 2027 as AI workloads increase. Training models like ChatGPT can consume more than 80 kW per rack, while Nvidia's latest chip, the GB200, combined with its servers, may require rack densities of up to 120 kW.

Such high energy demand and power density, along with the complexity of different AI workloads, are bringing about rapid change in three main areas in the construction of data centers: their location and the accompanying power infrastructure, the design of mechanical systems, and the design of electrical systems.

Data center location and power infrastructure

As more data centers are built and the amount of power they require grows, power supply is becoming an issue in markets that have traditionally attracted clusters of data centers, such as Northern Virginia and Santa Clara in the United States. Many utilities find they haven't been able to build out transmission infrastructure quickly enough, and there is concern that at some stage they may be unable to generate sufficient power.

This can slow data center expansion. For example, some utilities initially offer only relatively small chunks of power to data centers, which they then increase as they build out the power infrastructure—perhaps in 15- to 25-MW tranches for a 100-MW new data center campus. Additionally,

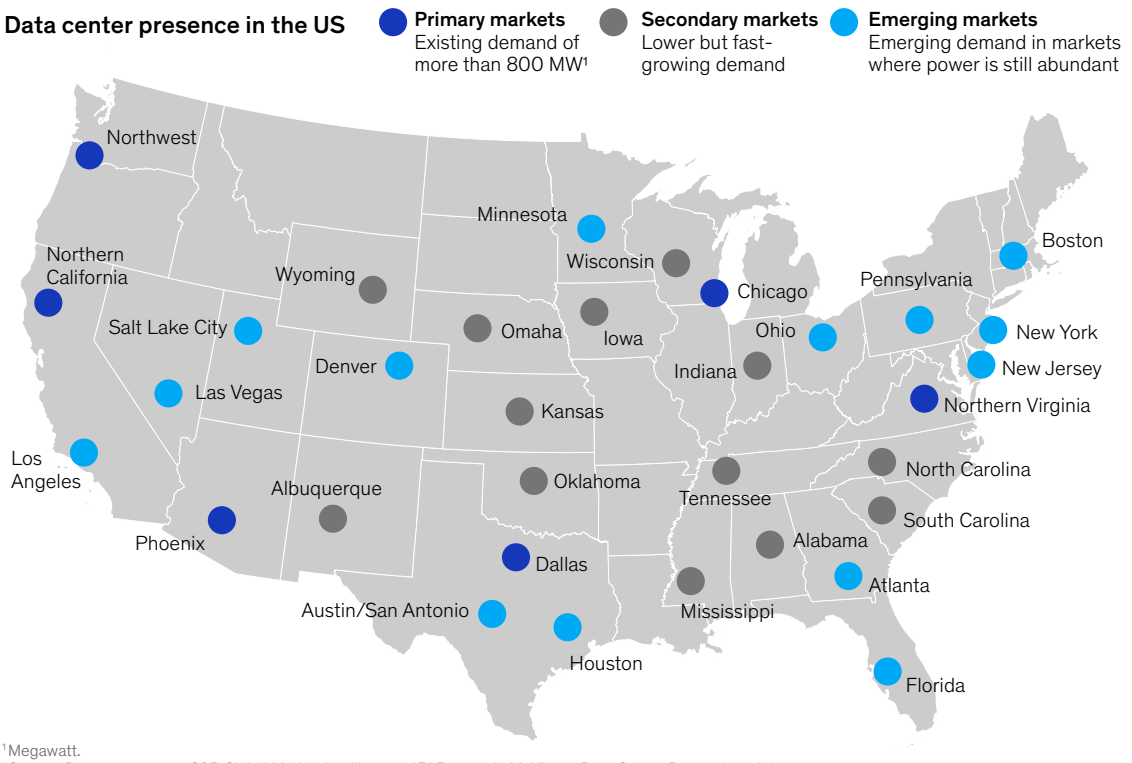
[3] Rod Walton and Matt Vincent, "CoreWeave, Chirisa tap Bloom Energy for Illinois AI data center project, lean into microgrids," Data Center Frontier, July 22, 2024.
[4] "North America data center trends H2 2023," CBRE, March 6, 2024; "Data centers 2024 global outlook," Jones Lang LaSalle, January 31, 2024.
[5] "Global data center trends 2024: Limited power availability drives rental rate growth worldwide," CBRE, June 24, 2024.

in some countries, concern about the pressure data centers exert on electricity grids as well as the impact on national climate targets have brought a complete halt to the building of new ones. Ireland, for instance, has stopped issuing new grid connections to data centers in the Dublin area until 2028. Ireland's transmission system operator estimates that data centers will account for 28 percent of the country's power use by 2031.[6]

The fact that not all AI workloads are equal has partly alleviated the power problem. Historically low latency has been one of several critical factors in determining data center location, often leading colocation providers to establish facilities near population centers. When AI models are being trained, typical performance factors such as low latency and network redundancy are less important. It is only when the model is put into operation—during the inferencing workload—that these factors become crucial for optimal performance. Hence, data centers dedicated to training AI models are being built in more remote locations in the United States, such as Indiana, Iowa, and Wyoming, where power is still abundant and grids are less strained (Exhibit 4). But given the lack of adequate power transmission infrastructure in these locations, power supply may still become an issue as demand grows.

