HyperFRAME
RESEARCH

Data Center Water Cooling: Debunking Misperceptions and Myths

As AI Datacenter Growth Explodes, The Doomer Narrative Has Gained Traction, HyperFRAME Research Puts The Record Straight

Authors:

Ron Westfall
VP and Practice Leader for
Infrastructure and Networking

Steven Dickens
CEO and Principal Analyst

FEBRUARY 2026

Executive Summary

Current public discourse regarding data center water usage is often characterized by sensationalism. We have all seen numerous news articles and viral posts on various social media platforms this year, decrying the rise of AI-centric datacenter build-outs. The stories center on the 'facts' that reductively equate AI queries to water consumption. These 'facts' have led to the ill-informed clamoring to pass legislation restricting or, at worst, halting AI expansion altogether. This report will look to put the record straight, with actual facts.

Mainstream narratives frequently overlook the fact that traditional power generation, such as coal, consumes significantly more water up to 20 liters per kWh (L/kWh), primarily used for steam generation, boiler makeup, and cooling tower, than modern data centers. In contrast, the average U.S. data center uses ~1.8 L/kWh for evaporative cooling and facility humidity control with advanced AI data centers using < 0.2 L/kWh for closed-loop liquid cooling and warm-water direct-to-chip systems. Plus, the 2024 European Commission's 2024 report on data centers in the European Union (EU), reported the EU had an operational average WUE of 0.58 L/kWh across the surveyed EU data center fleet with large scale data centers (>10MW) having an operational average of 0.7 L/kWh.

To improve transparency, the industry is advocating for a shift from total volume reporting to an output-based metric. One proposal is liters of water per token. This approach directly links environmental costs to AI productivity, highlighting how advanced silicon and closed-loop systems actually improve computational efficiency.

The data center industry is undergoing a fundamental transition as cooling evolves from a passive background utility into a strategic, AI-driven asset. Fueled by the explosive growth of generative AI, chip thermal design power is rapidly approaching 1,000W, pushing rack densities toward 50--100 kW and beyond. Consequently, traditional air cooling is reaching its physical limits, necessitating a shift toward thermal orchestration and high-efficiency liquid cooling solutions.

By 2026, direct liquid cooling is projected to reach 50% market adoption (according to Vertiv & Schneider Electric 2026 market outlooks). Because liquid is 25 times more efficient at heat transfer than air, DLC enables a 16% increase in compute density within the same power footprint (according to Schneider Electric White Paper 132). Major manufacturers such as Dell, Lenovo and HPE are pioneering closed-loop secondary circuits, such as the PowerCool eRDHx and Neptune systems, which use glycol or treated water to capture heat at the source. These systems eliminate the need to draw from public water supplies for evaporation, paving the way for a water-neutral future. Put simply, glycol-based systems don't draw on local reserves at all; they are closed systems.

Decision-makers should prepare for hybrid cooling environments where advanced air-cooling retrofits, like Rear Door Heat Exchangers, coexist with liquid-cooled AI clusters. These Rear Door Heat Exchanger solutions even became keynote worthy when Michael Dell announced his company's efforts in this space at Dell Tech World earlier in 2025. This allows existing facilities to support high-density workloads without total reconstruction. Ultimately, the industry is moving toward a circular energy economy, where the concentrated heat captured by liquid cooling is repurposed for district heating or agriculture, transforming data centers into integrated, sustainable components of the overall energy grid.

Introduction

Headlines about data centers and GPUs consuming water are often alarmist and misleading. The mainstream media often adopts an alarmist tone when discussing the water usage of AI data centers, sometimes lacking the necessary context. Headlines frequently focus on the staggering total volumes used by tech giants like Google, xAI, and Microsoft, which can sound enormous without comparison. This coverage often overlooks the scale of water consumption by other major sectors, such as agriculture or energy generation, where usage is significantly higher. Should we discuss recycled water as well?

For instance, recycled water serves as a sustainable alternative to potable water, significantly reducing a data center's reliance on local drinking water supplies for its cooling towers. By using treated wastewater for heat rejection, facilities can maintain the massive cooling capacities required for high-density servers without straining municipal resources or local ecosystems. This shift not only lowers the facility's Water Usage Effectiveness (WUE) metric but also enhances long-term operational resilience against droughts and water scarcity.

Critics note that the emotional framing of individual AI queries using a bottle of water can be misleading, as the cumulative effect is applied to a single location. Furthermore, the media frequently fails to differentiate between water withdrawal (water taken) and consumption (water lost to evaporation), which are distinct metrics. Many reports do not adequately highlight the innovations in cooling technologies like liquid immersion that can drastically reduce water dependence. Focusing narrowly on data center water use diverts attention from the larger, systemic water-stress issues in specific drought-prone regions. Ultimately, while AI's water footprint is a genuine concern, the mainstream narrative often prioritizes sensationalism over balanced reporting on the issue.

Many modern data center cooling systems do not actually use water at all, but rather closed-loop liquid coolants like glycol. These liquids continuously cycle through the system without being consumed or evaporated, meaning the data center is not drawing from the public water supply for its cooling needs.

When generating power and running data centers, water is used in a variety of ways. Although closed-loop liquid cooling technologies are used in data centers to reduce water use, there is persistent misunderstanding about how evaporative cooling works. A shift in how to measure water usage for AI could help clear up these misconceptions and improve public understanding.

For example, we expect that reflecting a strategic balance of control, performance, and flexibility, 36.6% of enterprises now use hybrid data architectures that integrate on-premises and cloud systems as the foundation of their AI stacks (according to HyperFRAME Research). This reliance on hybrid data architectures forces a shift toward exploring and adopting more innovation throughout cooling environments, as organizations must manage the relatively low heat density of traditional on-premises systems alongside the extreme thermal demands of cloud-bursting and AI GPU clusters. By balancing on-premises control with cloud scalability, enterprises are driving the adoption of hybrid cooling strategies that combine standard air-cooled aisles for legacy hardware with modular DLC for high-performance workloads.

To provide greater clarity and transparency, we anticipate that new metrics under consideration, such as litres of water per token and system-level reporting that balances tokens (a value metric) with physical work (kWh) through refined WUE targets, can make inroads and be used to provide new measurement aspects on water consumption across AI data center environments. This would move the conversation beyond just the technology and focus on the direct water cost of AI computations, helping to educate the public and counter common misconceptions about methods such as evaporative cooling.

