Title
Beyond Air: Energy Performance and Market Insights for Data Center Liquid Cooling Strategies

Journal
ACEEE Summer Study on Energy Efficiency in Buildings

Authors
Chakraborty, Subhrajit
Tano, Ines-Noelly
Rasouli, Erfan
et al.

Publication Date
2026-08-08

Data Availability
The data associated with this publication are available upon request.

Beyond Air: Energy Performance and Market Insights for Data Center Liquid Cooling Strategies

Subhrajit Chakraborty, Ines-Noelly Tano, Erfan Rasouli, Vinod Narayanan,
University of California Davis, Western Cooling Efficiency Center

ABSTRACT

The rapid expansion of artificial intelligence, machine learning, and cloud computing is driving unprecedented growth in datacenter energy demand. Traditional air-cooling systems, which can consume up to 40% of facility electricity, are increasingly inadequate for high density racks exceeding 100 kW. Liquid cooling technologies present a significant opportunity to slash datacenter electricity consumption and peak demand, but their performance has not been systematically evaluated. This paper presents insights from a market characterization and energy modeling study of liquid cooling technologies, comparing single-mode and hybrid approaches to baseline air-cooled datacenters. The energy modeling framework integrates component-level simulations with room scale computational fluid dynamics. Server fans, heat sinks, Direct to Chip (D2C) cold plates, and rear door heat exchangers were modeled to quantify airflow, pressure drop, and pumping power requirements. Each cooling strategy was evaluated for the percentage of rack power consumed by the secondary cooling loop while maintaining chip case temperatures below 70°C for the same facility water temperature of 30°C, keeping the primary side energy consumption percentage similar across the various strategies. Results show that D2C provides the lowest percentage of secondary loop energy use (~ 93% lower than baseline air-cooled strategy) and the highest compute density(~ 15 times of highest density baseline air-cooled strategy), although it comes with higher cost, greater maintenance complexity, and limited retrofitability. Hybrid liquid-cooling strategies offer a more balanced approach that improves retrofit feasibility while moderating cost and efficiency trade-offs. Market analysis highlights that commercial D2C systems are deployment-ready, while immersion technologies remain niche and 2-phase systems are emerging. The study demonstrates how advanced liquid cooling technologies can meet rising compute demands while impr30oving energy efficiency.

Introduction

The growth of artificial intelligence (AI), machine learning, and cloud computing is driving unprecedented demand for datacenter capacity. In California alone, utilities such as Pacific Gas and Electric (PG&E) anticipate spending $73 billion to support 10 gigawatts of new datacenter load in the next 10 years and Southern California Edison (SCE) has 16 new datacenter projects under active tracking and assessment in 2025 (Buffington, 2025; Skidmore, 2025). Traditional air-cooling systems, typically using computer room air conditioners (CRACs) or computer room air handlers (CRAHs), can account for 33 percent to 40 percent of a facility's electricity consumption (Offutt & Zhu, 2025). CRACs transfer heat from the datacenter to refrigerant undergoing direct expansion, whereas CRAHs are used with facility chilled water supply. As rack power densities rise, traditional air-cooling becomes increasingly inefficient and insufficient due to poor heat transfer characteristics, lower conductivity and lower heat capacitance of air. Liquid cooling technologies provide an efficient means of managing high thermal loads by targeting the high thermal dissipating server components: graphics processing unit (GPU) and central processing unit (CPU). Several categories exist within the broader liquid cooling umbrella, including single-phase and two-phase direct liquid cooling, single- and two-phase immersion cooling, and hybrid liquid-air systems. These solutions offer substantial performance benefits, such as reduced power usage effectiveness (PUE), lower water use, and potential for waste heat recovery. However, market adoption has been slow, hindered by concerns over cost, integration complexity, ease of retrofit ability, standardization, and long-term reliability.

Recent industry developments have renewed interest in liquid cooling. The Open Compute Project (OCP) has released guidelines and standards to streamline integration. Major hyperscale operators have begun piloting and deploying liquid-cooled infrastructure. Programs such as Advanced Research Projects Agency–Energy's (ARPA-E) COOLERCHIPS aim to develop energy-efficient, scalable, and cost-effective cooling strategies (ARPA-E, 2021). Key literature and past studies in California, including those by Lawrence Berkeley National Laboratory (Coles & Greenberg, 2014), the California Energy Commission (Asetek Inc, 2024), and white papers from the OCP community, provide essential technical foundations and early evidence of energy and cost savings.

There have been some studies in published literature on identifying and comparing liquid cooling technologies. Ohadi et al. (2012) focus the analysis on component level heat extraction comparing air cooled heat sink with liquid cold plates. Curtis et al. (2023) provide one of the more comprehensive comparisons of the performance of air cooled vs liquid cooled servers using algebraic thermal models with empirical performance numbers for components. Their analysis was focused on processors with 350 W thermal dissipation power. Azarifar et al. (2024) reviewed literature on liquid cooling methods, including hybrid systems with rear door heat exchangers, direct-to-chip (D2C) cold plates, and single- and two-phase immersion cooling. They recommended that a comprehensive analysis that considers all levels from chip to server to room and facility levels is needed to compare the different liquid cooling schemes.

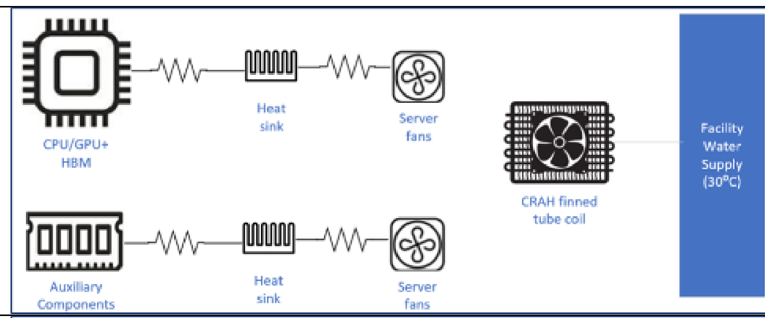
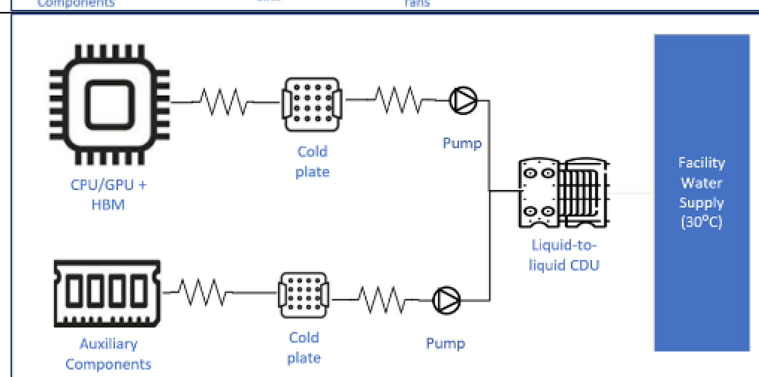
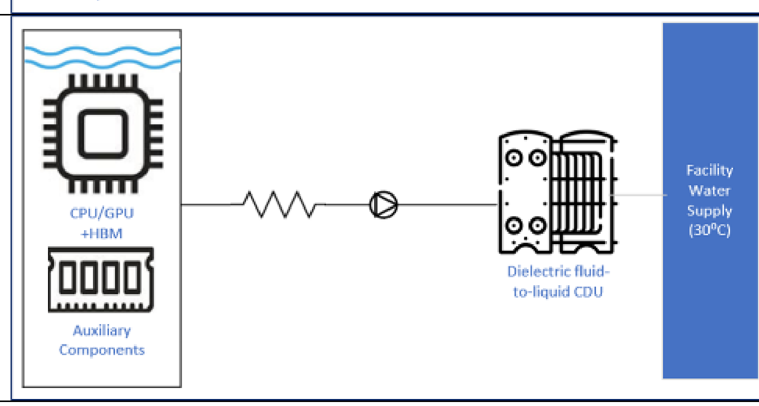
We address the gap identified by Azarifar et al. (2024) by performing a comprehensive comparison of the energy and non-energy aspects of different liquid cooling technologies and compare their performance with traditional air-cooled datacenter rooms. We begin with a classification of the different technologies based on the heat removal scheme and then quantify the energy use in the secondary side of each scheme. Non-energy benefits and market-related characteristics of the cooling strategies in terms of cost, ease of retrofit, and environmental impacts are researched and contrasted.

