Proceedings of the 4th International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/2026.38702

Selecting Cooling Architectures for High-Density AI Data Centres

Wanxiao Yu

Faculty of Science, University of Waterloo, Waterloo, Canada
yw_x@foxmail.com

Abstract. Rising power concentration in artificial intelligence servers has made cooling architecture a primary design constraint for high-density data centres. This paper compares recent peer-reviewed studies of air cooling, direct-to-chip cold plates, single- and two-phase immersion, and their environmental effects. The findings show that liquid cooling expands the thermal operating envelope of high-power hardware, although the size of the benefit depends on the measurement boundary. Direct-to-chip systems can retain a substantial air-side load, warmer coolant may increase server power, and immersion performance varies with flow distribution, fluid properties, and hardware design. Facility-level outcomes also depend on climate, electricity supply, water source, coolant production, and equipment lifetime. Cooling options should therefore be compared under a common workload and system boundary. Optimised air remains suitable for moderate-density or legacy facilities, direct-to-chip cooling provides a practical route for many high-density racks, and immersion is best suited to cases where exceptional heat density and purpose-built infrastructure justify changes to hardware and maintenance.

Keywords: AI data centres, liquid cooling, immersion cooling, energy efficiency

1. Introduction

Artificial intelligence has intensified the thermal problem faced by data centres because power is increasingly concentrated in a small number of accelerators and nodes. Measurements on an eight-GPU NVIDIA H100 system found a maximum demand of approximately 8.4 kW and a median of about 7.9 kW during one large-language-model workload, while the temporal pattern varied with workload configuration [1]. A rack containing several such nodes creates a cooling problem that differs from the relatively uniform loads assumed in many conventional facilities. Total computational electricity describes only part of this problem. Heat must also be removed from small chip areas, transported out of the server, rejected from the building, and managed without excessive fan, pump, chiller, or water demand.

Recent studies have consequently examined airflow management, cold plates, immersion tanks, free cooling, and combinations of these technologies. A broad review by Xu et al. shows that air-side performance depends on containment and airflow distribution, whereas liquid and free-cooling systems remain sensitive to coolant flow, system structure, and climate [2]. Li et al. further argue that non-uniform computing loads require a multiscale view: chip-level cooling may suppress a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38702

hotspot but add pressure-drop or pumping penalties, while room-level airflow still governs the large-space thermal environment [3].

This paper addresses that comparison through a review of recent peer-reviewed journal research. It asks under what load, facility, and environmental conditions advanced air cooling, direct-to-chip cold plates, and single- or two-phase immersion cooling are appropriate for high-density AI data centres. The aim is to identify the conditions under which each architecture is a reasonable choice.

2. Cooling requirements and architectural options

2.1. Matching the cooling boundary to the heat load

The thermal management system of a data centre comprises four interdependent levels: the chip, the server, the rack, and the facility. At the chip level, the relevant questions include junction temperature, heat flux, thermal resistance, and whether cooling prevents frequency reduction. At server and rack levels, heat-capture fraction, flow distribution, fan power, and serviceability become important. At the facility level, chillers, cooling towers, dry coolers, pumps, water use, climate, and heat reuse determine the final energy and environmental result. A technology can perform well at one level while shifting a burden to another. This distinction is especially important for AI systems because nameplate power alone does not describe operating demand. Measurements of an eight-GPU H100 system show that workload settings changed both average power and the time required to finish a training task [1]. In one comparison, a larger batch increased average power but reduced total energy because training was completed sooner. Cooling design must therefore account for peak heat rejection, transient behaviour, and task-level energy rather than assuming one constant rack load. A multiscale analysis similarly indicates that hotspot location and load non-uniformity can change the optimal geometry and control strategy [3].

2.2. Air cooling and direct to chip systems

Air cooling remains the most established architecture, and its suitability depends on the specific layout rather than a generic air-cooled baseline. Performance varies with hot- and cold-stream mixing, pressure distribution, cooling-unit placement, and chiller demand. Gupta et al. compared four air-cooled layouts using computational fluid dynamics and exergy analysis [4]. In their six-rack model, the chiller accounted for roughly 55-60% of total exergy destruction, while a rack-mountable cooling arrangement produced up to 23% less total exergy destruction than the under-floor configuration.

Direct-to-chip cooling places a liquid cold plate on high-power components while usually retaining air for the rest of the server. Its main advantage is compatibility with familiar rack-based hardware and maintenance practices. Because cold plates can operate with coolant warmer than room air, the captured heat may be rejected through free cooling for more hours or used in a heat-recovery loop [2]. A controlled CPU study comparing air, cold plates, and two-phase immersion found that the liquid configurations supported higher overclocked frequencies under the tested conditions [5].

