Tech & AI Systems | Jan 16, 2026 | 9 min read | By Bryan White

Infrastructure at the Boiling Point: An Analysis of AI's Cooling Problem

Climate Change * AI * Ethics

Introduction: The Convergence of AI Heat Density and Atmospheric Warming

Early 21st-century digital infrastructure is defined by the collision of two exponential curves: the surging computational power density of generative Artificial Intelligence (AI) and the rising baseline of global surface temperatures driven by anthropogenic climate change. For decades, data centers operated under a stable thermodynamic regime in which 2-to-10 kW server racks rejected waste heat into a reliably cooler atmosphere via Hot Aisle/Cold Aisle air containment and "free air" economization. As AI accelerators push rack densities past 100 kW while regional climatology shifts toward higher ambient dry-bulb and wet-bulb temperatures, legacy air cooling and evaporative water towers are hitting a physical thermal ceiling.[1][2][3]

The Physics of Silicon: Thermal Design Power and the HBM Heat Trap

As lithography packs tens of billions of transistors onto silicon dies, component Thermal Design Power (TDP) -- the sustained thermal flux a cooling loop must dissipate -- has escalated by an order of magnitude. Whereas NVIDIA's 2017 V100 (300-350 W) and 2020 A100 (400 W) accelerators remained manageable with high-static-pressure air fans, the Hopper H100 SXM5 reaches 700 W per GPU, generating over 10 kW per 8-GPU HGX chassis.[4] NVIDIA's Blackwell B200 architecture pushes chip TDP to 1,000 W+, packing 72 GPUs into a single NVL72 rack drawing 120 kW -- a density where air cooling would require hurricane-velocity airflow and prohibitive acoustic and parasitic fan power.[5][6]

Thermal margins are further compressed by 2.5D and 3D packaging that couples compute dies directly to vertically stacked High Bandwidth Memory (HBM3/HBM3e). While the H100 logic die tolerates a maximum temperature (Tmax) of 93 deg C (recommended below 85 deg C), adjacent HBM stacks suffer bit-flip errors and accelerated degradation above 95 deg C.[7] When elevated ambient intake temperatures reduce cooling headroom, protective thermal throttling drops clock frequencies and voltage. Because large language model (LLM) training relies on synchronous gradient updates across thousands of GPUs, thermal throttling in a single overheated rack stalls the entire supercomputing cluster.[4][8]

| Accelerator Model | | Release Year | | Architecture | | Max TDP (Watts) | | Typical AI Rack Density | | Thermal Constraint Focus |

| NVIDIA V100 | | 2017 | | Volta | | 350 W | | 20-30 kW | | GPU Core Temp (Air Viable) |

| NVIDIA A100 | | 2020 | | Ampere | | 400 W | | 30-40 kW | | Balanced Core/HBM2e (Containment) |

| NVIDIA H100 | | 2022 | | Hopper | | 700 W | | 40-70 kW | | HBM3 Junction Temp (Hybrid/DLC) |

| Google TPU v4 | | 2021 | | TPU | | ~192 W | | 30-50 kW | | Volumetric Pod Density |

| NVIDIA B200 | | 2024/25 | | Blackwell | | 1,000 W+ | | 100-120 kW | | HBM3e (Full Liquid Loop Mandatory) |

Thermodynamics of Heat Rejection: Air, Psychrometrics, and Liquid Cooling

Traditional air convection is fundamentally limited by fluid properties: dry air has a specific heat capacity of 1.006 kJ/kg * K and a thermal conductivity of 0.026 W/m * K, whereas liquid water exhibits a specific heat capacity of 4.18 kJ/kg * K and a thermal conductivity of 0.6 W/m * K -- providing roughly 24 times the thermal conductivity and over 3,000 times the volumetric heat capacity of air.[6]

At the facility boundary, heat rejection is constrained by psychrometrics and atmospheric water vapor. Evaporative cooling towers rely on the ambient wet-bulb temperature -- the lowest temperature achievable via adiabatic evaporation -- plus a 3 deg C to 5 deg C approach temperature.[9][10][11] When humid heatwaves drive wet-bulb temperatures toward 30 deg C-32 deg C, cooling towers cannot supply water below 35 deg C, causing chiller refrigerant head pressures to spike and trip safety cutouts.[3][12] Consequently, operators are migrating to two liquid-cooling architectures:

* Direct-to-Chip (D2C) Cold Plates: Circulating water-glycol mixes directly over CPU/GPU cold plates captures 70-80% of heat via sensible heat transfer, returning fluid at 40 deg C-60 deg C that can reject heat to dry ambient radiators even during hot summers.[6][13]

