Hot Water for Hot Chips

Thermodynamics, Climate and the Cooling Cost of 130 kW AI Racks

Abhay Chakra Sadineni
University of California, Berkeley
abhaychakra@berkeley.edu

November 2026
DOI: 10.5281/zenodo.24210933

Abstract

A rack of 72 Blackwell GPUs is provisioned for about 132 kW. Most of that heat leaves through cold plates in water that may enter at up to 45°C and return at up to 65°C, and the rest leaves through air. This paper works through the thermodynamics of that rack and asks how much of the heat-rejection energy and water a facility can avoid by running its water warm. Moving the rack's heat with air would take roughly 16,000 cubic feet of air a minute at a 15 K rise and up to 15 kW of fan power. Moving it with water takes about 130 litres a minute and well under 1 kW of pumping. The warm return stream carries an exergy of 11 to 14 percent of the heat, enough to matter for heat reuse but too little to generate useful power. Using 2025 hourly reanalysis weather for seven US data-centre markets, a dry-cooler plant feeding 45°C facility water would need a chiller for 0 hours a year in five markets, 15 hours in Dallas and 658 hours in Phoenix. If the 45°C limit applies at the rack and the coolant distribution unit needs a 5 K approach, the facility must supply about 40°C, and chiller hours rise to between 1 in Columbus and 1,715 in Phoenix, with little change in annual energy. At 17°C, the same plant would need a chiller for 4,400 to 8,000 hours. Blending in the rack's air-cooled share, annual heat-rejection energy falls from 0.07 to 0.15 kWh per kWh of IT load at 17°C to 0.035 to 0.055 at 40 to 45°C, with no water at all. An evaporative plant at 17°C saves some of that energy but uses 1.0 to 1.8 litres per kWh. At 45°C, the residual air-cooled share accounts for a quarter to two fifths of the cooling energy and nearly all the water. The cheapest kilowatt-hour of cooling in an AI hall is the one avoided by keeping the water hot and shrinking what still goes to air.

Keywords: data-centre cooling; direct liquid cooling; exergy; facility water temperature; water use; AI infrastructure.

1 Introduction

For two decades, data-centre racks drew a few kilowatts each and were cooled with air. In the 2024 Uptime Institute survey, the average facility still reported a typical rack density of about 8 kW [1]. In the 2025 survey, 82 percent of respondents said their densest rack drew under 30 kW [2]. NVIDIA's GB200 NVL72, by contrast, is built around racks provisioned for 132 kW through eight 33 kW power shelves [3]. Its compute trays are cooled by liquid pumped through cold plates, with a maximum inlet temperature of 45°C and a maximum return temperature of 65°C. About 115 kW of the rack's heat leaves in the liquid and 17 kW in air [4]. Vertiv's reference design for the platform, developed with NVIDIA, uses coolant at 45°C inlet and 65°C outlet [5].

Those temperatures are unusual for a data centre, and they change what a cooling plant has to do. A conventional chilled-water plant supplies water cold enough to need refrigeration for much

of the year in most of the United States. Water at 45°C can be cooled by blowing outside air over a radiator in almost any American summer. The question this paper asks is how much that is worth, in kilowatt-hours and in litres, and where the remaining costs sit.

Section 2 sets out why the rack cannot be cooled by air. Section 3 asks what the rejected heat is worth. Section 4 runs a year of hourly weather for seven US markets through a simple plant model. Section 5 draws out the design consequences.

2 Moving 132 kW

The heat a fluid stream carries is $\dot{Q} = \dot{m}\, c_p\, \Delta T$, where $\dot{m}$ is the mass flow, c_p the specific heat and ΔT the temperature rise. For air near 30°C, with $c_p = 1{,}006$ J/kg K and density 1.16 kg/m^3, carrying 132 kW at a 15 K rise needs 8.75 kg/s, or 7.5 m^3/s. That is about 16,000 cubic feet a minute through one rack. At a 10 K rise it is 24,000. An ASHRAE technical committee white paper put the air requirement of a 40 to 50 kW rack at up to 5,000 cubic feet a minute, and the supply from one of the best perforated floor tiles at about 1,900 [6].

