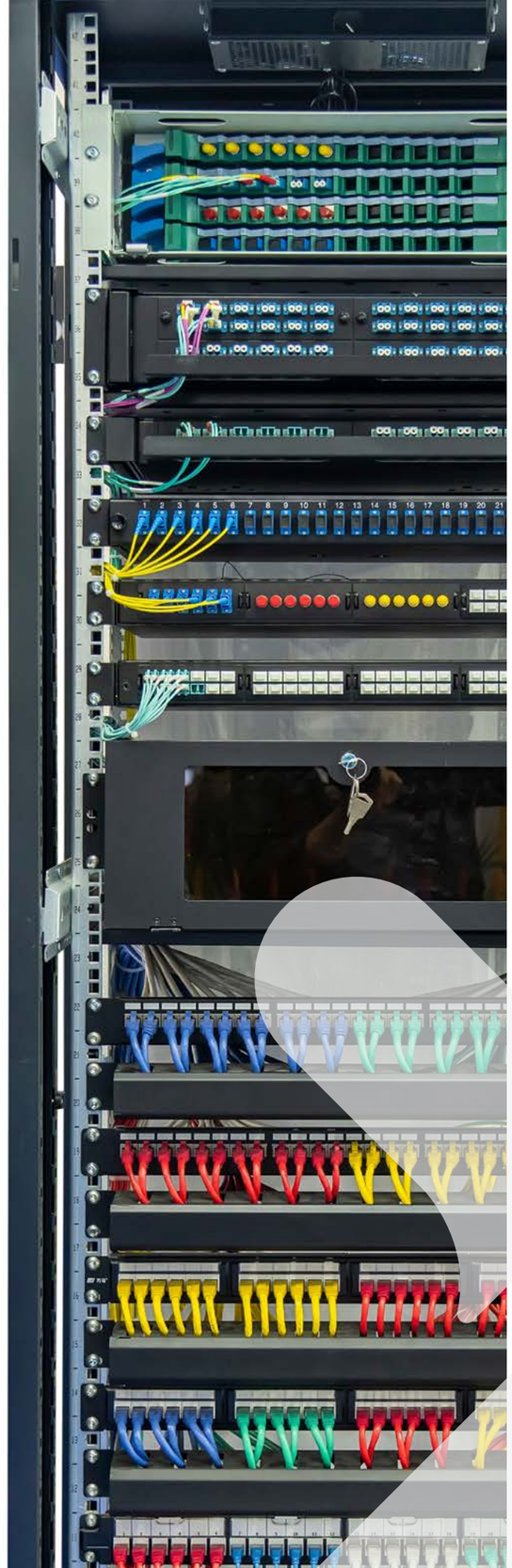
EBOOK

Overcoming power, heat, and scale hurdles

Innovations shaping data center expansion

flex

Table of contents



Executive summary

Artificial intelligence (AI) is set to dominate IT priorities during the second half of the decade, creating $15.7 trillion in economic opportunity by 2030.[1] The red-hot demand for AI-based compute capabilities is compelling companies to rethink data center infrastructure as they grapple with operational complexity, technical advancements, and aggressive build timelines. As a result, worldwide data center capex is projected to surpass $1 billion by 2029.[2] Billions more will be invested as incremental government spend, such as the $500 billion Stargate project backed by the U.S.

The infrastructure required to accommodate escalating compute requirements from AI is also motivating companies to address power, heat, and scale in innovative ways. Enterprise cloud service providers (CSPs) are leading generative AI large language model (LLM) development and exploring alternatives to traditional data center power sources, rack and room configurations, and cooling technologies. They're also refining their approach to global manufacturing, supply chain and inventory management, and data center lifecycle services to ease constraints on data center expansion. Integrated solutions are of particular interest, as are partner engagements that help them manage technology, operational, and regulatory complexity.

Flex, with its unique grid-to-chip capabilities, global manufacturing expertise, well-established supplier network, and comprehensive services, is poised to help design, build, and service data center infrastructure worldwide to meet the demands of the AI era. The company's focus on innovation and end-to-end view of data center requirements position Flex to deliver exceptional power and cooling solutions that help data center operators overcome the challenges inherent in today's fast-growth, high-density compute environments.