2.3. Single phase and two phase immersion

Immersion cooling removes conventional server fans and places hardware in a dielectric fluid. In a single-phase system, the circulating liquid remains liquid and transfers heat through a heat

DOI: 10.54254/2755-2721/2026.38702

exchanger. In a two-phase system, boiling at the heated surface and condensation elsewhere provide latent-heat transport. Two-phase operation can achieve strong local heat transfer, but it adds sensitivity to pressure, condensation, surface condition, and fluid selection. Kang et al. conclude that surface coatings and porous structures can enhance boiling, but note that coatings or foam may separate from the surface, large structures require economic assessment, and working fluids face environmental constraints [6].

The two immersion types impose different design constraints. Single-phase systems avoid boiling control and may use less complex tanks, yet their performance is strongly influenced by viscosity, bypass flow, inlet arrangement, and pump demand. Two-phase systems can accommodate higher local heat flux, while adding vapour containment, condensers, fluid management, and regulatory risk. Removing server fans still leaves pump and external heat-rejection energy. The net benefit depends on the complete loop and on the temperature at which heat leaves the data centre.

3. Comparative evidence on thermal and energy performance

3.1. Thermal performance and hotspot control

A controlled CPU experiment found that cold-plate and two-phase immersion configurations allowed maximum clock-frequency gains of about 41% and 51%, respectively, relative to air cooling in the same overclocking study [5]. Testing the same processor across cooling modes links the result directly to thermal throttling. It does not show that normal production servers will receive the same performance gain, since production systems may operate below their thermal limit and may be constrained by memory, networking, or software.

Immersion performance also depends on how the fluid reaches the heat source. Yuan et al. compared pure immersion with an immersion-jet arrangement for a single server [7]. The jet configuration produced a heat-transfer coefficient approximately 2.6 times that of pure immersion and lowered the heated-surface temperature from 43.8 °C to 30.3 °C under the reported condition. Increasing flow from 8 to 18 L/min produced further cooling, but greater flow also implies pump work.

A separate single-phase study using an FC-40 dielectric fluid tested a 1U server across different inlet arrangements, flow rates, heat loads, bypass gaps, and heat-spreader designs [8]. A T-shaped inlet-outlet arrangement reduced thermal resistance by 12.6% relative to a Z-shaped arrangement, and a vapour-chamber base reduced thermal resistance by 19% relative to a solid base. Two-phase systems present similar configuration effects. Liu and Yu combined laboratory tests with computational modelling and reported that, as server power increased from 1127 W to 1577 W, the cooling coefficient of performance rose from 19.0 to 26.7 and pPUE fell from 1.053 to 1.037 [9].

3.2. Residual air load and operating temperature

Direct-to-chip systems illustrate why a matched boundary is essential. Hnayno et al. measured five operational racks that combined close-coupled air cooling with direct liquid cooling [10]. Approximately 44% of rack heat entered the liquid loop and about 56% remained on the air side.

Higher coolant temperature creates another system-level trade-off. Stahlhut et al. tested ten direct-liquid-cooled servers at inlet temperatures from 25 °C to 50 °C and developed a facility model [11]. Benchmark performance was not reduced within their test range. However, increasing average coolant temperature from 30 °C to 50 °C raised server power by 8.5% and reduced the heat

DOI: 10.54254/2755-2721/2026.38702

transferred into the liquid by 14.6%. In the model, this shift increased the residual heat delivered to air.

3.3. Deployment cost and reliability

Thermal performance does not by itself determine whether a cooling architecture can be deployed. Haghshenas et al. synthesised evidence on energy, space, density, capital cost, operating cost, and maintenance [12]. Their literature-based comparison suggests that immersion cooling can substantially reduce energy use and floor space in favourable cases, but it also highlights difficult maintenance, incomplete reliability information, and the high cost of converting existing air-cooled facilities. Because the underlying data come from heterogeneous studies and industry reports, the reported percentages should be treated as indicative ranges rather than guaranteed outcomes.

Direct-to-chip cooling generally imposes less change on the rack and server service model, which makes it a plausible transition architecture for existing sites. Yet it adds manifolds, quick disconnects, leak detection, water-quality management, and CDUs, while retaining some air equipment. Immersion can remove server fans and support compact hardware, but technicians must handle wetted components, compatible materials, fluid contamination, and different lifting or maintenance procedures. Two-phase systems further require confidence in vapour containment and long-term fluid behaviour. Reliability analysis must therefore account for interfaces and supporting systems, rather than only the coolant near the chip, because material durability and the division of heat between liquid and air create additional system-level risks [6, 10].

4. Environmental and site constraints

4.1. Carbon and water accounting

The environmental balance of a cooling system depends partly on its location. Siddik et al. estimated that US data centres accounted for about 1.8% of national electricity use and 0.5% of greenhouse-gas emissions in their study period [13]. Their spatial analysis further showed that a meaningful share of direct water consumption and electricity supply was associated with water-stressed regions. Identical cooling equipment can therefore have different consequences across sites because grid carbon intensity and local water scarcity vary.

