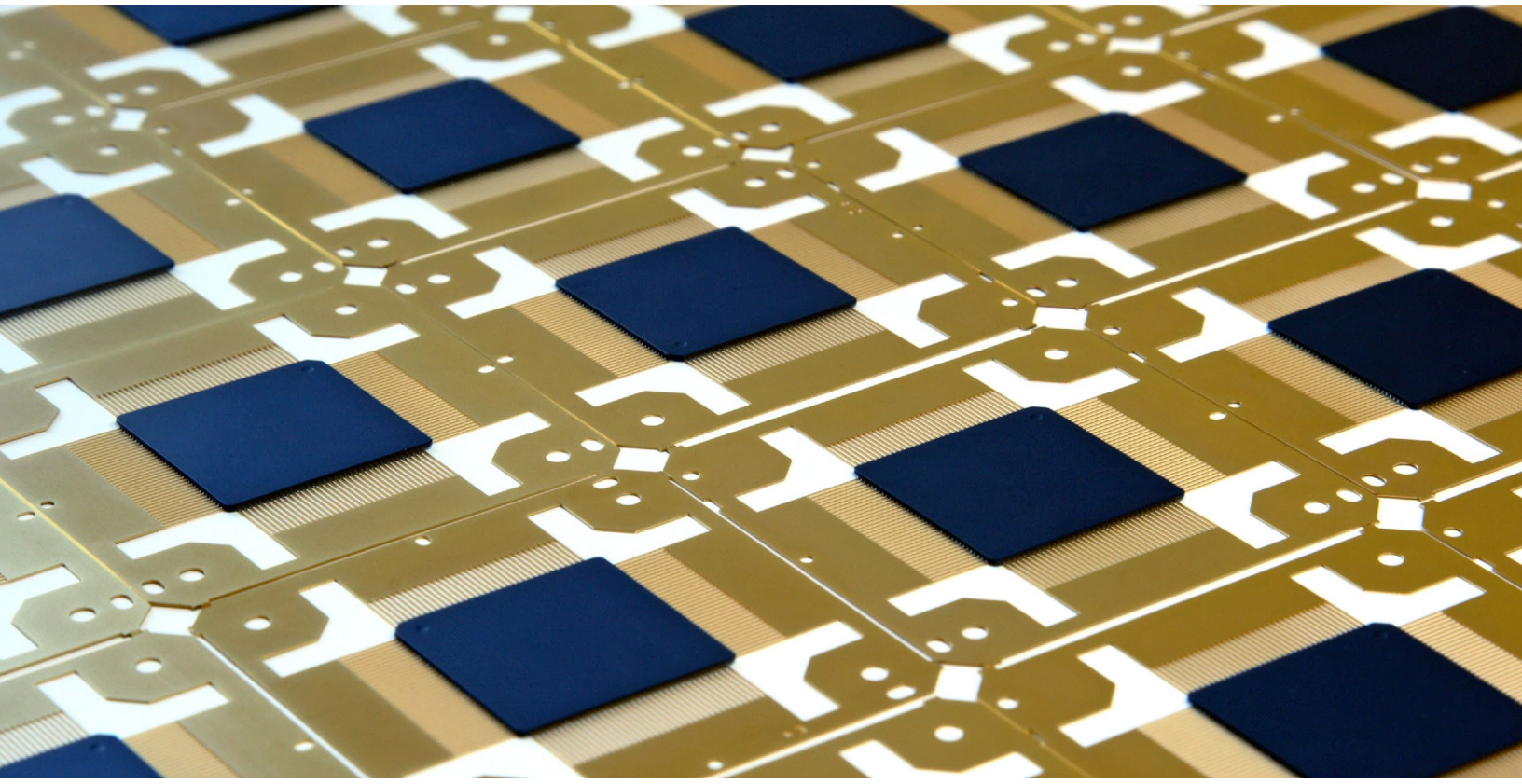
McKinsey & Company

McKinsey Direct

Beyond compute: Infrastructure that powers and cools AI data centers

Power and cooling equipment are the backbones of data center infrastructure. Innovations and on-time supply of this technology will become increasingly relevant as the demand for data centers grows.

This article is a collaborative effort by Ammanuel Zegeye and Pankaj Sachdeva, with Arjita Bhan, Kenza Bouhaj, Rishi Gupta, Vaibhav Chandak, and Wendy Zhu, representing views from McKinsey's Industrials & Electronics Practice.



October 2025

As AI use becomes continuously more widespread, the demand for data centers is surging in tandem. Data center demand is expected to grow at a CAGR of 22 percent and reach 220 gigawatts by 2030—nearly six times larger than demand in 2020 and driven predominantly by AI adoption and hyperscaler spending.[1] Moreover, McKinsey research shows that by 2030, data centers are projected to require $6.7 trillion in cumulative capital outlays worldwide to keep pace with the demand for compute.[2] A large part of that capital spending will be allocated to the underlying systems that deliver electricity to IT equipment and the cooling systems that remove the heat generated by the IT equipment.