Moving that air also costs power. The fan power is $\dot{V}\,\Delta p/\eta$. At a pressure rise of 250 to 1,000 Pa across servers, filters and coils, and a combined fan and motor efficiency of 0.5, the fans would draw 3.8 to 15 kW, or 2.9 to 11 percent of the rack's own load. That power becomes heat too, and has to be removed in turn.

Water has a volumetric heat capacity about 3,560 times that of air. The 115 kW that the GB200 rack sends to liquid, at the 130 litres a minute specified for the rack [4], implies a 12.8 K rise. At the full 20 K allowed between 45 and 65°C, the flow would fall to 83 litres a minute. Pumping 130 litres a minute against 100 to 150 kPa at an efficiency of 0.5 takes 0.43 to 0.65 kW, under half a percent of the rack. On transport alone, liquid wins by roughly an order of magnitude in power, 6 to 35 times across the ranges above, and by three orders in volume.

3 What the rejected heat is worth

A stream of water cooled from T_2 to T_1, in surroundings at T_0, can deliver at most a fraction

$$\frac{X}{Q} = 1 - \frac{T_0 \ln(T_2/T_1)}{T_2 - T_1} \tag{1}$$

of its heat as work, with all temperatures in kelvin [7]. For the GB200 return stream, from 65 down to 45°C, that fraction is 13.7 percent with surroundings at 10°C and 10.6 percent at 20°C. For a conventional liquid loop running from 45 to 32°C it is 9.1 and 5.9 percent, and for hot-aisle air at 40 to 25°C it is 7.3 and 4.1 percent.

The 115 kW leaving one rack in liquid therefore carries at most 12 to 16 kW of work potential. Real heat engines at these temperatures recover a small fraction of that, so generating power from the waste heat is not attractive. Using it directly as heat is a different matter. A 65°C stream can feed a district heating network or an industrial process with little or no heat-pump lift, where a 35°C stream needs a large one. Earlier work on hot-water cooling showed that supply temperatures of about 60°C could keep processors within limits while delivering usable heat [8], and a review of low-grade waste-heat recovery from data centres identifies the low temperature of the heat as the main barrier to using it [9]. The second law, in short, rewards a hot loop twice. It makes the heat easier to reject and easier to reuse.

4 Climate and the plant

4.1 Model

The facility water system takes heat from the racks' coolant distribution units and rejects it outdoors. ASHRAE's liquid-cooling classes name the facility water supply temperature that equipment must tolerate at full performance, W17, W27, W32, W40, W45 and W+ above that [10]. The model asks, hour by hour, which of three ways of rejecting heat can deliver water at the top of each class.

A dry cooler, a radiator with fans, can deliver water at the outdoor dry-bulb temperature plus an approach, taken here as 6 K. An evaporative cooling tower can deliver water at the wet-bulb temperature plus 4 K, at the cost of evaporating water. When neither suffices, a chiller makes up the difference. Its coefficient of performance is taken as half the Carnot value between an evaporator 2 K below the supply temperature and a condenser 8 K above the dry bulb (6 K above the wet bulb with a tower). Fans and pumps draw 2 percent of the heat rejected with a dry cooler and 1 percent with a tower, plus 1 percent for pumping. Evaporation uses 1.48 litres per kWh of heat rejected, from a latent heat of 2,430 kJ/kg, raised to 1.98 litres by blowdown at four cycles of concentration. Each of these values is an assumption typical of current equipment, not a measured property of any plant.

The rack's 17 kW of air-cooled heat is treated as going to air handlers served by water at W17, an assumption chosen to represent conventional coils. The blended result weights the liquid and air shares by 115 and 17 kW.