* Two-Phase Immersion and Microfluidics: Submerging motherboards in dielectric fluids that boil on the silicon surface exploits the latent heat of vaporization, eliminating server fans (saving 10-20% power) and supporting 200+ kW racks, while silicon-etched microfluidics cut thermal resistance directly inside the die.[6][14][28]

| Feature | | Air Cooling | | Direct-to-Chip (D2C) | | Immersion (Two-Phase) |

| Primary Medium | | Air | | Water/Glycol Mix | | Dielectric Fluid |

| Max Practical Rack Density | | ~20-30 kW | | ~100 kW+ | | >200 kW |

| Heat Capture Efficiency | | Low (room mixing) | | High (70-80% at cold plate) | | Near 100% (Latent phase change) |

| Facility Water Consumption | | High (evaporative towers) | | Low (closed loop) | | Minimal |

| High Ambient Sensitivity | | Critical failure risk | | Low (high return-loop temp) | | Very Low |

Regional Vulnerabilities and the Water-Energy-Carbon Feedback Loop

The intersection of compute heat flux and regional warming manifests distinctly across major global hubs:

* Northern Virginia (Ashburn): Hosting ~35% of global hyperscale capacity, Fairfax and Loudoun counties project days above 95 deg F (35 deg C) to jump from 7 per year historically to 36 by 2050 (Cooling Degree Days rising from ~1,300 to ~1,976).[1][15] Coincident summer air-conditioning peaks, local exhaust heat-island effects, and transmission bottlenecks prompted Loudoun County to end by-right data center zoning.[16][17]

* Phoenix, Arizona: Facing 47 projected days above 110 deg F (43.3 deg C) by 2050 and historic Colorado River drought, Phoenix enacted 2025 zoning restrictions curbing open evaporative cooling -- which can evaporate millions of gallons daily per campus and ~700,000 liters for a single GPT-3-scale training run.[2][18][19][23][24]

* Europe (FLAP-D): Engineered for temperate 32 deg C-35 deg C limits, London facilities operated by Google and Oracle suffered chiller trips and cloud outages when July 2022 temperatures breached 40 deg C.[3][21] Germany's Energy Efficiency Act (EnEfG) now mandates a Power Usage Effectiveness (PUE) of 1.2 and waste-heat reuse for post-July 2026 facilities, while Ireland requires on-site or proximate dispatchable generation and storage to prevent grid blackouts.[22][36]

| Region | | Primary Climate Stressor | | Infrastructure Constraint | | Regulatory Response | | Operational Impact |

| N. Virginia | | Heat Domes (+676 CDD by 2050) | | Transmission Capacity | | End of By-Right Zoning | | Delayed cluster energization |

| Phoenix, AZ | | Extreme Peaks (>110 deg F / 43.3 deg C) | | Colorado Basin Water Scarcity | | Closed-Loop Zoning Rules | | Higher chiller electricity OPEX |

| London / Frankfurt | | 40 deg C+ Spikes Beyond Design Spec | | Legacy Air-Cooling Ceilings | | German EnEfG (PUE 1.2) | | Forced liquid-cooling retrofits |

| Dublin, Ireland | | Grid Reserve Margins | | National Generation Deficit | | On-Site Dispatchable Mandate | | Grid decoupling via gas turbines |

Trading water conservation (Water Usage Effectiveness, WUE) for dry air-cooled chillers imposes an unavoidable thermodynamic penalty under the second law of thermodynamics: as ambient temperature rises from 35 deg C to 45 deg C, chiller Coefficient of Performance (COP) drops from ~5.0 to ~3.5 while transmission lines and gas peakers simultaneously derate, amplifying greenhouse gas emissions.[20][25][26][27]

Economic Bifurcation and Regulatory Sovereignty

Transitioning to liquid cooling reshapes data center economics: while Coolant Distribution Units (CDUs), piping manifolds, and structural floor reinforcements make retrofitting legacy air-cooled facilities cost-prohibitive, eliminating server fans and chillers lowers operating cost per FLOP for hyperscalers.[6][29][30] Although Nordic climates offer abundant hydro/geothermal power and natural free

cooling, European data sovereignty statutes (GDPR and the EU Data Act) anchor latency-sensitive inference clusters inside warming metropolitan hubs like Frankfurt and Paris -- imposing an efficiency penalty on the regional AI ecosystem while pushing latency-tolerant training toward remote or nuclear-colocated "AI factories" under ASHRAE Class H1 high-density guidelines.[31][32][33][34][35]

Conclusion: Engineering Beyond the Thermal Ceiling

The convergence of kilowatt-scale silicon accelerators and atmospheric warming has obsoleted air-cooled, water-intensive data center design. By late 2026, rack-scale GB200 NVL72 deployments and Vera Rubin platform reference architectures have further standardized 100% closed-loop warm-water (45 deg C) liquid cooling to cut facility evaporative water consumption while spotlighting acute grid baseload bottlenecks. Sustaining large-scale AI without exhausting municipal watersheds or overloading power grids requires treating the semiconductor die, the liquid manifold, the heat exchanger, and the regional climate as a unified thermodynamic system.