As data center designs evolve to keep up with increasing compute requirements, power, cooling, and IT components are no longer seen as separate entities; they must be codesigned and considered in the context of the holistic performance of the data center. More opportunities are opening for stakeholders across the data center value chain to design equipment with the overall architecture in mind, either reducing silos or allowing manufacturers to capture vertical integration opportunities to expand their market footprint. What's more, time to market is now one of the most important buying preferences for data center operators. Therefore, to deliver end-to-end solutions to customers, it's becoming increasingly important for stakeholders to provide services including equipment repair and maintenance, start-up, and power and cooling equipment commissioning.

This article presents the prospective advancements in power and cooling for data centers and the opportunities stakeholders have to advance this area of data center technology ahead of the curve.[3]

The big—and increasing—role of power and cooling equipment

Of the $6.7 trillion in cumulative capital spending through 2030, 60 percent ($4.2 trillion) will go to technology developers and designers, which produce the IT equipment for data centers, such as chips and computing hardware. Among all non-IT equipment, spending is highest on power and cooling equipment, which accounts for 47 to 58 percent of all non-IT equipment spending (see sidebar, "What is power and cooling equipment?").[4] AI-driven facilities push these costs even higher because these facilities require more advanced technology and additional pieces of equipment.

What is power and cooling equipment?

In this article, references to power equipment pertain specifically to distribution and backup systems within the data center, rather than grid or utility infrastructure.

Power equipment is responsible for safely distributing power from the utility (or substation) downstream to the IT equipment inside the data center. It includes distribution equipment (such as switchgear, power distribution units, transformers, and cabling) and backup equipment (such as generators and uninterruptible-power-supply systems). Distribution equipment transmits electricity from the utility source to the compute equipment, and backup equipment ensures that the electricity supply is continuous or with minimal disruption to protect against outages that might affect end users.

Cooling equipment (such as air cooling and liquid cooling) regulates the thermal environment for the IT equipment to ensure the latter operates efficiently, preventing overheating, and extending the lifespan of the hardware.

[1] "The data center balance: How US states can navigate the opportunities and challenges," McKinsey, August 8, 2025; "Scaling bigger, faster, cheaper data centers with smarter designs," McKinsey, August 1, 2025.
[2] "The cost of compute: A $7 trillion race to scale data centers," McKinsey Quarterly, April 28, 2025.
[3] This article focuses on the equipment that is used in data centers to safely receive, condition, and distribute electricity to IT equipment, which does not include broader power systems, such as the utility grid or on-site power generation.
[4] The share of power distribution and cooling spending is ~15 percent of all equipment, including IT.

Of the total non-IT capital expenditure, about 35 to 40 percent is spent on power equipment, and about 12 to 18 percent is spent on cooling equipment. Of the share spent on power equipment, about 45 percent is spent on power backup equipment, and 55 percent is spent on power distribution equipment (Exhibit 1).

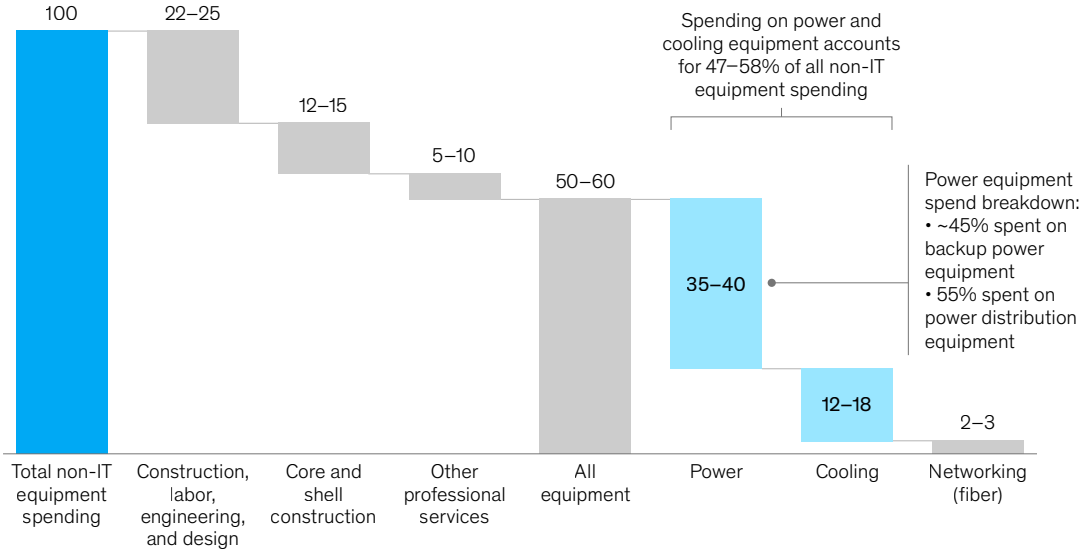
Meeting future data center demand will require significant changes in equipment areas to scale power backup capacity, redesign distribution systems for higher rack densities, and deploy technologies for cooling, such as liquid cooling. As such, these types of equipment are some of the largest areas of investment for data center development. Of the $6.7 trillion in cumulative capital spending through 2030, $5.2 trillion will be spend on AI workloads. Of the $5.2 trillion in capital expenditures needed to energize AI workloads in the next five years, about 14 percent (about $720 billion) of cumulative spending will be allocated to power and cooling equipment. While cumulative spending on power equipment will reach $500 billion, spending on cooling will reach $220 billion (Exhibit 2).