To get a clearer picture of efficient cooling, we examine liquid cooling technologies from major manufacturers. For example, Dell's rear-door heat exchangers, HPE's Adaptive Rack Cooling, and Lenovo's Neptune systems are excellent examples of highly effective, closed-loop liquid cooling solutions that are designed to minimize or eliminate water usage. Such innovations show that the industry is focused on sustainable designs, which directly counters the inaccurate, sensationalist claims often found in media reports.
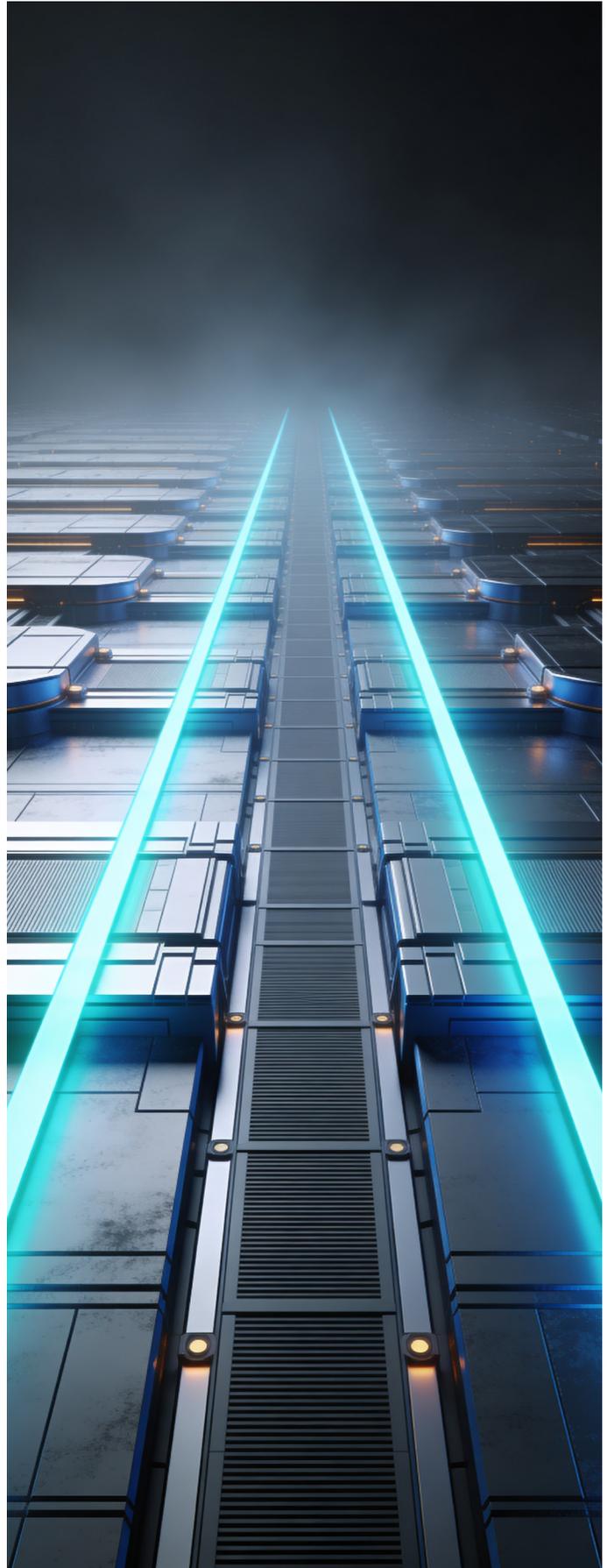
The Fundamentals of Data Center Cooling

Traditional power generation methods, particularly coal and natural gas, have a significant water footprint, using anywhere from 10 to 72 liters of water per kilowatt-hour, a consumption pattern that can only be eliminated by using water-free renewable energy sources. This issue is often misunderstood in the context of data centers, where many headlines are misleading. Most modern data centers use highly efficient closed-loop cooling systems that rely on liquids like glycol and consume little to no water.

Even with evaporative cooling, often cited as a major water consumer, a large portion of the water isn't evaporated but is instead collected and sent to a wastewater stream. To address these misconceptions and improve public understanding, the new metric liters of water per token should be used, directly linking AI's output to its water cost and highlighting how newer technologies and high-performance hardware are reducing fluid consumption to generate valuable AI outputs.

Water Consumption in Power Generation: It is critical to understand the water consumption metrics of different power generation methods. Specifically, on average, coal-based power plants use about 72 liters of water for every kilowatt-hour of electricity they produce. Natural gas plants are significantly more efficient, consuming approximately 10 liters per kilowatt-hour. These figures can fluctuate depending on the specific location and technology, but some water use is an unavoidable part of traditional power generation.

The only way to completely avoid this water consumption is by using renewable energy sources that don't rely on water for cooling or steam production. While the exact numbers vary, these metrics show that the choice of power source has a direct and significant impact on water resources.

Water Consumption in Data Centers: Closed-loop cooling systems used in data centers, like those that rely on refrigerant or other liquids, consume no water at all. For instance, liquid-cooled systems from companies such as Lenovo capture a significant amount of heat - around 80% - and transfer it into a liquid stream. This process is highly efficient and, because the liquid is contained within a closed loop, it results in no water consumption.

While some data centers use evaporative cooling, which can be a source of confusion, it is important to understand how they work. While these systems may use about 13 liters of water per kilowatt-hour, a large amount of that water is not evaporated into the air. Instead, it is directed to a wastewater stream because it has collected minerals and dirt from the cooling process. This distinction is crucial for understanding the true water footprint of these systems.

Addressing Misconceptions about Evaporative Cooling: Many people mistakenly believe that the water used in data centers, particularly with evaporative cooling, is simply flushed away and continuously consumed. As such, it is important to clarify that this is not the case. While evaporative systems do use water, it is not flushed in an open-loop system. The water is used for the cooling process, but a significant portion of it is often recycled or directed into a wastewater stream after it collects minerals and dirt.

Newer data center designs are moving away from evaporative cooling altogether. By tolerating higher fluid temperatures, some facilities can use dry coolers, which have no evaporation and therefore use no water. This approach is preferred whenever possible, but some locations with limited surface area or geographical constraints may still have to rely on evaporative coolers.

For instance, to accommodate the specialized requirements of Grok AI, X has aggressively expanded its computational power, a move that necessitates the adoption of high-density, water-efficient infrastructure. xAI's Colossus supercomputer and other massive training hubs have pivoted toward rack-scale liquid cooling, using Coolant Distribution Units (CDUs) to manage extreme thermal loads. This system effectively captures heat from high-performance GPUs and transfers it to a secondary loop, where it is dissipated via external dry cooling towers rather than traditional water-intensive methods.

By using this closed-loop architecture, the facility can support the dense installation of NVIDIA H100, H200, and Blackwell GPUs without the massive municipal water consumption typical of older evaporative swamp cooler designs. This shift not only protects local water resources but also eliminates the risk of thermal throttling, allowing the cluster to maintain peak performance during intensive AI training runs.

Measuring Water Consumption per AI Token: To make the public's understanding of AI's water usage more accessible, the current metric of liters per kilowatt-hour can be reframed. Instead, a new metric, such as liters of water per token, would be more relatable to the average person, as a token is analogous to a word. This would create a direct connection between the water consumed and the output of the AI, making the environmental impact more tangible and easier to grasp.

This new measurement would also highlight the efficiency of advanced technology. It could clearly demonstrate how high-performance silicon and modern closed-loop liquid cooling systems significantly reduce water consumption. By tying water usage directly to the value of the AI's output, it becomes easier to show that newer, more efficient solutions are consuming far less fluid to generate the same amount of useful information.

However, the industry has not yet formed a consensus to support the liters of water per token metric due to considerations such as tokens are not a physical measure of work, as it is a value proposition, and the metric is not normalized for the climatic impacts of data center location to data center water consumption. As a result, to address the limitations of the liters per token metric, researchers and standards bodies are shifting toward contextualized water footprints that use the ISO 14046 framework to normalize consumption based on local water-stress indices and climatic variables like Cooling Degree Days (CDD). Simultaneously, industry decision makers are integrating system-level reporting that balances tokens (a value metric) with physical work (kWh) through refined WUE targets, ensuring that efficiency claims account for the varying environmental cost of a gallon of water in a desert versus a temperate region.

This white paper takes a closer look at how the data center ecosystem is building toward this vision. The updated analysis incorporates the latest insights on the industry's most important developments and innovations, providing data center decision makers with the information needed to navigate this rapidly evolving landscape.

What is PUE?

Power Usage Effectiveness (PUE) is the fundamental metric used to measure the energy efficiency of a data center. It is calculated by dividing the total energy consumption of the facility by the energy consumed solely by the IT equipment (PUE = Total Facility Power / IT Equipment Power). An ideal PUE value is 1.0, indicating that all power consumed goes directly to running the servers with zero overhead for cooling, lighting, and power delivery losses. The current industry average PUE, however, remains around 1.58, meaning nearly 60% of energy is spent on non-computing functions. Data centers rely heavily on maintaining a low PUE to reduce operational costs (OPEX) and meet critical sustainability goals.

The massive global AI buildout for training and inference is dramatically increasing the IT power density of racks and total facility consumption. These high-density AI clusters (e.g., using specialized GPUs) place unprecedented thermal loads on existing cooling infrastructure, making it harder to maintain a low PUE. Consequently, achieving a PUE close to 1.0 is becoming even more vital, as any inefficiency (overhead) is multiplied by the enormous increase in core IT power. The economic and environmental pressure from the AI surge is forcing operators to adopt highly efficient solutions like liquid

cooling and advanced evaporative cooling to reject heat while keeping the overhead power negligible. Ultimately, the PUE metric serves as a crucial economic gatekeeper, as the viability of the energy-hungry AI infrastructure depends on maximizing computational output per unit of electricity consumed.

