VERTIV.

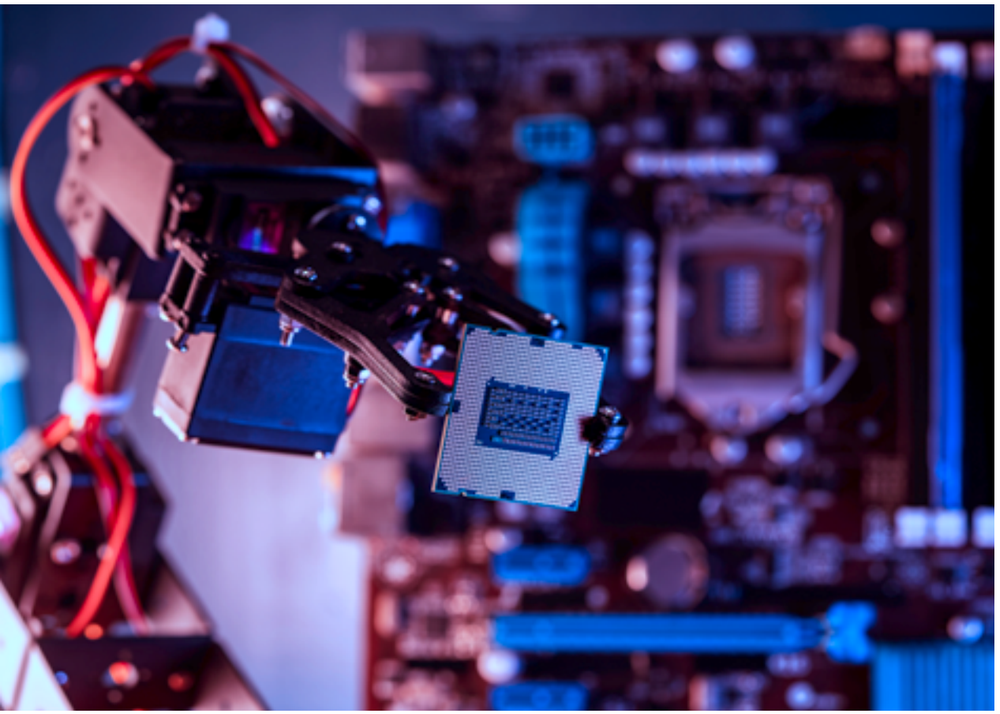
E-book

Infrastructure imperatives: Solving critical challenges for managing AI workloads

Table of Contents

Adapting to AI demands

The adoption of artificial intelligence (AI) is creating the most significant transformation in IT infrastructure in decades. Over the past 40 years, data centers have had to adapt quickly to new technologies that have transformed society, from the introduction of the personal computer and mainstream use of the internet to the shifts to cloud and edge computing.



In the years ahead, the “AI effect” on the digital infrastructure of data centers is more significant than any of those transformations. AI will touch every life and every industry on every level.

At the heart of AI and accelerated computing are extremely fast and power-intensive graphics processing units (GPUs). GPUs can compute much faster than central processing units (CPUs) because they are designed and built to run parallel operations all at the same time. By comparison, CPUs perform sequential processing one step at a time. That means calculations that once took hours now happen in seconds, or fractions of a second.

Initially designed for rendering graphics, GPUs are now used for training and processing inputs, known as inferences, for AI systems. GPUs process a single Chat GPT inference in one second—an operation that would take 32 hours with a CPU.[1] This example highlights the remarkable increase in computing power to take on anything an organization demands.

The remarkable processing power of GPUs drives new infrastructure requisites, including a significant upscaling of power and cooling systems from grid to chip. Air cooling was sufficient to meet the demands of traditional data centers with CPUs operating at a thermal design point (TDP) of 300W. But today’s GPUs have a TDP of 1,200W[2] in a single server, and that number will increase over time.