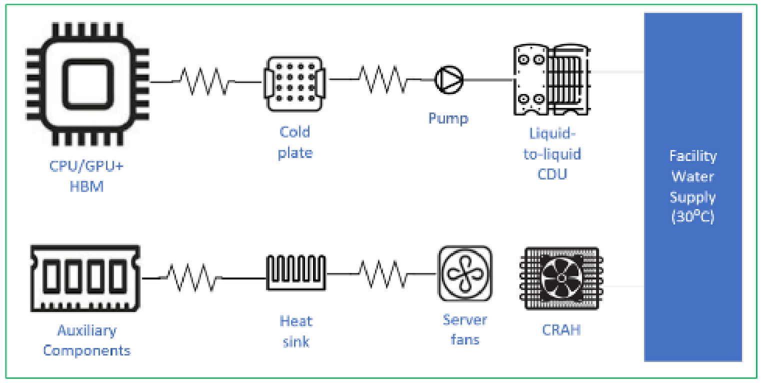
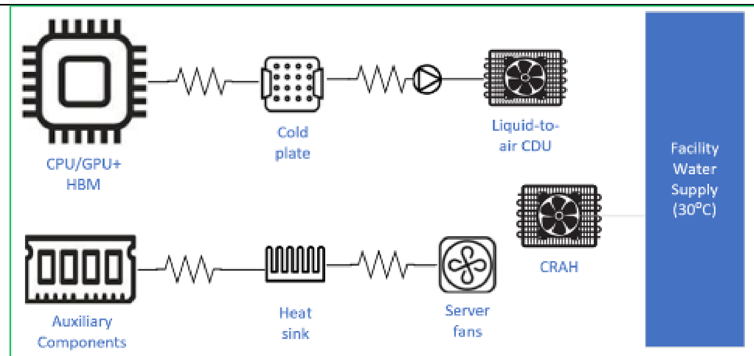
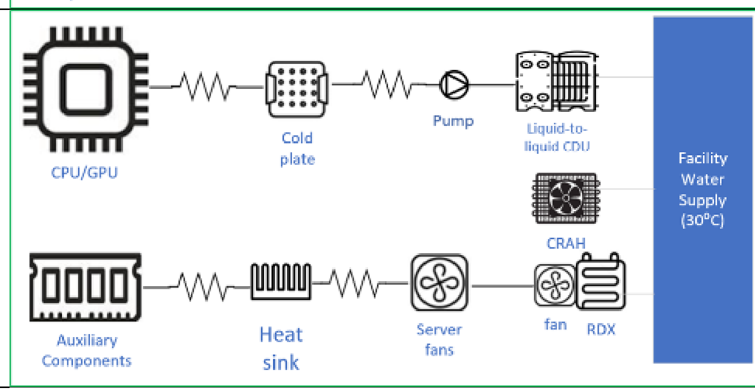
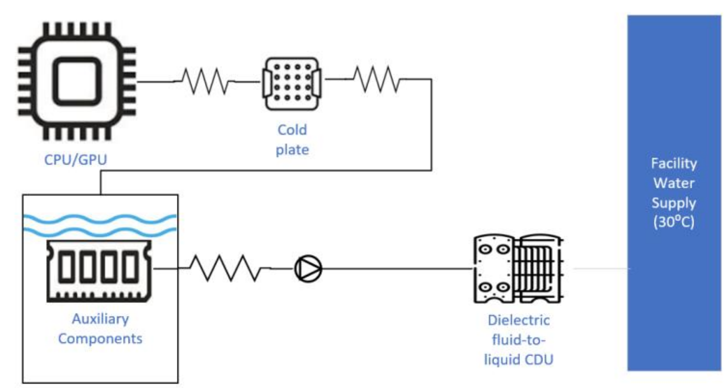
Classification of cooling strategies

Based on the different cooling strategies that exist, we have classified datacenter cooling into single-mode or hybrid approaches. Table 1 summarizes these strategies and provides schematics of each. Single-mode systems cool all compute components with one method and using a single fluid. The three single-mode strategies are (i) air-cooling (baseline), (ii) cold plate-based cooling for all components, and (iii) immersion cooling. Air-cooling using CRAC/CRAH units is the most common type of cooling and has advancements in efficiency with hot or cold aisle containment. All-cold-plate designs use direct-to-chip (D2C) liquid cooling for CPUs, GPUs, and auxiliary components (AUX), targeting heat sources directly and eliminating fan use. Full-immersion cooling uses a tank filled with di-electric fluid to submerge the servers instead of rack design in other approaches. Proprietary methods are used to direct more fluid to high thermal design power (TDP) components, such as CPUs and GPUs.

Hybrid cooling strategies are being increasingly explored as datacenter densities continue to rise. One hybrid strategy combines D2C with traditional air-cooling, with D2C handling loads from high TDP components (CPUs and GPUs), while AUX residual heat is removed through air systems. In this hybrid strategy, 26% of the total TDP is rejected through air-cooling and the rest through liquid cooling. A second hybrid strategy building on the first hybrid case is adding liquid-to-air coolant distribution units (CDU) which are useful in retrofitting fully air-cooled datacenters with liquid cooled racks. A third hybrid strategy consists of D2C combined with rear-door heat exchangers (RDX). The rear door heat exchangers extract heat dissipated by the AUX from air into the liquid, thereby reducing or eliminating the load on the CRAHs in the datacenter hall. The final hybrid strategy considered in this study combines D2C with immersion cooling. A dielectric fluid first flows through cold plates to cool the GPUs or CPUs and then circulates to immersion-cool the server's AUX and finally transferring that heat to the primary side through the dielectric fluid-to-liquid CDU. Each of these approaches balances energy efficiency, technical feasibility, cost, and operational complexity in different ways.