The ASHRAE classes describe facility water. The GB200's 45°C limit is stated for the liquid entering the rack, which in most designs comes from a coolant distribution unit fed by facility water a few kelvin colder. The W40 results below stand in for 45°C at the rack with a 5 K approach across that unit.

Several further simplifications are built in. In hours when the chiller runs, it carries the full load, with no partial pre-cooling by the dry cooler, which overstates chiller energy. The compressor's own heat is not added to the heat rejected, which slightly understates fan energy and water. The Carnot lift has a floor of 3 K. The tower is used only in hours when the dry cooler fails. Air-handler fan power is not counted.

Weather is the 2025 hourly dry-bulb and wet-bulb temperature from the ERA5-based Open-Meteo historical archive [11, 12] for seven sites in established or fast-growing data-centre markets. The sites are Ashburn, Virginia (39.04°N, 77.49°W), Dallas (32.90°N, 97.04°W) and Abilene (32.45°N, 99.73°W), Texas, Phoenix, Arizona (33.45°N, 112.07°W), Hillsboro, Oregon (45.52°N, 122.99°W), Columbus, Ohio (39.96°N, 82.99°W), and Council Bluffs, Iowa (41.26°N, 95.86°W). The model works on the annual distribution of each temperature in one-degree bins, evaluated at bin mid-points, with hours below −10°C pooled in one bin. For the evaporative plant, it assumes the hours too hot for a dry cooler are also the most humid hours. That overstates chiller use and energy in the evaporative plant. Water use does not depend on the assumption.

4.2 Results

Figure 1 shows the hours in which a dry cooler alone cannot reach each supply temperature. At W17 that is most of the year everywhere, from 4,442 hours in Hillsboro to 7,971 in Phoenix. The count falls steeply with each class. At W45 it reaches zero in five of the seven markets. Dallas needs a chiller for 15 hours, and Phoenix, where the temperature reached 46.9°C, for 658. At W40, which corresponds to 45°C at the rack after a 5 K approach, no market reaches zero. Columbus needs a chiller for 1 hour, Council Bluffs for 34, Hillsboro for 45, Ashburn for 50,

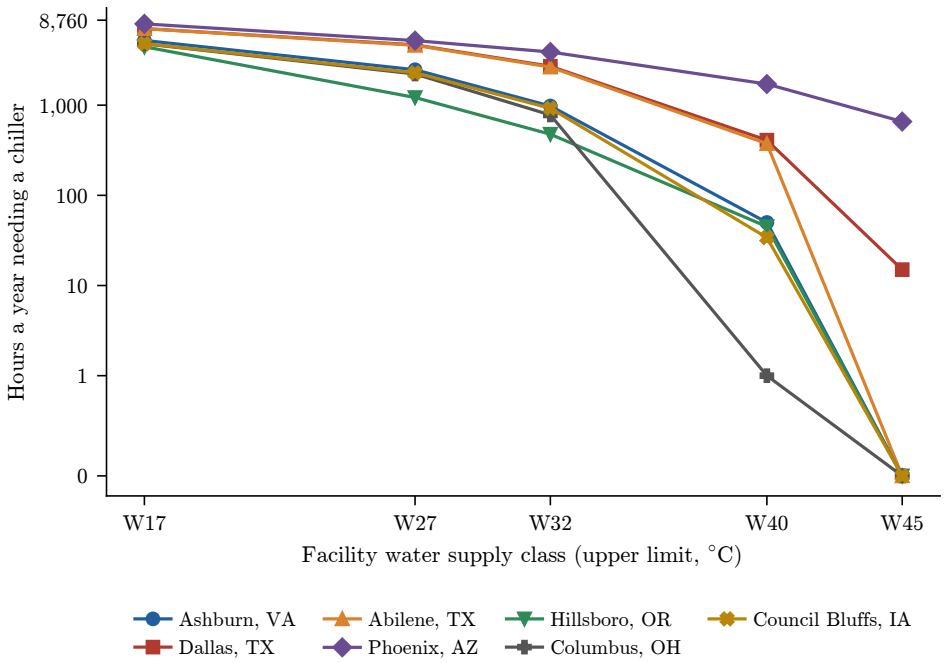


Figure 1. Hours in 2025 in which a dry-cooler plant could not deliver water at the top of each ASHRAE class, so a chiller would be needed. The vertical axis is logarithmic above one hour and linear below it, so that zero can be shown.