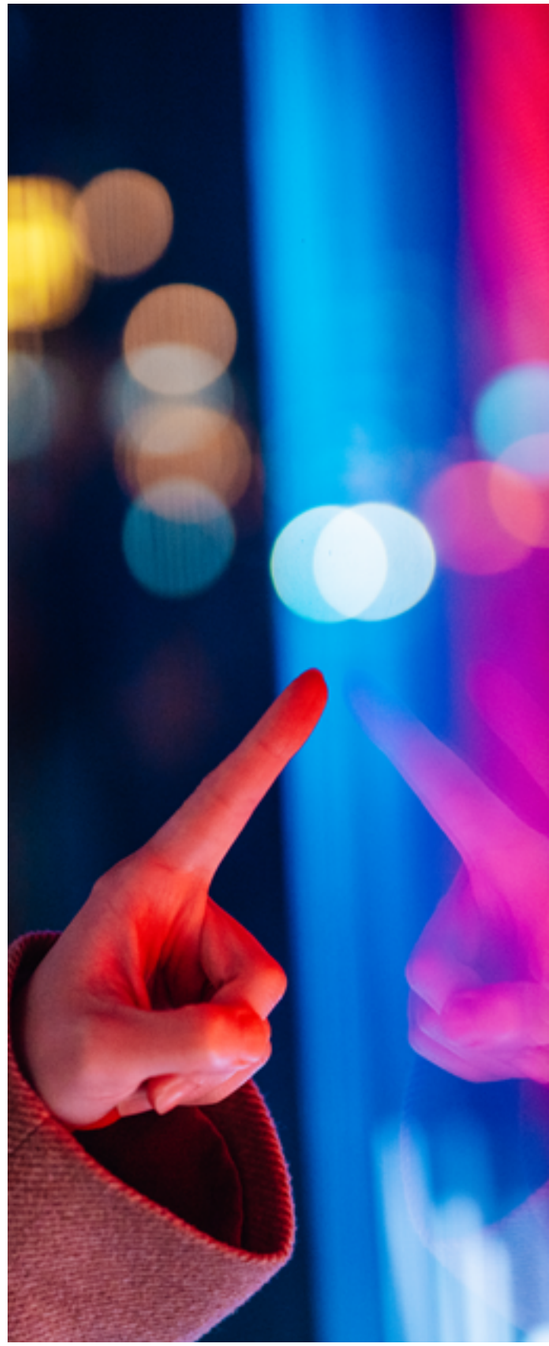
Today, some organizations are getting started by introducing an AI server or two into a conventional data center. It’s an important step to begin understanding AI’s potential and challenges. However, it creates a disparate system in which much higher levels of power and cooling are needed to support the AI rack compared to the rest of the facility. It’s not an environmentally responsible approach to addressing the increasing densification, scale, and speed of AI.

Instead of disparate systems, the AI factory data center will be fully integrated from the chip outward. The data center becomes the unit of compute capable of delivering the full promise of AI.

To help technology leaders meet the AI challenge, this e-book focuses on six physical infrastructure imperatives for successful AI factory implementations:

- Thermal imperatives for AI
- Power imperatives for AI
- Service imperatives for AI
- Rack imperatives for AI
- Design imperatives for AI
- System management imperatives for AI

That’s important because successful AI factory projects will not just need a different thought process but also a different, more collaborative approach.



Traditionally, roles have been siloed: the Facilities team built the data center, and the IT team installed whatever technology was needed. Consultation between the two teams was minimal, at best. But creating an integrated data center necessitates a fully integrated team that understands the imperatives of creating an AI factory. The team must include five key people:

- **IT systems** will be responsible for IT equipment, networking and applications.
- **Mechanical** will take charge of direct-to-chip cooling for servers as well as room cooling.
- **Electrical** will provide expertise to the complete power chain, including switchgear, UPS, and distribution at density.
- **Design/build** will integrate multiple infrastructure systems in proximity.
- **Operations** will coordinate IT/infrastructure for the life of the site.

Using these AI physical infrastructure imperatives, the right people can raise and discuss the right questions to bring an AI factory up and running as efficiently as possible and prepare for the future.

"With AI driving rack densities into three- and four-digit kWs, the need for advanced and scalable solutions to power and cool those racks, minimize their environmental footprint, and empower these emerging AI factories has never been higher."