AI workloads are fueling data center innovation

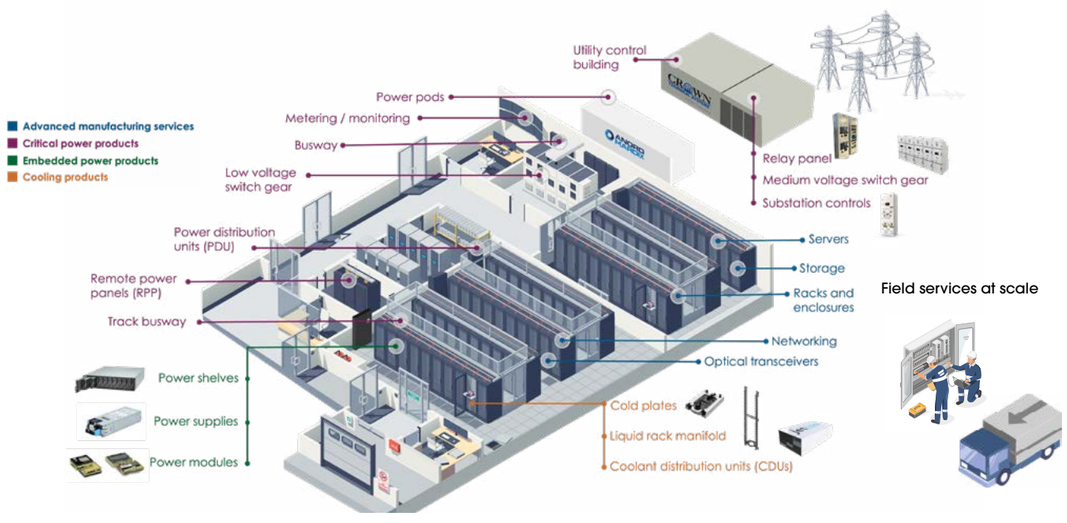
OpenAI shocked the world with the delivery of ChatGPT in 2022. Since then, generative AI models have advanced rapidly, as has their integration into mainstream applications such as copilots and chat assistants. Researchers are now making strides with agentic AI, which can make decisions, solve problems, and work with limited human supervision, well beyond the "learn and mimic" capabilities of generative AI. They're also exploring use cases for ambient agents that act without explicit prompts. With AI technology evolving quickly and adoption across industries following suit, hyperscalers are investing in AI-ready data centers at breakneck speed to fuel model training and refinement, both of which require massive compute capabilities. By 2030, nearly 65 percent of AI workloads in the U.S. and Europe will be hosted by hyperscalers, with private hosting by technology companies and smaller enterprises responsible for the rest.[3]

19% - 27% increase in global demand for data center capacity from 2023 to 2030, driven primarily by the thirst for advanced AI workloads[3]

The vector processing-based compute parallelism required of high-performance AI workloads relies on accelerated platforms that primarily use graphics processing units (GPUs) to perform complex calculations simultaneously for faster model training and inference. It's redefining every aspect of compute, from memory technology to data storage hierarchies to data center network topologies. High bandwidth memory (HBM) devices are supporting low-latency data access, a key requirement for training AI clusters. Innovative solid-state drives (SSDs) are being deployed in lieu of traditional hard disk drives (HDDs) to keep pace with the capacity hyperscalers demand, including the world's first 100+TB SSD. Standards-based ultra ethernet and ultra accelerator networking links are gaining credence to mitigate supply chain risk. Optical networking may soon take hold as hyperscalers seek to accelerate node-to-node data movement within large compute clusters.

Collectively, the technology advances required of the AI era are driving significant changes in operational technology, especially in data center power and cooling infrastructure.