History of Liquid Cooling

While AI is driving innovation in how server infrastructure is cooled, this field is not new, in fact, far from it. In the early days of high-performance computing, liquid cooling was a necessity rather than a luxury because the vacuum tubes and early transistors generated immense heat within dense frames. IBM pioneered this field in the 1960s with the System/360 Model 91, the precursor to the modern z17 mainframe systems that were launched in April of 2025, which utilized a complex internal refrigeration system to maintain operational stability.

This was further refined with the introduction of the Thermal Conduction Module (TCM) in the 3081 series, which used water-cooled cold plates to draw heat directly from the chip logic. Simultaneously, Seymour Cray pushed the boundaries of supercomputing by incorporating liquid Freon into the Cray-1 to manage its power-hungry Emitter-Coupled Logic (ECL) circuits. HPE remains rightly proud of the Cray legacy of innovation, and took a few Equities and Industry analysts to tour the Cray museum and its production facilities recently.

By the mid-1980s, the Cray-2 took a more radical approach by becoming the first commercially successful machine to use total liquid immersion. Its processors were completely submerged in Fluorinert, a non-conductive dielectric fluid from 3M that circulated in a visible waterfall to dissipate heat from its three-dimensional circuit stacks. This immersion technique allowed for unprecedented density, as the components were packed so tightly that air could no longer flow between them. While liquid cooling fell out of fashion in the 1990s as energy-efficient CMOS technology took over, these early innovations from IBM and Cray laid the essential groundwork for the modern liquid-to-chip systems used in today's AI data centers.

Types of Datacenter Cooling

When evaluating data center cooling, the primary focus should be the Total Cost of Ownership (TCO), balancing upfront infrastructure costs against long-term energy savings. Traditional air cooling remains the most common choice due to its simplicity and lower initial investment, but it struggles to manage the extreme heat generated by modern, high-density AI and GPU clusters.

For these high-performance environments, direct liquid cooling is becoming a necessity because liquids can transfer heat up to 25 times more efficiently than air, allowing for much greater server density in a smaller footprint (according to Lenovo research).

Meanwhile, evaporative cooling offers a highly sustainable middle ground by using the natural evaporation of water to cool the air, significantly reducing electricity consumption in the right climates. However, decision makers must also weigh local environmental factors, such as water scarcity and humidity levels, which can limit the effectiveness of evaporative systems. Ultimately, the choice depends on finding the sweet spot between your hardware's specific thermal demands and your organization's sustainability goals.

Direct Liquid Cooling

Direct liquid cooling (DLC), commonly known as direct-to-chip cooling, represents a paradigm shift in thermal management by targeting heat at its point of origin. Rather than attempting to chill the air surrounding a server, this method utilizes specialized cold plates that are physically attached to high-heat components like CPUs and GPUs. Inside these plates, microchannels facilitate the flow of a liquid - either treated water or a dielectric fluid - which absorbs thermal energy through direct conduction, preventing the heat from ever leaking into the room's ambient environment.

The mechanical framework of a DLC deployment operates through a sophisticated, multi-tiered hierarchy designed for maximum reliability. At the individual server level, flexible hoses transport fluid to and from the cold plates, connecting them to a vertical rack manifold. This manifold aggregates the heated liquid from every server in the rack and routes it to a CDU. Acting as the bridge between the IT equipment and the

building's infrastructure, the CDU uses a heat exchanger to pass the energy from the sensitive internal loop to a robust facility water system for final rejection.

This technology is becoming a fundamental requirement for modern facilities because it overcomes the physical limitations of traditional air cooling. As specialized AI workloads drive rack densities toward 50-100 kW, the sheer volume of air required to keep chips from throttling becomes impossible to move. Because liquid is roughly 25 times more efficient at heat transfer than air, DLC enables operators to consolidate massive amounts of compute power into a fraction of the floor space, effectively retiring the need for massive fans and expansive aisle containment systems.

From a sustainability perspective, DLC drastically improves a data center's Power Usage Effectiveness (PUE) by allowing for much higher approach temperatures. Because liquid captures heat so effectively, the facility can often use warm water for cooling, which bypasses the need for energy-intensive mechanical refrigeration or chillers. Furthermore, because the heat is concentrated in a liquid loop rather than dissipated into the air, it becomes a high-grade energy source that can be easily repurposed for district heating or greenhouse climate control, turning waste into a valuable resource.

Traditional Air Flow - Hot Aisle/Cold Aisle

Hot aisle/cold aisle containment is a standard, essential practice in data center design used to maximize cooling efficiency and maintain optimal operating temperatures. The fundamental strategy involves arranging rows of server racks so that their air intakes face each other, creating cold aisles, while their exhausts face each other, creating hot aisles. This separation ensures that the cold air supply (typically 18 °C to 27 °C) from the Computer Room Air Handlers (CRAHs) is delivered directly to the equipment inlets.

The equipment then draws this cold air through the servers, where it absorbs the waste heat and is expelled into the hot aisle. This hot aisle contains the resulting exhaust air, often exceeding 30 °C, and directs it back to the cooling units for conditioning. Containment is achieved by physically isolating the two air streams using blanking panels, aisle-end doors, and ceiling barriers. Cold aisle containment encloses the cold air supply, while hot aisle containment encloses the exhaust air return.

By preventing the mixing of hot and cold air, the system eliminates recirculation and hot spots, which are the primary cause of cooling inefficiency. This methodology raises the return air temperature to the CRAHs, allowing the cooling

units to run more efficiently or use "free cooling" more often. Ultimately, implementing hot aisle/cold aisle containment is a simple yet powerful way to reduce cooling energy consumption and significantly lower the facility's Power Usage Effectiveness (PUE).

Evaporative Cooling

Evaporative cooling is a cornerstone of energy-efficient data center thermal management, relying on the thermodynamic principle of latent heat absorption to dissipate heat. As water evaporates, it absorbs significant energy (latent heat of vaporization) from the surrounding air, lowering the air's dry bulb temperature toward its wet-bulb temperature. This process is highly efficient, as the main power consumers are only fans and pumps, unlike energy-intensive mechanical chillers. Utilizing evaporative cooling can cut cooling costs by up to 70%.

Architectural Solutions and Climatic Constraints

Evaporative cooling is deployed in three main architectures to manage the critical conflict between energy efficiency and humidity control:

- **Direct Evaporative Cooling (DEC):** Outside air passes through a wet medium and is introduced directly into the data center. While the most efficient, it adds humidity, restricting its use primarily to dry climates to prevent condensation and equipment damage.

- **Indirect Evaporative Cooling (IEC):** This system uses a heat exchanger to isolate the data center air from the evaporatively cooled outside air. The heat is transferred across the exchanger, minimizing humidity and contamination risk. This isolation makes IEC the preferred, high-integrity strategy for a broader range of climates, despite a slightly higher energy profile than DEC due to requiring two fans.

- **Hybrid Systems & Water-Side Economizers:** Hybrid systems combine indirect pre-cooling with a controlled direct stage. Water-side economizers use evaporative heat rejection (via cooling towers) to chill a fluid, which then cools the data center air via internal coils. This is essential for high-density infrastructure as it facilitates warmer chilled water supply temperatures, optimizing liquid cooling systems.