This scale of spending and the equipment's pivotal role in the success of data center infrastructure builds creates significant opportunities for market players. Equipment suppliers that can deliver higher efficiency systems will be well-positioned to meet surging demand. Equipment manufacturers that can design, configure, and deploy these systems quickly and at scale will play a critical role in meeting aggressive build timelines. And data center operators that can integrate advanced power and cooling into their facilities will reduce their time to market, improve efficiency, and reduce operating costs. More efficient power equipment also minimizes energy losses between distribution layers, allowing scarce grid capacity to be converted into additional compute output and increasing a site's overall value.

Exhibit 1

On-site equipment that manages and transports energy accounts for nearly half of all non-IT data center capital expenditures.

Capital expenditure breakdown of non-IT spending for data centers, %



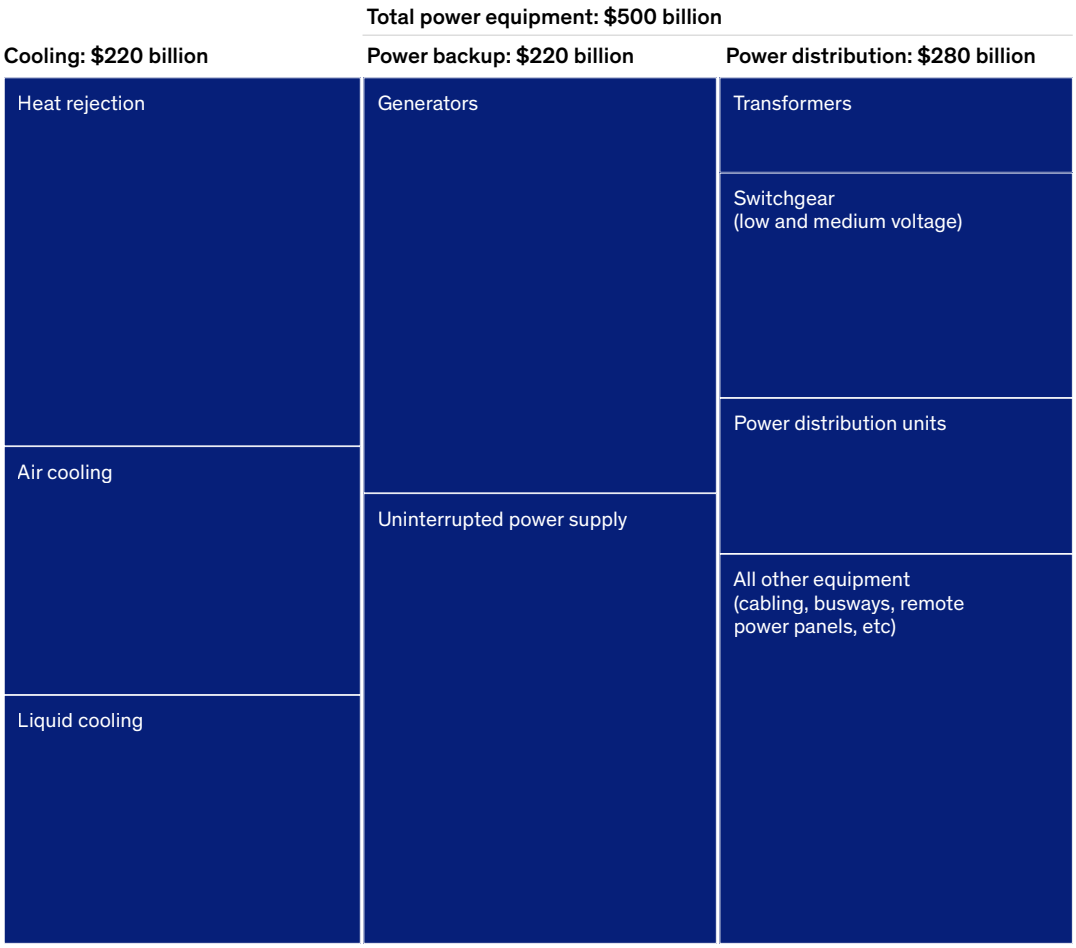
Note: Figures may not sum to 100%, because of rounding.

