IOSR Journal of Applied Physics (IOSR-JAP)
e-ISSN: 2278-4861.Volume 18, Issue 5 Ser. 1 (Sep. – Oct. 2026), PP 29-35
www.iosrjournals.org

AI Data Centres Under Heat: The Physics and Limits of Advanced Thermal Management

Ramoliya Monil Ashish kumar

Abstract
The rapid expansion of artificial intelligence (AI) has increased the computational and thermal demands placed on modern data centres, creating challenges for conventional cooling systems. This paper examines the extent to which advanced thermal management systems can overcome the heat-transfer and energy-efficiency challenges associated with high-density AI computing. It applies principles of thermodynamics and heat transfer, including conduction, convection, heat flux, thermal resistance, specific heat capacity and latent heat, to evaluate the physical limitations of air cooling and the potential of direct-to-chip liquid and immersion cooling. The analysis finds that advanced cooling technologies can substantially improve heat removal by reducing thermal resistance and increasing heat-transfer capacity, although they introduce constraints including pumping requirements, hydraulic resistance and fluid properties. At the facility level, improved cooling can reduce energy overhead but cannot eliminate the underlying electricity consumption and heat generation of AI computation. Advanced thermal management therefore mitigates, rather than fully resolves, AI data centres' energy-efficiency challenges.

Keywords: *Artificial Intelligence; Data Centres; Thermal Management; Heat Transfer; Liquid Cooling*

Date of Submission: 03-09-2026 Date of Acceptance: 13-09-2026

I. Introduction

The computational demands on modern data centres have rapidly increased due to artificial intelligence (AI). Data centres are crucial to this infrastructure and are where processors execute the training and inference workloads required by modern AI systems (AI, 2025). Different from conventional data centres dominated by central processing units (CPUs), AI facilities increasingly rely on graphics processing units (GPUs) that perform large numbers of calculations in parallel but concentrate considerably greater electrical power within individual servers and racks. As a result, data centre infrastructure has growing energy requirements, with the International Energy Agency identifying AI as a major driver of future growth in data centre electricity demand (IEA, 2025). Therefore, data centres are not mere digital spaces: they are physical systems in which electrical energy is supplied to computing hardware and ultimately dissipated largely as thermal energy. Hence, as AI computation expands, managing this heat has become an increasingly important engineering challenge.

The problem here arises because improvements in computational capability are increasing the amount of heat that cooling infrastructure must continuously remove while maintaining processors within safe operating temperatures. While conventional data centres have traditionally relied on air-based cooling, increasing thermal loads have encouraged the development of alternatives such as direct-to-chip liquid cooling and immersion systems. These technologies promise greater heat-transfer capability; however, their ability to cool increasingly powerful hardware does not necessarily eliminate the wider energy demands associated with thermal management. Evaluating their effectiveness therefore requires consideration of both the physics behind heat transfer and the energy required to operate the cooling system. As a result, this paper aims to answer the following question: To what extent can advanced thermal management systems overcome the heat-transfer and energy-efficiency challenges created by high-density AI centres?

This paper argues that advanced thermal management can overcome the immediate heat-transfer constraints created by high-density AI hardware to a considerable, but inherently limited, extent.

AI Heat Generation and the Physical Limits of Air Cooling

The thermal challenge created by AI begins with the increasing concentration of computational power. Training and inference for AI models rely heavily on GPUs and specialised accelerators capable of performing large numbers of calculations simultaneously. As a result, electricity consumption is not only higher overall, but also concentrated at substantially higher power densities. Data Centre Knowledge reports a rapid increase in average rack densities and identifies AI workloads as an important contributor to configurations considerably above conventional levels (Kleyman, 2024). Similarly, Goldman Sachs identifies increasing power density as a major constraint on the infrastructure needed to support AI (Long, 2025). These industry observations align with ASHRAE's broader technical assessment, which identifies increasingly high-density racks as a defining feature of AI-oriented facilities (Hassen et al., 2026). The difference between total power and power density is fundamental because the difficulty of cooling a processor depends not only on how much heat is produced but on how concentrated that heat is.