Table 1: List of cooling categories, type and schematics

Cooling category	**Type of cooling**	**Schematic**
Single mode	Traditional CRAC/CRAH air-cooling	
Single mode	D2C with cold plates	
Single mode	Full-immersion cooling	

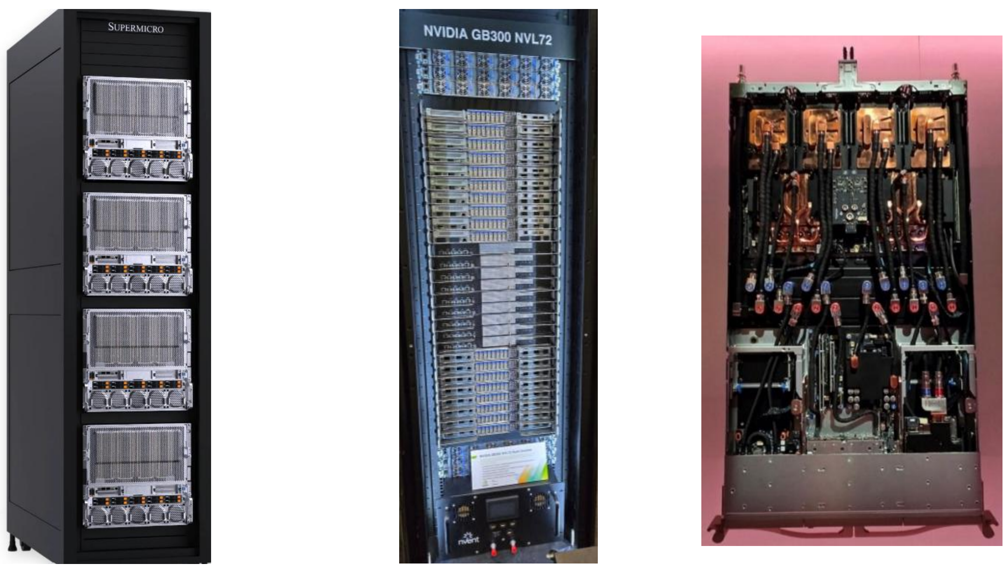
Hybrid mode	D2C cold plate + traditional air-cooling	
Hybrid mode	D2C cold plate with liquid-to-air CDU + traditional air-cooling	
Hybrid mode	D2C cold plate + Rear door HX for traditional air-cooling	
Hybrid mode	D2C cold plate + immersion cooling	

Methodology and Approach

We use a four step process to analyze the cooling energy use for various single-mode and hybrid cooling strategies: (1) develop reference system design for the compute servers and racks (2) develop reference datacenter facility layout (3) identify metrics for evaluation (4) develop component level models that combine into system models.

Reference system design

We develop normalized air-cooled and liquid-cooled rack reference designs based on NVIDIA's latest air- and liquid-cooled AI rack offerings, as NVIDIA is the dominant AI chip supplier and collaborates broadly on server and rack implementations built around its reference cooling solutions. These designs standardize load density, component layout, and thermal interface materials, and inlet temperature of coolant from the facility loop, so that observed performance differences arise solely from the cooling method. Using these reference designs, we apply a consistent evaluation framework centered on secondary-loop energy use, compute power density, and retrofit feasibility. The cooling architectures and thermal models developed here are generic and can apply to high-density AI racks from any vendor (Chakraborty, et al., 2025).



(a) **(b)** **(c)**

Figure 1: Reference architectures: (a) Air-cooled design: Supermicro HGX B200 rack with four 10U compute server. (b) Liquid cooled design: NVIDIA GB300 rack layout with (c) 1U compute server tray.

Air-cooled rack reference design. We model an air-cooled rack using NVIDIA Blackwell B200 GPUs (1,000 W TDP each). A server is 10 rack units (U) in height and consists of eight GPUs, two CPUs, NVlink switches, DPUbluefield, and memory has a total TDP of about 10.3 kW. Four such servers fit in a 48U rack as illustrated in Figure 1(a) giving ~41 kW silicon TDP. Including power supply losses (~7%, or 2.9 kW), rack TDP is ~44 kW.

Liquid-cooled rack reference design. We model the liquid-cooled system on NVIDIA's GB300 architecture, a fully liquid-cooled, high-density design. Each superchip combines two 1,400 W GPUs with a 350 W CPU. Two superchips along with NVlink switches, DPU, and memory form a 1U server tray with a ~7 kW TDP. The rack contains 18 compute trays and 9 switch trays (each with ~2kW TDP) in an alternating pattern, as shown in Figure 1(b) & (c). Total silicon power is ~126 kW. Including switch trays and power supply losses, total rack TDP is ~154 kW.

Datacenter Facility and Layout

We choose a fixed 30°C facility water supply (FWS) temperature for energy comparisons (OCP, 2024). A reference datacenter facility is modeled to evaluate different cooling systems. For air-cooled capacities above ~35 kW/rack, a fan-wall CRAH layout is recommended instead of a raised floor design. The fan wall design was modeled using Coolsim (Applied Math Modeling, 2025) for a datacenter room (26×25 m, 4.85 m height, 1.3 m ceiling plenum), as shown in Figure 2. The baseline air-cooled design has 80 racks (44 kW each) with hot-aisle containment in between two rows of server racks. A total of four such double rows make up the arrangement of racks in the datacenter room. A total of 16 CRAHs, each rated at 228.6 kW and supplied with 30°C chilled water are located on the sides of the datacenter room. In the liquid-plus-air hybrid design, racks dissipate ~40.5 kW to air from auxiliary components and power supply. The same configuration of four double rows of racks and 16 CRAHs supply air as in the baseline case. In addition, there are 16 heat-dissipation units (HDUs) that extract heat from the server racks directly to the liquid. Each of these HDUs dissipates 1.875 kW parasitic heat to the datacenter room that also needs to be cooled by the CRAHs. The total air-side load is ~3.27 MW. CRAHs are sized at 211 kW each, providing 3.37 MW total cooling capacity.

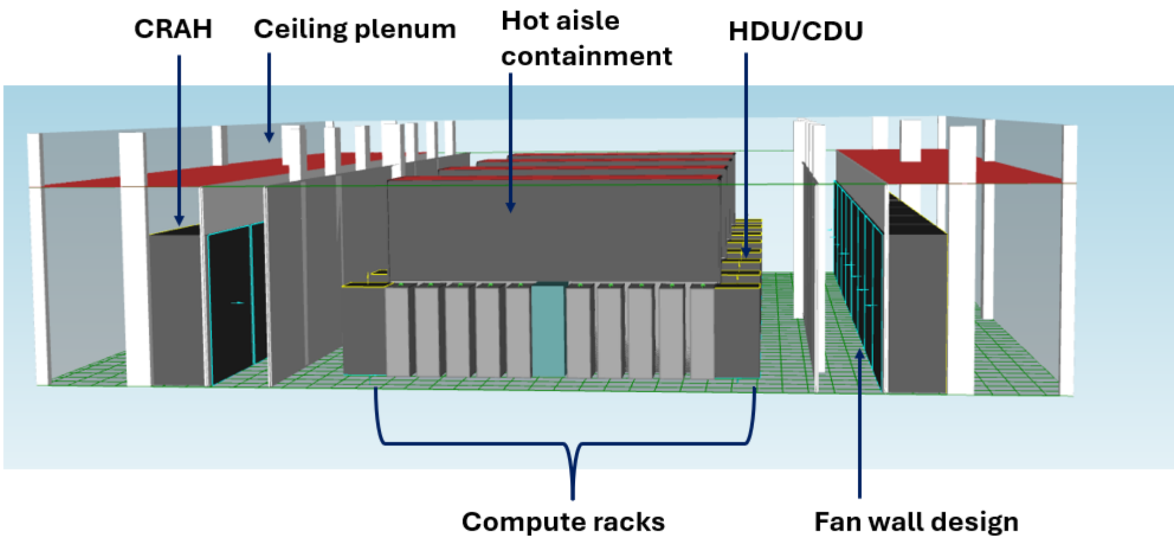


Figure 2: Air and Liquid hybrid cooled datacenter room layout in Coolsim (Applied Math Modeling) for modeling secondary cooling loop energy consumption and compute power density

Metrics of Evaluation

Datacenter cooling consists of two loops: a primary loop that rejects heat outdoors and a secondary loop that removes heat from IT equipment, as shown in Figure 3. Server heat is first absorbed by the secondary loop, which may use air, liquid, or hybrid cooling (combination of two coolants). This heat is transferred to the primary loop through a CRAH unit (air-cooled) or a CDU (liquid-cooled). The primary loop rejects heat using chillers, cooling towers, dry coolers, or adiabatic systems. Because primary-loop performance depends on external conditions and is typically designed for a 30°C Facility Water Supply (OCP, 2024), this study focuses on the secondary loop, which varies across cooling technologies. Future studies will evaluate the impact of choosing a higher FWS temperature.