McKinsey & Company

Exhibit 2

Cumulative spending on power backup, power distribution, and cooling equipment is expected to reach $720 billion by 2030.

Cumulative spending on on-site power and cooling equipment, 2025–30



McKinsey & Company

The evolution of electrical and cooling design

Data center facilities consume massive amounts of power. In 2030, about 11.7 percent of total US power consumption will come from data centers, up from less than 1 percent in 2020.[5] Some facilities are already being designed to reach gigawatt-scale to enable the future compute equipment needed to run AI workloads. Data center designs are evolving in response to technological advances, including GPUs that can process larger amounts of power. As the demand for AI increases, GPUs are used more and more, requiring massive amounts of power, increasing rack densities, and inducing changes in power and cooling architecture.

Within power equipment, power distribution equipment is responsible for safely directing and distributing the electricity to the IT equipment. This equipment consists of medium-voltage switchgear, low-voltage switchgear, transformers, busways, and power distribution units (PDUs). Backup equipment is responsible for ensuring a continuous supply of electricity during utility outages and consists of

[5] "Scaling bigger, faster, cheaper data centers with smarter designs," McKinsey, August 1, 2025.

generators and uninterrupted power supply (UPS). While generators provide long-duration backup electricity, they start after a short duration, which exposes the IT equipment to disruptions. To bridge the gap, the UPS provides instantaneous backup electricity until the generators start (Exhibit 3).

The power topology of a data center is shaped by four factors that influence both equipment demand and placement:

— Rack density, measured in kilowatt (kW) per rack, defines how much compute is concentrated in a physical footprint, affecting per-data hall and per-rack power requirements.

— Power redundancy measures system resilience, with higher redundancy enabling greater uptime and minimizing downtime during maintenance or failures.

— Latency reflects how quickly workloads must respond to inputs, with low latency designs essential for real-time inference in use cases such as autonomous vehicles.

— Scalability measures how easily a facility can expand compute, power, cooling, and space to meet future demand.

Together, these factors dictate the architecture, layout, and investment priorities in data center electrical systems. As AI becomes the dominant workload in data centers, rack densities will increase, power systems will require greater redundancy, latency will need to be minimized, and scalability will become essential. Such advances are driven by the innovative designs of leading chipmakers and hyperscalers' rapid adoption of these innovations. Together, these stakeholders are establishing the standards that shape the broader market, not only for chips but also for the supporting infrastructure.

How electrical equipment will change over time

By 2030, AI inference workloads—the process through which a trained model is applied to respond to new data—are expected to grow faster than AI training workloads (the process to train new models). Today, mainstream IT rack deployments are shifting from a range of under 30 to 50 kW per rack to a range of 70 to 100 kW per rack because denser racks enable more concentrated and faster compute. While inference workloads generally require 20 to 40 kW per rack, training workloads require up to 100 kW per rack due to the intensity of compute required to train models. These denser racks require more robust power equipment, such as higher-capacity switchgear, PDUs, and busways to safely manage the added electrical capacity. Air cooling is no longer sufficient at these densities, requiring a shift to liquid cooling.

For power backup equipment, the demand for backup generators and UPS systems will increase to meet the latency and redundancy requirements of AI use. However, multiple companies are reviewing their use of diesel-based generators to meet sustainability requirements.[6] GPU-heavy servers, alternatively, are more sensitive to electricity sags or interruptions, further increasing the importance of robust and reliable backup electricity.

Power distribution equipment could change in more ways, which affects investments in different types of equipment. In higher-density data centers, spending on medium-voltage switchgear equipment will increase to ensure lower power losses as power is delivered to denser racks, whereas spending on low-voltage switchgear equipment will grow at a slower rate because power is delivered to fewer, more concentrated high-density endpoints within a data center. Moreover, bus bars, which are more modular and embedded in distribution architectures, are emerging as competitive alternatives to rack PDUs: They can deliver higher currents efficiently, support flexible rack layouts, and simplify scalability as densities increase, making them more attractive options.