Water accounting must distinguish direct and indirect use. Mytton explains that data centres may consume water on-site through evaporative cooling and indirectly through electricity generation [14]. The paper also identifies limited operator disclosure and argues that source-level water-use efficiency should accompany PUE. Dry heat rejection may reduce direct water consumption while increasing electricity use during hot conditions; evaporative systems may do the reverse.

4.2. Life cycle performance and coolant choice

A recent cradle-to-grave assessment provides a broader comparison. Alissa et al. modelled air-cooled, cold-plate, single-phase immersion, and two-phase immersion data-centre designs using a functional unit of computation per year [15]. Relative to the modelled air-cooled baseline, the three advanced cooling designs reduced greenhouse-gas emissions by at least 15%, energy demand by at least 15%, and blue-water consumption by at least 31% per Vcore-year. Highly optimised cold-plate and single-phase immersion designs could perform on par with two-phase immersion.

DOI: 10.54254/2755-2721/2026.38702

The same assessment reports uncertainty in inventories for cooling-fluid production, data-centre-specific equipment, and electronics life cycles [15]. Two-phase dielectric fluids also face evolving environmental regulation, while single-phase oils differ in viscosity, flammability, compatibility, and re-refining options. Coolant choice should consequently be evaluated as part of the system life cycle rather than as a thermal property table detached from procurement and disposal.

5. Discussion

5.1. Conditions for adopting each architecture

The reviewed evidence supports a conditional transition toward liquid cooling. Optimised air cooling remains reasonable for moderate rack density, mixed legacy equipment, or sites where rapid serviceability and low retrofit complexity dominate. Its limits become more severe as chip heat flux and rack power rise, because fan power, acoustic constraints, airflow maldistribution, and chiller demand grow together. Nevertheless, comparing an advanced liquid system with a poorly arranged air baseline exaggerates the apparent improvement. Consequently, the air-side architecture and external plant should be specified before the technologies are compared [4].

Direct-to-chip cooling is the most plausible bridge for many high-density installations. It targets the processors and accelerators that create the largest local heat flux while preserving a rack-based operating model. Operational measurements and modelling show that its performance depends on both heat-capture fraction and coolant temperature [10, 11].

Immersion cooling becomes most credible for purpose-built facilities, unusually high density, compact deployment, or hardware designed for tank operation. Experimental evidence confirms strong thermal capability but also shows sensitivity to flow delivery, spacing, and internal construction [7-9]. Life-cycle analysis indicates that an optimised single-phase or cold-plate design may match two-phase performance on the studied environmental metrics [15]. Selection should thus follow requirements rather than technological novelty.

5.2. Evaluation framework and research gaps

A practical comparison should use the same workload and a complete energy boundary. At the chip and server level, it should report maximum junction or case temperature, temperature uniformity, transient response, throttling, IT power, fan power, and pump power. Facility accounting should add CDUs, room fans, chillers or dry coolers, water-treatment equipment, and the heat-rejection conditions. Operational evidence should cover leaks, corrosion, material compatibility, maintenance time, hot swapping, monitoring, and retrofit disruption. Environmental and cost assessments should reflect local grid and water conditions, include coolant production and loss and equipment end-of-life, and cover the expected service life rather than purchase price alone.

The evidence base still has important limitations. Many thermal studies use one processor, one server, or a short experimental campaign. CFD studies often adopt idealised boundary conditions, while facility simulations depend on climate files and assumed operating schedules. pPUE is frequently reported without all infrastructure loads, and proprietary reliability data remain scarce. A broad life-cycle model also identifies missing or confidential fluid and electronics inventories [15].

Future work should prioritise matched experiments on modern multi-GPU racks and should report both workload completion energy and facility cooling energy. Long-duration field studies are needed to quantify leaks, corrosion, connector failure, fluid degradation, maintenance time, and hardware reuse. Comparative assessments should also test cooling architectures under identical

DOI: 10.54254/2755-2721/2026.38702

climates and heat-rejection options, with transparent PUE and water boundaries. Finally, workload-aware control should be evaluated jointly with cooling control.