The performance of all evaporative systems is fundamentally limited by the ambient wet-bulb temperature. In hot and humid regions, the high moisture content of the air diminishes the cooling effect, necessitating the activation of high-energy
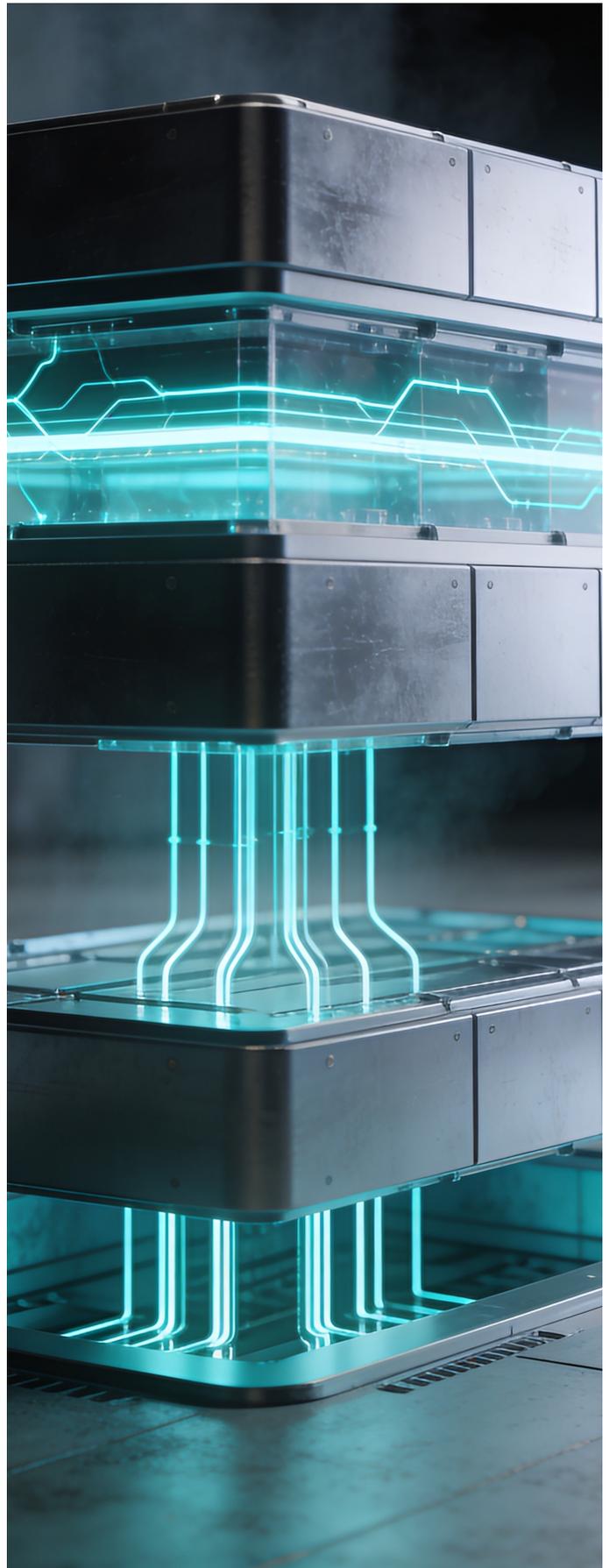
auxiliary cooling. This geographical constraint is addressed by advanced solutions like Liquid Desiccant (LD) assisted IEC, which pre-dehumidifies the air, significantly lowering the wet-bulb temperature. Case studies show LD-assisted systems can improve temperature drop by over 72% and reduce annual auxiliary cooling operation time from thousands of hours to negligible levels, dramatically expanding the technology's geographic viability.

Performance Metrics: The PUE-WUE Trade-Off

The viability of evaporative cooling is assessed using two metrics:

- **Power Usage Effectiveness (PUE):** The ratio of total facility power to IT equipment power. Evaporative cooling is the primary enabler of low PUE values (close to the ideal 1.0), with hyperscalers achieving an average of 1.1.

- **Water Usage Effectiveness (WUE):** The ratio of annual site water usage (liters) to annual IT energy consumption (kWh). High efficiency targets are 0.2L/kW or less.

A critical dynamic is the inverse relationship (paradox) between PUE and WUE. Strategies to achieve a low PUE (using evaporation instead of mechanical chilling) inherently increase water consumption, raising the WUE. Strategic balancing is required based on local conditions: prioritizing low PUE in areas with a high-carbon power grid, or prioritizing low WUE in water-stressed regions. A holistic view, considering the water consumed to generate electricity for high-PUE mechanical

systems, suggests low-PUE evaporative systems may offer a superior overall water footprint.

Operational Management and Future Trends

Implementing evaporative cooling shifts operational complexity from energy management to water management. Essential requirements include:

- **Water Chemistry Control:** Continuous monitoring of the cycles of concentration (COC) and use of automated conductivity controllers to manage blowdown and prevent corrosion, scale, and biological fouling (including Legionella risk).

- **Redundancy:** Robust design (e.g., N+1, 2N) is essential for resilience against failures, requiring detailed O&M plans for switchover sequences between evaporative and mechanical backup.

A critical trade-off exists between PUE and WUE in data center design. Strategies to achieve low PUE (such as evaporative cooling) typically increase water consumption, resulting in higher WUE values. Strategic balancing is required based on local conditions: prioritizing low PUE in regions with high-carbon power grids, or prioritizing low WUE in water-stressed areas. A holistic analysis that accounts for the water consumed in electricity generation for high-PUE mechanical systems may show that low-PUE evaporative systems offer a better overall water footprint in certain contexts.
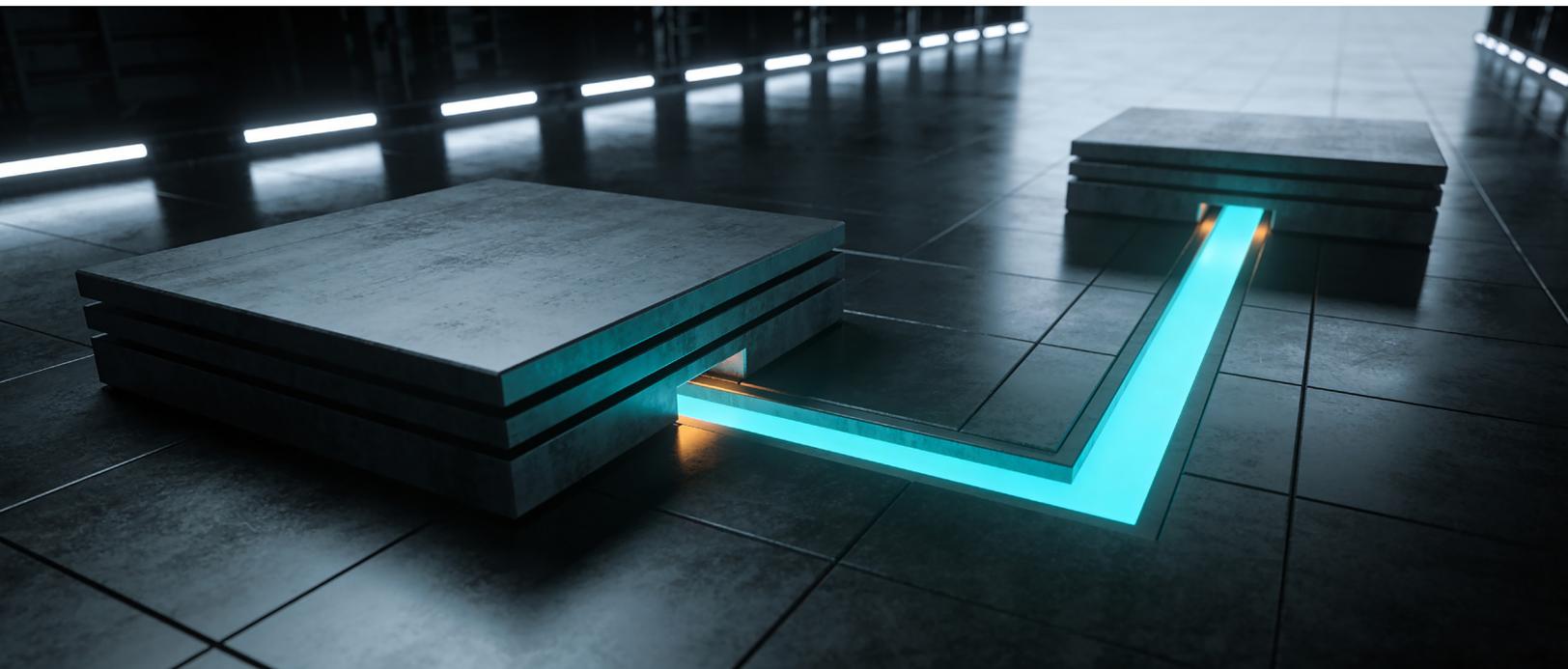
Rear Door Heat Exchangers

A Rear Door Heat Exchanger (RDHx) functions as a specialized thermal barrier that intercepts heat at the rack's exit point, preventing it from ever heating up the broader data center floor. By swapping a conventional perforated rear door for a liquid-cooled radiator coil, the system uses the server's existing internal fans to push hot exhaust through the chilled medium. This process transfers thermal energy into a circulating fluid - usually water - resulting in room-neutral air that exits the rack at the same temperature as the surrounding environment, effectively isolating the cooling load to the rack itself.