Abilene for 374, Dallas for 409 and Phoenix for 1,715. The annual energy at W40 is still only 0.035 to 0.055 kWh per kWh of IT load, because those hours are few.

	Dry bulb (°C)		Chiller hours, dry plant			kWh per IT kWh, dry			L per IT kWh, wet	
Site	Mean	Max	W17	W32	W45	W17	W32	W45	W17	W45
Ashburn, VA	13.5	37.6	5,219	978	0	0.087	0.042	0.037	1.18	0.15
Dallas, TX	20.1	40.7	7,064	2,701	15	0.124	0.057	0.042	1.59	0.21
Abilene, TX	20.0	39.0	7,060	2,642	0	0.123	0.056	0.042	1.59	0.21
Phoenix, AZ	24.4	46.9	7,971	3,887	658	0.150	0.075	0.048	1.80	0.36
Hillsboro, OR	12.2	37.4	4,442	476	0	0.069	0.037	0.035	1.00	0.13
Columbus, OH	11.8	34.2	4,792	779	0	0.081	0.040	0.037	1.08	0.14
Council Bluffs, IA	11.5	38.6	4,820	924	0	0.082	0.041	0.037	1.09	0.14

Table 1. Results by site for 2025. Chiller hours are for the liquid loop alone. Energy and water are for the whole rack, blending 115 kW of liquid heat at the stated class with 17 kW of air-cooled heat served at W17. The dry plant uses dry coolers and chillers. The wet plant adds an evaporative tower before the chiller.

Table 1 and Figure 2 turn hours into energy and water. With a waterless plant, running the liquid loop at W45 rather than W17 cuts heat-rejection energy for the whole rack from between 0.069 and 0.150 kWh per kWh of IT load to between 0.035 and 0.048. In Phoenix the saving is about 0.10 kWh for every kWh of computing, about two thirds of the W17 figure. Most of the gain arrives by W32 in mild climates and only at W40 to W45 in hot ones.

An evaporative plant at W17 recovers part of the gap, bringing energy down to between 0.037 and 0.074 kWh per kWh, but evaporates 1.0 to 1.8 litres for every kWh of computing. Across the sites, each litre evaporated at W17 saves 0.027 to 0.055 kWh of electricity. At W45 the tower is almost never needed for the liquid loop. Phoenix is the exception, needing it for 658 hours.

At W45, what is left of both energy and water belongs mostly to the air-cooled share. That 13 percent of the rack's heat accounts for 25 to 40 percent of the remaining heat-rejection energy,