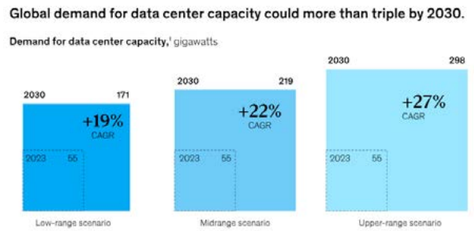
> "We've set ourselves up to scale computing and develop software at a level that nobody's ever imagined before. Over the next 10 years, our hope is that we could double or triple performance every year at scale—not at chip, at scale… we're going to be on some kind of hyper-Moore's law curve.[4]
>
> - Jensen Huang, CEO, NVIDIA



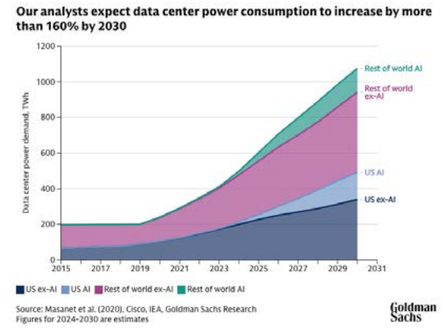
Power acquisition is critical to data center growth

The buildout of power-hungry accelerated compute platforms is placing intense demand on data center power consumption, which could reach 1,070 TWh globally by the end of the decade, more than triple the requirements of 2020.[6] Hyperscalers are investing in the construction of new data centers to fuel more capacity, but bringing data centers architected for accelerated computing online can take years. In the interim, existing data centers are under incredible strain regarding in-rack and critical power infrastructure. For instance, compute density is driving rack power requirements to historic thresholds of 100kW and beyond.

Data center operators need access to more power sources and once secured, the ability to distribute power effectively from the grid, through the facility and rack, and ultimately to the chip. Renewable sources such as nuclear power generation are gaining interest[5], and upgrades to power infrastructure, migration to DC power delivery, and other techniques are being used to improve efficiency from grid to chip.



Source: McKinsey & Co.[3]

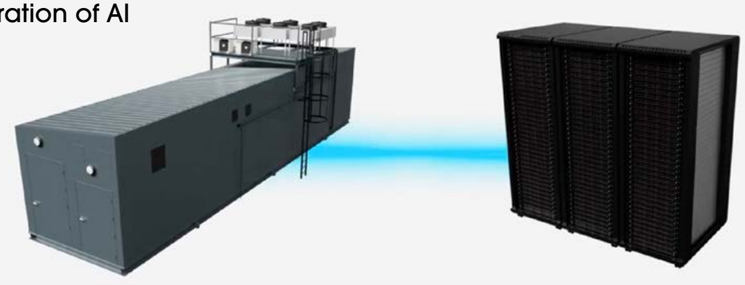


Source: Goldman Sachs[6]

From grid to rack

The challenges start at the facility level, and that begins with delivery of power from the grid. Long-term power purchase agreements are used to secure access from renewable sources to deliver incremental power to existing facilities or establish a reliable supply for new data centers—but that doesn't mean that the critical power infrastructure is in place to take advantage of it. Data center operators are revamping power delivery to racks to support AI workloads such as those that will be required of the NVIDIA GB200 NVL72 exascale computer. Acting as a single, massive GPU and requiring an estimated 120kW of power per rack, its widespread commercial deployment is on the horizon.[7] Data center operators are also partnering with experts in critical power to deliver next-generation power distribution units (PDUs) that can support 500kW (with roadmaps to 1GW) to meet forecasted requirements.

Advanced grid-to-rack solutions power the next generation of AI innovation



In-rack power solutions

Once facility power distribution is in place, operators must focus on in-rack requirements. Flex has designed custom power shelves to deliver up to 125kW for growing AI requirements, driving technology innovation to support evolving compute infrastructure demands. Furthermore, collaborations between Flex and hyperscalers has led to innovations such as the capacitive energy storage system that address the 1MW to 2MW power surges that GPUs generate every few seconds during AI training and inference calculations. Solving this challenge has been critical, because power fluctuations can prevent data centers from connecting to the grid based on utility permit requirements.