6. Conclusion

This review examined how air, direct-to-chip, and immersion cooling respond to the thermal demands of high-density AI computing. Liquid cooling can lower local thermal resistance, control hotspots, and expand the operating envelope of high-power hardware. These advantages are strongest at the chip and server level, but they do not automatically produce an equal reduction in facility energy. Direct-to-chip systems may retain a substantial air-side load, warm coolant can increase server power, and immersion performance depends on internal flow, spacing, coolant properties, and supporting equipment. Environmental outcomes also vary with grid electricity, water scarcity, coolant production, equipment lifetime, and end-of-life management. In practice, this favours a gradual and site-specific transition. Optimised air remains appropriate in moderate-density and legacy settings; direct-to-chip systems provide a practical transition for many high-density racks; and immersion is most suitable where density, hardware standardisation, and purpose-built operations justify greater changes to infrastructure and maintenance. This study is limited by heterogeneous experimental scales, inconsistent energy boundaries, and scarce long-term field reliability data. Future research should compare modern AI racks under matched workloads and complete facility boundaries, while reporting direct and indirect water use and life-cycle effects. Until those data are available, architecture decisions should state their workload, facility boundary, climate, and coolant assumptions explicitly.

References

[1] Latif, I., Newkirk, A. C., Carbone, M. R., Munir, A., Lin, Y., Koomey, J., Yu, X., & Dong, Z. (2025). Single-node power demand during AI training: Measurements on an 8-GPU NVIDIA H100 system. *IEEE Access, 13*, 61740–61747. https://doi.org/10.1109/ACCESS.2025.3554728

[2] Xu, S., Zhang, H., & Wang, Z. (2023). Thermal management and energy consumption in air, liquid, and free cooling systems for data centers: A review. *Energies, 16*(3), 1279. https://doi.org/10.3390/en16031279

[3] Li, Y., Zhu, C., Li, X., & Yang, B. (2025). A review of non-uniform load distribution and solutions in data centers: Micro-scale liquid cooling and large-scale air cooling. *Energies, 18*(1), 149. https://doi.org/10.3390/en18010149

[4] Gupta, R., Asgari, S., Moazamigoodarzi, H., Pal, S., & Puri, I. K. (2020). Cooling architecture selection for air-cooled data centers by minimizing exergy destruction. *Energy, 205*, 117625. https://doi.org/10.1016/j.energy.2020.117625

[5] Ramakrishnan, B., et al. (2021). CPU overclocking: A performance assessment of air, cold plates, and two-phase immersion cooling. *IEEE Transactions on Components, Packaging and Manufacturing Technology, 11*(10), 1703–1715. https://doi.org/10.1109/TCPMT.2021.3106026

[6] Kang, D., Lee, J.-Y., Chakraborty, A., Lee, S.-E., Kim, G., & Yu, C. (2022). Recent advances in two-phase immersion cooling with surface modifications for thermal management. *Energies, 15*(3), art. 1214. https://doi.org/10.3390/en15031214

[7] Yuan, L., Wang, Y., Kosonen, R., Yang, Z., Zhang, Y., & Wang, X. (2024). Comparative study on heat dissipation performance of pure immersion and immersion jet liquid cooling system for single server. *Buildings, 14*(9), art. 2635. https://doi.org/10.3390/buildings14092635

[8] Muneeshwaran, M., Lin, Y.-C., & Wang, C.-C. (2023). Performance analysis of single-phase immersion cooling system of data center using FC-40 dielectric fluid. *International Communications in Heat and Mass Transfer, 145*, art. 106843. https://doi.org/10.1016/j.icheatmasstransfer.2023.106843

[9] Liu, C., & Yu, H. (2021). Evaluation and optimization of a two-phase liquid-immersion cooling system for data centers. *Energies, 14*(5), art. 1395. https://doi.org/10.3390/en14051395

[10] Hnayno, M., Chehade, A., Klaba, H., Bauduin, H., Polidori, G., & Maalouf, C. (2022). Performance analysis of new liquid cooling topology and its impact on data centres. *Applied Thermal Engineering, 213*, art. 118733.

DOI: 10.54254/2755-2721/2026.38702

https://doi.org/10.1016/j.applthermaleng.2022.118733
[11] Stahlhut, M., Sekulla, M., Großöhme, P., Auweter, A., & Urbaneck, T. (2025). Data centers with direct liquid-cooled servers: Experimental analysis and modeling. *Energy Science & Engineering, 13*, 3605–3619. https://doi.org/10.1002/ese3.70116
[12] Haghshenas, K., Setz, B., Blosch, Y., & Aiello, M. (2023). Enough hot air: The role of immersion cooling. *Energy Informatics, 6*, art. 14. https://doi.org/10.1186/s42162-023-00269-0
[13] Siddik, M. A. B., Shehabi, A., & Marston, L. (2021). The environmental footprint of data centers in the United States. *Environmental Research Letters, 16*, art. 064017. https://doi.org/10.1088/1748-9326/abfba1
[14] Mytton, D. (2021). Data centre water consumption. *npj Clean Water, 4*, art. 11. https://doi.org/10.1038/s41545-021-00101-w
[15] Alissa, H. A., et al. (2025). Using life cycle assessment to drive innovation for sustainable cool clouds. *Nature, 641*, 331–337. https://doi.org/10.1038/s41586-025-08832-3