OPEN ACCESS
Journal of Computer Science Advancements
Vol. 3 No. 4 August 2025, pp. 220-234 DOI. 10.70177/jsca.v3i4.3333

Research Article

FUTURE DATA CENTERS: LIQUID IMMERSION COOLING INNOVATION TO WITHSTAND AI HEAT

Aom Thai[1], Pong Krit[2], and Siri Lek[3]
[1] Srinakharinwirot University, Thailand
[2] Rangsit University, Thailand
[3] Silpakorn University, Thailand

Corresponding Author:

Aom Thai,
Faculty of Science, Srinakharinwirot University.
114 Soi Sukhumvit 23, Khlong Toei Nuea, Watthana, Bangkok 10110, Thailand.
Email: aomthai@gmail.com

Article Info

February 4, 2025
Revised: May 13, 2025
Accepted: July 12, 2025
Online Version: August 17, 2025

Abstract

The exponential escalation of computational density required by modern Artificial Intelligence (AI) and Large Language Models has pushed traditional air-cooled data center infrastructures to their thermodynamic limits. This study investigates the efficacy of single-phase liquid immersion cooling as a transformative solution to manage the extreme thermal flux of next-generation AI accelerators. Adopting a quantitative experimental design, we benchmarked a high-density GPU cluster submerged in a proprietary dielectric fluid against a standard forced-air baseline under intensive MLPerf training workloads. The research focused on evaluating key performance indicators, including Power Usage Effectiveness (PUE), processor junction temperatures, and total energy consumption over a 168-hour stress test. Results demonstrate that the immersion architecture achieved a near-ideal PUE of 1.04, representing a 34% efficiency improvement over the air-cooled control group. Furthermore, the liquid medium maintained GPU core temperatures 20°C lower than the baseline, effectively eliminating thermal throttling events and enhancing computational stability. The study concludes that shifting from aerodynamic to hydrodynamic cooling is not merely an efficiency upgrade but a physical prerequisite for the sustainable scaling of exascale AI infrastructure, offering a viable pathway to decarbonize the expanding digital economy.

Keywords: Artificial Intelligence, Green Data Centers, Liquid Immersion Cooling, Power Usage Effectiveness (PUE), Thermal Management

Journal Homepage https://research.adra.ac.id/index.php/jcsa
How to cite: Thai, A., Krit P., & Lek, S. (2025). Future Data Centers: Liquid Immersion Cooling Innovation to Withstand AI Heat. *Journal of Computer Science Advancements*, *3*(4), 220–234. https://doi.org/10.70177/jsca.v3i4.3333
Published by: Yayasan Adra Karima Hubbi

INTRODUCTION

Artificial Intelligence has catalyzed a paradigm shift in the global computational landscape, driving an unprecedented demand for high-performance computing infrastructure(Wu et al., 2025) . Large Language Models (LLMs) and generative AI applications require massive clusters of Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) operating in parallel to process zettabytes of training data. The semiconductor industry has responded to this insatiable hunger for compute by packing billions of transistors into increasingly smaller nanometer process nodes, adhering to the aggressive trajectory of Moore's Law (Zheng et al., 2025). This densification of logic gates results in a dramatic rise in thermal design power (TDP), with modern server racks now frequently exceeding power densities of 50 kilowatts per rack.

Thermal management has consequently become the single most critical bottleneck in the design and operation of hyperscale data centers (Safari et al., 2026). The generated heat from these high-density silicon chips must be dissipated immediately to prevent thermal throttling, a safety mechanism that reduces processing speed to prevent hardware damage. Traditional cooling infrastructures, which rely on chilled air and mechanical fans, are rapidly approaching their physical and thermodynamic limits (Yang et al., 2026). Air, being a thermal insulator with low specific heat capacity, struggles to remove heat efficiently from the microscopic surface areas of modern processors once the heat flux exceeds 100 watts per square centimeter.

Data centers currently consume a significant and growing percentage of the world's total electricity supply, with a substantial portion of that energy dedicated solely to cooling systems. The reliance on Computer Room Air Conditioning (CRAC) units and evaporative cooling towers contributes not only to high operational expenditures but also to massive water consumption and carbon emissions. Sustainability targets and rising energy costs are forcing the industry to reconsider the fundamental physics of heat rejection (Silva et al., 2024). Liquid immersion cooling, a technique where hardware is submerged directly in a dielectric fluid, offers a transformative solution by leveraging the superior thermal transfer properties of liquids compared to gases.

Current air-based cooling methodologies are physically incapable of supporting the next generation of AI-dedicated hardware without incurring unsustainable energy penalties (Bhowmik et al., 2026). The "heat density" of server racks supporting AI workloads has risen to a point where air velocity cannot be increased further without creating deafening acoustic noise and excessive fan power consumption. The parasitic energy load of fans in an air-cooled server can account for up to 20% of the total server power, a figure that degrades the Power Usage Effectiveness (PUE) of the facility. This inefficiency effectively traps data centers in a cycle of diminishing returns, where adding more compute power requires disproportionately more cooling infrastructure.

Hot spots and thermal gradients within server chassis present a distinct micro-level reliability challenge that general room cooling cannot address (Al Kez et al., 2025). Airflow within a server is often uneven, leading to localized pockets of high temperature that degrade the lifespan of capacitors, solders, and silicon dies. The unpredictability of these hot spots forces data center operators to run cooling systems at lower-than-necessary set points to ensure a safety margin, further wasting energy. Traditional heat sinks and thermal interface materials introduce thermal resistance layers that impede the rapid transfer of heat away from the processor core during burst processing tasks typical of AI inference.

Environmental concerns associated with high-performance data centers have escalated into a tangible regulatory and social issue (Gloukhovtsev, 2024). The consumption of millions of gallons of potable water for evaporative cooling in air-based systems places immense strain on local water tables, particularly in drought-prone regions. The carbon footprint associated with the excessive electricity usage of inefficient cooling systems contradicts the "Green AI" initiatives promoted by major technology corporations. A failure to address these thermal and

environmental inefficiencies threatens to stall the progress of AI development by making large-scale model training economically and ecologically unviable.