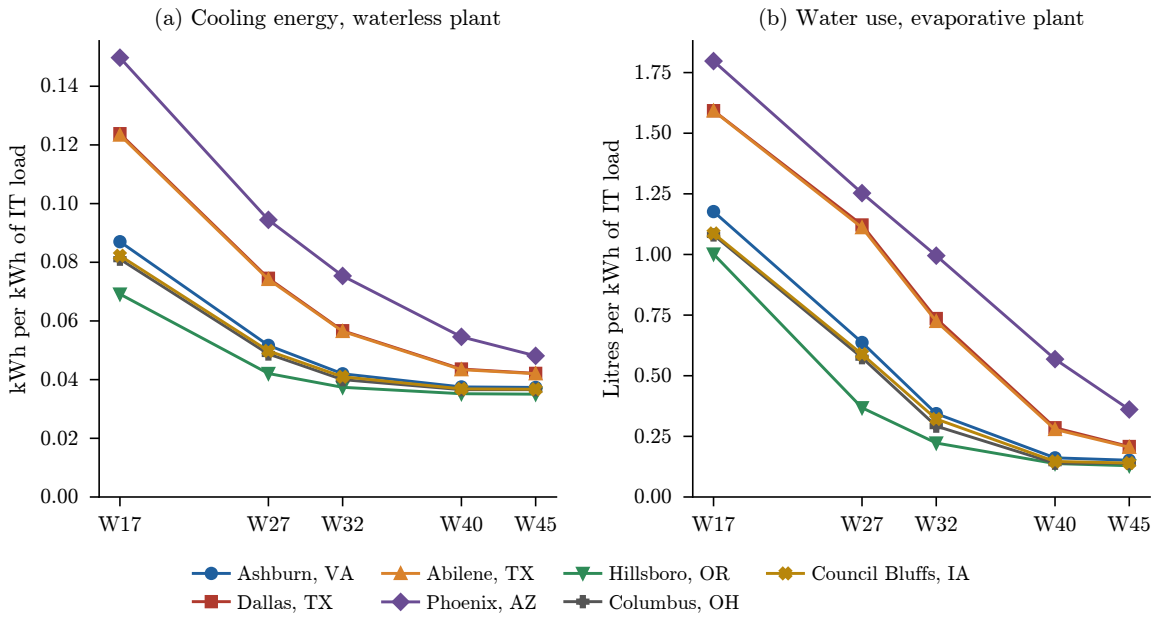


Figure 2. Annual heat-rejection energy for the waterless plant (a), and water use for the evaporative plant (b), per kWh of IT load. Both blend the liquid and air-cooled shares of the rack.

because it still needs 17°C water. It accounts for nearly all of the remaining water use outside Phoenix.

5 Design consequences

Three consequences follow for anyone building or financing an AI hall.

First, the facility water temperature is the most valuable number in the cooling design, and it is worth being exact about which loop a stated limit refers to. It sets the hours of refrigeration, and with it most of the energy and water the plant uses. Equipment that tolerates 45°C water at the rack lets a facility in most US markets run with a small trim chiller sized for tens to a few hundred hours, and with none at all where the coolant distribution unit's approach is small. Equipment that needs 17 to 27°C water does not. Reports from the field suggest that some operators have asked for colder water as chip power has risen [13]. The trend is understandable for keeping junction temperatures down, but in every market studied here it costs energy, water or both.

Second, the residual air-cooled heat is now the expensive part. Networking, power supplies, memory and storage that still need air push the plant back towards cold water. Moving those loads to cold plates, or serving them with rear-door exchangers on warmer water, matters more than further gains on the GPUs.

Third, siting and design interact. Phoenix with W45 water needs less heat-rejection energy than Hillsboro with W17 water. A developer weighing a hot, cheap-power site against a cool, expensive one should compare them at the water temperature the hardware will actually accept, not at a generic PUE.

These results connect to the economics of the chips in the racks. In a power-limited site, every kilowatt spent on cooling is a kilowatt not available to computing, and an earlier paper by the author found that rent per kilowatt is what decides when a GPU should be replaced [14]. At 2025 US industrial electricity prices of 8.62 cents per kWh [15], the 0.10 kWh per kWh saved in Phoenix is worth about $77 a year per kW of IT load in electricity alone. The kilowatts it frees

for computing are worth more.