Power represents the rate of energy transfer,

$$P = \frac{E}{t}$$

and the vast majority of electrical energy consumed by processors is ultimately dissipated as thermal energy that must be removed. The local cooling can accordingly be represented using heat flux,

$$q^n = \frac{\dot{Q}}{A}$$

Where $\dot{Q}$ is the heat transfer rate and A is the area through which the heat passes. As processor power increases without a proportional increase in surface area, heat flux rises. As a result, two systems dissipating identical total power may present very different thermal challenges if one concentrates that energy across a smaller area. This helps explain why AI creates a problem that facility electricity consumption alone cannot adequately characterise: heat must be transferred away from individual high-power components rapidly enough to prevent excessive junction temperatures.

Within a conventional server, this process begins through conduction. Heat travels from the semiconductor junction through the processor package and thermal-interface materials towards a heat sink. Fourier's law describes conductive heat transfer as

$$\dot{Q} = -kA\frac{dT}{dx}$$

Where k represents thermal conductivity, A transfer area and dT/dx the temperature gradient. The pathway can also be expressed using thermal resistance,

$$R_{th} = \frac{\Delta T}{\dot{Q}}$$

For a given heat-generation rate, greater thermal resistance produces a larger temperature difference between the processor and its cooling medium. Wu et al.'s (2025) review of cold-plate technology identifies that increasing thermal loads in integrated circuits is a major heat-dissipation challenge. It also identifies coolant properties, heat-sink structure and internal flow as primary factors affecting thermal performance. This demonstrates that lowering room temperature is not a complete solution, since reduced ambient temperature does not necessarily eliminate the thermal resistance in the pathway immediately surrounding the chip.

After reaching a heat sink, thermal energy is transferred to moving air through forced convection:

$$\dot{Q} = hA(T_s - T_f)$$

Where h is the convective heat-transfer coefficient and $T_s - T_f$ represents the temperature difference between the surface and fluid. Therefore, engineers can increase the rate of heat removal by using a larger heat-sink area, higher airflow or a greater temperature difference. Existing research provides substantial evidence that significant improvements remain possible in air cooling. Chu and Wang's (2019) review of airflow management identifies strategies, such as hot- and cold-aisle arrangements, containment and better airflow distribution as mechanisms for reducing mixing and recirculation. More recently, Jiang's (2024) CFD comparison found that aisle containment increased air-supply efficiency from 65.69% to as high as 90.25% for an underfloor precision-air-conditioning configuration. Inter-column cooling improved from 71.29% to over 92%. These results are significant because they show that the transition to liquid cooling should not be taken as evidence that air cooling is inherently ineffective. Its performance depends heavily on fluid-flow organisation.

Nonetheless, improving airflow does not alter the underlying thermophysical properties of the cooling medium. The sensible heat transported by a flowing fluid can be represented as

$$\dot{Q} = \dot{m} c_p \Delta T$$

Where $\dot{m}$ is the mass flow rate, $c_p$the specific heat capacity and ΔT the fluid's temperature rise. In the case of increasing heat loads, an air-cooled system must therefore increase airflow, allow a larger temperature rise or combine both. The former requires additional fan work while the latter risks higher component temperatures. Research comparing conventional and advanced cooling architectures additionally identifies the mixing of warm exhaust and cold supply air, reliance on air as the heat-transfer medium and cooling-hardware power consumption as important limitations of conventional systems (Hnayno et al., 2022).

The resulting trade-off is visible at the facility level. Cooling represents a significant component of non-IT energy demand, although its share varies with climate, facility efficiency and architecture. A CRS report (Lawson et al., 2026) and a UK POST (Williams & Brawley, 2026) assessment both identify cooling as a crucial component of the wider resource demands created by data centres. IEA's (2025) analysis of AI and energy shows why reducing infrastructure overhead becomes increasingly important as overall data-centre electricity consumption expands. The objective must therefore not be to achieve the lowest possible processor temperature regardless of energy cost. A cooling architecture that controls junction temperature by requiring progressively greater fan and refrigeration power may improve thermal stability while undermining facility efficiency.