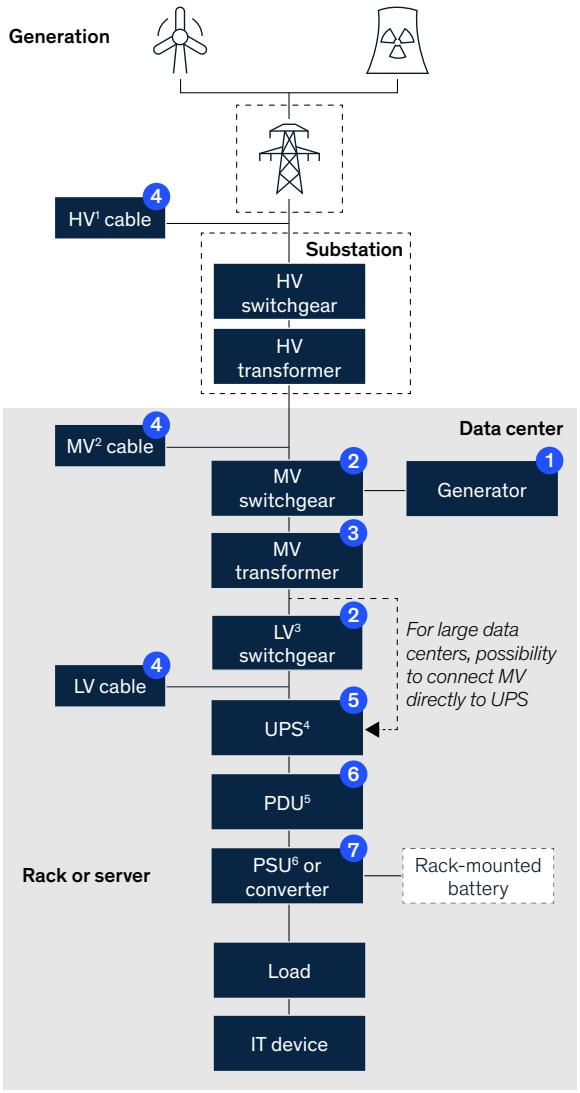
As densities push above 100 kW, the next major shift in data center topology is the adoption of power shelves and direct AC-to-DC architectures. At these levels, traditional power distribution with multiple conversion stages between different voltage levels becomes increasingly inefficient, takes up considerable space, and is difficult to manage. Alternatively, power shelves,

[6] For example, see *Serverless*, "Progress on our commitment to sustainable backup power in datacenters by 2030," blog entry by Noelle Walsh-Elwell, Microsoft Azure, June 1, 2021.

Exhibit 3

In data centers, switchgear equipment ensures that medium-voltage utility feeds deliver a reliable source of power to low-voltage racks.

Simplified illustration of key electrical components within, or adjacent to, a data center, nonexhaustive



Description of component

1. **Backup generator**
   - Supplements power supply, allowing nonstop power, providing redundancy, and reducing risk
2. **Switchgear or switchboard**
   - Electrical equipment that receives electricity from main grid and transfers the electricity to transformers
   - Controls, protects, and isolates electrical equipment
   - Can be an intermediate point among UPS, generators, and the rest of IT infrastructure
3. **Transformer**
   - Steps down incoming voltage
4. **Cables (busways)**
   - Different mediums for wire components, depending on voltage level (ie, copper vs fiber optic)
   - Can also be a power busway
5. **UPS**
   - Provides backup power to critical equipment when utility power fails long enough for backup generators to start (in the event of a utility failure) or for critical equipment to shut down safely so no data is lost
6. **PDUs**
   - Distribute, control, and monitor the critical power from the upstream UPS system to IT racks
   - Usually contain a main input circuit breaker, branch circuit panelboards, a power transformer, output power cables, a surge arrester, and the monitoring and communication modules
7. **PSU or converter**
   - Converts AC power from the electrical grid into DC power that can be used by the IT equipment in the server
   - Assists in improving the efficiency and reliability of the power supply for the servers

[1]High-voltage.
[2]Medium-voltage.
[3]Low-voltage.
[4]Uninterrupted power supply.
[5]Power distribution unit.
[6]Power supply unit.

McKinsey & Company

typically integrated within the rack, bring centralized DC conversion closer to the IT load, reducing conversion steps and improving efficiency. Direct AC-to-DC designs go further by eliminating intermediate conversions, feeding high-voltage DC directly to rack-level power supply units that deliver regulated DC power to servers. This approach reduces electrical losses, improves responsiveness to rapid load changes, and enables modular, high-density deployments. This shift is being accelerated by NVIDIA's 800V DC design, which is early in its development but expected to gain traction in the next few years.[7]

Cooling increasingly dense racks with new technologies

Cooling equipment is essential to prevent compute equipment from overheating and data center performance from degrading. As data centers become more packed with high-density compute chips to handle AI workloads, traditional air-cooling systems (such as computer room air conditioners that supply air directly or through raised floors) struggle to remove heat efficiently at densities above 50 kW per rack.[8] Consequently, liquid cooling has become a necessity, not an option. This solution offers far more efficient heat removal, which can support dense, high-performance racks while reducing energy consumption and the facility's footprint.