6 Limits

The model is a lumped, steady-state energy balance. It does not resolve flow distribution in the cold plates or temperatures inside the chips, which would need a coupled thermal and fluid model of the rack. The approach temperatures, chiller efficiency, fan and pump fractions, and blowdown ratio are assumptions, and real plants vary around them. One year of weather is used. 2025 may be warmer or cooler than a design year, and design should use a longer record and extremes. The evaporative calculation pairs the hottest dry-bulb hours with the most humid hours, which is conservative for energy and neutral for water. The air-cooled share is taken from one integrator's specification and will differ between systems. Power-conversion losses, lighting and other loads are outside the scope, so the energy figures are a part of PUE, not the whole of it. None of these simplifications changes the order of magnitude of the gap between W17 and W45, which is set by the weather and the second law.

Data and code

The script that fetched the hourly weather, the annual exceedance counts derived from it, the plant model and the figure scripts that reproduce every number in this paper are published with the paper on the author's research page.

Funding and interests

No external funding was received. The author has no financial interest in any company named.

References

[1] Jain, A. (2026). Air cooling vs. liquid cooling in high-density data centers. Credence Research, 24 October 2026, citing the Uptime Institute Global Data Center Survey 2024. https://www.credenceresearch.com/blog/air-cooling-vs-liquid-cooling-data-centers

[2] Snider, S. (2026). For high-density AI, available data center space may not be usable. Data Center Knowledge, 18 August 2026. Reporting the Uptime Institute Global Data Center Survey 2025. https://www.datacenterknowledge.com/cooling/for-high-density-ai-available-data-center-space-may-not-be-usable

[3] Super Micro Computer, Inc. (2025). SuperServer SRS-GB200-NVL72 specifications. https://www.supermicro.com/en/products/system/gpu/48u/srs-gb200-nvl72

[4] Quanta Cloud Technology (2025). QCT QoolRack stand-alone: Advanced liquid cooling for NVIDIA GB200 NVL72 systems. https://blog.qct.io/wp-content/uploads/2025/04/QCT-Qoolrack-Stand-Alone_Advanced-Liquid-Cooling-for-NVIDIA-GB200-NVL72-Systems.pdf

[5] Liu, P. (2024). Vertiv unveils its reference designs for NVIDIA's GB200 NVL72 platform. 451 Research, S&P Global Market Intelligence, 4 November 2024. Reprint hosted by Vertiv. https://prod.vertiv.cn/4ad4c4/globalassets/campaigns/ai-hub/451research_reprint_vertiv_04nov2024.pdf

[6] ASHRAE Technical Committee 9.9 (2021). Emergence and expansion of liquid cooling in mainstream data centers. White paper. Atlanta: ASHRAE. https://www.ashrae.org/File%20Library/Technical%20Resources/Bookstore/Emergence-and-Expansion-of-Liquid-Cooling-in-Mainstream-Data-Centers_WP.pdf

[7] Bejan, A. (2016). *Advanced Engineering Thermodynamics* (4th ed.). Hoboken, NJ: Wiley.

[8] Zimmermann, S., Meijer, I., Tiwari, M. K., Paredes, S., Michel, B., and Poulikakos, D. (2012). Aquasar: A hot water cooled data center with direct energy reuse. *Energy*, 43(1), 237-245.

[9] Ebrahimi, K., Jones, G. F., and Fleischer, A. S. (2014). A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renewable and Sustainable Energy Reviews*, 31, 622-638.

[10] ASHRAE (2021). *Thermal Guidelines for Data Processing Environments* (5th ed.). Atlanta: ASHRAE.

[11] Zippenfenig, P. (2023). Open-Meteo.com Weather API [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.7970649

[12] Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999-2049.

[13] DatacenterDynamics (n.d., accessed November 2026). Hot water, cold water: What is the right temperature for liquid cooling? https://www.datacenterdynamics.com/en/analysis/hot-water-cold-water

[14] Sadineni, A. C. (2026). The economic life of an accelerator: Rental prices, power and the six-year depreciation of data-centre GPUs. Preprint, November 2026. https://doi.org/10.5281/zenodo.24210935

[15] US Energy Information Administration (2026). Electricity explained: Prices and factors affecting prices. https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php