Air cooling consequently faces a scaling constraint rather than an absolute physical limit. Jiang (2024) has shown that improved fluid management can substantially enhance conventional cooling, while the broader airflow literature confirms that containment and configuration remain valuable optimisation strategies (Chu & Wang, 2019). However, these measures cannot indefinitely compensate for rising heat flux because the heat-transfer medium itself remains air. This is why engineering interest is growing in changing both the coolant and the physical distance between the coolant and the processor (Al Kez et al., 2025; Wu et al., 2025). Thus, the question is not whether air can remove heat, but whether it can continue doing so at increasing AI power densities without unacceptable spatial and energetic penalties.

Can Advanced Cooling Overcome the Physical Limits?

The limitations of air as a heat-transfer medium at increasing power densities have driven the development of liquid-based cooling systems, of which direct-to-chip (D2C) liquid cooling is one of the most established. In this, a cold plate containing coolant channels is placed directly against high-power components like CPUs and GPUs. A 2025 review by Wu et al. describes cold-plate cooling as the most commonly utilised liquid-cooling approach due to its combination of thermal performance, cost, compatibility and maintainability. The advantage follows directly from the heat-transport relationship established in Section 1: replacing air with a coolant of substantially greater heat capacity allows greater thermal loads to be transported at manageable flow rates. D2C cooling also shortens the thermal pathway by bringing the coolant closer to the heat source. This reduces effective junction-to-coolant thermal resistance.

This theoretical advantage is supported by experimental research. A study by Heydari et al. (2024) evaluated direct-to-chip cold-plate liquid cooling under high-heat-density data-centre conditions and provided evidence beyond purely numerical modelling. Similarly, another study by Hnayno et al. (2022) tested a liquid-cooled topology across five operational IT racks under different thermal conditions. Their system combined direct-to-chip cooling with close-coupled cooling and successfully validated a rack temperature difference of 20 K. Importantly, approximately 44% of the thermal load was absorbed through direct-to-chip cooling while the remaining 56% was handled by close-coupled cooling. Rather than demonstrating a complete technological replacement of air by liquid, the experiment illustrates how different heat-transfer mechanisms can operate within the same thermal architecture. This suggests that near-term AI data centres may optimise cooling by assigning liquid cooling to the highest-flux components while retaining other methods for lower-density thermal loads.

That said, increasing coolant flow introduces its own physical cost. Pump power can be approximated as

$$P_{pump} = \frac{\Delta p \dot{V}}{\eta}$$

Where Δp represents pressure drop, $\dot{V}$ volumetric flow rate and η pump efficiency. Increasing flow velocity may improve convective heat transfer, but it also raises frictional losses through pipes and small cooling channels. Wu et al. (2025) consequently identify microchannel geometry, coolant properties, inlet/outlet configuration and internal flow patterns as important determinants of cold-plate performance. The researchers also identify a research limitation: much of the microchannel literature still relies on numerical simulations with insufficient in-situ visualisation and validated predictive models. Thus, maximising heat-transfer coefficient is not equivalent to maximising system efficiency. Additional heat removal must be weighed against the mechanical work which is required to produce it.

Immersion cooling changes the thermal pathway more radically by placing hardware directly within a dielectric fluid. Al Kez et al.'s 2025 review comparing immersion with direct liquid, rear-door heat exchangers, spray/jet systems and other emerging technologies concluded that different approaches present distinct trade-offs in thermal efficiency, water consumption, scalability and commercial readiness. In single-phase immersion, the dielectric fluid remains liquid and absorbs sensible heat. In two-phase systems, however, coolant vaporisation introduces latent heat:

$$Q = mL_v$$

Where L_v represents the latent heat of vaporisation. During boiling, substantial thermal energy can be absorbed through phase change without requiring a proportionately large increase in coolant temperature. Vapour then moves towards a condenser, releases its latent heat and returns to the liquid phase. This boiling-condensation cycle makes two-phase systems attractive where local heat flux becomes difficult to manage through sensible heating alone.