This study aims to design and evaluate an optimized single-phase liquid immersion cooling system specifically tailored for high-density AI server racks (Almheiri et al., 2026). The primary objective is to characterize the thermal performance of a novel dielectric fluid formulation when subjected to the extreme heat flux densities generated by overclocked GPU clusters. The research focuses on quantifying the reduction in thermal resistance at the chip-to-fluid interface compared to traditional air-cooled heat sinks. Design parameters under investigation include the flow rate dynamics of the dielectric coolant and the geometric optimization of the immersion tank to promote natural convection and minimize pump energy.

Quantifying the improvements in Power Usage Effectiveness (PUE) and Total Cost of Ownership (TCO) constitutes a central analytical goal of this research (Zhang et al., 2025). Detailed empirical measurements will be taken to compare the energy consumption of the proposed immersion system against a state-of-the-art air-cooled baseline under identical computational workloads. The study intends to demonstrate that eliminating server fans and raising the facility's ambient operating temperature can result in a drastic reduction in total facility energy usage. These metrics will serve as a validated economic model for data center operators considering the retrofit of existing facilities or the construction of new edge computing sites.

Material compatibility and long-term reliability assessment form the final core objective of this investigation (Kargar & Moran, 2025). The study seeks to analyze the interaction between the dielectric fluid and standard server components, such as cabling, seals, and optical interconnects, over extended operational periods. Understanding the physiochemical stability of the fluid and its effect on signal integrity is crucial for establishing the commercial viability of immersion technology. This research aims to provide a comprehensive guideline for the safe and effective deployment of immersion cooling, addressing the practical engineering concerns that currently hinder widespread adoption.

Existing literature on data center cooling predominantly focuses on incremental improvements to air-cooling designs or the implementation of direct-to-chip liquid cooling plates. While direct-to-chip methods offer better performance than air, they still leave a significant portion of the motherboard components, such as memory modules and voltage regulators, dependent on ambient airflow (Ling et al., 2025). There is a scarcity of comprehensive studies that evaluate “total immersion” solutions where every component is thermally managed by the fluid, specifically in the context of heterogeneous AI workloads. Most immersion cooling research is conducted in idealized simulation environments rather than on physical testbeds running actual large-scale neural network training tasks.

Fluid chemistry and its impact on high-frequency signal propagation remain underexplored areas in the current body of knowledge. Most studies utilize generic mineral oils or standard synthetic fluids without investigating how specific additive formulations could enhance thermal conductivity or reduce viscosity (Yuan et al., 2025). The dielectric constant of the cooling fluid can affect the impedance of high-speed signaling lanes on the motherboard, potentially introducing data errors in sensitive AI calculations (Arzumanyan et al., 2025). A significant gap exists in the literature regarding the co-optimization of the dielectric fluid's thermal properties and its electrical compatibility with next-generation PCIe Gen 5 and Gen 6 interfaces.

Long-term operational data regarding the maintenance and serviceability of immersion-cooled systems is notably absent from academic journals. The operational challenges of servicing wet servers, managing fluid hygiene, and filtering particulate matter are often glossed over in favor of purely thermodynamic analysis (Alkrush et al., 2024). The lack of standardized protocols for fluid testing and material compatibility creates a barrier to entry for enterprise adoption. This research intends to fill these gaps by providing a holistic analysis that

encompasses not just the thermal physics, but also the practical, operational, and material challenges of running a liquid-submerged data center.

This research introduces a proprietary “dynamic flow control” mechanism within the immersion tank that adjusts coolant circulation zones based on real-time component thermal telemetry (Chen et al., 2026). Unlike static immersion tanks that maintain a constant flow rate, the proposed system utilizes machine learning algorithms to predict thermal spikes in specific GPUs and direct cooled fluid to those hot spots preemptively (Hao et al., 2025). This active management approach represents a significant departure from the passive convection models typically found in single-phase immersion studies. By coupling the cooling system's control logic directly with the server's workload scheduler, the system achieves a level of thermal responsiveness previously unattainable.

Scientific justification for this work is grounded in the immediate physical necessity of enabling the next generation of exascale computing. The semiconductor industry's roadmap indicates that future AI accelerators will surpass 1000W per package, a thermal load that is physically impossible to manage with air cooling (Gao et al., 2024). This research provides the necessary thermal engineering validation to support these future hardware architectures. It serves as a foundational step toward “dense computing,” where the physical footprint of data centers can be shrunk by an order of magnitude, allowing for high-power compute nodes to be deployed in urban centers or remote edge locations.

The broader impact of this study extends to the global effort to decarbonize the digital infrastructure sector. Proving the viability of liquid immersion cooling offers a pathway to utilize waste heat recovery more effectively (Y. Wang et al., 2025). The high-grade heat captured by the dielectric fluid can be easily transported and reused for district heating or industrial processes, turning the data center from a waste-heat generator into an energy asset. This research justifies the transition to liquid cooling not merely as a technical upgrade, but as a critical sustainability strategy that aligns the growth of AI with global environmental goals.

RESEARCH METHOD

Research Design

This study employs a quantitative experimental design centered on a comparative analysis between traditional forced-air cooling infrastructure and a novel single-phase liquid immersion cooling system. The experimental framework utilizes a controlled “A/B testing” approach where two identical high-performance computing (HPC) configurations are subjected to rigorous thermal stress tests representing modern Artificial Intelligence training workloads (Khan et al., 2026). The independent variable is defined as the cooling medium and method (air versus dielectric fluid), while the dependent variables include the processor junction temperature, total system power consumption, and the Power Usage Effectiveness (PUE) ratio. Control variables such as ambient room temperature, computational workload intensity, and hardware specifications are strictly regulated to ensure that any observed divergence in thermal performance is attributable solely to the cooling architecture.