Dell spotlighted this concept at Dell Technologies World 2025 by unveiling the PowerCool Enclosed Rear Door Heat Exchanger (eRDHx). This solution is specifically engineered to support the massive thermal output of AI-driven hardware, offering 80 kW of cooling capacity per individual rack. Unlike open-loop systems, the enclosed architecture of the eRDHx creates a self-contained airflow environment that recirculates air internally; this design lowers fan energy consumption and allows the system to function efficiently without relying on costly facility-wide chillers.

The adoption of such technology can prove vital for organizations looking to modernize their infrastructure for high-density AI without a total facility reconstruction. By using warmer facility water - typically between 32 °C and 36 °C - the eRDHx can slash cooling-related energy expenses by up to 60% compared to traditional air-conditioning methods. This approach not only optimizes PUE but also allows for a 16% increase in compute density within the same energy budget, providing a scalable path for expanding AI capacity in existing data centers.

What Are The Major Vendors Doing In this Space

The industry-wide shift toward high-density AI infrastructure has transformed data center cooling from a facility utility into a core architectural requirement. Leading technology providers are standardizing on DLC and warm-water cooling to manage the extreme thermal loads of next-generation chips like NVIDIA's Blackwell. While hardware manufacturers such as Lenovo (Neptune) and HPE focus on achieving near-100% heat removal and fanless designs, cloud giants like Google, AWS, and Microsoft Azure are integrating proprietary sidekick heat exchangers and AI-driven autonomous optimization to reach ultra-low PUE.

A parallel focus on sustainability has emerged, with companies like Oracle, Equinix, and Azure pivoting toward closed-loop, water-neutral systems that repurpose waste heat for local communities. We believe that the market is moving toward a hybrid-ready, modular future where flexible cooling architectures from vendors like Dell, Cisco, and Vertiv enable operators to scale from traditional air to high-performance liquid cooling as their AI workloads evolve.

Lenovo

The Lenovo NeptuneTM line of servers is the company's flagship high-performance computing (HPC) and AI solution, leveraging direct water cooling to manage immense thermal loads. These systems use warm-water cooling technology, piping liquid directly to heat-generating components like the CPU, GPU, and memory. This method can remove up to 100% of the heat and allows data centers to operate with significantly reduced or even eliminated air conditioning. Consequently, Neptune servers can deliver up to a 40% reduction in data center energy consumption and allow processors to run in continuous turbo mode for a % performance boost.

Lenovo's capability in liquid cooling is a direct inheritance from IBM's pioneering history in thermal management. In 2012, this expertise was extended to x86 based systems for HPC and later AI. By acquiring IBM's x86 server business, Lenovo gained decades of expertise and foundational patents in robust, large-scale liquid cooling design.

Neptune modernizes this legacy with innovations like high-temperature water tolerance (up to 45 °C), eliminating the need for expensive chillers. The use of custom copper loops and aerospace-grade dripless connectors ensures reliability for today's high-density AI racks.   Lenovo has extended their Neptune direct cooling to other liquid cooling technologies that utilize liquid in air cooled systems. Neptune Core uses an open cooling loop for CPUs and memory in standard 1U/2U systems. Neptune Air uses a closed glycol loop to remove CPU heat without having to add plumbing.

HPE

HPE's data center strategy focuses on a high-density, energy-efficient architecture that integrates DLC to manage the extreme thermal demands of modern AI and high-performance computing. Leveraging over 50 years of expertise and 300 patents, HPE has pioneered "100% fanless" systems that pump coolant directly to the hottest components like the NVIDIA Blackwell GPUs and next-generation CPUs.

This approach is highly efficient because liquid can have roughly 25 times the thermal conductivity of air, allowing HPE to remove heat up to 3,000 times more effectively than traditional fans. By eliminating the need for bulky air-cooling infrastructure, HPE's modular data center designs can reduce required physical space by up to 77.5% (according to research cited in HPE's "The Benefits and Implementation of Modular Data Centers").

Sustainability is a core pillar of this strategy, as DLC enables the capture of waste heat that can be recycled for district heating or agricultural projects, such as greenhouses. Furthermore, these systems can reduce data center energy consumption by up to 87% and operational costs by 86% compared to traditional air-cooled setups. Through platforms like the HPE Cray Supercomputing GX5000, the company provides a scalable, modular framework that allows organizations to deploy massive AI clusters with a significantly lower carbon footprint.

Dell

Dell's strategy is built on a hybrid cooling philosophy that balances the extreme efficiency of liquid with the simplicity of air to maximize rack density. A cornerstone of this approach is DLC, which uses factory-installed cold plates to remove heat directly from high-TDP components like CPUs and GPUs, significantly reducing the energy required by chassis fans. To extend the life of air-cooled facilities, Dell also utilizes Smart Flow designs, which rearrange internal server components to increase airflow by up to 52% and allow for higher-performance chips without a full liquid transition.

For large-scale AI deployments, Dell offers the Integrated Rack 7000 (IR7000) and the PowerCool Enclosed Rear Door Heat Exchanger (eRDHx), a self-contained system that can capture 100% of IT-generated heat. This enclosed loop is particularly innovative because it can utilize warmer facility water, reducing the data center's reliance on energy-intensive chillers and cutting cooling power consumption by as much as 60%. These systems are managed through the OpenManage Power Manager software, which uses AI-driven insights and real-time telemetry to adjust thermal policies across rows and racks. By integrating these technologies, Dell enables organizations to support high-density AI Factories with a Power Usage Effectiveness (PUE) as low as 1.14, as seen in co-designed projects like the Dawn AI supercomputer.
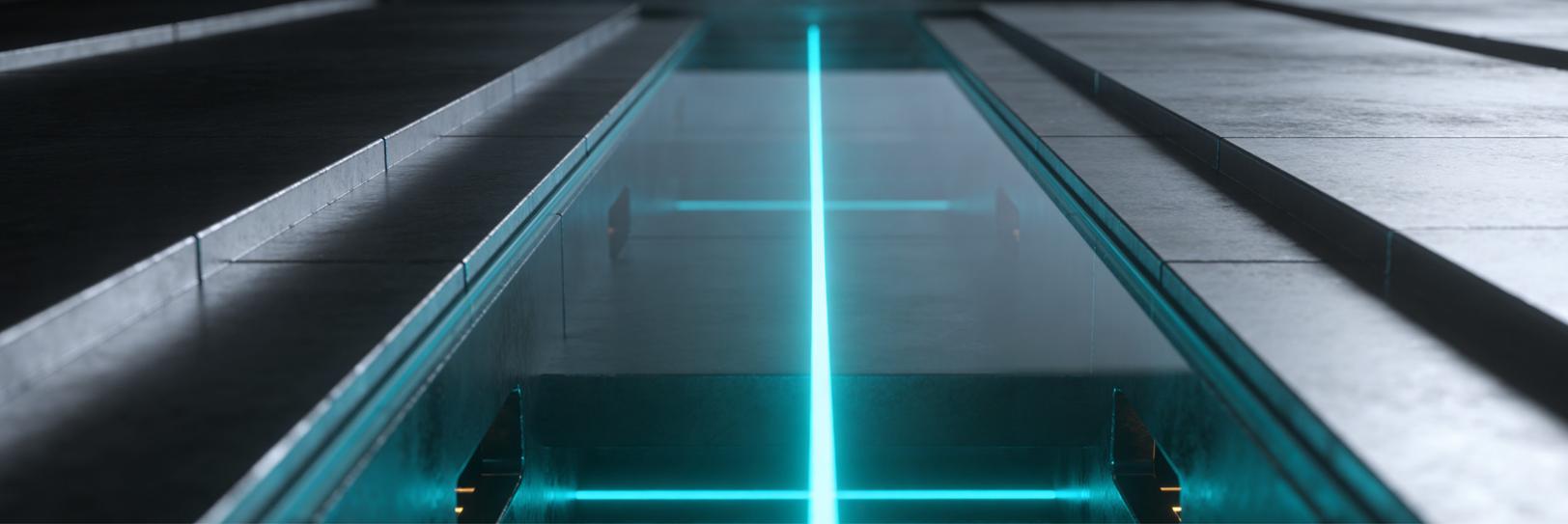
AWS