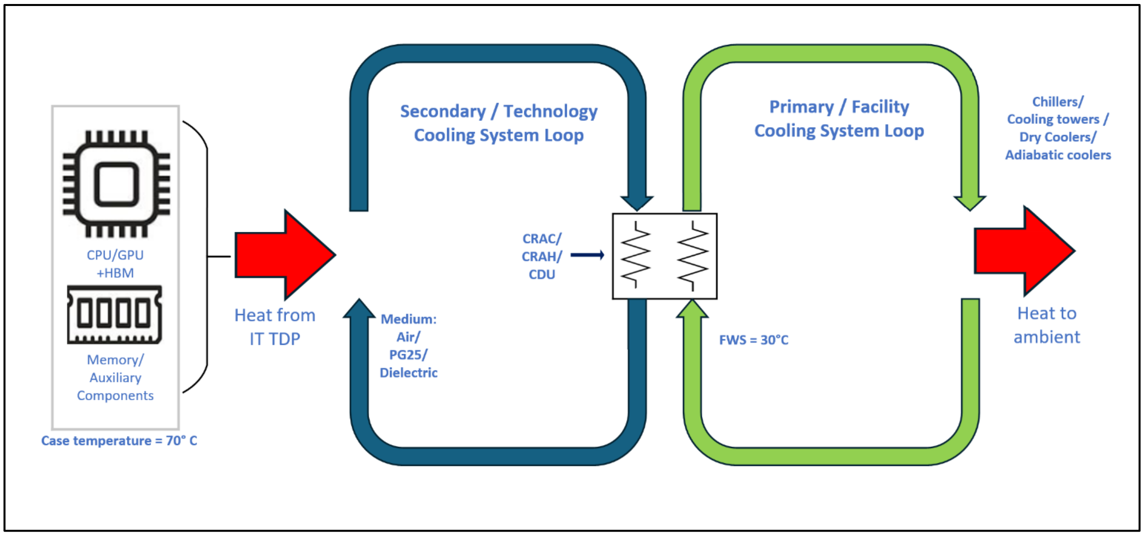


Figure 3: Datacenter cooling plant primary and secondary loops.

We benchmark liquid-cooling configurations against a state-of-the-art air-cooled datacenter baseline using four criteria:
(1) *Secondary-loop cooling energy:* the electrical power required to keep chip temperatures at or below a specified case temperature, chosen to be 70°C. This threshold represents a conservative estimate for 1.4kW GPU used in this study based on OCP liquid cooling guidelines, suggesting 75°C case temperature for a 1kW chip (OCP, 2023). The facility water temperature is kept constant at 30°C to keep the primary side energy consumption percentage same across all cooling strategies. The electrical power use includes fans in various components (Server, CRAH, RDX and liquid-to-air CDU), pumps, and other heat-transport components in the secondary loop. We use the $SCCP_{ratio}$, defined as secondary-loop electrical power divided by total IT compute TDP,

$$SCCP_{ratio} = \frac{Secondary\ loop\ input\ electrical\ power}{Total\ IT\ compute\ power}$$

(2) *Compute density:* compute power per unit floor area occupied by the secondary cooling system ($C_{density}$), indicating how much compute a cooling technology can support per square meter of the datacenter room,

$$C_{density} = \frac{IT\ compute\ power\ supported}{Total\ floor\ area\ of\ the\ data\ center\ room}$$

(3) *Retrofit feasibility:* infrastructure and operational changes needed to transition from air to liquid cooling.
(4) *Reliability, environmental impacts, and cost:* qualitative factors tied to hardware, plumbing, and cost drivers.

Component and system level energy modeling

The secondary loop energy consumption and performance metrics for each cooling strategy were calculated based on energy and heat transfer modeling of the individual components. The components modeled in this study are generic representations informed by the design principles of commercially available products. They are only intended for comparative analysis of various cooling strategies and do not replicate any specific product from any individual manufacturer. Table 2 identifies the key component models that are developed and combined together to estimate system level $SCCP_{ratio}$ and $C_{density}$ metrics. The key components that have been modeled for air and air-based hybrid cooling are: CRAH with 30°C FWS and 35°C air outlet temperature, the datacenter room with fan wall configuration, and server fan with heat sinks. The key components enabling liquid cooling that have been modeled are: D2C cold plates with PG25 and di-electric fluid, liquid-to-air CDU, liquid-to-liquid CDU, and rear door heat exchangers. Full-immersion cooling strategy is not modeled since it uses proprietary technologies to direct the flow at the hot spots on the chips. Instead, product brochures and discussions with immersion cooling providers and manufacturers are used to calculate energy metrics for this scenario. We restrict our modeling to single-phase technologies; future work will evaluate two-phase cold plates and immersion cooling. The different component level models identified in Table 2 are described further in this section.

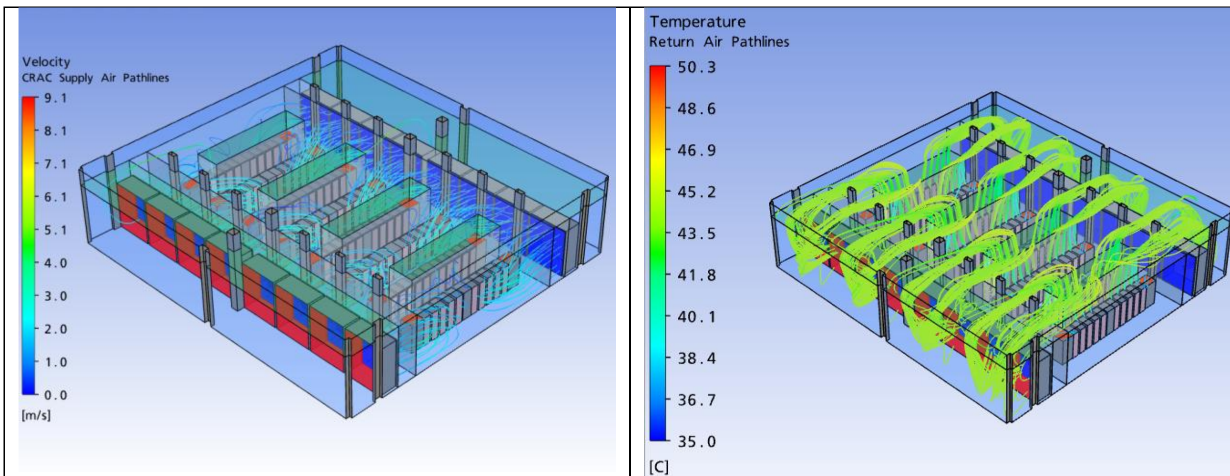
Table 2: Components models used for each cooling strategy