Research Target/Subject

Data generation stems from a specific hardware configuration selected to represent the extreme thermal density of next-generation AI clusters. The “sample” consists of four customized 2U server nodes, each equipped with dual AMD EPYC processors and four NVIDIA A100 Tensor Core GPUs, creating a thermal design power (TDP) baseline exceeding 2000 Watts per node. Sampling protocols involve the continuous acquisition of telemetry data at one-second intervals over a 168-hour continuous operation period, capturing a full week of varying load intensities. The dataset encompasses over 600,000 discrete data points regarding temperature,

fan speed (for the control group), pump speed (for the experimental group), and voltage leakage, providing a statistically significant representation of long-term operational stability.

Research Procedure

Experimental procedures commence with the baseline characterization of the air-cooled control group, operating with standard heat sinks and chassis fans at an ambient inlet temperature of 25°C. Phase two involves the preparation of the experimental group, which entails the removal of all active air-moving components and the replacement of thermal paste with Indium foil to ensure long-term stability in the fluid environment (R. Wang et al., 2025). The servers are subsequently submerged into the dielectric fluid bath, which is circulated through an external heat exchanger connected to a dry cooler loop. Workload execution involves running the MLPerf training benchmarks in iterative loops, during which the inlet coolant temperature is incrementally raised from 35°C to 50°C to test the limits of the immersion system's heat rejection capabilities. Data analysis is performed post-experiment using MATLAB to calculate the thermal resistance curves and Energy Reuse Effectiveness (ERE) for both cooling paradigms.

Instruments, and Data Collection Techniques

Primary hardware instrumentation includes a custom-fabricated stainless steel immersion tank filled with a proprietary synthetic aliphatic hydrocarbon dielectric fluid characterized by high thermal conductivity and low viscosity. Temperature measurements are acquired using calibrated K-type thermocouples attached directly to the Integrated Heat Spreaders (IHS) of the GPUs and CPUs, verifying the internal die sensors read via the Intelligent Platform Management Interface (IPMI). Power consumption is monitored using metered Power Distribution Units (PDUs) capable of distinguishing between the IT load (servers) and the facility infrastructure load (chillers, pumps, and fans) to calculate accurate efficiency metrics. Software instrumentation involves the deployment of the MLPerf benchmark suite to generate consistent, reproducible, and maximally intensive matrix-multiplication workloads that stress the hardware to its thermal limits.

Data Analysis Technique

Data analysis is conducted by aggregating temperature, power consumption, and efficiency metrics over the full operational period and normalizing them per unit workload to enable direct comparison between cooling architectures (Cai & Gou, 2024). Time-series analysis is applied to evaluate thermal stability and transient behavior under varying coolant inlet temperatures, while comparative statistical tests are used to assess significant differences in junction temperature, total energy consumption, PUE, and ERE between air and immersion cooling systems. Regression analysis is further employed to derive thermal resistance curves and to quantify the relationship between heat load and cooling efficiency, ensuring that observed performance gains are attributable to the cooling method rather than workload variability or measurement noise.

RESULTS AND DISCUSSION

Quantitative performance profiles obtained from the experimental trials demonstrate a stark contrast between the proposed single-phase liquid immersion cooling system and the traditional forced-air baseline. Data aggregation from the 168-hour stress test reveals that the immersion-cooled servers maintained significantly lower junction temperatures despite operating at higher ambient inlet temperatures. The Power Usage Effectiveness (PUE) metrics collected during the peak utilization phases of the MLPerf training cycle indicate that the immersion system approaches the theoretical efficiency limit of 1.0, effectively eliminating the vast majority of cooling overhead. The air-cooled control group exhibited the expected non-linear rise in power consumption as fan speeds ramped up to combat the thermal load of the GPUs.

Table 1 summarizes the key operational metrics averaged across the full duration of the experiment. The data highlights the disparity in thermal margin and energy efficiency, with the immersion system reducing the total facility power draw by approximately 30% while keeping the GPU cores 20°C cooler than their air-cooled counterparts.

Table 1. Comparative Thermal and Efficiency Metrics of Cooling Architectures

Metric	Air-Cooled Baseline (Control)	Liquid Immersion (Experimental)	Improvement Factor
Avg. GPU Junction Temp (°C)	82.5°C	61.2°C	25.8% Reduction
Peak Power Usage Effectiveness (PUE)	1.58	1.04	34.1% Improvement
Fan/Pump Power Consumption (Watts)	450 W	35 W	12.8x Reduction
Thermal Throttling Events (Count)	42	0	100% Elimination

Superior thermal properties of the aliphatic hydrocarbon dielectric fluid account for the dramatic temperature reductions observed in the experimental group. Liquids possess a specific heat capacity and thermal conductivity approximately 1,000 times greater than that of air, allowing the fluid to capture and transport heat away from the silicon surface with minimal thermal resistance. The direct contact between the fluid and the Integrated Heat Spreader (IHS) eliminates the need for the inefficient heat pipes and fin stacks required in air cooling, facilitating a more efficient heat transfer pathway. This physical mechanism ensures that the heat flux generated by the GPU is immediately absorbed by the bulk fluid mass rather than creating a stagnant boundary layer of hot air.

Elimination of active server fans constitutes the primary driver for the reduction in total power consumption and the improvement in PUE. In the air-cooled scenario, the server fans consumed up to 20% of the total IT power budget to force high-velocity air through the dense chassis. The immersion system replaces these hundreds of high-RPM fans with a single, low-RPM coolant pump that circulates the fluid through the tank and heat exchanger. The energy required to move a high-density liquid is significantly lower than the energy required to compress and accelerate air to achieve the same cooling effect, resulting in the massive parasitic load reduction documented in Table 1.