This does not, however, imply that immersion is automatically superior to D2C cooling. The advanced-cooling review (Al Kez et al., 2025) stresses differences in technological maturity and implementation requirements while the cold-plate literature (Wu et al., 2025) highlights the comparatively favourable compatibility and maintainability of D2C systems. Immersion requires changes to hardware servicing, fluid selection and facility architecture. Coolant viscosity and material compatibility can themselves affect the thermal performance, demonstrating that replacing air with liquid does not remove the relevance of fluid properties. More broadly, reviews of next-generation cooling favour technology selection according to heat density, infrastructure and operating conditions rather than a universal cooling hierarchy.

The comparison therefore reveals a progression in which physical constraint dominates. Air cooling is increasingly constrained by volumetric airflow and weak convective heat transport. Cold plates reduce thermal resistance but introduce hydraulic resistance and pumping requirements. Immersion exposes more hardware directly to coolant while phase-change systems exploit latent heat, but these approaches introduce greater dependence on fluid behaviour, material compatibility and condensation.

Advanced cooling can therefore overcome the immediate heat-transfer bottleneck to a considerable extent, but it does so by shifting rather than eliminating physical constraints. As thermal resistance falls and heat-carrying capacity rises, higher-density AI hardware becomes feasible, while hydraulic resistance, pumping work and phase behaviour become increasingly important design variables. Whether this improvement translates into greater efficiency, however, depends on the energy performance of the wider facility.

From Thermal Performance to System Efficiency: Where Does the Heat Go?

Being able to maintain safe processor temperatures does not necessarily indicate that a data centre has become more energy efficient. This distinction is particularly important because thermal performance describes how effectively heat is removed from computing hardware while facility efficiency considers the additional energy required to operate cooling, power conversion and any other infrastructure. Power usage effectiveness (PUE) expresses this relationship:

$$PUE = \frac{P_{total\ facility}}{P_{IT\ equipment}}$$

A PUE that approaches 1 indicates that a greater proportion of facility electricity reaches IT equipment rather than supporting infrastructure. Research examining data-centre energy and cooling performance demonstrates that cooling architecture can therefore influence overall facility efficiency (Chu & Wang, 2019; Zhao et al., 2023). Hnayno et al. (2022), for example, found that configurations combining liquid cooling with indirect free and evaporative cooling reduced cooling partial PUE by at least 16% compared with the mechanical-cooling reference case. This provides evidence that improvements in heat-transfer architecture can translate beyond processor temperature into measurable reductions in cooling-related energy overhead.

However, PUE can obscure an important distinction between relative efficiency and absolute energy demand. If we consider an illustrative facility whose computing equipment consumes 100 units of electrical energy while supporting infrastructure consumes another 30, its PUE is 1.30. Reducing infrastructure demand to 10 units improves PUE to 1.10 and total consumption to 110 units, but the original 100 units consumed by computation remain. If greater cooling capability subsequently allows twice as much AI hardware to operate, then absolute electricity consumption may rise despite the more efficient facility. This matters because current institutional assessments expect data-centre demand to continue growing substantially. The US Congressional Research Service reports that projections place US data centre electricity consumption between approximately 6.7% and 12% of national electricity use by 2028 (Lawson et al., 2026). Therefore, improved thermal management can reduce the energy cost of cooling each unit of computation without guaranteeing that the sector's overall electricity consumption falls.

The limitation follows directly from the first law of thermodynamics:

$$\Delta U = Q - W$$

Energy is conserved rather than destroyed. In steady operation, the electrical energy consumed by computing equipment and cooling infrastructure must ultimately leave the facility predominantly as thermal energy. A cold plate therefore does not eliminate the heat produced by a GPU. It transports that energy into coolant, after which a heat exchanger, cooling tower, dry cooler or other system must transfer it elsewhere. This means that the thermodynamic boundary of the problem must extend beyond the processor: improvements in component-level heat transfer ultimately remain dependent on how efficiently the facility handles and rejects the captured thermal energy.