Updates... (Nov 24, 2026): Replaced the HTTP 403 Guardian URL in Ref. [2] with Lawrence Berkeley National Laboratory's 2024 United States Data Center Energy Usage Report (Shehabi et al., Dec 2024, doi:10.71468/P1WC7Q), added the exact verified publication date (Jul 2, 2025) to Ref. [19], corrected punctuation typos in Refs. [12] and [35], and removed the mismatched meteorological wiki link on "hurricane-velocity" airflow (consolidating prior Nov 23, 2026 updates).

References

* [1] Southern Environmental Law Center (Aug 8, 2025). Overhyped data center growth is shaping our energy future. -- https://www.selc.org/news/overhyped-data-center-growth-is-shaping-our-energy-future/

* [2] Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J., Masanet, E., & Sartor, D. (Dec 2024; DOI issued 2025). 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory. -- https://doi.org/10.71468/P1WC7Q

* [3] Data Center Knowledge (Jul 20, 2022). Google, Oracle Data Centers Knocked Offline by London Heat. -- https://www.datacenterknowledge.com/cooling/google-oracle-data-centers-knocked-offline-by-london-heat

* [4] NVIDIA Corporation (Oct 17, 2026). Introduction to NVIDIA DGX H100/H200 Systems -- User Guide. -- https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html

* [5] Vast.ai (Aug 10, 2026). NVIDIA H200 vs. B200: Comparing Datacenter-Grade Accelerators. -- https://vast.ai/article/nvidia-h200-vs-b200-comparing-datacenter-grade-accelerators

* [6] Park Place Technologies (May 14, 2025). Data Center Liquid Cooling vs Air Cooling -- Which is Best? -- https://www.parkplacetechnologies.com/blog/data-center-liquid-cooling-vs-air-cooling/

* [7] NVIDIA Corporation (Dec 2, 2022). NVIDIA H100 PCIe GPU Product Brief (PB-11133-001). --

https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf

* [8] DataCenter Forum (Aug 26, 2025). Learning from Downtime: What Recent Data Center Outages Reveal. -- https://datacenter-forum.ro/en/learning-from-downtime-what-recent-data-center-outages-reveal/

* [9] U.S. Department of Energy (Sep 20, 2016). Psychrometric Bin Analysis for Alternative Cooling Strategies in Data Centers. -- https://www1.eere.energy.gov/buildings/publications/pdfs/rsf/psychrometric_bin_analysis_alternative_cooling_strategies_data_cente

* [10] Delta Cooling (Mar 11, 2021). Understanding Wet Bulb Temperatures and How It Affects Cooling Tower Performance. -- https://deltacooling.com/resources/news/understanding-wet-bulb-temperatures-and-how-it-affects-cooling-tower-performance

* [11] Brentwood Industries (Jun 20, 2024). Approach & Range, Explained. -- https://www.brentwoodindustries.com/resources/learning-center/cooling-tower/approach-range-explained/

* [12] Vecellio DJ, Wolf ST, Cottle RM, et al. (Feb 1, 2022). Evaluating the 35 deg C wet-bulb temperature adaptability threshold for young, healthy subjects (PSU HEAT Project). Journal of Applied Physiology. -- https://doi.org/10.1152/japplphysiol.00738.2021

* [13] Introl (Aug 4, 2025). Liquid Cooling vs Air: The 50kW GPU Rack Guide. -- https://introl.com/blog/liquid-cooling-gpu-data-centers-50kw-thermal-limits-guide

* [14] TRG Datacenters (Mar 3, 2025). Liquid Cooling vs Immersion Cooling: What's the Difference? -- https://www.trgdatacenters.com/resource/liquid-cooling-vs-immersion-cooling/

* [15] Fairfax County Government (Feb 17, 2022; updated Nov 7, 2024). 5 Key Findings of the Fairfax County Climate Projections Report. -- https://www.fairfaxcounty.gov/news/5-key-findings-fairfax-county-climate-projections-report

* [16] IT Brew (Aug 26, 2024). Northern Virginia's "data center alley" is thirstier than ever. -- https://www.itbrew.com/stories/2024/08/26/northern-virginia-s-data-center-alley-is-thirstier-than-ever