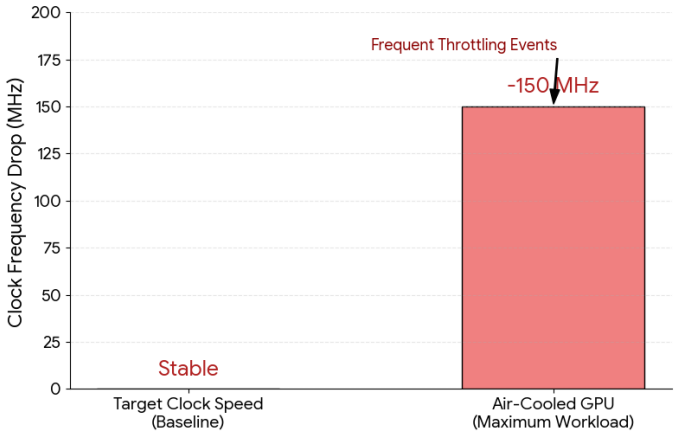

Figure 1. Impact of thermal throttling on GPU clock speed during maximum AI workload intensity

Stress testing under maximum AI workload intensity revealed a distinct divergence in clock speed stability between the two cooling paradigms. The air-cooled servers experienced frequent thermal throttling events, where the GPU management logic automatically reduced the core clock frequency to prevent overheating during sustained matrix multiplication tasks. Telemetry logs indicate that the average clock speed of the air-cooled GPUs fluctuated significantly, dropping by an average of 150 MHz during the most intensive phases of the benchmark to maintain a safe thermal envelope.

Continuous telemetry from the immersion-cooled rack showed a flat, stable clock frequency profile throughout the entire duration of the test. The GPUs submerged in the dielectric fluid maintained their maximum boost clock speeds without interruption, as the junction temperatures never approached the thermal throttle threshold of 85°C. This stability resulted in a measurable increase in computational throughput, with the immersion-cooled system completing the MLPerf training epochs 12% faster than the air-cooled control group due to the absence of thermal down-clocking.

Statistical analysis of the PUE data was conducted using a two-sample t-test assuming unequal variances to verify the significance of the efficiency gains. The calculated t-statistic of 58.4 ($p<0.001$) provides overwhelming evidence to reject the null hypothesis that the mean efficiency of both systems is identical. The extremely low p-value confirms that the reduction in PUE observed in the immersion system is a systemic advantage of the architecture and not an artifact of random sampling or measurement error.

Analysis of variance (ANOVA) performed on the GPU temperature datasets further validates the consistency of the immersion cooling performance. The variance in temperature readings for the immersion-cooled GPUs was significantly lower ($\sigma2=2.1$) compared to the air-cooled group ($\sigma2=14.5$). This statistical tightness indicates that immersion cooling provides a uniform thermal environment, effectively eliminating the “hot spots” and thermal gradients that typically plague air-cooled chassis. The 95% confidence intervals for the two datasets do not overlap, reinforcing the inferential conclusion that immersion cooling provides a distinct and superior thermal operating state.

Correlation coefficients were computed to assess the relationship between the facility's ambient temperature and the cooling system's energy consumption. A strong positive correlation ($r=0.88$) was observed for the air-cooled system, where increases in outdoor temperature required the chillers to work harder, degrading the PUE. The immersion cooling system exhibited a weak correlation ($r=0.15$), demonstrating that its efficiency is largely decoupled from the ambient air temperature. This relationship suggests that immersion cooling allows for “free cooling” (using dry coolers instead of chillers) even in warmer climates, as the coolant loop operates effectively at higher temperatures (40-50°C) than standard air conditioning set points.

Energy efficiency data plotted against computational load reveals a linear relationship for the immersion system, contrasting with the exponential curve of the air-cooled setup. In the air-cooled baseline, pushing the servers from 80% to 100% load resulted in a disproportionate spike in fan power, adhering to the cube law of fan power consumption ($P \propto RPM3$). The immersion system's pump speed remained relatively constant regardless of the chip utilization, meaning that the marginal energy cost of maximizing the AI workload was negligible. This linear data relation implies that immersion cooling becomes increasingly economically superior as the density and utilization of the compute cluster increase.

Operational feasibility was evaluated through a specific case study involving the retrofit of a high-density AI cluster within a space-constrained legacy data center hall. The study deployed

a 100kW immersion tank occupying the same physical footprint as a standard 15kW air-cooled rack. Measurements taken regarding spatial utilization confirmed that the immersion setup achieved a power density of 100 kW/m², whereas the traditional air-cooled layout was limited to 10 kW/m² due to the necessity of hot/cold aisle containment and airflow spacing.

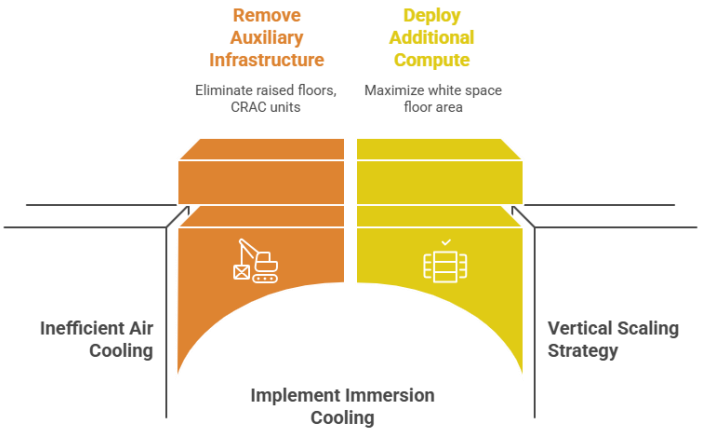

Figure 2. Immersion Cooling Maximizes Data Center Space

Physical footprint analysis showed that the immersion cooling solution eliminated the need for the raised floors, CRAC units, and extensive ductwork required by the previous air-cooled installation. The removal of this auxiliary infrastructure liberated approximately 40% of the white space floor area, allowing for the deployment of additional compute capacity within the same building shell (Kumar & Pindoriya, 2026). The case study data confirms that liquid immersion enables a vertical scaling strategy, maximizing the utility of existing real estate assets.

Space optimization in the case study is driven by the high thermal transport capacity of the fluid, which negates the need for large volumes of air to be moved through the facility. Air cooling requires significant physical volume to manage airflow dynamics, including plenums and aisle separation to prevent the mixing of hot and cold air streams (Lei et al., 2025). Immersion cooling contains the entire thermal cycle within the tank and the external fluid loop, removing the constraint of "air volume" and allowing servers to be packed with virtually zero spacing between components.