This final heat rejection may also create local environmental consequences. A 2026 study using satellite-derived land-surface temperatures to investigate what the authors termed a 'data heat island effect' estimated an average land surface temperature increase of approximately 2°C following the start of operations across the AI data centres examined (Marinoni et al., 2026). This is in line with the physical expectation that concentrated electricity consumption ultimately produces heat that must be transferred to the surroundings. It also connects datacentre thermal management with the broader urban heat island mechanism described by the EPA (2024), where changes in surfaces, vegetation and energy flows can alter local thermal conditions.

The same thermodynamic principle that creates this problem also interestingly suggests a partial solution: waste-heat reuse. If thermal energy cannot be destroyed, capturing it in liquid at useful temperatures can make it available for district heating or other nearby thermal demands. The advanced cooling review by Al Kez et al. (2025) identifies energy reuse as one potential advantage of next-generation liquid systems. Yet heat reuse is constrained by temperature quality. Low-grade heat might require a heat pump to raise its temperature before use, introducing additional work, while the economic and energetic value of recovery depends on having a nearby and sufficiently consistent demand. Waste-heat reuse therefore improves utilisation of energy already consumed rather than reversing the original computational energy requirement.

Water introduces a further system-level trade-off. Some cooling architectures might reduce electrical consumption through evaporated heat rejection, but they do so by consuming water. Broader sustainability discussions emphasise that cooling technology should not be evaluated through electricity consumption alone (Directorate-General for Energy, 2025; Punch, 2026; Williams & Brawley, 2026). A system that lowers cooling electricity but sharply increases water consumption may simply shift its environmental burden from one resource to another. Conversely, warmer-water liquid cooling may expand opportunities for free cooling and heat recovery, demonstrating that component-level thermal design can influence facility-level resource use.

Advanced thermal management therefore solves only part of the physical problem which is created by high-density AI. Evidence from experimental liquid cooling systems shows that it can reduce thermal resistance and cooling overhead, while broader reviews conclude that liquid, immersion and hybrid technologies provide viable pathways towards higher-density computing. Yet institutional energy assessments, emerging local temperature evidence and wider resource considerations demonstrate why cooling capacity cannot be treated as equivalent to sustainability. The most effective thermal-management system is ultimately not the one that removes the greatest quantity of heat at any cost but the one that maintains hardware within safe operating limits while minimising the work and resources required to transport that heat and, where possible, recovering its remaining usefulness.

II. Conclusion

This paper aimed to examine to what extent advanced thermal management can address the heat-transfer and energy-efficiency challenges created by high-density AI data centres. The evidence in the analysis suggests that these technologies can overcome the immediate heat-transfer challenge to a considerable extent. With increasing computational power concentrating greater thermal loads within smaller areas, conventional air cooling faces a scaling constraint resulting from its heat-transfer properties and the increasing energy required to sustain greater airflow. Direct-to-chip and immersion cooling can respond to this constraint by bringing higher-capacity cooling media closer to the heat-generating components, while two-phase systems further exploit latent heat to accommodate high heat fluxes.

That said, as the analysis establishes, these technologies do not eliminate physical constraints. Instead, they change which constraints become most significant. Direct-to-chip systems introduce hydraulic resistance and pumping requirements while immersion and phase-change cooling rely on fluid properties, material compatibility and increasingly complex thermal architectures. Therefore, their effectiveness must be evaluated as an optimisation problem instead of through maximum cooling capacity alone.

More importantly, improved heat transfer does not fully resolve the energy-efficiency challenge. Advanced cooling may reduce cooling-related overhead and improve facility efficiency; however, the electrical energy required for AI computation remains and must ultimately be dissipated as heat. Waste-heat rejection, water consumption and emerging evidence of local thermal effects remain relevant at the facility level. Overall, advanced thermal management is increasingly capable of enabling higher-density AI computing, but it cannot make continued growth in computational demand thermodynamically consequence-free. Its greatest value lies not in eliminating those consequences but rather in managing them more efficiently.