Exhibit 4

Data centers are emerging in more remote locations, where power is still abundant and grids less strained.



[1] Megawatt.
Source: Datacenters.com; S&P Global Market Intelligence 451 Research; McKinsey Data Center Demand model

McKinsey & Company

[6] *Ireland capacity outlook 2022–2031*, EirGrid and SONI, October 2022.

Against this backdrop, some data center operators are acquiring facilities built close to power plants to help overcome transmission issues (for example, the Talen Energy data center powered by a nuclear power plant).[7] And some have started generating their own off-grid power using behind-the-meter solutions, such as fuel cells, batteries, or renewables. In the longer term, small modular reactors (SMRs) might be an option.

Mechanical system design

AI servers consume so much energy that they get hot—so much so that air-based cooling systems, which circulate cold air around them, often can't keep up. The upper limit to their effectiveness is generally considered to be power densities of up to 50 kW per rack—a level that might be adequate for AI inferencing workloads that have lower power densities, but not for training workloads.

This has prompted a shift to an approach that removes heat directly from racks by using liquid, which is significantly more effective in absorbing and transferring heat than air. There are three such rack-based technologies that differ both in their application and in the extent to which they depart from conventional data center cooling systems:

— *Rear-door heat exchangers* (RDHX), which are the closest to conventional cooling technology, combine cold air that is forced to the racks with liquid-cooled heat exchangers installed at the back of the rack. They tend to be used in data centers, where space is constrained and rack density is in the range of 40 to 60 kW.

— *Direct-to-chip* (DTC) technology uses a liquid (generally antifreeze coolant or a mix of water and glycol) that circulates through a cold plate in direct contact with the most power-dense electronic components, such as GPUs and certain central processing units. Of the three technologies, DTC is the one most commonly deployed to date, as it can handle power densities of 60 to 120 kW and can be integrated relatively easily within the existing infrastructure of a data center.

— *Liquid immersion cooling* entails placing the servers in a tank filled with dielectric fluid. There are two variations of this cooling method: single-phase immersion and dual-phase immersion. Both can cool racks with a power density of 100 kW, though dual-phase immersion has been used for racks with power densities upward of 150 kW per rack. The pace of adoption of liquid cooling in data centers has been slow, however, limited largely to crypto-mining applications that tend to be more open to experimentation. Additionally, there is concern about the health and environmental impact of the per- and polyfluoroalkyl substances (PFAS) chemicals used in dual-phase cooling.[8]

Another important benefit of liquid cooling systems is that they can keep electronics at a more consistent temperature by targeting the hottest spots. This can increase the life of the hardware and allow it to operate at higher speeds than those originally intended by manufacturers. Also, because the liquid extracts heat directly from the electronic components, capital and operational costs and power usage effectiveness (PuE)—a measure of how efficiently a data center uses energy—are improved.[9] Some data centers have seen a 10 percent reduction in PuE using liquid cooling systems rather than air cooling ones.

Electrical system design

AI workloads call for larger power distribution units to cope with higher power densities, leading many data center operators to install larger switchgear and floor-mounted power distribution units. This reduces the complexity as well as the capital and operational costs of installing and maintaining multiple smaller units.

[7] "Amazon buys nuclear-powered data center from Talen," Nuclear Newswire, March 7, 2024.
[8] "Will PFAS be the death of two-phase cooling?," *Electronics Cooling*, June 11, 2024.
[9] Higher circulation temperature reduces the size of the cooling system required and therefore the energy required and PuE.

Operators are also rethinking the power architecture at rack level. Because of the increasing power of AI chips, some hyperscalers and OEMs are considering installing servers with a 48-volt power supply unit rather than the traditional 12-volt unit, thereby reducing energy loss and improving system efficiency. In tests, these units have been shown to reduce energy loss by at least 25 percent.[10] Power-hungry AI workloads also require bigger, higher-capacity centralized uninterruptible power supply systems, leading to more complex designs.

Backup systems are also changing, as some AI-focused data centers reassess the amount of backup power capacity required. Traditional data centers that run mission-critical business applications for clients have backup generators to guard against any interruption due to a power outage. However, since training workloads are less critical to business operations, they can operate with lower power redundancies.