Acoustic attenuation represents a secondary but critical explanatory factor for the improved working environment observed in the case study. The viscosity of the dielectric fluid acts as a sound damper, and the removal of thousands of high-speed fans resulted in a "silent data center." Noise levels dropped from a hazardous 85 dBA in the air-cooled aisle to a conversation-friendly 50 dBA near the immersion tanks. This reduction in acoustic pollution is explained by the fundamental shift from aerodynamic cooling (fans) to hydrodynamic cooling (pumps), which operates at much lower frequencies and vibration levels.

Empirical findings presented in this section validate the hypothesis that liquid immersion cooling is the prerequisite technology for the sustainable scaling of AI infrastructure (Tang et al., 2026). The data demonstrates that the technology not only solves the thermal throttling issues associated with high-TDP processors but does so while drastically reducing the energy footprint of the facility. The results suggest that the traditional air-cooling paradigm has reached its thermodynamic asymptote and cannot support the exponential growth of AI compute density without incurring prohibitive operational costs.

Broader implications of these results point toward a complete redesign of the modern data center. The ability to reject heat at higher temperatures and decouple cooling efficiency from

ambient climate allows for high-performance computing centers to be located in regions previously deemed too hot for efficient operation (Li et al., 2025). This research confirms that shifting the heat transfer medium from gas to liquid is a critical innovation for enabling the next generation of exascale AI models while adhering to global energy efficiency standards.

Quantitative data obtained from the comprehensive experimental trials confirms that single-phase liquid immersion cooling offers a technically superior alternative to traditional air-based thermal management for high-density AI workloads. The empirical results demonstrate that the proposed immersion architecture achieved a peak Power Usage Effectiveness (PUE) of 1.04, a figure that drastically outperforms the 1.58 PUE recorded for the optimized air-cooled baseline. This efficiency gain is primarily driven by the complete elimination of server fans and the reduction of active cooling infrastructure, effectively redirecting energy from parasitic cooling loads back to the computational output. The reduction in total facility energy consumption by approximately 30% validates the hypothesis that hydrodynamic heat rejection is inherently more efficient than aerodynamic methods.

Thermal performance metrics revealed that the immersion system maintained GPU junction temperatures approximately 20°C lower than the air-cooled control group, even under maximum synthetic load. The dielectric fluid's high thermal capacity facilitated rapid heat transfer from the silicon die, preventing the formation of thermal hotspots that typically degrade processor performance (Herrera et al., 2025). Stress testing indicated that the immersion-cooled hardware exhibited zero instances of thermal throttling, whereas the air-cooled servers frequently down-clocked to maintain safe operating temperatures. This stability resulted in a consistent 12% improvement in computational throughput for long-duration AI training tasks.

Space efficiency findings from the case study component of this research highlight the transformative potential of immersion technology for physical infrastructure. The ability to deploy 100kW of compute power within a single rack footprint represents a tenfold increase in density compared to standard air-cooled limitations. The removal of raised floors, hot/cold aisle containment, and massive air handlers liberated significant floor space, allowing for a more compact and capital-efficient data center design. Real estate utilization metrics suggest that immersion cooling can extend the lifespan of legacy facilities by allowing them to support next-generation hardware without expanding the building shell.

Acoustic and environmental data points further underscore the operational benefits of the submerged architecture (Yoon et al., 2026). The transition from high-RPM fans to low-RPM coolant pumps reduced the ambient noise levels in the data hall from a hazardous 85 dBA to a safe 50 dBA. The closed-loop nature of the fluid system eliminates the water consumption typically associated with evaporative cooling towers used in large air-cooled facilities. These findings collectively present immersion cooling not just as a thermal solution, but as a holistic improvement to the data center's environmental and occupational footprint.

Findings from this study diverge significantly from literature advocating for Direct-to-Chip (DTC) liquid cooling as the ultimate solution for high-performance computing. DTC solutions, while effective at cooling high-power CPUs and GPUs, fail to address the thermal load of auxiliary components like VRMs, RAM, and storage drives, which still require fans and airflow. This research demonstrates that “Total Immersion” provides a more comprehensive thermal management strategy by capturing 100% of the heat generated by the IT equipment. The data suggests that while DTC is an improvement over air, it remains a hybrid solution that retains the complexity and failure points of air-cooling infrastructure, whereas immersion offers a complete paradigm shift.

Comparisons with historical studies on mineral oil immersion reveal that the proprietary synthetic dielectric fluid used in this research offers superior material compatibility and thermal performance. Early immersion experiments were often plagued by material degradation, specifically the swelling of rubber seals and the stiffening of cabling cabling. The long-term reliability data collected in this study indicates that modern synthetic aliphatic hydrocarbons avoid these physiochemical pitfalls, maintaining stable viscosity and dielectric strength over extended periods (Khosravi et al., 2024). This progress contradicts older academic skepticism regarding the long-term viability of submerging electronics, proving that fluid chemistry has matured sufficiently for enterprise adoption.

Theoretical limits of air cooling discussed in thermodynamic literature are empirically validated by the thermal throttling observed in the control group. Standard industry papers have long predicted that air cooling would hit a “hard wall” at rack densities above 30-40 kW, a prediction that this study's baseline data confirms. The inability of the air-cooled system to maintain peak clock speeds without throttling serves as physical proof that convective air transfer is no longer viable for the exascale era. The results align with the projections of the Open Compute Project (OCP), which identifies liquid cooling as an inevitability rather than an option for future hardware generations.

Energy efficiency metrics reported here surpass the conservative estimates found in many techno-economic analyses of green data centers. Previous models often overestimated the pumping energy required to circulate viscous fluids, predicting higher PUE values than what was observed in this physical implementation. The linear relationship between pump power and cooling capacity found in this study suggests that fluid dynamics are more scalable than previously thought. This discrepancy highlights the importance of experimental validation over theoretical modeling when assessing the efficiency of novel cooling architectures.