— Giordano Albertazzi, CEO, Vertiv

"New data centers are built for accelerated computing and generative AI with architectures that are significantly more complex than those for general-purpose computing... With Vertiv's world-class cooling and power technologies, NVIDIA can realize our vision to reinvent computing and build a new industry of AI factories that produce digital intelligence to benefit every company and industry."[3]

– Jensen Huang, CEO, NVIDIA

Imperative 1: Thermal imperatives for AI

Adopt liquid cooling

Traditional air-cooling systems alone are not adequate to manage the heat levels produced by high-performance GPUs. To support high densities, liquid cooling must be delivered over and around racks to every server. Key issues:

Liquid cooling becomes a requirement. Liquid cooling becomes essential once the thermal design power per chip exceeds 700-800W 4 or sooner, depending on the server. The heat is then transferred to a chilled water plant or vented through the air-handling system.

Liquid availability and distribution are mission-critical. To support performance, liquid-cooled servers need a constant coolant flow.

Coolant distribution units (CDUs) become the engines of the thermal chain. CDUs are compact, flexible, and energy-saving devices that provide consistent flow of coolant fluid to liquid-cooled IT while removing heat from the liquid-cooled IT coolant fluid loops. CDUs are now considered critical load and cannot use power, so UPS systems are used to deliver continuous operation in addition to redundant pumps and power.

Liquid cooling fluid is the fuel of liquid cooling systems. These materials are different than what has been used in data centers to date. To enable optimal performance, special expertise is key at every step in the fluid lifecycle: filling, commissioning, storing, maintaining, and disposing.

Air cooling remains a "must," working in tandem with liquid cooling. Liquid cooling systems do not replace air cooling. In fact, liquid-cooled systems increase rack air-cooling requirements. The air and liquid cooling systems must work in parallel to optimize efficiency and effectiveness. There is no one-size-fits-all solution—the type of IT equipment that is deployed drives the design of air and liquid cooling systems.

> "The rise of liquid cooling should raise quite a few questions for those who run data centers. For example, 'How do I make sure the liquid loop is stable? How do I make sure it is redundant? How do I put a rack into the system? How do I take a rack out of the system?' These are all things people need to think about with the changes that are coming."
>
> **— Steve Madara , Vice President, Global Cooling, Vertiv**

Featured solutions

Vertiv™ CoolChip CDUs

A family of CDUs, suitable for direct-to-chip and rear-door applications, designed for high-density environments.

Learn More

Vertiv™ CoolPhase CDU

Liquid-to-refrigerant CDU for direct-to-chip and rear door cooling applications removes barriers to liquid cooling in air-cooled environment.

Learn More

Vertiv™ Liebert® XDU070 Rack CDU

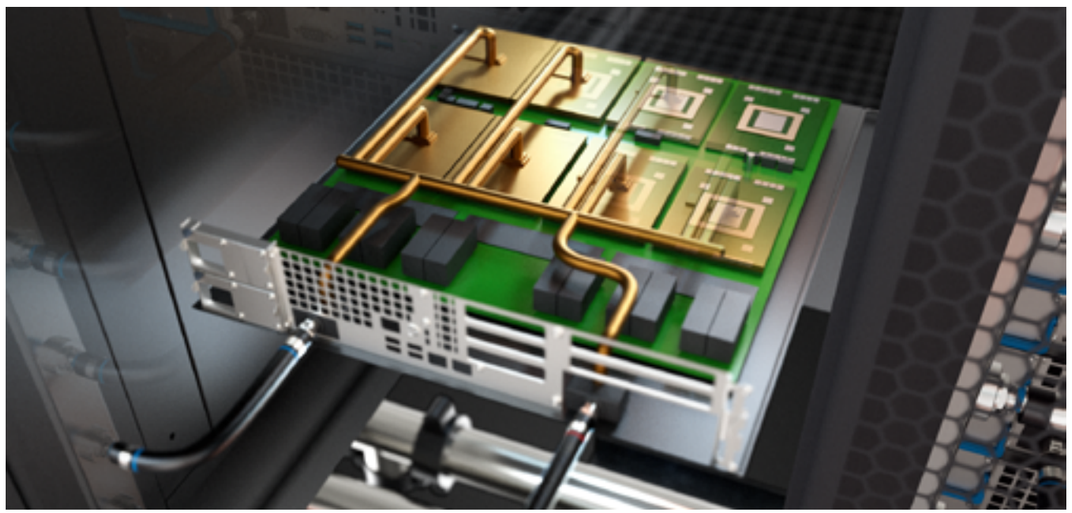
Liquid-to-air heat CDU supports efficient deployment of liquid-cooled servers in any environment.