As hyperscalers and enterprises scale AI infrastructure, the growth rate of spending on liquid cooling and associated mechanical systems is rising rapidly and is projected to outpace the overall spending on data centers over the next five years. McKinsey analysis shows that between 2025 and 2030, spending on air-cooling equipment is estimated to climb from between $6 billion and $8 billion to between $11 billion and $13 billion, growing 10 to 15 percent annually. Spending on the liquid-cooling market is expected to grow 45 to 50 percent annually, growing from between an estimated $2 billion and $3 billion in 2025 to between $15 billion and $17 billion in 2030.

Three types of liquid-cooling solutions will become more popular as investments in liquid cooling increase: rear-door heat exchangers (RDHx), direct-to-chip (DTC), and immersion cooling.[9] These new cooling methods are reshaping data center mechanical design, requiring integrated piping, coolant distribution units (CDUs), and thermal monitoring, and opening new markets for cooling component and system integration innovation.

In addition to cooling systems that are used to cool the IT equipment, heat rejection systems are also required to remove the extracted heat from the data hall and discharge it into the external environment through chillers, cooling towers, or dry coolers. This market's estimated value is between $3 billion and $5 billion in 2025, growing to between $12 billion and $14 billion in 2030 at 30 to 35 percent CAGR by 2030, according to McKinsey analysis.

The opportunity in DTC liquid cooling

DTC liquid cooling may be the most attractive liquid-cooling solution in the near term, offering investment opportunities across various components. DTC liquid cooling delivers high thermal efficiency by placing cold plates directly on heat-generating components such as GPUs, which enables precise, localized heat extraction. It supports higher rack densities (often upward of 70 kW[10] and reaching up to 250 kW per rack[11]) without requiring full immersion, which often requires structural changes, making it ideal for AI and high-performance computing deployments within conventional data center footprints.

DTC systems are also modular and scalable, allowing data centers to incrementally adopt liquid cooling (for example, node by node or rack by rack) and integrate with existing air-cooled infrastructure or RDHx systems. Overall, DTC strikes the best balance between performance, serviceability, and deployment flexibility, which is why it is being prioritized by hyperscalers, OEMs, and co-location providers for next-generation infrastructure. McKinsey analysis shows that by 2030, DTC liquid cooling will account for 30 percent of the cooling market.

[7] Mathias Blake et al., "NVIDIA 800 VDC architecture will power the next generation of AI factories," NVIDIA Developer, May 20, 2025.
[8] "AI power: Expanding data center capacity to meet growing demand," McKinsey, October 29, 2024.
[9] "AI power: Expanding data center capacity to meet growing demand," McKinsey, October 29, 2024.
[10] Shaun Lee, "Liquid cooling enters the mainstream in data centers," Jones Lang LaSalle, accessed September 29, 2025.
[11] This figure does not represent a theoretical limit.

As data center design evolves, mechanical, electrical, and IT systems may be increasingly codesigned to support extreme-density AI workloads at scale.

Five components are most relevant to DTC solutions and should be recognized as growth areas by stakeholders:

- *Cold plates* are metal blocks (typically copper) mounted directly on CPUs, GPUs, and memory to absorb heat and transfer it to circulating coolant.
- *CDUs* are central pump and heat exchange systems that circulate coolant through the loop and interfaces with facility water.
- *Coolant distribution manifolds* are rack-mounted distribution pipes that route coolant to and from each cold plate in the server.
- *Piping and quick connects* are flexible hoses and leak-proof connectors used to link servers, manifolds, and CDUs within the rack.
- *Sensors and controls* are thermal, pressure, and flow monitors for tracking loop health, integrated with data center management systems.

As rack densities continue to rise, single-phase cooling may see limitations with coolant flow management, although at a much smaller scale. As a result, two-phase technologies, including DTC and two-phase immersion cooling, are gaining traction. As data center design evolves, mechanical, electrical, and IT systems may be increasingly codesigned to support extreme-density AI workloads at scale, and innovations including direct-to-facility loop integration, cooling-integrated racks, and liquid-native architectures may be used more.

The convergence between power and cooling

Data center suppliers have traditionally focused on distinct categories along the data center infrastructure value chain, such as power (including switchgear and UPS) and cooling (including air cooling and facility chillers). However, recent M&A activity and new-product launches suggest that the industry is blurring the line between these categories as more suppliers deliver integrated solutions instead of stand-alone components.

The convergence between power and cooling equipment is driven by data center market growth; customers increasingly preferring fast deployments, which has created a market pull for more integrated solutions; and AI workloads that benefit from IT, power, and cooling components existing inside the rack as they push higher rack densities (Exhibit 4).