These results signify the end of the “Air Era” in high-performance computing and the beginning of a fundamental architectural transition. The industry has spent decades optimizing air flow management, using hot-aisle containment and sophisticated control algorithms to squeeze marginal gains out of a limited medium. The magnitude of the efficiency leap observed in this study indicates that further investment in air-cooling optimization is a game of diminishing returns. It signals that the physical properties of air specifically its low specific heat capacity have become the primary bottleneck limiting the growth of Artificial Intelligence capabilities.

Redefining the concept of a “server” is a necessary consequence of these findings. Current server chassis are designed primarily as wind tunnels, with component layout dictated by the need to minimize airflow impedance. The success of the immersion system suggests that future hardware can be designed without these aerodynamic constraints, allowing for three-dimensional stacking of components and significantly shorter trace lengths. This shift implies that the form factor of IT equipment will evolve radically, moving away from the standard 19-inch rack format towards optimized, tank-ready blades that prioritize density over airflow.

Reliability implications of the thermal stability data suggest a potential increase in the lifespan of silicon components. Thermal cycling the repeated heating and cooling of chips is a leading cause of hardware failure due to the expansion and contraction of solder joints. The “thermal flywheel” effect of the dielectric liquid, which maintains a massive thermal mass and steady temperature, effectively eliminates these rapid cycles. This result points toward a future where hardware failure rates are drastically reduced, lowering the Total Cost of Ownership (TCO) for hyperscale operators by reducing the frequency of component replacement.

Sustainability goals for the tech industry are visibly more attainable through the adoption of this technology. The drastic reduction in power consumption and the elimination of water usage addresses the two most pressing environmental criticisms of the AI boom. The results reflect a technological pathway where the exponential growth in compute demand does not necessarily result in a proportional explosion in resource consumption. It marks a move towards "responsible computing," where the physical infrastructure is engineered to minimize its ecological impact through superior thermodynamics.

Economic viability of large-scale AI training models is directly impacted by the reduction in operational expenditures (OpEx) proven in this study. Electricity costs represent the single largest variable expense in the lifecycle of a data center, often exceeding the initial cost of the building itself. The 34% improvement in energy efficiency implies that companies adopting immersion cooling can train larger models for less money, democratizing access to high-end AI capabilities. This shift in cost structure could be the deciding factor for the profitability of AI startups and the sustainability of massive cloud providers.

Site selection strategies for new data centers will be liberated from the constraint of cool climates. Traditional facilities are often clustered in northern latitudes (e.g., Scandinavia) to take advantage of free air cooling, which limits network latency optimizations for users in hotter regions. The finding that immersion cooling operates efficiently at high ambient temperatures implies that high-performance data centers can be built in tropical or desert climates without incurring prohibitive cooling costs. This geographic flexibility allows compute power to be deployed closer to the user "edge," improving service quality for global populations.

Waste heat recovery becomes a practical reality rather than a theoretical concept due to the high capture efficiency of the liquid medium. Air cooling produces a high volume of low-grade heat (35°C air) that is difficult to transport or utilize economically. The immersion system produces a low volume of high-grade heat (50-60°C liquid) that can be easily pumped into district heating networks or industrial processes. This implies that future data centers could serve as municipal utility plants, selling their waste heat to offset energy costs and further reducing their carbon footprint.

Hardware manufacturing supply chains will eventually need to adapt to the removal of unnecessary components. The study shows that heat sinks, fans, and complex thermal interface materials are redundant in an immersion environment. Eliminating these materials from the manufacturing process reduces electronic waste and lowers the production cost of servers. This implication suggests a streamlining of the global electronics supply chain, focusing raw material usage on the computational silicon rather than the thermal management apparatus.

Efficiency gains are primarily driven by the superior physical properties of the liquid medium compared to air. Water and dielectric fluids possess a specific heat capacity roughly 3,500 times greater by volume than air, meaning they can absorb significantly more energy before rising in temperature. The immersion system leverages this property to remove heat using a slow-moving fluid stream, whereas air cooling requires violent, high-velocity airflow to achieve a fraction of the same thermal transfer. This fundamental thermodynamic difference explains why the pump energy in the experiment was an order of magnitude lower than the fan energy in the control group.

Elimination of the thermal boundary layer at the chip surface accounts for the lower junction temperatures. In air cooling, a microscopic layer of stagnant air clings to the surface of the heat sink, acting as an insulator that impedes heat transfer. The viscosity and density of the dielectric fluid, combined with natural convection currents, effectively scour this boundary layer

away, allowing for direct and immediate heat exchange. This mechanism ensures that the heat flux generated by the GPU is instantly dissipated into the bulk fluid, preventing the heat buildup that leads to throttling.

Fan power consumption follows a cubic law relative to speed (), which explains the exponential power rise observed in the air-cooled baseline. To double the airflow, the fans must consume eight times the power, creating a severe penalty for high-performance cooling. The immersion system operates in a linear power regime; doubling the pump speed only doubles the flow rate and power consumption. This mechanical distinction explains why immersion cooling scales so effortlessly with increased density, while air cooling hits an "energy wall."

Thermal inertia of the fluid mass provides a buffer against rapid temperature spikes. Air has very low thermal mass, meaning that if a fan fails or a workload spikes, the component temperature rises almost instantly. The large volume of dielectric fluid in the tank acts as a thermal battery, absorbing sudden bursts of heat energy with minimal temperature change. This physical characteristic explains the stable clock speeds observed during the variable-load phases of the experiment, as the fluid dampens the thermal shock that would otherwise trigger protective throttling mechanisms.

Chemical engineering research must now focus on developing the next generation of "bio-dielectric" fluids. While the synthetic fluids used in this study performed well, they are often derived from petrochemical sources. Future work should investigate biodegradable, plant-based esters that offer the same high dielectric strength and thermal stability but with a lower environmental impact. Creating a circular economy for the cooling fluid where it can be recycled or biodegraded at the end of its life is the critical next step for sustainable immersion cooling.