Chip-level power management

Chip-level power management is essential for accelerated compute platform deployment. Vendors and data center operators regularly partner with embedded power product companies early in the design process, often years before the official launch of products into market. Many of these companies work with Flex to develop power modules that increase efficient power distribution to GPU, central processing unit (CPU), field programmable gate array (FPGA), and custom application-specific integrated circuit (ASIC) designs. Accelerated computing is driving innovation in this space as well, with some of the latest power module designs for intermediate bus converters delivering continuous power of 750W and up to 1,500W peak power with over 98 percent peak efficiency.

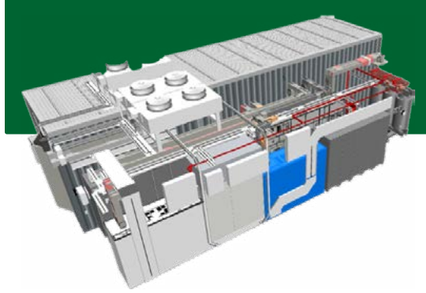
Data centers are feeling the heat

Accelerated computing not only comes with an increase in power consumption, it creates a huge challenge for rack-level cooling solutions as well. For example, a 100kW rack generates over 340,000 BTU/hour, equivalent to 34 standard home furnaces working at capacity. While this is a relatively small use case today—the average rack power requirements are still in the 12kW range—hyperscalers are driving the need for innovative cooling technologies to support dense AI server clusters.

Liquid cooling has been used for decades in the automotive and aerospace industries. Its deployment in data centers has been historically focused on high-performance compute clusters used by governments and academic institutions. As the need for cooling solutions that surpass the capabilities of traditional air cooling accelerates, it is driving innovation in liquid cooling solutions.

In-rack power solutions

Chip-level power management



Direct-to-chip liquid cooling

Air cooling, the traditional preference for cooling data center infrastructure, is effective at up to approximately 50kW per rack. But AI and high-performance computing are different, with rising rack power densities signaling the need to transition to liquid cooling. As hyperscalers deploy more GPU clusters, there is a growing recognition that liquid-cooled racks and coolant distribution units (CDUs) are best suited to remove heat in high-density environments.[8]

CDUs are specialized devices within a closed-loop liquid cooling system that precisely manages coolant temperature and flow rates, ensuring optimal cooling efficiency. By controlling the flow of coolant to IT equipment and returning it to the facility's water for recooling, CDUs stabilize temperatures and reduce the risk of overheating. Where facility water is available, a liquid-to-liquid CDU can transfer heat from the IT equipment's coolant loop to the facility's water loop for recooling, further stabilizing temperatures, reducing overheating risk, and isolating equipment from liquids.

JetCool SmartSense CDU

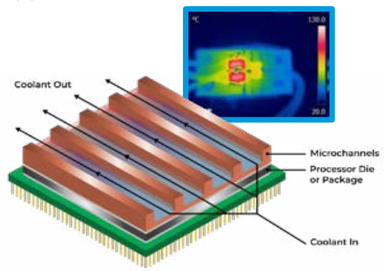
The JetCool SmartSense CDU is a high-performance, rack-mounted liquid-to-liquid 6U CDU built to handle the intense thermal loads of GPU-dense, high-power racks. With a cooling capacity of up to 300kW per rack, or neighboring racks, and scalability to row-based configurations delivering 2MW+ cooling capacity, it ensures efficient, reliable cooling for data centers. Paired with JetCool's advanced cold plates, the SmartSense CDU provides a complete, end-to-end solution for even the most compute-intensive environments. The SmartSense CDU, combined with JetCool SmartPlate cold plates, provides an industry-leading cooling solution for high power processors, including those over 1,500W.

JetCool microconvective cooling® technology

Microconvective technology has proven effective for high power density application due to its efficient heat transfer that directly targets hotspots with an array of fluid jets. This approach delivers up to 40 percent lower thermal resistance over microchannel cold plate technology and can be used in direct-to-die and direct-to-package configurations without changes to chip assembly. JetCool has patented its microconvective cooling technology tailored for the demanding requirements of today's data centers, high-performance computing, and AI applications. Flex works directly with chipmakers to customize cold plates for their respective processor families. The company recently launched an Open Compute Project (OCP)-compliant line of liquid cooling-ready server designs for hyperscalers seeking customized accelerated computing platforms.