Discover

Vertiv™ CoolPhase Flex

Industry's first-of-its kind hybrid air and liquid cooling technology designed to support AI.

Explore



Imperative 2: Power imperatives for AI

Meet growing, variable power demands

As GPUs deliver the speed needed for AI, the power needed by these high-density chips impacts the entire electrical infrastructure of the data center. Power demand is expected to escalate to an astonishing 1 MW per rack. To keep pace, data centers must address four key issues:

The power train, from chip to grid, must account for the growing and unique GPU load profiles. With traditional CPU servers, power demands tend to be consistent. However, GPU demands are highly variable. As AI compute engages, many processors act in unison, which results in massive power draws. For instance, loads can increase to 150% of theTDP in fractions of a second before dropping. Without proper management, these repeated surges can damage power systems. A new generation of UPS systems now functions as a power manager, drawing energy from multiple sources to smooth out peaks and valleys. In sizing the UPS, it is critical to anticipate the peaks.

Accelerating rack densities have become the norm. Distributing power to high-density racks is mission-critical. Cabling systems must be designed to enable scalability and optimized cable management. UPS equipment must play a bigger role than ever by combining energy storage, management, and distribution. Power infrastructure will need to be automated and reimagined for ease of scalability.

Cooling distribution units (CDUs) need uninterruptible power supplies. Liquid cooling is an essential element of managing the heat levels created by AI equipment (see Imperative 1). Coolant distribution units (CDUs) play a critical role in delivering cooling fluid directly to servers in the rack. Uptime is critical; servers shut down if there is even a one-second delay in coolant delivery. UPS systems are critical in facilitating the continuous operation of CDUs. This is why CDUs are built with redundant pumps and power inputs to enable continuous operations, as well as monitoring the system for early fault detection.

Bring your own power (BYOP) is becoming more common as higher power loads are needed. While AI is adopted at a blistering pace, electric utility providers struggle to keep up. It can take five years to bring a new power plant online. Data centers can also be vulnerable to prolonged power outages. A BYOP strategy can either be connected to a larger grid to reduce utility outage risks or stand independently when necessary. The BYOP approach uses distributed energy resources (DERs) such as UPS systems, battery energy storage systems (BESS), and fuel cells to create an always-on microgrid that enables a constant power supply.

> "What is the secret to meeting these power challenges? I think a big part will be anticipating the more specific changes in power. You need to know where the chip makers are headed, and you need to know what the IT industry insiders think."
>
> **— Peter A. Panfil, Vice President of Global Power, Vertiv**

Featured solutions

Power Conversion/Distribution

Vertiv™ Trinergy™

Next-generation UPS designed to support high-capacity, high-availability AI power demands in room and prefabricated deployments.

Learn More

Vertiv™ PowerNexus

Integrated system that combines the robust power of Vertiv™ Trinergy™ and Switchboard to reduce equipment footprint, cabling materials, and installation labor costs for hyperscale and colocation data centers.

Discover

Vertiv™ EnergyCore

Lithium-ion battery storage that supports high-density computing by saving floorspace in increasingly crowded data centers.

Explore

Room and Row Power Distribution

Vertiv™ PowerBar iMPB

Intelligent medium-power busway is ideal for data centers of any size that have frequent, planned power configuration changes.

Learn More

Vertiv™ PowerBar (HPB)

The enclosed busduct offers a complete power solution for transformer to main panel board connections, designed to move large amounts of power.

Explore

Vertiv™ Liebert® RXV

Remote power distribution cabinet provides dense power distribution in a small footprint with accurate monitoring and multiple configuration possibilities.