Cooling Technology	Model 1	Model 2	Model 3	Model 4
Traditional CRAC/CRAH air-cooling	Datacenter room and CRAH simulations	Server fan with heat sink analysis		
D2C with cold plates	D2C cold plate simulation for all components	Liquid-liquid CDU		
Full-immersion cooling	Manufacturer and published research data used for comparison			
D2C cold plate + traditional air-cooling	D2C cold plate simulation for GPU and CPU	Liquid-to-liquid CDU for GPU and CPU	Datacenter room and CRAH for AUX	Server fan with heat sink for AUX
D2C cold plate + liquid-to-air CDU for traditional air-cooling	D2C cold plate simulation for GPU and CPU	Liquid-to-air CDU for GPU and CPU	Datacenter room and CRAH for AUX	Server fan with heat sink for AUX
D2C cold plate + Rear door HX for traditional air-cooling	D2C cold plate simulation for GPU and CPU	Liquid-to-liquid CDU for GPU and CPU	RDX model for AUX	Server fan with heat sink for AUX

D2C cold plate + immersion cooling	D2C cold plate for GPU and CPU with dielectric fluid	Dielectric oil flow through AUX with heat sinks	Liquid-to-liquid CDU	

Air-Cooled room and CRAH Simulation

Eighty 44-kW racks were simulated in Coolsim (Applied Math Modeling, 2025) for the fully air-cooled design. As shown in Figure 4(a), CRAHs supply 35°C air from both sides of the datacenter room, and exhaust air at ~45.5°C is returned via the hot-aisle containment through the ceiling plenum and is circulated back as return air to the CRAH (Figure 4(b)). The airflow path across the room produces a 36-Pa pressure drop. A finned-tube CRAH coil sized for 228 kW adds ~56.4 Pa, giving a total air-side pressure drop of ~92.4 Pa. The CRAH coils are sized with a flow area of 3 x 3 m^2, with 0.625 in outer diameter of copper tubes and 14 fins per inch of 0.006 in thick aluminum fins. The same methodology is used to estimate secondary-side energy demand for air-based hybrid systems, where each rack rejects 40.5 kW to air and each CRAH delivers 211 kW of cooling.



(a) (b)

Figure 4: Full air-cooled datacenter simulation pathlines (a) from CRAH to racks and (b) from racks to CRAH.

Server Fan and Heat Sink Analysis

Server-fan energy for air-cooling was estimated using a simplified model. Tray-level fans supply airflow through heat sinks, and the fan–chip–heat-sink assembly was modeled in Ansys Icepak (Ansys by Synopsis , 2025). A 40×40 mm fan cooling 35-W chips of two adjacent NV link cards in a 1U server was analyzed using a Sunon fan curve (SUNON, 2025) to determine the airflow and static pressure needed to keep chip case temperature at 70°C with 35°C inlet air.
The chips had a rectangular straight fin heat sink with thermal interface material (TIM) in between the heat sink and the chip. Simulations show the heat sink requires 172 CFM to maintain this temperature of 70ºC. With an assumed 35% combined motor and fan efficiency, fan power is approximately 0.005 kW per kW of thermal TDP. This value of fan power was used for all the auxiliary components in a hybrid system with air-cooling (Model 4 in Table 2).

D2C Cold Plate Simulation

D2C cold plates can be a solution in single-mode liquid cooling approach and are also the most common component in hybrid liquid-cooled designs, supplied with PG25 (assumed at 35°C) by a liquid-to-liquid CDU. This CDU separates the primary and secondary cooling loops and uses pumps and filters to provide chilled PG25 to the server's liquid-cooled components. This study's reference design, which incorporates the NVIDIA GB300, employs an all-parallel tubing configuration, making it crucial to maintain a consistent pressure drop across all parallel lines as shown in Figure 1(c).

In this study, cold plates were specifically designed to handle 1,400 W, the highest TDP of the GPUs, and this same design is uniformly applied to all other server components. The most prevalent cold plate design in the industry features skived fins, where the chilled fluid (PG25) enters the cold plate from the middle of the shorter dimension and flows outward to minimize pressure drop, as shown in Figure 5 (a). Through parametric simulations, the team identified a cold plate dimension for high heat transfer rate and reduced pressure drop, consisting of 0.2 mm thick fins with 180-micron spacing and a height of 1.5 mm. These dimensions were subsequently used to calculate performance metrics for various cooling strategies.

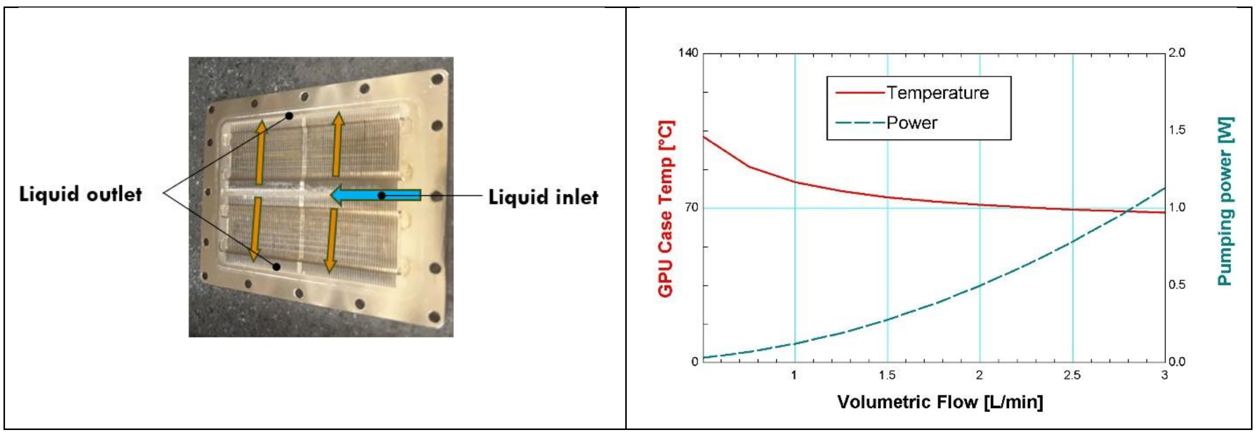


Figure 5: (a) Example of Skived-fin cold plate with liquid flow directions. (Alphacool, 2025) (b) Skived cold plate simulation of case temperature and pumping power for 1400 W GPU

Figure 5 (b) shows how coolant flow rate affects GPU case temperature and pump power for a 1,400-W GPU cooled with 35°C PG25. Increasing flow rate sharply reduces case temperature at low flow rates, then yields diminishing returns, while pumping power rises exponentially with flow rate increase. An optimal trade-off is therefore required between thermal performance and pump energy. A flow rate of 2.25 LPM keeps the GPU below the 70°C case-temperature limit while minimizing pumping power, providing an efficient operating point for the liquid-cooling design. For hybrid immersion cooling, dielectric oil was used as the coolant.

Liquid-to-Liquid CDU Performance

Liquid-to-liquid CDUs are the central infrastructure for any datacenter liquid-cooling system, managing coolant circulation, temperature control, and pressure regulation. They isolate the facility water loop from the server-side secondary loop: the primary side interfaces with the heat-rejection plant, while the secondary side delivers coolant directly to IT equipment. A typical CDU includes pumps, a plate-and-frame heat exchanger, filtration, and an expansion/pressurization tank.

For this study, a simplified in-row CDU with 2,300 kW capacity—consistent with commercial designs (CoolIT Systems 2025, Vertiv Group Corp. 2025)—was modeled to serve 20 hybrid racks along a hot-aisle containment. Secondary coolant modeled for D2C cold plate is PG25 and for immersion hybrid is dielectric oil. Flow requirements derived from D2C cold-plate modeling were used to estimate secondary-loop pressure drops for each cooling technology. Plate-heat-exchanger sizing software provided secondary-side pressure-drop estimates, while pressure losses from filtration and valves were assumed to be twice the cold-plate pressure drop, based on prior work by the group (Milan, 2023; ARPA-E, 2021). Total system pressure drop was then used to calculate CDU pump power for the fully liquid-cooled and hybrid configurations.