The implications for stakeholders along the value chain

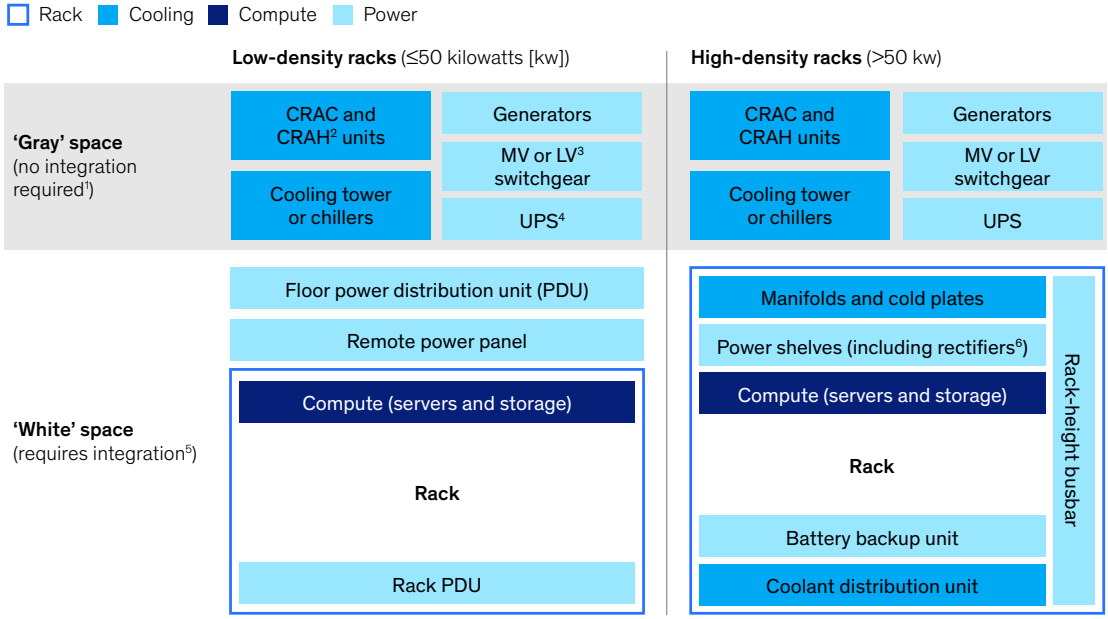
In this moment of advancement, stakeholders along the data center value chain can assess the areas in which they can have the biggest impact, innovate diligently to keep up with the evolving nature of this market, and consider the potential trade-offs each of each move.

Investors. High-density AI workloads are driving a massive market opportunity for investors, with power and cooling equipment representing a disproportionate share of non-IT spending. Companies that can innovate to stay ahead of chip technology, codesign their individual pieces

Exhibit 4

With increasing rack densities, more and more components live within the rack, requiring greater integration and codesign of white space equipment.

Illustration of components within 'gray' or 'white' space



Note: <50-kw racks could also contain busbars or battery backup units.
[1]Components remain separate and do not need to be co-engineered and bundled together.
[2]Computer room air conditioner and computer room air handler.
[3]Medium-voltage or low-voltage.
[4]Uninterrupted power supply.
[5]Because the rack becomes not just a mechanical shelf but a power distribution and cooling component, many electrical and cooling components are shifted inside the rack.
[6]Power conversion device that transforms AC input power into regulated DC output power.

McKinsey & Company

to fit into the overall data center architecture, and offer fast deployments that can fit into a variety of contexts will have more-favorable returns. Moreover, consolidation opportunities are emerging as more suppliers seek to offer integrated solutions. Investors that back companies with both technical readiness and robust supply chains will be well positioned to benefit from this market shift.

Traditional power OEMs. The rise of AI workloads is accelerating the need for equipment that can be codesigned with advances in chip technology and the associated cooling requirements. In response, many traditional power OEMs are expanding beyond their primary product focus into adjacent lanes, such as prefabricated, modular power and cooling assemblies. While this diversification marks an important first step, long-term success will depend on OEMs' ability to deepen these capabilities through in-house innovation, strategic partnerships, or targeted acquisitions. OEMs that effectively align product development with changing power and cooling needs while sustaining technical leadership in their core offerings will be best positioned to capture growth in this evolving market.

Cooling OEMs. AI workloads are accelerating the shift from conventional air systems to advanced liquid cooling. Cooling OEMs, which tend to be smaller and more specialized, will need to scale and innovate quickly to meet integration requirements. However, these stakeholders must also be aware of the increasing competition from larger incumbents moving into liquid cooling and from new entrants with niche technologies

such as immersion cooling. This dynamic creates significant M&A opportunities as smaller OEMs look for capital, manufacturing scale, and broader market reach, while larger players seek to acquire specialized expertise to deliver thermal solutions.