AWS uses a flexible cooling strategy that blends free outside air cooling, direct evaporation, and proprietary direct-to-chip liquid cooling technologies to manage the thermal loads of hyperscale AI infrastructure. To support the extreme power density of the most powerful AI chipsets, AWS developed the In-Row Heat Exchanger (IRHX), a modular liquid-cooling system that places cold plates directly on high-performance chips. Unlike standard industry solutions, the IRHX features a decoupled design that separates the pumping unit from fan-coil modules. Unlike many publicly available options, the IRHX can scale efficiently by adding or subtracting fan coil units depending on the amount of cooling needed by the row. This allows AWS to scale rapidly, retrofitting existing air-cooled data centers without major structural renovations, meaning it can be deployed in existing data centers as well as new builds.

This flexible cooling approach effectively reduces heat from custom silicon, ensuring they maintain peak performance during complex AI training and inference. AWS's IRHX can support a wide range of racks requiring liquid cooling, uses 9% less water than fully-air cooled sites, and offers a 20% improvement in power efficiency compared to off-the-shelf solutions. AWS has committed to becoming water positive by 2030, a goal supported by the use of recycled wastewater at over 120 data center sites. As such, AWS can achieve a global PUE as low as 1.15, significantly outperforming traditional enterprise benchmarks.

Google

Google's data center cooling strategy is built on a geographically aware, hybrid model that prioritizes extreme energy efficiency and high-density performance for its AI infrastructure. A pioneer in the field, Google utilizes a custom DLC system, specifically

for its latest TPU v7 (Ironwood) chips, which can support power outputs in the 10 MW range while maintaining temperature stability within ±2 °C. To manage these massive thermal loads, Google employs "direct-to-chip" cooling loops that circulate fluid through cold plates, removing heat up to 4,000 times more effectively than air and allowing for significantly denser server racks.

In addition to hardware innovations, Google leverages DeepMind-powered AI algorithms to autonomously optimize its cooling parameters in real-time, a strategy that has historically reduced cooling energy consumption by up to 40% (according to the Google DeepMind report). This technological stack is paired with a strict sustainability commitment: Google aims to be water positive by 2030, replenishing 120% of the freshwater it consumes. To achieve this, the company adapts its cooling method to the local environment, using recycled wastewater and evaporative cooling in temperate zones, while shifting to advanced air-cooling in water-scarce regions like Texas. By integrating these advanced thermal designs, Google maintains a global fleet-wide PUE of approximately 1.10, far outperforming the industry average.

Microsoft

Microsoft Azure's data center cooling strategy is built on a zero-water commitment, standardizing on closed-loop DLC for all new AI infrastructure to eliminate evaporative water loss. To support the immense 1,200W+ thermal demands of NVIDIA Blackwell and the newly announced Vera Rubin platform, Azure utilizes a sidekick cooling unit, a rack-adjacent heat exchanger that circulates fluid directly to cold plates on high-heat components. Beyond standard cold plates, Microsoft is pioneering microfluidic cooling, which etches tiny channels directly into the silicon to remove heat up to three times more effectively than traditional methods.

The company has also successfully moved two-phase immersion cooling into production environments, where servers are submerged in a dielectric fluid that boils and condenses to manage extreme heat densities. These innovations are critical to Azure's goal of becoming water positive by 2030, as they allow facilities to operate efficiently even in water-stressed regions without relying on traditional chillers. To ensure industry-wide scalability, Microsoft has open-sourced its Heat Exchanger Unit designs through the Open Compute Project, enabling other providers to retrofit existing air-cooled data centers for liquid-cooled AI clusters.

Microsoft needs to move its microfluidic R&D from lab-scale testing to wide-scale production for its custom Maia and Cobalt silicon to gain a distinct performance-per-watt advantage over generic cloud offerings. Following the success of its European projects, Azure should expand its waste heat recovery initiatives to North American hubs, piping the captured thermal energy from liquid-cooled racks into local district heating grids.

Equinix

Equinix's cooling strategy centers on a high-flexibility, liquid-ready architecture designed to support the rapid transition from traditional air-cooled racks to high-density AI workloads. At the core of this transition is a commercial commitment to support Direct-to-Chip liquid cooling across more than 100 of its global International Business Exchange (IBX) data centers, enabling racks to reach power densities exceeding 100 kW.

This approach allows customers to choose their preferred cooling method, including single-phase and two-phase direct-to-chip systems, while utilizing Equinix's controlled facility water loops as a demarcation point. To maintain extreme efficiency, the company employs a multimodal strategy that includes

augmented air cooling through Rear-Door Heat Exchangers and high-temperature chilled water set points, which collectively reduced its global PUE to 1.39 in 2024.

Sustainability is deeply integrated into this thermal strategy; Equinix aims for a global Water Usage Effectiveness (WUE) of 0.95, utilizing non-potable water and deep-lake cooling in specific regions to minimize environmental impact. Furthermore, Equinix is a leader in heat export initiatives, capturing waste heat from its cooling loops to provide low-carbon heating for thousands of homes and local community facilities, such as Olympic-sized swimming pools. By standardizing on a 30 °C coolant temperature through its work with the Open Compute Project, Equinix ensures its cooling infrastructure remains viable for next-generation silicon while driving toward a long-term goal of chiller-free operations.

Cisco

Cisco's data center cooling strategy is defined by a modular, liquid-ready approach that enables a transition from air-cooled environments to high-density AI infrastructure. At the center of this strategy is the UCS X-Series modular system, which is engineered to support future liquid cooling upgrades without requiring a complete hardware overhaul.

For high-performance networking, Cisco has introduced a breakthrough liquid-cooled 51.2T switch that utilizes a hybrid thermal design to dissipate nearly 80% of the heat generated by the CPU, NPU, and high-speed optics. This system is designed for efficiency, supporting warm-water cooling with inlet temperatures up to 45 °C, which can reduce overall data center power consumption by up to 28% by minimizing the need for energy-intensive chillers (according to the 2024 Cisco technical announcement and performance report released at SC24).

HyperFRAME RESEARCH

Cisco also leverages a robust ecosystem of Engineering Alliances with specialists like DeepCoolAI and LiquidStack to offer a variety of solutions, including RDHx and advanced two-phase immersion cooling. Through its Secure AI Factory with NVIDIA, Cisco provides validated reference architectures that integrate these cooling technologies to maintain the signal integrity and thermal stability required for massive GPU clusters. This comprehensive strategy supports Cisco's goal of reaching net-zero emissions by 2040 by optimizing power usage at the component, rack, and facility levels.

Oracle

Oracle's data center cooling strategy is centered on a radical shift toward liquid-ready infrastructure to support the unprecedented thermal demands of AI superclusters. For its latest deployments, such as the NVIDIA GB200 NVL72 racks, Oracle has fully adopted direct-to-chip cold plate liquid cooling, which targets the hottest components, including CPUs, GPUs, and NVLink switches. This transition is essential for Oracle's AI Factories, where rack densities are climbing toward 100 kW and beyond, requiring heat removal that is thousands of times more efficient than air.