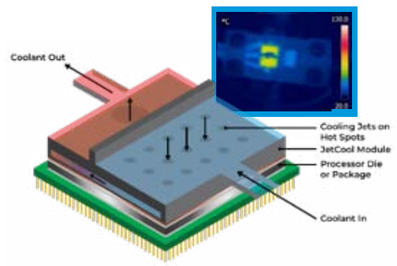
Microchannel Cooling

Microchannel liquid cooling relies on parallel, uniform heat spreading via small internal fluid channels.



JetCool's Cooling

Microconvective liquid cooling uses targeted, perpendicular jets to directly cool processor hot spots, **driving better efficiency and performance .**



Integrated rack designs solve for power, heat, and scale

With the increasing complexity and scale of computing in the AI era, integrated system rack designs that bring together compute, storage, network, power, and cooling technologies are imperative. Balancing rack standardization and customization, OCP introduced the Open Rack v3 (ORv3) specification to drive efficiency into the design and delivery of integrated rack solutions that meet several hyperscaler design points.[9] Innovative high-density racks deliver breakthroughs in compute performance per square foot, increasing data center capacity return on investment (ROI) while supporting scalability.

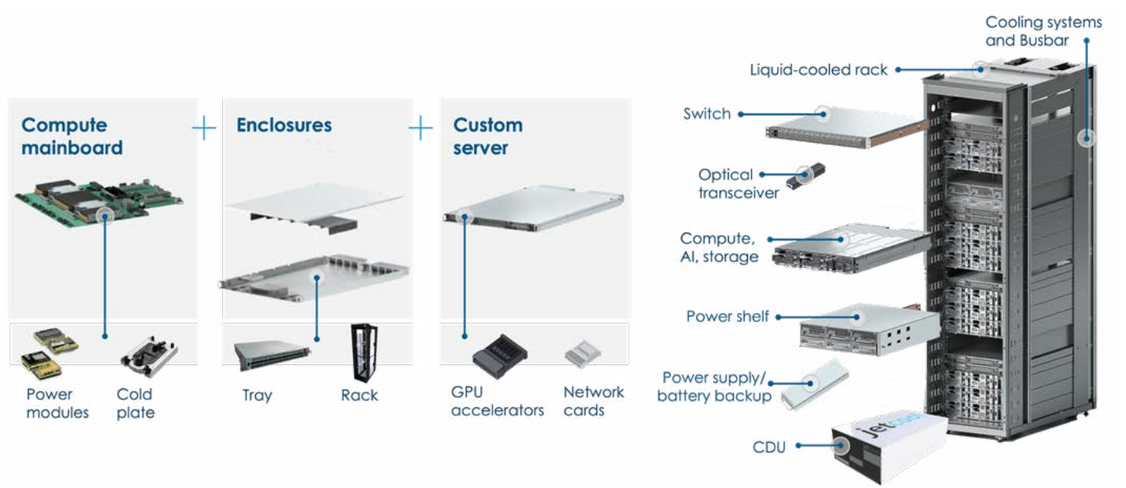
In partnership with Flex, leading hyperscalers are deploying customized ORv3-based rack designs at scale. They are also tapping Flex to manage vertical integration of data center rack solutions, from the fabrication of sheet metal frames and enclosures to the design and manufacture of servers, storage, racks, cabling, switches, busbars, power shelves, battery backup, and liquid cooling systems. For example, Flex's ORv3-compatible rack is currently integrated with single-phase direct-to-chip liquid cooling and can support two-phase liquid cooling, which provides options for customers.

But just as compute supply shortages have constrained data center growth, lack of availability of critical power equipment such as transformers, switches, and generators has caused multiyear delays in power delivery, according to a recent CBRE report.[10] Hyperscalers also note that an inability to scale is impacting broader deployment of emerging liquid cooling solutions.