Learn More

Vertiv™ Liebert® PPC

Power conditioning and distribution cabinet offers the benefits of a custom-tailored system while offering the convenience and cost savings of a pre-packaged, factory-tested solution.

Discover

Distributed Energy Sources

Vertiv™ DynaFlex Battery Energy Storage System

Supports BYOP strategies by providing an always-on energy solution that helps organizations increase power reliability and strengthen operational resilience, while reducing costs and carbon emissions.

Learn More

Vertiv™ Liebert® EXL S1 UPS

Enables data center operators to effectively engage in demand management and response, enhancing overall energy and environmental responsibility management.

Learn More

Imperative 3: Service imperatives for AI

Provide specialized AI expertise

End-to-end lifecycle services are key to streamlining the deployment, commissionin, and maintenance of AI infrastructure. In creating a support team for AI, keep in mind:

Specialized service expertise is important for maintaining an AI environment. Given the complexity of AI technology, service personnel must have the AI-specific knowledge and skills.

Ongoing training and new processes are essential for both vendors and internal staff. The technology will continue to evolve rapidly in the years ahead, so it is essential that the entire support team has up-to-date knowledge and skills.

Featured solutions

Vertiv™ Liquid Cooling Services

Expert turnkey services that facilitate operational efficiency and enhanced system availability as heat levels rise due to AI applications.

Explore

Vertiv™ Global Services

A full array of maintenance and performance services that increase efficiency and reduce complexity anywhere in the world.

Learn More

Imperative 4: Rack imperatives for AI

Use racks that support high density

AI requires power-to-rack ratios that are 10 times those of traditional IT racks. In addition, racks must be designed to accommodate and efficiently manage liquid cooling infrastructure. Key issues:

IT racks must allow for high-density power and cooling. Current AI chips from NVIDIA require up to 140 kW per rack. Yet by 2029, rack densities may reach 1 MW.[5] To support these loads:

- Multiple rack PDUs (rPDUs) are needed if using conventional AC power. Four or more rPDUs per rack may be needed in a combination of horizonal and vertical units.
- Higher voltages up to 277/480V and increased amperage levels—ranging from 60A to 100A—are becoming increasingly common.
- Tailored designs help minimize footprint and maximize power delivery.

Self-sensing racks are important for mitigating system failure. Racks must become more intelligent as liquid cooling is introduced to cool expensive IT equipment. Racks must have drip pans with sensors that communicate to a central management controller. Pressure and flow sensors detect breaks in supply lines and trigger control systems to respond. Fluid quality and flow sensors monitor for irregularities and alert control systems to take corrective action in the rack.

Data centers will need to fit more computing into higher, wider, deeper, and heavier IT racks. Stronger racks are needed to handle the additional weight of IT equipment and liquids. Racks must have higher dynamic and static ratings for shipping and operation. Larger racks must accommodate rPDUs, manifolds, larger cables, and inlet/outlet hoses. All this makes retrofitting existing data centers more challenging, given the expansion of rack footprint.

> "The technologies essential to IT infrastructure will be significantly different from what we employ today, which means keeping one eye on the developmental pipeline is an essential part of being prepared for the future."
>
> **— Greg Ratcliff, Chief Innovation Officer, Vertiv**

Featured solutions

Vertiv™ PowerDirect Rack

Delivers resilience to even the most demanding AI and high-performance computing environments with scalable 50V DC power for IT racks

Learn More

Vertiv™ Geist™ Rack PDUs

High-capacity power distribution to support AI servers and network infrastructure

Discover

Imperative 5: Design imperatives for AI

Design for the AI revolution

The magnitude of AI demands is driving fundamental changes in how data centers are planned and built. Traditional approaches can limit scalability, create long lead times, and reduce ability to efficiently support AI. To sidestep these issues, data center planners must address four critical issues:

Design power and cooling as an integrated system. Traditionally, data center planning was siloed, with power and cooling systems designed separately. AI, however, creates a new paradigm in which planning starts at the chip level and power and cooling are designed together to operate as a holistic unit. A key reason: AI hardware takes up a lot of space. By integrating power and cooling, planners optimize the share of space and energy dedicated to AI processing and support future scalability.