To address the environmental impact often associated with large-scale cooling, Oracle is pioneering the use of closed-loop, non-evaporative cooling systems. Unlike traditional evaporative methods that consume millions of gallons of water daily, Oracle's closed-loop design requires only an initial fill, keeping annual water consumption on par with a standard office building. This "water-neutral" approach is a key part of Oracle's broader goal to achieve 100% renewable energy for its global data center operations by 2025.

Furthermore, the company utilizes advanced real-time telemetry and OCI-native APIs to monitor thermal health at the rack level, allowing for predictive optimizations that prevent hotspots before they cause performance throttling. By integrating these liquid loops with specialized CDUs and fault-tolerant plumbing, Oracle can scale its OCI Superclusters to over 131,000 GPUs while maintaining a market-leading Power Usage Effectiveness (PUE).

NVIDIA

NVIDIA's cooling strategy has shifted from an optional enhancement to a core architectural requirement, specifically designed to handle the 1,200W+ thermal design power (TDP) of its latest Blackwell GPUs. At the center of this strategy is the GB200 NVL72, a rack-scale system co-engineered from the ground up for direct-to-chip liquid cooling to prevent thermal throttling in trillion-parameter AI models. By utilizing

specialized cold plates and a blood circulation network of coolant distribution units (CDUs), NVIDIA can maintain GPU temperatures between 46 °C and 54 °C - significantly lower than the 71 °C common in air-cooled setups.

A key pillar of NVIDIA's approach is warm-water cooling, which supports inlet temperatures as high as 45 °C (113 °F). This high thermal tolerance allows data centers to utilize free cooling from ambient outdoor air, potentially eliminating the need for energy-intensive mechanical chillers and reducing cooling costs by up to 25x. Furthermore, NVIDIA claims that Blackwell-based AI factories can achieve 300x greater water efficiency than air-cooled architectures, a critical metric for hyperscalers aiming for water-positive goals.

By integrating these liquid loops with its NVLink Switch System, NVIDIA enables a single rack to act as a unified, liquid-cooled super GPU with 72 interconnected processors. This strategy not only slashes the physical footprint of AI infrastructure by 75% but also optimizes PUE toward an industry-leading 1.10. Looking forward, NVIDIA's newly announced Vera Rubin platform (slated for 2026) is expected to push these boundaries even further, potentially operating in chiller-free environments to maximize the performance-per-watt of next-generation AI superfactories.

Vertiv

Vertiv's data center cooling strategy is currently centered on a rapid pivot toward liquid cooling to support the extreme heat densities of AI-driven gigawatt-scale facilities. By moving beyond traditional air cooling, they are prioritizing direct-to-chip and immersion cooling technologies to handle rack densities that now frequently exceed 100kW. A core pillar of their approach is the hybrid-ready design, exemplified by products like the CoolPhase Flex, which enables operators to transition from air to liquid cooling seamlessly as their workloads evolve. Furthermore, Vertiv is integrating digital twin technology to simulate thermal performance virtually, significantly reducing the time-to-token for new AI deployments.

The company is also doubling down on adaptive liquid cooling, using AI itself to monitor and predict potential system failures or fluid inefficiencies in real time. We see that Vertiv should further integrate its cooling systems with higher-voltage DC power architectures to minimize energy loss during the power-to-cooling conversion process. Also, following the recent acquisition of PurgeRite, the company should focus on globalizing their specialized liquid cooling maintenance services to ensure the long-term reliability of complex, closed-loop systems.

Vultr

Vultr's cooling strategy is built on a location-aware, hybrid approach that leverages the specific climatic advantages of its 32 global regions. To support its high-density NVIDIA and AMD GPU clusters, Vultr partners with leading infrastructure providers such as Digital Realty and Sabey Data Centers to utilize specialized cooling environments. In cooler climates, such as their Seattle (SDC Columbia) site, the strategy focuses on free air cooling and sustainable hydropower, achieving an impressive annualized PUE of 1.15. For its most intensive AI workloads, Vultr is increasingly shifting from traditional air cooling to DLC and rear-door heat exchangers to manage the extreme heat of modern chips.

This transition enables the company to support power densities that far exceed the limits of legacy air-cooled systems while maintaining a smaller physical footprint. By integrating these systems into PlatformDIGITAL, Vultr ensures that thermal management is optimized for both energy efficiency and operational reliability. Furthermore, Vultr's commitment to sustainability includes a push toward net-zero carbon emissions by selecting facilities that prioritize renewable energy and minimal water waste. Collectively, we see this strategy as enabling Vultr to deliver high-performance cloud compute that is both scalable and environmentally responsible.

Key Strategy Comparison

Company	Core Strategy	Signature Technology	Sustainability Goal
Lenovo	Warm-water efficiency	Neptune™ Direct Water Cooling	40% energy reduction
HPE	100% Fanless architecture	Cray Supercomputing GX5000	Waste heat recycling
Dell	Hybrid/Modular flexibility	Smart Flow & IR7000 Rack	PUE of 1.14
AWS	Retrofit-ready DLC	In-Row Heat Exchanger (IRHX)	Water positive by 2030
Google	AI-optimized hybrid	TPU-specific custom DLC	1.10 global PUE
Microsoft	Zero-water immersion	Two-phase immersion/ Sidekicks	Water positive by 2030
NVIDIA	Integrated "Super GPU"	GB200 NVL72 Liquid Loop	300x water efficiency
Equinix	Liquid-ready colocation	100+ liquid-ready IBX centers	Community heat export
Cisco	Modular networking	Liquid-cooled 51.2T Switch	Net-zero by 2040
Oracle	Water-neutral clusters	Non-evaporative closed-loops	100% renewable by 2025
Vertiv	Infrastructure-as-a-Service	Digital Twin & CoolPhase Flex	High-voltage DC integration
Vultr	Location-aware hybrid	DLC & Rear-Door Heat Exchangers	Net-zero carbon by 2029

Source: HyperFRAME Research

Key Trends And Observations

As the data center landscape evolves to meet the demands of AI workloads such as generative AI, cooling has shifted from a background utility to a central pillar of infrastructure strategy. Below are the key trends and observations shaping the industry's path forward.

Data center decision makers must prepare for a definitive shift toward thermal orchestration, a paradigm where cooling evolves from a passive utility into a strategic, AI-driven asset. This transition is essential for managing unprecedented rack densities that traditional infrastructure can no longer sustain. By integrating real-time intelligence into the cooling loop, operators can transform thermal management into a dynamic system that anticipates heat spikes and optimizes airflow or fluid delivery with surgical precision.

This shift is largely driven by the explosive growth of high-performance workloads, with AI chip Thermal Design Power (TDP) rapidly climbing toward 1,000W per chip. To meet these thermal demands, single-phase direct-to-chip liquid cooling is projected to reach nearly 50% market adoption by 2026. This technology is quickly becoming the industry standard, providing the necessary efficiency to cool high-density clusters while significantly reducing the energy overhead associated with traditional air-cooling methods.

Specifically, we find that while data modernization is a work in progress, 36.2% of organizations are already actively upgrading their data architectures to meet the intensive infrastructure demands of AI workloads (according to HyperFRAME Research).This data point signifies a critical shift toward high-density infrastructure, as traditional air-cooling systems are largely ineffective for the 40kW–150kW rack densities

required by modern AI data architectures. Consequently, these modernization efforts are the primary driver for the industry's transition to DLC and hybrid thermal management, enabling organizations to maintain operational stability while supporting the extreme heat output of next-generation GPU clusters.

What are the key things decision makers should be looking out for?

The primary development to watch is the transition from broad-room cooling to precision thermal orchestration. As AI chips push rack densities toward 50-100 kW, look for hybrid cooling designs that integrate both air and liquid systems. These setups typically use liquid to pull heat from the highest-intensity components while maintaining traditional air containment for secondary hardware, ensuring that legacy facilities can be retrofitted without a total reconstruction.