As compute, power, and cooling architectures evolve, data center operators will need to rely on companies that can collaborate and scale to support roadmaps and ensure that data center upgrades and capacity come online at targeted timelines—companies with design, product, manufacturing, supply chain, systems integration, and open ecosystem experience. Reliable delivery of innovative power and cooling solutions is required for rapid deployment of existing and new data centers. Flex's ability to manufacture products in North America, Europe, and Asia increases capacity, shortens lead times, and enables rapid deployment of innovative data center power and cooling solutions at scale worldwide.

Key features
21-inch Open Rack v3

- ORv3 design customized to your requirements, including 19-inch footprints
- 29% more front space available
- Optimized front-to-back airflow



Maximizing ROI at scale with data center services

Data center operators have their hands full transitioning to AI-era capabilities. A single supply chain delay can represent billions in lost revenue and put them behind in the race for AI leadership. But supply chain disruptions—whether geopolitical, physical (such as a blocked port), or labor-related—aren't rare. Neither are constraints caused when demand outstrips production. In 2024, the industry experienced supply limitations for GPU accelerators and high bandwidth memory, and the market continues to face critical power infrastructure shortages.

As demand spikes for new technologies, data center operators can expect much of the same. Under pressure to scale and maximize ROI, they are seeking a competitive edge in infrastructure optimization and control from design through deployment and end of life. As a result, deployment services are expected to grow to more than $110 billion annually by 2030, highlighting the critical nature of compute capacity delivery at the time of grid and facility availability.[11]

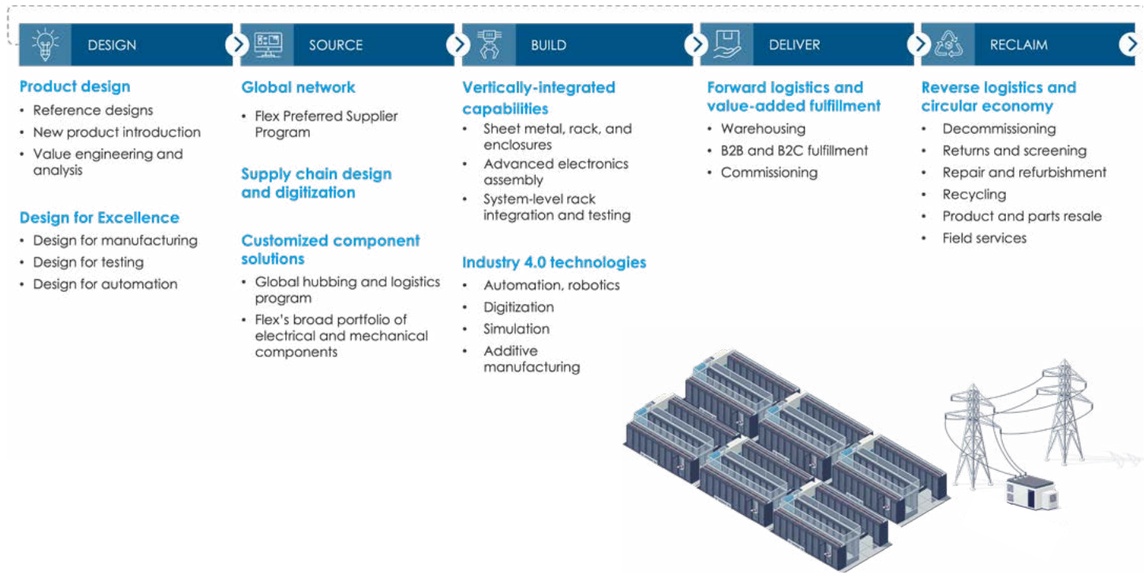
The strategic value of simulation

Expert analysis of the operational, financial and environmental impacts of various fulfillment options can uncover significant opportunities for improvement. Using its proprietary simulation tools, Flex found that establishing a single site for vertical integrated rack fulfillment could help one hyperscaler:

- **Reduce costs by 20%**
- **Shorten delivery time by 2 days**
- **Eliminate 7M kg of CO_2**