Adopt pre-engineered, pre-built solutions to reduce deployment time and improve ROI. To meet the rapidly growing demand for AI data center capacity, it is critical to reduce design and construction timelines. Prefabricated modular data center (PMDC) solutions can be configured and scaled to customer requirements while enhancing deployment speed. The factory-built units reduce construction and labor costs to lower the total cost of ownership (TCO) while simplifying scalability.

Plan AI data center infrastructure to be resilient to future challenges. The demands of AI technology are certain to grow and evolve. To avoid designing for future obsolescence, it is essential to build a strong foundation today that will deliver the scalability and adaptability important to meeting future demands for power, cooling, and other mission-critical solutions.

Collaborate across disciplines. In the past, design roles were siloed: the Facilities team built the data center, and the IT team installed whatever technology was needed. Consultation between the two teams was minimal, at best. With AI, however, it is essential to start with IT needs, then build the data center around those. That will entail a team approach that brings together the right people who understand infrastructure imperatives (see Figure 1).

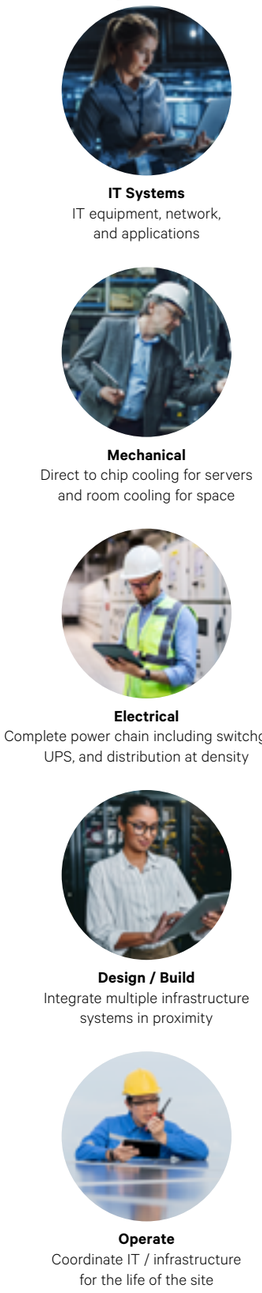


Figure 1. *Building an AI data center will entail a new paradigm in design and operations through collaboration.*

"It's important to look at power and cooling together as one along with the controls and the software around it. It needs to operate as one unit of compute."

— Martin T. Olsen, Senior VP, Products and Solutions, Vertiv

Featured solutions

Vertiv™ 360AI

A complete portfolio of power, cooling, and service solutions that solves complex challenges to support success in the AI revolution.

Learn More

AI Infrastructure Reference Designs

Co-developed by Vertiv and NVIDIA, the 30 reference designs provide practical blueprints for designing, building, and deploying AI factories.

Explore Designs

Prefabricated Modular Solutions

Alternatives to traditional brick-and-mortar data centers that provide the flexibility, scalability and rapid deployment to meet AI challenges.

Discover

Imperative 6: System management imperatives for AI

Establish a comprehensive IT system management

Given the intense demands associated with AI, there is a critical need for an open, scalable management platform that provides centralized, remote visibility across the entire solution along with total control of power and cooling technologies.

Monitoring and management systems must be holistic. It's more important than ever to eliminate silos in system management with comprehensive monitoring software that collects all the data needed to run critical infrastructure and delivers an integrated, real-time view. Key elements of a complete AI management system:

- Alarm systems throughout the digital infrastructure.
- Capacity management that provides visibility to stranded power, cooling, and IT while simplifying energy usage reporting.
- Advanced predictive services that anticipate issues and prescribe actions to optimize AI infrastructure.
- 3D visual incidents visibility to troubleshoot issues quickly.

Controls must be integrated at the infrastructure component level to protect servers from catastrophic events. Leak detection, once a minor consideration, is now a must-have given the sensitivity and high value of AI equipment in combination with liquid cooling. While a traditional server might cost in the low six figures, AI servers cost millions of dollars. As a result, controls are now integrated not just into the rack but into the servers, valving, and electric outlets. Any interruption in coolant flow should trigger the shutdown of servers , so automation and controls are mission-critical. Management platforms can proactively—and immediately—address issues based on pre-defined parameters that assess whether it is a slow leak or a more serious one.