Rear-Door Heat Exchanger Modeling
The rear-door heat exchanger (RDX) in a hybrid cooling scenario absorbs the heat from the auxiliary components while using liquid cooling to extract heat from GPUs and CPUs. Modern RDX can extract 100 percent of the heat of the auxiliary components, which means the air temperature (35°C) leaving the RDX is the same as the temperature entering the rack. To achieve the cooling capacity of 40.5 kW/rack generated by the auxiliary components RDX is modeled with 48U height and 600 mm width to fit a standard datacenter rack. The CoilDesigner model of the RDX as a micro-channel heat exchanger consists of two tube banks, each containing 266 tubes. Each tube has an overall length of 0.6 m, width of 60 mm, and height of 3 mm. The tubes are microchannel-type, featuring rectangular channels with a width of 4.3 mm and a height of 1.8 mm. The RDX with a flow rate of 6950 CFM required 115 Pa of pressure head for the fans. This equates to electrical power consumption of 0.76 kW for the RDX fans with a combined motor and fan efficiency of 50 percent (Ziehl-Abegg AG, 2021).

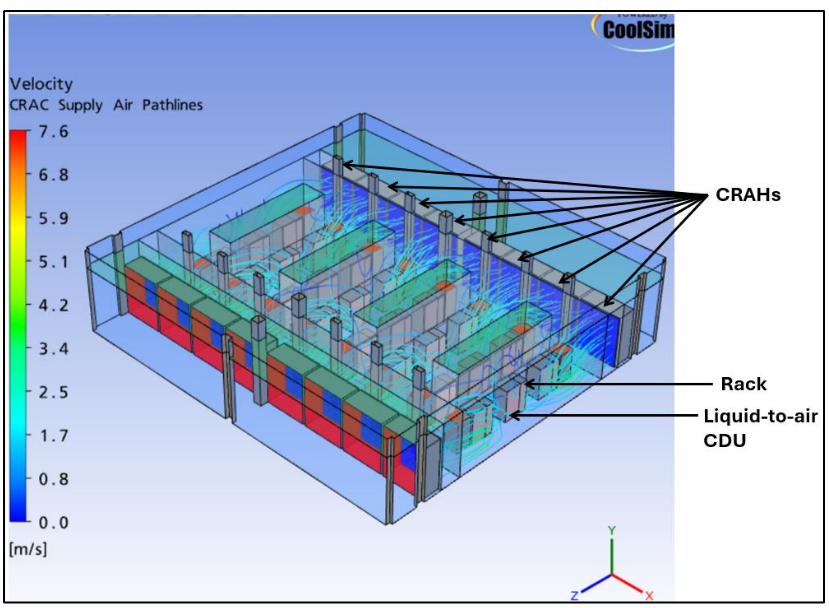
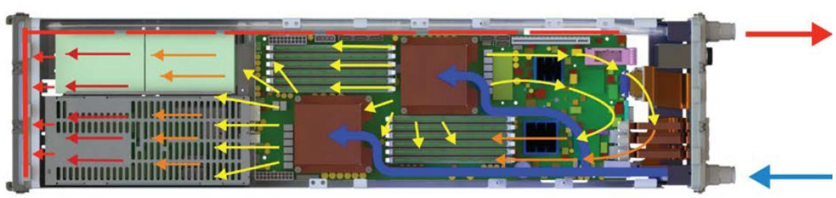


Figure 6: Coolsim modeling of liquid-to-air CDU of (a) supply air path lines, and (b) temperature distribution

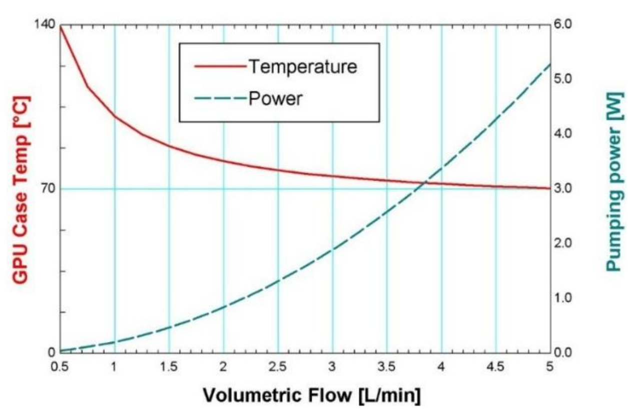
Liquid-to-Air CDU Performance
Several manufacturers offer liquid-to-air CDUs with varying capacities. For this study, the CDU's air-coil heat exchanger—responsible for rejecting CPU and GPU heat to room air—was modeled in CoilDesigner (Optimized Thermal Systems, Inc, 2025). The required air-side

pressure head for the CDU fans is 94.5 Pa at 15,600 CFM, corresponding to 1.39 kW of fan power assuming 50% combined motor-plus-fan efficiency (Ziehl-Abegg AG, 2021).
Because the CDU transfers heat from liquid to air before it is removed by room CRAHs, the additional thermal resistance raises the secondary-loop liquid supply temperature to 43°C (an 8°C approach over the 35°C CRAH supply air). This higher inlet temperature requires greater PG25 flow through the D2C cold plates to maintain GPU temperatures below 70°C. Modeling showed that even at 4 LPM, the GPU could not sustain its full 1,400-W TDP and had to be throttled to 1,300 W. An air-side layout study was conducted to determine achievable compute density within the same datacenter room used for the fully air-cooled baseline. Only 24 racks paired with liquid-to-air CDUs fit within the cooling capacity of the existing 16 CRAHs (228 kW each), compared to 80 racks in the fully air-cooled configuration, as shown in Figure 6.

Hybrid Cold Plate and Immersion Analysis
In a hybrid cold plate and immersion cooling solution, a dielectric fluid first flows through cold plates to cool the GPUs or CPUs and then circulates to immersion-cool the server's auxiliary components, as shown in Figure 7 (a) (Kozubal, 2017). To ensure adequate GPU cooling with this setup, the skived-fin cold plate was modeled using Opticool 87252 dielectric fluid, with an increased flow rate up to 5 LPM. However, due to the heat transfer properties of the coolant, the GPU compute power needs to be throttled to 1,100 W even with increased flow rate.
As shown in Figure 7 (b), the 4.75 LPM flow rate is required to keep the GPU maximum case temperature below 70°C, which also led to increased pumping power as compared to PG25 in D2C cold plate solutions. Using the outlet temperature from the cold plates (44°C), the heat transfer from the AUX was modeled using the same heat sink design as an air-cooled server, confirming that these components also remained within the 70°C case temperature limit.



(a)



(b)

Figure 7: Hybrid immersion modeling and simulation. (a) Dielectric coolant flow pattern and (b) cold plate modeling of the case temperature.

Results and discussion

The component models, developed using first principles, discussed in the prior section were used together to simulate overall cooling technologies to compare across strategies and validated with manufacturer published performance tables. Table 3 shows a summary of the different component models (described above) that were used to simulate secondary cooling energy consumption ($SCCP_{ratio}$) and compute density ($C_{density}$) for each datacenter cooling strategy. The energy modeling results along with the market insights (ease of retrofit, environmental impact, and capital cost) for each cooling strategy are shown in Figure 8. Single-mode strategies are compared in Figure 8a while hybrid strategies are presented in in Figure 8b. The spider charts are arranged in a way where the desirable outcomes of the metrics are put at the center and thus smaller the area under each cooling strategy the better it is compared to others.