Bibliography

[1]. Al Kez, D., Foley, A. M., Hasan Wong, F. W. B. M., Dolfi, A., & Srinivasan, G. (2025). AI-driven cooling technologies for high-performance data centres: State-of-the-art review and future directions. *Sustainable Energy Technologies and Assessments*, *82*, 104511. https://doi.org/10.1016/j.seta.2025.104511

[2]. Chu, W.-X., & Wang, C.-C. (2019). A review on airflow management in data centers. *Applied Energy*, *240*, 84–119. https://doi.org/10.1016/j.apenergy.2019.02.041

[3]. Directorate-General for Energy. (2025, November 17). *In focus: Data centres – an energy-hungry challenge*. Energy. https://energy.ec.europa.eu/news/focus-data-centres-energy-hungry-challenge-2025-11-17_en

[4]. Hassen, M., Rosenstock, S., Anand, M., Bock, P. De, & Neudorfer, J. (2026). *Energy and thermal efficiency | AI data center energy performance framework*. Ashrae.Org. https://www.ashrae.org/technical-resources/ai-data-center-framework/energy-and-thermal-efficiency

[5]. Heydari, A., Gharaibeh, A. R., Tradat, M., soud, Q., Manaserh, Y., Radmard, V., Eslami, B., Rodriguez, J., & Sammakia, B. (2024). Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers. *Applied Thermal Engineering*, *239*, 122122. https://doi.org/10.1016/j.applthermaleng.2023.122122

[6]. 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*, 118733. https://doi.org/10.1016/j.applthermaleng.2022.118733

[7]. IBM. (2025, February 21). *AI data center*. Ibm.Com. https://www.ibm.com/think/topics/ai-data-center

[8]. IEA. (2025, April 10). *Energy and AI – analysis - IEA*. IEA. https://www.iea.org/reports/energy-and-ai

[9]. Jiang, Di. (2024). Effects and optimization of airflow on the thermal environment in a data center. *Frontiers in Built Environment*, *10*. https://doi.org/10.3389/fbuil.2024.1362861

[10]. Kleyman, B. (2024). *Data center rack density has doubled. And it's still not enough*. Datacenterknowledge.Com. https://www.datacenterknowledge.com/cooling/data-center-rack-density-has-doubled-and-it-s-still-not-enough

[11]. Lawson, A. J., Offutt, M. C., Ortiz, N. R., & Zhu, L. (2026). *Data centers and their energy consumption: Frequently asked questions*. Congress.Gov. https://www.congress.gov/crs-product/R48646

[12]. Long, F. (2025, May 7). *1,000 Homes of power in a filing cabinet - rising power density disrupts AI infrastructure*. Goldmansachs.Com. https://www.goldmansachs.com/insights/articles/rising-power-density-disrupts-ai-infrastructure

[13]. Marinoni, A., Lio, P., Cambria, E., & Horton, B. (2026). The data heat island effect: Quantifying the impact of AI data centers in a warming world. *ResearchGate*, *3*. https://doi.org/10.48550/arXiv.2603.20897

[14]. Punch, J. (2026, August 19). *There's a better way to cool data centres that cuts their huge thirst for water*. The Conversation. https://theconversation.com/theres-a-better-way-to-cool-data-centres-that-cuts-their-huge-thirst-for-water-289547
[15]. United States Environmental Protection Agency. (2024, December 10). *What are heat islands?* US EPA. https://www.epa.gov/heatislands/what-are-heat-islands
[16]. Williams, A., & Brawley, S. (2026, March 16). *What are data centres and how sustainable are they?* POST. https://post.parliament.uk/research-briefings/post-pn-0762/
[17]. Wu, Z., Zhang, G., Lu, S., Leng, P., Yu, Y., Deng, J., & Huang, W. (2025). A comprehensive review of cold plate liquid cooling technology for data centers. *Chemical Engineering Science*, *310*, 121525–121525. https://doi.org/10.1016/j.ces.2025.121525
[18]. Zhao, R., Du, Y., Yang, X., Zhou, Z., Wang, W., & Yang, X. (2023). A critical review on the thermal management of data center for local hotspot elimination. *Energy and Buildings*, *297*, 113486. https://doi.org/10.1016/j.enbuild.2023.113486