An open, scalable management platform is critical for future-ready data centers. By designing with scalability in mind, leaders can avoid costly "rip-and-replace" upgrades as technology evolves. An open system should support multiple protocols, integrate with systems across facilities, and interoperate seamlessly with a broad range of vendor solutions and situations as technology advances.

"AI (is) not a chip problem. It's a reinvention-of-computing problem. You can't solve this new way of computing by just designing a chip. Every aspect of the computer has fundamentally changed."[6]

— Jensen Huang, NVIDIA CEO

Featured solutions

Vertiv™ Environet™ Connect

Cloud-based monitoring platform that allows you to manage and monitor your physical infrastructure anytime and anywhere

Learn More

Liebert® iCOM™-S

Thermal control and monitoring solution providing a single point for managing your entire cooling infrastructure and gaining quick access to actionable data, system diagnostics and trends

Discover

Vertiv™ Unify

An integrated energy and power management platform that simplifies data center operations by consolidating power, thermal, and management systems into a single interface

Explore

Selecting the right partner

Looking ahead, AI will create vast opportunities to drive innovation, transform customer experiences, accelerate the creation of new products and services, and streamline organizational efficiency. To achieve these goals, it is important to completely rethink data center design and operations. It is essential to choose the right partner who can provide the expertise and solutions to enable data centers to address AI imperatives.

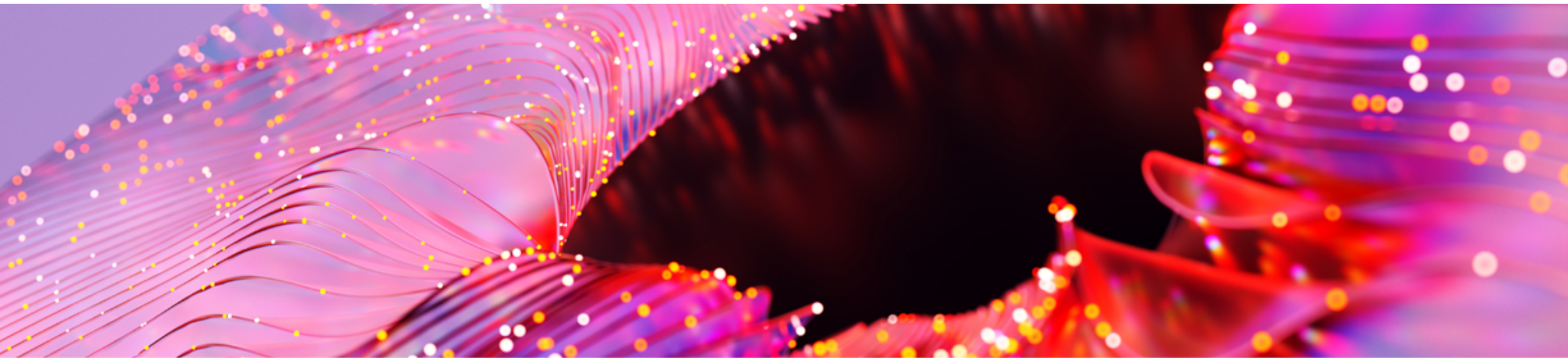
As a respected leader in data center infrastructure, Vertiv offers proven performance and a complete portfolio of solutions designed for the challenges of AI and advanced computing.

To learn more, visit the **Vertiv™ AI Hub.**

More AI Insights from Vertiv

A companion e-book, **Strategic Imperatives: Prioritizing Your AI Transformation,** provides a framework for establishing priorities and selecting projects and technologies that position IT to drive business success.

> "Virtually every industry is exploring opportunities to drive business value through AI, but there are more questions than answers around how to deploy the infrastructure. A recognized infrastructure provider like Vertiv is valuable to businesses building an AI strategy and looking for a single source for information."[7]
>
> **— Sean Graham, Research Director, Data Centers at IDC**



References

[1] Appenzeller, G., Bornstein, M., & Casado, M. (2023). *Navigating the High Cost of AI Compute.* AH Capital Management, L.L.C. https://a16z.com/navigating-the-high-cost-of-ai-compute/.