Traditional Air-Cooled (Baseline CRAH System). As the study's baseline, the fully air-cooled configuration, shown in blue shading in Figure 8(a), employs traditional CRAC/CRAH systems. It requires high cooling energy, resulting in an $SCCP_{ratio}$ of 2% - highest among the strategies evaluated in this study. Despite advancements in aisle containment, the inherent low heat capacity of air limits this architecture to rack densities to 44 kW, which corresponds to a $C_{density}$ of 5.4 kW/m². This mature technology is well understood and easy to maintain. However, it is less efficient at high densities and prone to thermal hotspots if airflow is poorly managed. While this architecture maintains a low environmental impact, it may require higher land use. As the baseline for capital cost and retrofitting, it represents the most economical option. Waste heat recovery potential is low and would typically need higher water and energy usage.

Full D2C Cold Plate. The all-D2C scenario, shown in red shading in Figure 8(a), achieves the highest rack power density of 154 kW ($C_{density}$ of 82.7 kW/m²) and yields the lowest secondary side cooling power consumption with an $SCCP_{ratio}$ of 0.14%. However, retrofitting is difficult because it necessitates specialized server designs for liquid cooled components, internal tubing management, and rack level CDUs. This transition involves replacing fans with pumps and results in higher capital costs for cold plates, hoses, and fittings. The system offers predictable performance, though it requires robust leak detection protocols. Environmental impact remains low when using PG25. Waste heat recovery can be attained easily with this strategy due to higher liquid temperatures returning from the racks and would also allow use for dry-coolers lowering overall energy and water usage.

Full-Immersion Cooling. Full-immersion cooling, shown in green shading in Figure 8(a), submerges all components in dielectric liquid, supporting moderate densities of 100 kW or more per tank. It results in an $SCCP_{ratio}$ of 1.08%, largely due to the high pump power requirements of dielectric oil. The $C_{density}$ is 16.3 kW/m² for a 32°C FWS, though values fluctuate based on manufacturer design and fluid selection. While lower coolant temperatures can increase density, such adjustments lead to higher primary loop energy consumption. This cooling strategy is suitable for new construction due to its distinct architecture requirements, though it presents a risk of vendor lock-in and a moderate to hard retrofitting path depending on the existing rack form factor. Maintenance is more involved, requiring fluid monitoring and periodic replacement of dielectric oils which require proper recycling and can only be achieved with proper training of the datacenter operations team. The system carries high capital costs for specialized immersion baths and CDUs. Waste heat recovery can be attained easily with this strategy due to higher liquid temperatures from the tank and would also allow use for dry-coolers lowering overall energy and water usage.

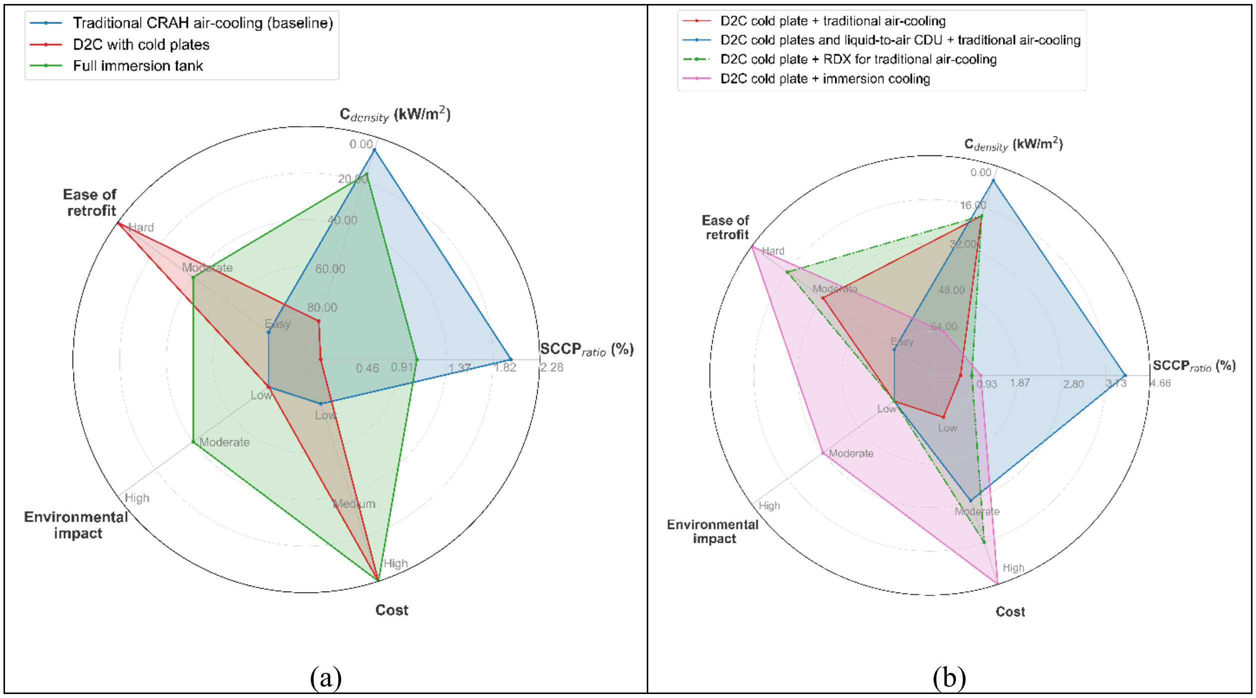


Figure 8: Spider charts showing energy and market insights for (a) single-mode datacenter cooling strategies and (b) hybrid cooling strategies

D2C cold plate + traditional air-cooling. This strategy, shown in blue shading in Figure 8(b), integrates D2C for primary components with traditional air-cooling for AUX, resulting in an SCCPratio of 0.65%. This hybrid configuration achieves a Cdensity of 18.9 kW/m². Retrofitting is considered moderate as existing CRAH units remain usable. Reliability is supported by a high level of understanding of traditional air cooling, though it requires careful thermal zoning and dual maintenance protocols for the combined liquid and air systems. Environmental impact is low because of PG25 fluid. Costs are moderate due to cold plates and liquid-to-liquid CDUs. Waste heat from the liquid cooled components (GPUs and CPUs) can be recovered easily due to higher liquid temperatures and would also lower overall energy and water usage.

D2C cold plate with liquid-to-air CDU + traditional air-cooling. The strategy, shown in red shading in in Figure 8(b), utilizes a liquid-to-air CDU to extract heat using D2C from the primary components, and requires the highest secondary cooling energy with an $SCCP_{ratio}$ of 4.14%. In this mode, heat from liquid-cooled racks is rejected into the room air, placing an additional load on CRAH units. Consequently, the $C_{density}$ is limited to 5.4 kW/m², due to the fixed CRAH capacity for the datacenter room. The advantage of this strategy is that it allows for liquid cooling deployment without installing new facility liquid lines. The system features a small, contained liquid loop and coordination between hybrid airflow and liquid components. While environmental impact is low, it may involve higher land use for the same compute capacity. Capital costs are higher than the baseline due to the liquid-to-air CDUs. Recovering any waste heat for this technology would be very hard due to lower return air temperatures and would result in higher water and energy usage.