The benefits of the circular economy

One hour of server downtime is now estimated to be equivalent to more than $300,000 in lost revenue.[12] Fleet management—maintenance, repair, refurbishment, and responsible asset disposition—is not only critical to maximizing uptime, it also lends itself well to the opportunities presented by the circular economy, which aims to keep materials and products in circulation for as long as possible. A partner with expertise in fleet management as well as circular economy disciplines can help hyperscalers recover reusable parts and materials, achieve corporate sustainability goals, stay in compliance with environmental regulations, and even tap into new revenue streams. In the simulation showcased above, the hyperscaler would also have reduced the total cost of receiving and disassembling racks 35 percent by choosing a single location for fulfillment.



Overcome data center power, heat, and scale challenges with Flex

With competitiveness and data center capacity tightly intertwined, choosing the right partner to deliver next-generation power and cooling solutions at scale is central to data center operators' growth strategies. Flex enables companies to expand data center capacity faster and more cost-effectively with advanced manufacturing, a robust portfolio of critical and embedded power and cooling solutions, vertically integrated system rack manufacturing services, and end-to-end lifecycle services available in every major region of the world. Flex offers:

ADVANCED MANUFACTURING SERVICES
that support the mass deployment of vertically integrated data center racks, from materials sourcing and private-label components to the design, fulfillment, manufacture, and maintenance of servers, storage, racks, cabling, switches, busbars, power shelves, cooling technologies, and battery backup

POWER PRODUCTS
that enable data center operators to more efficiently manage power through innovative critical power infrastructure, such as switchgear and PDUs, and embedded power at the server and rack levels, such as power modules and power shelves

COOLING TECHNOLOGIES
including direct-to-chip liquid cooling modules and coolant distribution units that provide liquid cooling technologies to address the challenges of thermal density and increasing rack power

SPECIALIZED END-TO-END CAPABILITIES
to optimize and streamline the product lifecycle, and to seamlessly and sustainably design, build, deliver, and service products at scale for customers with increased quality, productivity, and speed across a global footprint

Accelerate data center infrastructure expansion at scale with advanced manufacturing services, innovative power and cooling products, and data center services from Flex.

Learn more about Flex data center solutions

Resources

1. PWC, Sizing the Prize: What's the real value of AI for your business and how can you capitalise?, Accessed February 19, 2025
2. Dell'Oro Group, Data Center Capex to Surpass $1 Trillion by 2029, According to Dell'Oro Group, February 5, 2025
3. McKinsey & Co., AI power: Expanding data center capacity to meet growing demand, October 29, 2024
4. No Priors: AI, Machine Learning, Tech, & Startups, Episode 89, Accessed February 19, 2025
5. Is Nuclear Energy the Answer to AI Data Centers' Power Consumption?, Goldman Sachs, January 23, 2025
6. Goldman Sachs, Generational Growth: AI, data centers and the coming US power demand surge, C. Davenport, B. Singer, N. Mehta, et. al., April 28, 2024
7. DataCrunch.io, NVIDIA GB200 NVL72 for AI Training and Inference, Aug 30, 2024
8. Uptime Intelligence, Capacity planning for liquid-cooled data centers, J. Williams-George, 3 May 2024
9. Open Compute Project, Open Rack V3 Based Specification, Revision 1.0, G. Charest, S. Mills, L. Vorreiter, 24 August 2022
10. CBRE, North America Data Center Trends H1 2024, August 19, 2024
11. Research and Markets, Data Center Services Market by Services, Data Center Size, Deployment Model, End User Verticals, Global Forecast, 2025-2030, November 7, 2024
12. Information Technology Intelligence Consulting, ITIC 2024 Hourly Cost of Downtime Report, L. DiDio, September 3, 2024

For more information, visit **flex.com/connect**

Flex (Reg. No. 199002645H) is the manufacturing partner of choice that helps a diverse customer base design and build products that improve the world. Through the collective strength of a global workforce across 30 countries and responsible, sustainable operations, Flex delivers technology innovation, supply chain, and manufacturing solutions to various industries and end markets.