Secondly, keep an eye on the shift toward closed-loop secondary circuits. Manufacturers such as Dell, with their PowerCool eRDHx, HPE CDU/CLLC Cray EX & ProLiant Compute Gen12 solutions, and Lenovo's Neptune systems, are setting a new standard by using sealed loops filled with glycol or treated water. These systems operate as thermal bridges that capture heat at the source and transport it to the facility exterior without ever exposing the coolant to the open air, a move that drastically simplifies maintenance and protects sensitive IT equipment from contamination.

What is fact, what is fiction?

The industry is moving toward a water-neutral future where cooling efficiency is no longer tied to local water supplies. By leveraging technologies that can tolerate higher fluid temperatures - often between 32 °C and 36 °C - data centers are increasingly able to use dry coolers. These systems reject heat using ambient air and large radiator coils, completely

eliminating the need for evaporative towers and the billions of gallons of water they typically consume.

Furthermore, we are heading toward a more transparent reporting era defined by the liters per token metric. Rather than focusing on staggering total facility volumes, this metric directly links the environmental cost to the AI output. This shift will help educate the public on how advanced silicon and high-performance hardware actually reduce the fluid consumption required to generate valuable information, turning the focus from gross consumption to computational efficiency.

Headlines suggesting that AI data centers are guzzling local drinking water are often sensationalist and lack critical context. In reality, most modern cooling systems are closed-loop, meaning the liquid continuously cycles through the system without being consumed or evaporated. Many reports also fail to differentiate between water withdrawal (water taken and returned to the source) and consumption (water lost to evaporation), leading to a persistent misunderstanding of the true environmental footprint.

It is also a misconception that evaporative cooling flushes water away in an open-loop waste stream. While these systems do use water, a significant portion is often recycled multiple times or sent to wastewater treatment after it has collected minerals and dirt from the cooling process. As high-density liquid cooling becomes the norm, the industry is proving that it can scale AI performance while actually reducing its reliance on public water utilities compared to older, less efficient facilities.

The massive global AI buildout is making the pursuit of a 1.0 PUE more than a goal; it is a business necessity. As energy costs and thermal loads multiply, every watt of overhead power spent on cooling is a watt that cannot be used for computation. This economic pressure is the ultimate driver of innovation, forcing the adoption of direct-to-chip cooling that is 25 times more efficient than air, allowing for 16% more compute density in the same power footprint.

Ultimately, progress in AI cooling is inextricably linked to energy recovery. Because liquid captures heat at higher concentrations than air, it is much easier to recycle that thermal energy for secondary uses, such as district heating or greenhouse climate control. This turns the waste heat of an AI factory into a valuable resource for the surrounding community, transforming the data center from a resource consumer into an integrated component of a sustainable urban energy grid.
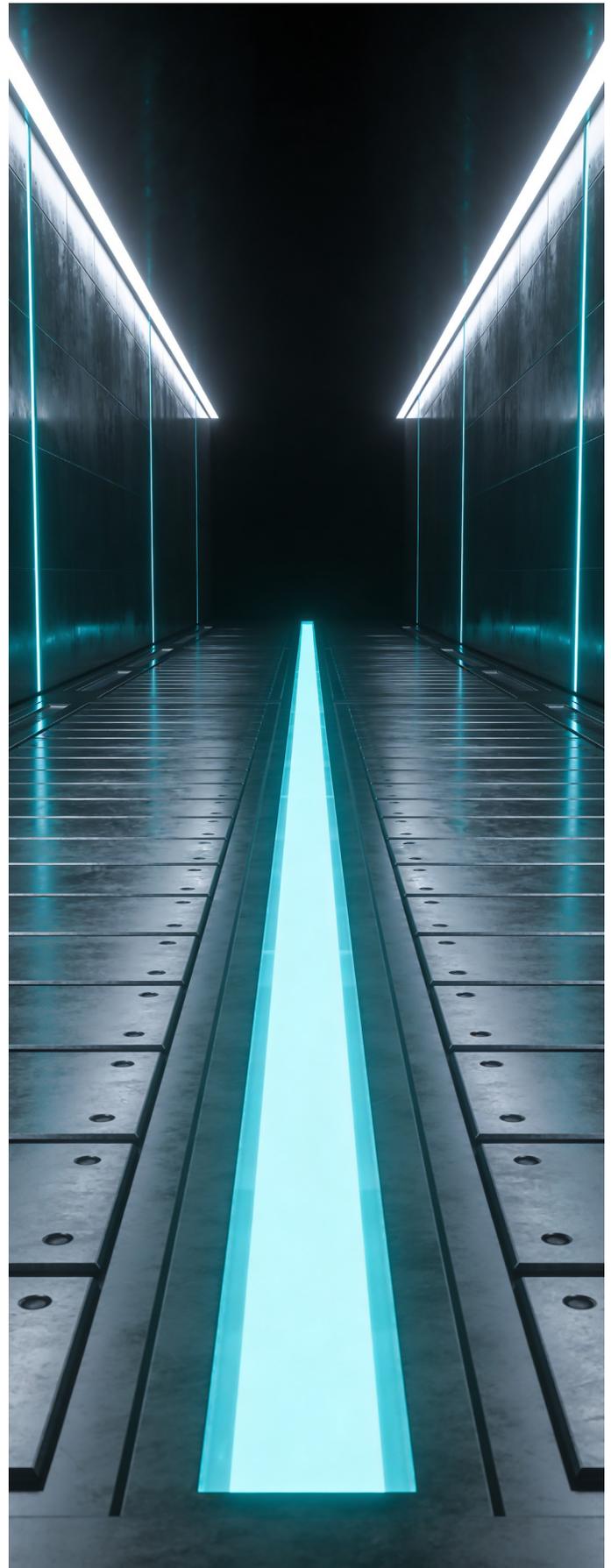
Looking Ahead

In 2026, data center decision makers should prioritize a shift in perspective from sensationalized "total volume" headlines toward a more nuanced, output-based transparency. A key development to watch is the adoption of the liters of water per token metric, which directly links water consumption to AI output. This shift aims to counter misleading narratives that overlook the scale of water use in other sectors, such as agriculture or traditional coal-fired power generation, which can consume up to 72 liters per kilowatt-hour. By focusing on computational efficiency, operators can better demonstrate how high-performance silicon and modern cooling architectures actually reduce the fluid consumption required to generate valuable information.

The industry is moving rapidly toward a water-neutral future by standardizing closed-loop secondary circuits and dry cooling technologies. Solutions like Lenovo's Neptune systems and Dell's PowerCool eRDHx utilize sealed loops of glycol or treated water, functioning as thermal bridges that eliminate the need to draw from public water supplies for evaporation. By leveraging systems that tolerate higher fluid temperatures (typically 32 °C to 36 °C), decision makers can utilize dry coolers that reject heat using ambient air and radiator coils. This transition not only simplifies maintenance and protects equipment from contamination but also effectively decouples data center scaling from local water scarcity concerns.

Moreover, data center decision makers should prepare for a definitive shift toward thermal orchestration, where cooling evolves from a passive utility into a strategic, AI-driven asset capable of managing unprecedented rack densities. With AI chip TDP climbing toward 1,000W per chip, liquid cooling - specifically single-phase direct-to-chip - is projected to reach nearly 50% market adoption as the industry standard for high-performance workloads.

However, the year's hallmark will be the rise of hybrid cooling environments, as operators leverage advanced air-cooling retrofits (like Rear Door Heat Exchangers) to support legacy hardware alongside new liquid-cooled AI clusters. Beyond hardware, digital twins and real-time AI controls will become essential for navigating the energy-water trade-off, enabling facilities to optimize WUE in response to tightening global environmental regulations and grid constraints.

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research
Email Address:
steven.dickens@hyperframeresearch.com
Telephone Number:
+1 845 505 1678
X: @StevenDickens3
LinkedIn: Steven Dickens
BlueSky: Steven Dickens

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.