Article

Cool for the summer: Why liquid cooling is the next AI imperative

Advanced processors are pushing thermal boundaries in data centers. To keep up, enterprises are turning to liquid cooling to balance performance, cost, and sustainability.

AI is driving some of the most demanding workloads in the enterprise. It's diagnosing disease in real time, optimizing logistics in global supply chains, and driving the next wave of product innovation. But behind every insight and automation is a server running at full power — and getting hot.

Today's high performance CPUs and GPUs consume more power and generate more heat than traditional air-cooling systems were ever designed to handle. Fans and heatsinks can't keep up, and data centers are paying the price in performance throttling, wasted energy, unsafe sound levels, and mounting infrastructure costs.

Liquid cooling — once reserved for supercomputers and specialized environments — is now being adopted more broadly as a practical response to AI's thermal demands. Here's a look at how direct liquid cooling systems are being used to manage heat more efficiently, reduce energy waste, and keep AI infrastructure viable at scale.

The AI heat crisis

The rise of large language models and agentic AI has pushed compute demands into uncharted territory. High-end CPUs now routinely exceed 500 watts of power consumption, and some GPUs are approaching the 1,000-watt mark. At the same time, thermal tolerances are dropping. Where chips once operated safely at 90° to 100° Celsius, some of today's most advanced silicon now maxes out at just 60°.[1]

Traditional air cooling simply isn't built for this new thermal reality. Managing the heat from modern processors requires massive heatsinks, powerful fans, and more physical space, all of which drive up costs and reduce server density. It's an inefficient cycle that undermines performance and scalability.

It's also causing energy usage to skyrocket. U.S. data center electricity consumption jumped from 76 terawatt-hours in 2018 to 176 TWh in 2023, and could hit as much as 580 TWh by 2028.[2] Heat has become AI's hidden bottleneck and the stakes are rising fast.

How liquid cooling works

Liquid cooling works by circulating coolant over the hottest components in a server — namely CPUs and GPUs — to absorb and carry away heat. Coolant circulates through cold plates attached to key components, drawing heat away and carrying it to an external manifold, where it's routed to a heat exchanger for removal. The result is faster, more efficient thermal transfer than air cooling can provide.

Hewlett Packard Enterprise offers three primary liquid cooling options. Closed-loop systems are fully contained within the server chassis, making them ideal for incremental upgrades to existing infrastructure. The alternative option is liquid to air cooling which uses facility water to cool down an air cooling system which is connected to the racks, bringing cool air more precisely to the hottest IT and transfers the hot air heat back to the facility water. Finally, direct liquid cooling operates at the rack level, distributing coolant to multiple nodes and offering superior heat management for dense, high performance workloads.

There's also a combined approach, which uses liquid cooling for the most power-hungry components and air for the rest. This balanced design verifies that every component is cooled efficiently without overengineering the system.

Compared to air cooling, liquid systems deliver better performance, allow for higher rack density, and significantly reduce energy use, cutting operational costs along with carbon emissions.

[1] "HPE ProLiant Gen11 Servers with Direct Liquid Cooling," HPE, 2024.

[2] 2024 United States Data Center Energy Usage Report, Lawrence Berkeley National Laboratory, U.S. Department of Energy, December 2024

How HPE is deploying liquid cooling at scale

HPE has engineered its liquid cooling solutions to meet the scale, density, and efficiency demands of modern AI. At the center of this effort are HPE ProLiant Compute Gen12 servers, which support liquid cooling configurations. Closed-loop systems offer a self-contained cooling solution ideal for high-wattage CPUs in existing rack setups. Direct liquid cooling goes further, distributing coolant across an entire rack to manage heat at scale — perfect for dense, GPU-heavy workloads.

For environments with mixed workloads, HPE's Adaptive Cascade Cooling offers a unified approach. This patented system dynamically switches between air and liquid cooling in real time, depending on the thermal load of connected components. The result is greater cooling efficiency with less energy use and infrastructure complexity.

Regardless of which approach you choose, the cost and sustainability gains are substantial—offering a way to cut energy consumption without sacrificing performance, making it a strategic option for organizations balancing operational demands with environmental goals. In a 10,000-server deployment, HPE's direct liquid cooling can save over $2 million per year in cooling costs while cutting 17 million pounds of CO_2 emissions annually.[3]

These systems are already in use across 100% fanless HPE Cray supercomputers and HPE ProLiant XD platforms, supporting high-density deployments where traditional cooling would fall short. As AI workloads continue to expand, liquid cooling offers a practical way to meet thermal demands without redesigning the entire data center, as liquid cooling doesn't require a complete overhaul. Organizations can phase in advanced cooling where it's needed most — such as on high-density or high-heat workloads — making it a practical step toward meeting environmental, social, and governance targets.

[3] Liquid Cooling: A Cool Approach for AI, HPE Newsroom Blog, August 2024.

Designing for growth: HPE's approach to scalable AI

HPE's liquid cooling innovation is just one part of a broader strategy to future-proof AI infrastructure. Solutions like HPE Private Cloud AI combine liquid-cooled HPE ProLiant Compute systems with NVIDIA accelerated computing, integrated networking, and AI-ready software in a turnkey platform. Designed for production-scale AI, these systems deliver faster time to value while simplifying deployment and scaling.

Because these are on-premises solutions, organizations retain full control over their data, an advantage for enterprises with sovereignty, compliance, or low-latency requirements. At the same time, predictable economics and built-in cooling efficiency make it easier to manage operational costs as AI demands grow.

HPE AI Services can help organizations plan and implement infrastructure that supports both the performance and thermal demands of AI workloads. That includes selecting the right cooling approach to match deployment scale and workload type. This kind of alignment — among compute, cooling, and operations — can reduce complexity and make it easier to scale efficiently over time.

Staying ahead of the heat

AI isn't slowing down and neither are the demands it places on infrastructure. As chips run hotter and workloads grow more complex, traditional cooling can no longer keep pace. Liquid cooling changes the way enterprises must approach infrastructure planning. By managing heat more efficiently, it supports higher-density deployments, reduces energy use, and helps extend hardware life — all essential for scaling AI workloads sustainably.

The future of AI is hot — but with liquid cooling, your infrastructure doesn't have to be.

Learn more at

HPE.com/ai/insights

Visit HPE.com

Chat now

a50013353ENW

HEWLETT PACKARD ENTERPRISE

hpe.com