* [17] Holland & Knight (Apr 4, 2025). Loudoun County, Virginia, Eliminates By-Right Data Center Development. -- https://www.hklaw.com/en/insights/publications/2025/04/loudoun-county-virginia-eliminates-by-right-data-center-development

* [18] Environmental and Energy Study Institute (Jun 25, 2025). Data Centers and Water Consumption. -- https://www.eesi.org/articles/view/data-centers-and-water-consumption

* [19] City of Phoenix (Jul 2, 2025). City of Phoenix Updates Zoning to Safeguard Health and Safety as Data Center Growth Accelerates. -- https://www.phoenix.gov/newsroom/pdd-news/city-of-phoenix-updates-zoning-to-safeguard-health-and-safety-as.html

* [20] Schneider Electric Blog (Nov 18, 2025). Air vs. liquid cooling: Finding the right strategy for AI-ready data centers. -- https://blog.se.com/datacenter/2025/11/18/air-vs-liquid-cooling-finding-right-strategy-ai-ready-data-centers/

* [21] PC Gamer (Jul 19, 2022). Google, Oracle datacenters melt down in extreme European heatwave. -- https://www.pcgamer.com/google-oracle-datacenters-melt-down-in-extreme-european-heatwave/

* [22] Watson Farley & Williams (Jun 5, 2025). Data Centres: An International Legal and Regulatory Perspective Spotlight on Germany. -- https://www.wfw.com/articles/data-centres-an-international-legal-and-regulatory-perspective-spotlight-on-germany/

* [23] Li, P., Yang, J., Islam, M. A., & Ren, S. (Jun 27, 2025). Making AI Less 'Thirsty'. Communications of the ACM. -- https://doi.org/10.1145/3724499

* [24] Cornell Chronicle (Nov 10, 2025). "Roadmap" shows the environmental impact of AI data center boom. -- https://news.cornell.edu/stories/2025/11/roadmap-shows-environmental-impact-ai-data-center-boom

* [25] Microsoft Cloud Blog (Dec 9, 2024). Sustainable by design: Next-generation datacenters consume zero water for cooling. -- https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/12/09/sustainable-by-design-next-generation-datacenters-consume-zero

* [26] Scientific Systems (Nov 29, 2022). What to Consider When Choosing an Industrial Chiller. -- https://scientificsystem.com/choosing-a-chiller/

* [27] California Public Utilities Commission (Jan 26, 2023). Proposal for Derating Thermal Power Plants based on Ambient Temperature. -- https://www.cpuc.ca.gov/-/media/cpuc-website/divisions/energy-division/documents/resource-adequacy-homepage/r21-10-002/4_ed

* [28] Microsoft Source (Sep 23, 2025). A microfluidics breakthrough goes straight to the silicon to cool up to three times better. -- https://news.microsoft.com/source/features/innovation/microfluidics-liquid-cooling-ai-chips/

* [29] Data Center Knowledge (Oct 31, 2025). Data Center Cooling Methods: Costs vs. Efficiency vs. Sustainability. -- https://www.datacenterknowledge.com/cooling/data-center-cooling-methods-costs-vs-efficiency-vs-sustainability

* [30] Digital Realty (Mar 7, 2025). Guide to Data Center Cooling Technologies. -- https://www.digitalrealty.com/resources/blog/future-of-data-center-cooling

* [31] NorthC Datacenters (Jan 6, 2026). Data sovereignty: why a regional data center is more strategic than ever. -- https://www.northcdatacenters.com/en/blogs/data-sovereignty-why-a-regional-data-center-is-more-strategic-than-ever/

* [32] Piwik PRO (Mar 19, 2025). Norwegian DPA warns against EU-US data transfers. -- https://piwik.pro/blog/norwegian-dpa-warns-against-eu-us-data-transfers/

* [33] Boston Consulting Group (Jan 15, 2025). Breaking Barriers to Data Center Growth. -- https://www.bcg.com/publications/2025/breaking-barriers-data-center-growth

* [34] CoreSite (Mar 31, 2023). As Partners, We Can Reduce Data Center Energy Consumption. -- https://www.coresite.com/blog/as-partners-we-can-reduce-data-center-energy-consumption

* [35] Cui, R., Zhao, P., Luo, Q., et al. (Oct 16, 2025). The Artificial Intelligence Paradox: Does Digital Progress Fuel Environmental Injustice via Transboundary Pollution? Sustainability, 17(20), 9169. -- https://doi.org/10.3390/su17209169

* [36] Mason Hayes & Curran (Jun 11, 2026). New data centre connection rules: EirGrid and ESB Networks publish data centre connection processes. -- https://www.mhc.ie/latest/insights/new-data-centre-connection-rules