[2] NVIDIA. (2025). *NVIDIA Blackwell: The engine of the new industrial revolution, data sheet*. https://resources.nvidia.com/en-us-blackwell-architecture/datasheet?lx=AJq6FY&ncid=em-webi-775169.

[3] Vertiv. (2024). *Vertiv Codevelops with NVIDIA Complete Power and Cooling Blueprint for NVIDIA GB200 NVL72 Platform*. https://investors.vertiv.com/financial-news/news-details/2024/Vertiv-Codevelops-with-NVIDIA-Complete-Power-and-Cooling-Blueprint-for-NVIDIA-GB200-NVL72-Platform/default.aspx

[4] Moss, S. (2024). *Nvidia's CEO confirms upcoming system will be liquid cooled. Data Center Dynamics.* https://www.datacenterdynamics.com/en/news/nvidias-ceo-confirms-next-dgx-will-be-liquid-cooled/.

[5] Swinhoe, D. (2025). *Hyperscalers prepare for 1MW racks at OCP EMEA; Google announces new CDU. Data Center Dynamics.* https://www.datacenterdynamics.com/en/news/hyperscalers-prepare-for-1mw-racks-at-ocp-emea-google-announces-new-cdu/.

[6] New York Times. (2023). *NVIDA CEO Not Worried About Rising AI Competition.* https://www.nytimes.com/2023/11/29/business/dealbook/nvidia-ai-dealbook-chips.html.

[7] Vertiv. (2024). *Vertiv Launches New AI Hub, Featuring Industry's First AI Reference Design Portfolio for Critical Digital Infrastructure*. https://investors.vertiv.com/financial-news/news-details/2024/Vertiv-Launches-New-AI-Hub-Featuring-Industrys-First-AI-Reference-Design-Portfolio-for-Critical-Digital-Infrastructure/default.aspx.

Appendix

AI imperatives for data center infrastructure

Challenge	Imperatives
Design. Traditional approaches can limit scalability, create long lead times and reduce ability to efficiently support AI.	**Design for the AI revolution** • Design power and cooling as an integrated system. • Adopt pre-engineered, pre-built solutions to reduce deployment time and improve ROI. • Enable AI data center infrastructure to be resilient to future challenges. • Collaborate across disciplines.
Power. The energy demands of high-density AI chips are impacting the entire electrical infrastructure of the data center.	**Meet growing power demands** • From chip to grid, the power train must account for growing and unique GPU load profiles. • Accelerating rack densities have become the norm. • Cooling distribution units (CDUs) require uninterruptible power supplies (UPS). • Bring your own power (BYOP) is becoming more common as higher power loads are delivered to sites.
Cooling. Traditional air-cooling systems alone are not adequate to manage the heat levels produced by high-performance computing.	**Adopt liquid cooling** • Liquid cooling becomes a requirement. • Liquid availability and distribution are mission-critical. • CDUs become the engines of the thermal chain. • Liquid cooling fluid is the fuel of the liquid cooling systems. • Air-cooling remains a "must," working in tandem with liquid cooling.
Racks. AI requires power-to-rack ratios that are 10x traditional IT racks.	**Leverage racks that support high density** • IT racks must allow for high-density power and cooling. • Self-sensing racks are needed to mitigate system failure. • Data centers must fit more computing into higher, wider, deeper, and heavier IT racks.
System monitoring and management. Traditional systems do not offer the level of scalability, visibility and control AI needs.	**Establish AI-ready system management** • Monitoring and management systems must be holistic. • Controls must be integrated at the infrastructure component level to protect servers from catastrophic events. • An open and scalable management platform is essential.
Services. AI needs specialized services to maintain and optimize operations.	**Provide expert support services** • Specialized service expertise is essential to maintain an AI environment. • Ongoing training and new processes are indispensable for both vendors and internal staff.

Vertiv.com | **Vertiv Headquarters, 505 N Cleveland Ave, Westerville, OH 43082, USA**

SL-80214 (06/25)