Opportunities abound

Understanding the requirements of AI-ready data centers makes clear the wide range of opportunities that exist for companies and investors in such a high-growth market. A number of them follow:

— *Owners and operators of data centers.* Colocation providers can retrofit existing data centers and build more new ones, particularly to lease capacity to hyperscalers that might struggle to keep pace with demand despite their current investments. Colocation providers that are able to offer build-to-suit development services—that is, those able to build and operate data centers customized to the specific needs and designs of each hyperscaler—might prove to be particularly attractive partners. There are also opportunities for GPU cloud providers, whose business model is gaining traction with investors, as Nvidia, which previously sold its GPUs mainly to hyperscalers, broadens its customer base.

— *Data center construction companies and equipment suppliers.* The supply crunch raises demand for modularized construction, which not only speeds up the build-out of data centers but also promotes sustainable construction practices. And there is high demand for all types of mechanical and electrical equipment within data centers. Capital spending on procurement and installation of mechanical and electrical systems for data centers is likely to exceed $250 billion by 2030, according to McKinsey estimates.

— *Across the energy and power supply value chain.* A variety of players in this space can take advantage of the AI data center building boom. These players include businesses generating and distributing more energy, particularly green energy; developing on-site, sustainable power solutions such as fuel cells, solar power, and small modular reactors; or promoting ways to reuse the heat generated at data centers in residential or other commercial buildings.

Companies and investors keen to pursue such opportunities may have to consider modifying their usual approach, however, if they are to win in an AI era. Here are some alternative approaches:

— *They may have to move more quickly than they have in the past, given the pace of change in the sector.* The race is on, for example, for data center owners and operators to seek and secure new sites with access to reliable power, for cooling-system manufacturers to innovate and offer solutions that tackle rapidly increasing power densities, and for other equipment providers such as transformer manufacturers to scale up capacity.

— *They may have to collaborate more, as constant innovation is a defining feature of today's data center value chain.* Collaboration between companies in the value chain or with those in other sectors can speed up the pace of innovation and help scale it. Collaboration can

[10] Paul O'Shea, "48V: The new standard for high-density, power efficient data centers," *Power Electronics News*, August 7, 2016.

also help tackle the capacity constraints the sector faces, whether it's between utilities and hyperscalers or large colocation providers to help coordinate plans for grid investments and data center capacity expansion, or between chip and server manufacturers and liquid cooling providers to design and scale efficient, easily operable cooling solutions.

— *They will certainly have to invest more too.* Scaling data center infrastructure at an unprecedented pace is capital intensive and will, we estimate, require more than a trillion dollars in investment across the ecosystem. Although investment funds globally are already backing growth in the sector, significantly more growth opportunities exist.

Many companies are already making moves. To name but a few: Blackstone and Digital Realty entered into a $7 billion deal in 2023 to build new AI-ready data centers in Frankfurt, Paris, and Northern Virginia.[11] Super Micro Computer, a US manufacturer of servers, is investing in additional sites in its home market as well as in Asia,[12] while tech company HCL Technologies is collaborating with Schneider Electric to develop solutions for managing energy consumption in data centers in the Asia–Pacific region.[13]

Still, the range of opportunities is not limited to large, established players. Some large companies already have considerable order backlogs. There are OEMs with one- to two-year backlogs for customized electrical switchgear and power distribution equipment, for example. As a result, smaller or newer manufacturers have a real chance to bridge the gap, particularly when investors are willing to help scale their production.

Demand for AI-ready data centers is surging, and with it the potential for a serious supply deficit. The extent to which companies and investors in the value chain are able to speed the build-out of those centers could determine their fortunes. And ultimately, it could determine the pace at which AI is deployed in all our lives.

[11] "Digital Realty and Blackstone announce $7 billion hyperscale data center development joint venture," Blackstone press release, December 7, 2023.
[12] "Supermicro announces expansion of Silicon Valley corporate headquarters and groundbreaking for new 800,000-square foot building in Taiwan," Super Micro Computer press release, April 29, 2019.
[13] "HCLTech and Schneider Electric collaborate to develop sustainability solutions for data centers in APAC," HCLTech press release, July 19, 2023.

Bhargs Srivathsan is a partner in McKinsey's Bay Area office; **Marc Sorel** is a partner in the Boston office, where **Arjita Bhan** is a knowledge expert; **Pankaj Sachdeva** is a senior partner in the Philadelphia office; **Haripreet Batra** is an associate partner in the New York office; **Raman Sharma** is an alumnus of the Toronto office; **Rishi Gupta** is a consultant in the Chicago office; and **Surbhi Choudhary** is a consultant in the Seattle office.

The authors wish to thank Artem Shitov, Dave Sutton, Jesse Noffsinger, Matt Cherry, Nicholas Shaw, Ruchika Dasgupta, Satyam Taneja, Shibashish Chakraborty, Vijaya Mulakalapalli, and Wendy Zhu for their contributions to this article.

This article was edited by Daniel Eisenberg, an executive editor in McKinsey's New York office.

Designed by McKinsey Global Publishing