INTELLIGENCE UPDATE

Next-gen GPUs may not need chillers — but data centers do

Daniel Bizo 3 Feb 2026

At the start of 2026, a casual remark by Nvidia chief executive Jensen Huang created turbulence in the data center cooling industry. In a speech at the annual Consumer Electronics Show (CES), Huang said Nvidia's next generation of rack-scale compute systems set to debut in 2026 will accept water coolant temperatures as high as 45°C (113°F) despite using double the power. This means, Huang said, that: "No water chillers are necessary, we're basically cooling [...] with hot water."

Such a remark about a detail of facility infrastructure should amount to no more than a casual talking point within engineering circles. But everything Nvidia does is scrutinized microscopically, and every detail seems huge. In the hours following Huang's presentation, shares of chiller vendors fell, in some cases by as much as 10%. Most have yet to fully recover.

What Huang said was not news. Nvidia's current rack-scale systems (marketed as DGX GB200/GB300) already support 45°C inlet water coolant temperatures. Yet, most Nvidia customers and their facility operators are not ditching chillers, instead they prefer to operate at lower temperatures.

The concept of a chiller-free data center is not an new one either. Uptime analysts identified chiller-free cooling as a disruptive technology trend as far back as 2013, and since then several air-cooled data centers have been built that operate without compressors. These facilities are in climates where evaporative/adiabatic coolers can keep air supply temperatures under 27°C (80.6°F) — the upper end of ASHRAE's recommended envelope — with only a few hours of excursion per year.

Despite this, mechanical refrigeration remains predominant in data centers — even among operators that could theoretically do without it — and water chillers are becoming more popular than ever. This is due to clear technical and business reasons:

- **Direct liquid cooling (DLC):** After a period of moving toward direct expansion (DX) and air-to-air heat exchange systems with adiabatic and/or evaporative cooling, the data center industry has turned its attention back to chiller plants. Data centers are expected to install more chilled facility water systems than ever before during the current construction boom

in support of current and future DLC installations. While large-scale GPU compute is a major force behind this, it is not the only one: high-performance technical computing and dense virtualized environments will also see more liquid-cooled hardware.

- **IT performance:** Modern CPUs and GPUs respond in real time to factors such as silicon hot spots and total thermal power as they seek to maximize speeds. Lower coolant temperatures — enabled by mechanical refrigeration — can therefore improve performance. Lower coolant temperatures also offer some safety margin against issues such as degradation in thermal contact with the cold plate and debris accumulation in the fluid. These are seemingly marginal benefits, but IT infrastructure costs an order of magnitude more than chillers and delivers direct business value. Running at higher temperatures may thus be a false economy, particularly in temperate climates.
- **Resiliency:** When systems operate near to their temperature tolerance limits, even a minor cooling disturbance can affect IT hardware. Simply switching from utility power to on-site engine generators may cause a temperature drift that is sufficient to trigger IT throttling or shutdown. While this risk may be acceptable for AI training applications, most organizations do not expect — or permit — facility infrastructure events to affect IT operations. High operating temperatures may require larger thermal energy storage tanks and/or continuous cooling supported by UPS or DRUPS (diesel-rotary UPS) systems to guarantee thermal stability.
- **Pooled capacity:** Operators already tend to overprovision IT room cooling capacity to anticipate future technical changes, often installing air handlers and coolant distribution units (CDUs) with a combined total of more than 140% of nominal IT load capacity. Even so, the preference is often to keep a shared facility water system across all loads — air- or liquid-cooled — to avoid the cost and complexity of duplicating the entire fluid network and mechanical plant. At 45°C coolant supply, however, AI training racks would need an entirely dedicated cooling infrastructure — dry coolers, pumps, tanks, pipes, valves and controls — in addition to air-cooling and any other (non-Nvidia) liquid-cooled loads, as many hardware configurations cannot readily accept such high coolant temperatures.
- **Facility lifecycle:** Data center infrastructure investments have a horizon of at least 10 years and are expected to be viable for 20-25 years. Only a handful of frontier AI startups and their leased providers will build for specific IT hardware with an expected useful life of less than five years — and even then, it is pragmatic to design a facility that can be upgraded. Limiting the design to high inlet temperatures may prove a significant technical constraint in the years ahead, reducing the value of the facility.

Data centers are undeniably edging toward higher operating temperatures to reduce infrastructure costs and cooling energy. Modern facility water temperatures typically range from 16°C to 22°C (around 61-72°F), already well above traditional chilled water loop temperatures of 7-10°C (44.6-50°F). As DLC loads increase relative to air-cooled loads, Uptime Intelligence expects the next temperature step to be approximately 26-29°C (79-84°F) for a separate cooling infrastructure for all DLC loads. This shift can already significantly reduce compressor sizing and energy.

While technically feasible, a chiller-free design is unlikely to be the right choice for most operators in the foreseeable future.

ABOUT THE AUTHOR

Daniel Bizo

4 Feb 2026

Over the past 15 years, Daniel has covered the business and technology of enterprise IT and infrastructure in various roles, including industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

dbizo@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com