Rack original design manufacturers (ODMs) and electronic manufacturing services (EMS) providers. The increase in AI workloads is creating a new opportunity for rack ODMs and EMS providers to move closer to end customers by delivering fully integrated rack-level systems that combine compute, liquid cooling, and power distribution. Capturing this opportunity would allow these companies to have a hand in a greater portion of the data center infrastructure value chain and achieve higher margins. However, realizing this opportunity requires capabilities many ODMs and EMS providers currently lack, including advanced power and cooling integration, enterprise-grade service, and life cycle support. Moreover, moving upstream could strain relationships with existing OEM and integrator customers that depend on ODMs and EMS providers as suppliers. As a result, these companies face a strategic trade-off: Expanding into integrated rack solutions can lead to higher margins and market control, but companies would risk overstretching their capabilities and creating channel conflict.

Server OEMs. For server OEMs, the shift to high-density AI infrastructure may shift key suppliers into competitors if they offer fully integrated racks directly to end customers. At the same time, in addition to buying from OEMs, hyperscalers and other large buyers are also procuring racks directly from component OEMs or ODMs. To stay competitive, server OEMs may have to own distinctive, innovative power and cooling capabilities or build a partnership to have these offerings, enabling them to deliver complete solutions. However, expanding into these areas may force OEMs to either protect and grow their market share or erode high margins tied to their branded server offerings.

Service providers. Service providers will be expected not only to maintain equipment but also to take on a broader role in optimizing system performance, managing life cycle operations, and ensuring seamless integration across power, cooling, and IT infrastructure. This orchestration will require a cross-disciplinary skill set that spans power and cooling, from managing voltage distribution and liquid-cooling loops to conducting advanced thermal diagnostics. As equipment OEMs expand further into life cycle services, independent service providers can differentiate themselves by supporting multivendor environments and offering consultative engineering expertise, particularly for system retrofits and performance upgrades.

Chipmakers. The growth of AI data centers brings continued opportunity for chipmakers to continue leading in innovations and shape rack-level designs by defining power and cooling requirements for their systems and certifying suppliers to those specifications. This influence allows them to drive performance optimization and ecosystem alignment, but it also carries the risk of vendor lock-in or deployment delays if the supply chain cannot deliver to spec at scale. A key strategic choice lies in whether to push prescriptive infrastructure, such as standardized racks for flagship chips (such as the Blackwell GPU used in the GB200 system), or remain agnostic to encourage broad adoption across diverse data center environments.

Customers. Customers, including hyperscalers, face a choice between building best-of-breed solutions or buying end-to-end products. A best-of-breed approach allows them to select the strongest components in each category but adds complexity to procurement and integration. By contrast, end-to-end solutions simplify the buying process and consolidate vendor relationships. They can also deliver stronger economies of scale in pricing, volume, vendor commitment, and solution optimization.

A worldwide supply chain

Currently, most end customers for data centers, including hyperscalers and leading co-location centers, are based in the United States, with increasing interest in sovereign AI infrastructure present in other regions. Yet the innovation needed to sustain the ongoing development of AI markets is global, and the supply chain is spread across geographies.

For example, innovations in server, rack, power, and cooling designs largely originate in the Taiwan market, where contract manufacturers and ODMs are deeply embedded. Moreover, power and cooling component OEMs are present across Asia, Europe, and North America, and each brings unique strengths in innovative designs and manufacturing capabilities.

As rack densities continue to evolve, no single geography owns the end-to-end solution; international collaboration across these ecosystems is essential to integrate individual pieces into cohesive systems.

The ever-growing demand of AI is pushing the need to data centers to evolve, with power and cooling equipment presenting a large opportunity for investment. By 2030, increasing rack densities, more-efficient cooling systems, and scaled backup capacity will catalyze stakeholders along the value chain to widen their scope—and starting now will help them keep a foothold in the market.

Ammanuel Zegeye is a partner in the Bay Area office, of which **Wendy Zhu** is an alumna; **Pankaj Sachdeva** is a senior partner in the Philadelphia office; **Arjita Bhan** is a senior expert in the Boston office; **Kenza Bouhaj** is a consultant in the New York office; and **Rishi Gupta** and **Vaibhav Chandak** are consultants in the Chicago office.

The authors wish to thank Grace Guan, Graham Healy-Day, Jesse Noffsinger, Jules Kentgens, Julien Deschamps, Kevin Lin, Kevin Sachs, Marco Contini, Maria Goodpaster, Nicholas Shaw, Pranjali Kumar, Puneet Puri, Riya Garg, Rocío Marazuela, Shih-Yung Huang, and Shraddha Kumar for their contributions to this article.