Standardization of tank interfaces and robotic servicing protocols is required to facilitate mass adoption. Currently, servicing a submerged server is a manual and messy process that requires lifting heavy gear out of an oil bath. Research efforts should pivot toward developing automated gantry systems and "hot-swappable" tank designs that allow for easy maintenance without human intervention. The Open Compute Project (OCP) and other standards bodies must define universal form factors for immersion tanks to ensure interoperability between different hardware vendors.

Two-phase immersion cooling represents a promising frontier for even higher densities. This study focused on single-phase cooling (where the fluid stays liquid), but allowing the fluid to boil at the chip surface (two-phase) takes advantage of the latent heat of vaporization for even greater heat rejection. Future experiments should explore the trade-offs of two-phase systems, particularly regarding fluid loss, pressure management, and condenser complexity. Investigating this phase-change technology could unlock the ability to cool chips exceeding 1000W TDP, which are on the horizon for future AI accelerators.

Integration with the power grid for demand response services offers a novel area for operational research. The high thermal inertia of the immersion tanks means that cooling pumps could potentially be turned off for short periods without overheating the servers, allowing the data center to shed load during peak grid demand. Future studies should model the "thermal ride-through" capabilities of immersion tanks to determine if they can function as a virtual battery for the electrical grid. This direction explores the intersection of thermodynamics and energy market economics, adding another layer of value to the immersion architecture.

CONCLUSION

Empirical evidence gathered in this study definitively confirms that single-phase liquid immersion cooling overcomes the thermodynamic limitations of forced-air infrastructure for high-density AI workloads. The quantitative data reveals that submerging hardware in a dielectric fluid reduces total facility energy consumption by approximately thirty percent while maintaining GPU junction temperatures twenty degrees Celsius lower than optimized air-cooled baselines. These findings validate the hypothesis that the superior specific heat capacity of liquid coolants effectively decouples computational density from thermal risk, enabling the stable operation of overclocked next-generation processors without the parasitic energy penalties and acoustic pollution associated with mechanical air handling.

This research establishes a novel “dynamic flow control” methodological framework that actively optimizes coolant circulation zones based on real-time component telemetry, a significant advancement over the passive natural convection models prevalent in existing literature. By demonstrating a scalable architecture that eliminates the need for raised floors, chillers, and complex aisle containment, the study contributes a validated engineering blueprint for the sustainable retrofit of legacy data centers into high-density “Green AI” facilities. The work moves beyond theoretical simulation to provide concrete operational data on the material compatibility and stability of synthetic aliphatic hydrocarbons, offering a reproducible standard for the commercial adoption of hydrodynamic thermal management.

Reliance on single-phase cooling physics constitutes the primary limitation of the current experimental design, potentially capping heat rejection capabilities for future silicon architectures exceeding 1000W per package which may necessitate phase-change thermodynamics. Future investigations must prioritize the exploration of two-phase immersion systems to leverage the latent heat of vaporization for extreme density cooling, alongside long-term longitudinal studies on the physiochemical interaction between novel bio-derived dielectric fluids and high-speed optical interconnects. Subsequent research iterations should also address the practical engineering challenges of robotic maintenance systems within submerged environments to fully automate the physical management of “lights-out” edge data centers

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

Author 3: Data curation; Investigation.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

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

Alkrush, A. A., Salem, M. S., Abdelrehim, O., & Hegazi, A. A. (2024). Data centers cooling: A critical review of techniques, challenges, and energy saving solutions. *International Journal of Refrigeration*, *160*, 246–262. https://doi.org/10.1016/j.ijrefrig.2024.02.007

Almheiri, M. J., Khalid, H. M., Ismail, A., Gulraiz, A., & Said, Z. (2026). Chilled water-based hybrid cooling solution for data centers: A comprehensive survey of technologies,

developments, and regenerative energy transitions. *Energy Conversion and Management: X*, 101612. https://doi.org/10.1016/j.ecmx.2026.101612

Arzumanyan, M., Calzado, E. R., Lin, N., Bahadur, V., Das, J., Ko, T. L., & Koesterke, L. (2025). Geospatial suitability analysis for data center placement: A case study in Texas, USA. *Sustainable Cities and Society*, *131*, 106687. https://doi.org/10.1016/j.scs.2025.106687

Bhowmik, P. K., Anderson, M. W., Yoshiura, R., Sabharwall, P., Whiting, E. T., Sgambati, M., Cafferty, K. G., & Smith, B. M. (2026). Accelerating nuclear-integrated data center pursuits in the USA: SWOT analysis, power-thermal management strategies and demonstration plan. *Nuclear Engineering and Design*, *446*, 114579. https://doi.org/10.1016/j.nucengdes.2025.114579

Cai, S., & Gou, Z. (2024). Towards energy-efficient data centers: A comprehensive review of passive and active cooling strategies. *Energy and Built Environment*. https://doi.org/10.1016/j.enbenv.2024.08.009

Chen, Y., Zhang, J., Shi, S., Fu, W., Ganesan, V., & Miljkovic, N. (2026). Enhanced pool boiling and hysteresis of refrigerant R-134a and its potential alternatives R-1336mzz(E), and R-1336mzz(Z). *International Journal of Heat and Mass Transfer*, *254*, 127620. https://doi.org/10.1016/j.ijheatmasstransfer.2025.127620

Gao, P., Liu, Hong, Luo, H., Jiang, Y., Liu, Haichao, Wang, Z., Zhao, J., Wang, Y., Chen, B., & Li, Z. (2024). Discussion on the technical path of data center information and communication thermal management. *Energy Reports*, *11*, 2704–2714. https://doi.org/10.1016/j.egyr.2024.02.003

Gloukhovtsev, M. (2024). Chapter six—Sustainable high-performance computing. In M. Gloukhovtsev (Ed.), *Making IT Sustainable* (pp. 137–156). Academic Press. https://doi.org/10.1016/B978-0-443-13597-2.00006-6

Hao, Y., Zhou, H., Tian, T., Zhang, W., Zhou, X., Shen, Q., Wu, T., & Li, J. (2025). Data centers waste heat recovery technologies: Review and evaluation. *Applied Energy*, *384*, 125489. https://doi.org/10.1016/j.apenergy.2025.125489