D2C cold plate + Rear door HX for traditional air-cooling. Utilizing D2C for primary components with RDX for AUX yields an $SCCP_{ratio}$ of 0.89%, as shown in green shading in in Figure 8(b). The $SCCP_{ratio}$ is slightly higher than the D2C+tradional air-cooling option due to the energy required by RDX fans. This strategy provides a $C_{density}$ of 18.9 kW/m² and serves as an effective retrofit when existing CRAH capacity is insufficient for AUX heat. RDX integration improves cooling redundancy, but the presence of multiple cooling paths increases complexity and requires coordinated control across three subsystems. Costs are moderately high, driven by the combination of cold plates, RDXs, and liquid-to-liquid CDUs. Waste heat from all components can be recovered easily and would also lower overall energy and water usage.

D2C cold plate + immersion cooling. The hybrid D2C and immersion strategy, shown in pink shading in in Figure 8(b), results in an $SCCP_{ratio}$ of 1.08%. Due to the thermal properties of the single-phase dielectric fluid, the GPU power still needs to be throttled from 1400 W to 1100 W to maintain a maximum temperature of 70°C. Despite lower secondary side efficiency, this strategy offers the highest $C_{density}$ among hybrid options at 63.4 kW/m² due to removal of air-side infrastructure. It is a hard retrofit that can be fitted into a rack architecture but typically results in low serviceability and risks of vendor lock-in. Immersion cooling offers uniform thermal management, yet the integration of two distinct liquid systems requires careful isolation and essential fluid monitoring. The environmental impact is medium because of the dielectric oils used. This configuration represents the highest cost tier among the hybrid options analyzed. Waste heat recovery potential is high and will have lower energy and water consumption.

Conclusions

This study provides a unified evaluation of datacenter cooling strategies by comparing secondary loop energy use, compute density, retrofit feasibility, cost, and environmental factors across air, liquid, and hybrid configurations. The baseline air-cooled system remains the most economical option but is constrained by the low thermal capacity of air and limited ability to support rising rack power densities. The results show that single-mode D2C achieves the lowest SCCP ratio (0.14%) and the highest compute density (82.7%), a ~14× reduction in secondary cooling energy and ~15× increase in compute density relative to the air-cooled baseline.

Hybrid strategies present practical transition paths for operators preparing for higher density workloads. D2C combined with traditional air-cooling offers meaningful reductions in secondary loop energy with moderate retrofit difficulty. D2C paired with RDX provides additional cooling capacity in facilities where CRAH performance is constrained. Hybrid D2C and immersion configurations can reach high compute density, although dielectric fluid limitations require GPU throttling and introduce integration challenges.

This study has been limited to single-phase liquid cooling solutions. As next-generation GPUs continue to increase heat flux, two-phase liquid cooling technologies will become increasingly important. Future work should evaluate two-phase cold plates and two-phase immersion systems using the same metrics applied here. Extending the analysis to higher FWS temperatures will also clarify how primary cooling systems without compressors can support rapid AI-driven load growth while maintaining energy efficiency and high reliability.

References

Alphacool. (2025). *Alphacool ES Jet LGA 4677 2U CPU Cooler*. Retrieved from https://www.titanrig.com/alphacool-es-jet-lga-4677-2u-cpu-cooler.html

Ansys by Synopsis . (2025). *Ansys Icepak: Electronics Cooling & PCB Thermal Simulation and Analysis*. Retrieved from https://www.ansys.com/products/electronics/ansys-icepak

Applied Math Modeling. (2025). *Coolsim: Application-Specific, Easy to Use, and Cost-Effective CFD*. Retrieved from coolsimsoftware.com: https://coolsimsoftware.com/about/

ARPA-E. (2021). *Cooling Operations Optimized for Leaps in Energy, Reliability, and Carbon Hyperefficiency for Information Processing Systems. COOLERCHPS*. Retrieved from https://www.arpa-e.energy.gov/technologies/programs/coolerchips

Asetek Inc. (2024). *Demonstration of Low-Cost Data Center Liquid Cooling.* California Energy Commission. Retrieved from https://www.energy.ca.gov/sites/default/files/2024-06/CEC-500-2024-061.pdf

Azarifar, M., Arik, M., & Chang, J.-Y. (2024). Liquid Cooling of Data Centers: A Necessity Facing Challenges. *Applied Thermal Engineering*, 247:123112. doi:https://doi.org/10.1016/j.applthermaleng.2024.123112

Buffington, B. (2025). *SCE Data Center Tracking and Forecasting.* Sacramento: California Energy Commission, 25-IEPR-03.

Chakraborty, S., Narayanan, V., Eorwyn, E., Rasouli, E., Tano, I.-N., & OutCault, S. (2025). *Datacenter Liquid Cooling Market Characterization.* CalNEXT. Retrieved from https://calnext.com/wp-content/uploads/2025/12/ET24SWE0065_Datacenter-Liquid-Cooling-Market-Characterization_Final-Report.pdf

Coles, H. C., & Greenberg, S. (2014). *Direct Liquid Cooling for Electronic Equipment.* Lawrence Berkeley National Laboratory. Retrieved from https://eta.lbl.gov/publications/direct-liquid-cooling-electronic

Curtis, R., Shedd, T., & Clark, E. B. (2023). Performance Comparison of Five Data Center Server Thermal Management Technologies. *9th IEEE Semi-THERM Symposium.* IEEE. Retrieved from https://ieeexplore.ieee.org/document/10267908

Kozubal, E. (2017). *Laboratory Study and Demonstration Results of a Directed-Flow, Liquid Submerged Server for High Efficiency Data Centers.* U.S. Department of Energy. Golden, CO: National Renewable Energy Laboratory (NREL).

Milan, J. (2023, July 25). *HoMEDUCS Project's Unique Approach to Keeping Modular Data Centers Cool*. Retrieved from https://www.datacenterknowledge.com/cooling/homeducs-project-s-unique-approach-to-keeping-modular-data-centers-cool

OCP. (2023). *Open Accelerator Infrastructure (OAI) System Liquid Cooling.* Open Compute Project.

OCP. (2024). *30°C COOLANT - A DURABLE ROADMAP FOR THE FUTURE.* Open Compute Project. Retrieved from https://www.opencompute.org/documents/oai-system-liquid-cooling-guidelines-in-ocp-template-mar-3-2023-update-pdf

Offutt, M. C., & Zhu, L. (2025). *Data Centers and Their Energy Consumption.* Washington D.C: Congressional Research Service. Retrieved from https://www.congress.gov/crs_external_products/R/PDF/R48646/R48646.3.pdf

Ohadi, M. M., Dessiatoun, S. V., Choo, K., & Pecht, M. (2012). Comparison Analysis of Air, Liquid, and Two-Phase Cooling of Data Centers. *28th IEEE Semi-THERM Symposium*, (pp. 58-63). doi:https://doi.org/10.1109/STHERM.2012.6188826

Optimized Thermal Systems, Inc. (2025). *CoilDesigner: A highly customizable software tool for the design, simulation and optimization of air-to-refrigerant heat exchangers*. Retrieved from https://optimizedthermalsystems.com/coildesigner/

Skidmore, Z. (2025, September 30). *PG&E announces $73bn grid infrastructure upgrade plan to meet surging data center demand*. Retrieved from https://www.datacenterdynamics.com/en/news/pge-announces-73bn-grid-infrastructure-upgrade-plan-to-meet-surging-data-center-demand/

SUNON. (2025, 08 25). *DC Brushless fan & blower.* Retrieved from https://www.sunon.com/en/MANAGE/Docs/PRODUCT/286/503/DC%20Fan_20250825(255-A)_%E7%B6%B2%E9%A0%81%E7%94%A8.pdf

Ziehl-Abegg AG. (2021). *Centrifugal fans with energy-saving EFF1/IE2 motor*. Retrieved from https://www.ziehl-abegg.com/