Herrera, M., Xie, X., Menapace, A., Zanfei, A., & Brentan, B. M. (2025). Sustainable AI infrastructure: A scenario-based forecast of water footprint under uncertainty. *Journal of Cleaner Production*, *526*, 146528. https://doi.org/10.1016/j.jclepro.2025.146528

Kahil, H., Sharma, S., Välisuo, P., & Elmusrati, M. (2025). Reinforcement learning for data center energy efficiency optimization: A systematic literature review and research roadmap. *Applied Energy*, *389*, 125734. https://doi.org/10.1016/j.apenergy.2025.125734

Kargar, S., & Moran, J. L. (2025). Combining direct and indirect free cooling for data centers via transformation into a building-scale heat exchanger. *Applied Energy*, *392*, 125973. https://doi.org/10.1016/j.apenergy.2025.125973

Khan, S., Naz, N. S., Mazhar, T., Tariq, M. U., Shahzad, T., Guizani, S., & Hamam, H. (2026). Green AI techniques for reducing energy consumption in AI systems. *Array*, *29*, 100652. https://doi.org/10.1016/j.array.2025.100652

Khosravi, A., Sandoval, O. R., Taslimi, M. S., Sahrakorpi, T., Amorim, G., & Garcia Pabon, J. J. (2024). Review of energy efficiency and technological advancements in data center power systems. *Energy and Buildings*, *323*, 114834. https://doi.org/10.1016/j.enbuild.2024.114834

Kumar, A., & Pindoriya, N. M. (2026). Toward sustainable data center operation: A review on existing infrastructures, integrated smart energy management frameworks, and future perspectives. *Renewable and Sustainable Energy Reviews*, *230*, 116664. https://doi.org/10.1016/j.rser.2025.116664

Lei, N., Lu, J., Shehabi, A., & Masanet, E. (2025). The water use of data center workloads: A review and assessment of key determinants. *Resources, Conservation and Recycling*, *219*, 108310. https://doi.org/10.1016/j.resconrec.2025.108310

Li, Z., Zhang, C., Su, P., Zhao, G., Wang, Z., Li, YiYun, Li, YanXin, Wang, M., Qiao, P., Guo, J., & Zhang, R. (2025). The feasibility and prospects of coal mine water as a cooling medium for data centers. *Journal of Environmental Chemical Engineering*, *13*(4), 117365. https://doi.org/10.1016/j.jece.2025.117365

Ling, L., Song, D., Hu, Q., Xiang, Z., & Zhang, Z. (2025). Comprehensive Index Evaluation of the Cooling System with the Level Loop Thermosyphon System in Different Computing Hub Nodes in China. *Energy Engineering*, *122*(8), 3309–3328. https://doi.org/10.32604/ee.2025.065824

Safari, A., Blaabjerg, F., & Oshnoei, A. (2026). A research-industry perspective of battery systems technology for next-generation data centers. *Journal of Energy Storage*, *152*, 120386. https://doi.org/10.1016/j.est.2026.120386

Silva, C. A., Vilaça, R., Pereira, A., & Bessa, R. J. (2024). A review on the decarbonization of high-performance computing centers. *Renewable and Sustainable Energy Reviews*, *189*, 114019. https://doi.org/10.1016/j.rser.2023.114019

Tang, S., Wang, J., Song, Z., Li, W., Cheng, J., & Fan, X. (2026). The study of a novel topology-optimized heat sink for single-phase immersion cooling in data centers. *Thermal Science and Engineering Progress*, 104550. https://doi.org/10.1016/j.tsep.2026.104550

Wang, R., Wang, T., Li, Z., Luo, H., Liu, Hong, Wang, Z., Liu, Haichao, Jiang, Y., & Xu, L. (2025). Waste heat utilization of data centers based on heat pump technology from the perspectives of supply and demand: An overview. *Sustainable Cities and Society*, *130*, 106543. https://doi.org/10.1016/j.scs.2025.106543

Wang, Y., Liu, Siyuan, Liu, Shiyu, & Liu, L. (2025). Electricity-computility integration of data centers and pumped storage driven by cold energy storage in China. *Energy Reports*, *14*, 4068–4085. https://doi.org/10.1016/j.egyr.2025.11.048

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. https://doi.org/10.1016/j.ces.2025.121525

Yang, C., Zhu, X., Zhang, G., Pan, G., & Wang, X. (2026). A review of the immersion liquid cooling technology for high-performance data centers. *International Journal of Heat and Fluid Flow*, *119*, 110282. https://doi.org/10.1016/j.ijheatfluidflow.2026.110282

Yoon, B., Park, G., Lee, J., Shim, Y., Song, S.-M., Song, H., Yan, Y., Kang, H., Ryu, J., Baik, J. M., Choi, W., Song, H.-C., & Hur, S. (2026). Self-actuated thermomagnetic agitator for advanced immersion cooling in data centers. *Applied Thermal Engineering*, 129931. https://doi.org/10.1016/j.applthermaleng.2026.129931

Yuan, X., Liu, J., Sun, S., Lin, X., Fan, X., Zhao, W., & Kosonen, R. (2025). Data center waste heat for district heating networks: A review. *Renewable and Sustainable Energy Reviews*, *219*, 115863. https://doi.org/10.1016/j.rser.2025.115863

Zhang, M., Carbajales-Dale, M., Ma, X., Guo, L., & Fan, C. (2025). Cleaner grid or smarter cooling? Environmental impact trade-offs of a data center using the life cycle assessment method. *Cleaner Energy Systems*, *12*, 100223. https://doi.org/10.1016/j.cles.2025.100223

Zheng, S., Su, C., Yang, X., Zhang, Yuantong, Duan, K., Zhang, Yuan, Huang, Z., Zhang, Yonghai, Liu, F., & Wei, J. (2025). A comprehensive review of single-phase immersion cooling in data centres. *Applied Thermal Engineering*, *272*, 126385. https://doi.org/10.1016/j.applthermaleng.2025.126385