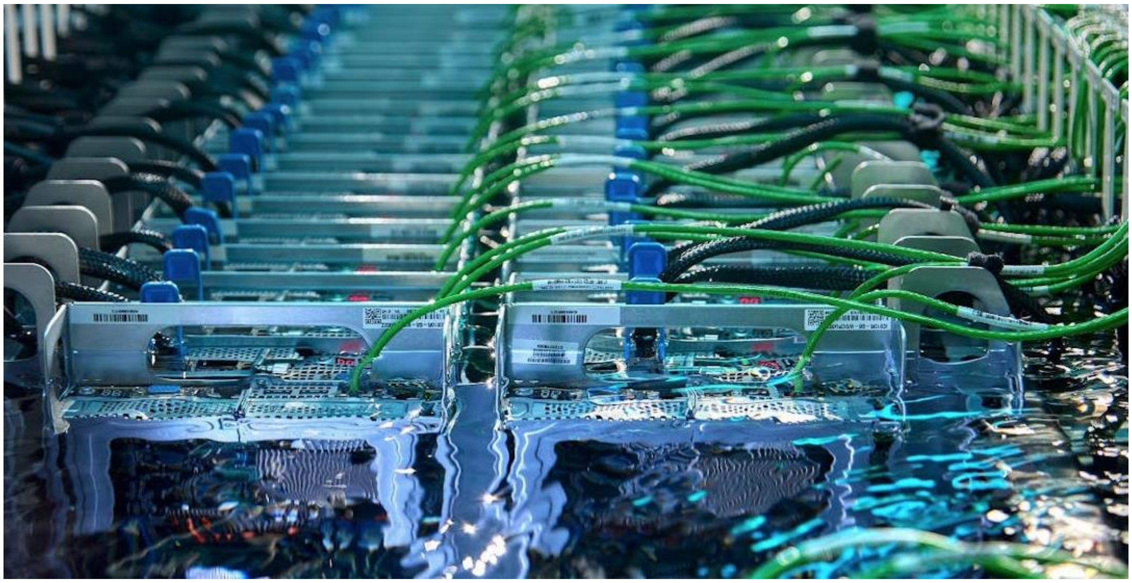
Datacenter Liquid Cooling Market Characterization

Final Report

ET24SWE0065



Source: datacenterdynamics.com

Prepared by:

Subhrajit Chakraborty,
Vinod Narayanan,
Ellian Eorwyn,
Erfan Rasouli,
Ines-Noelly Tano,
Sarah Outcault

WCEC, UC Davis

December 12, 2025

Acknowledgements

The authors gratefully acknowledge the funding received from the CalNEXT program in support of this research. Additionally, we greatly appreciate the willingness of the stakeholders—including datacenter cooling solution providers, utilities, and energy providers—to engage in thoughtful discussion.

Disclaimer

The CalNEXT program is designed and implemented by Cohen Ventures, Inc., DBA Energy Solutions ("Energy Solutions"). Southern California Edison Company, on behalf of itself, Pacific Gas and Electric Company, and San Diego Gas & Electric® Company (collectively, the "CA Electric IOUs"), has contracted with Energy Solutions for CalNEXT. CalNEXT is available in each of the CA Electric IOU's service territories. Customers who participate in CalNEXT are under individual agreements between the customer and Energy Solutions or Energy Solutions' subcontractors (Terms of Use). The CA Electric IOUs are not parties to, nor guarantors of, any Terms of Use with Energy Solutions. The CA Electric IOUs have no contractual obligation, directly or indirectly, to the customer. The CA Electric IOUs are not liable for any actions or inactions of Energy Solutions, or any distributor, vendor, installer, or manufacturer of product(s) offered through CalNEXT. The CA Electric IOUs do not recommend, endorse, qualify, guarantee, or make any representations or warranties (express or implied) regarding the findings, services, work, quality, financial stability, or performance of Energy Solutions or any of Energy Solutions' distributors, contractors, subcontractors, installers of products, or any product brand listed on Energy Solutions' website or provided, directly or indirectly, by Energy Solutions. If applicable, prior to entering into any Terms of Use, customers should thoroughly review the terms and conditions of such Terms of Use so they are fully informed of their rights and obligations under the Terms of Use, and should perform their own research and due diligence, and obtain multiple bids or quotes when seeking a contractor to perform work of any type.

Executive Summary

California’s datacenter energy demand is expected to surge by 10 gigawatts over the next 10 years, fueled by the rapid expansion of artificial intelligence, machine learning, and cloud computing. This growth presents a direct challenge to the state’s clean energy goals and grid reliability, particularly as conventional air-cooling systems—often responsible for 33 percent to 40 percent of a facility’s electricity use—prove inadequate for modern high-density server environments. This market study evaluates emerging liquid cooling technologies and hybrid systems, offering a strategic framework for utilities, policymakers, and datacenter operators to navigate the transition toward more efficient thermal management.

Using NVIDIA’s B200 (air-cooled, 44-kilowatt rack) and GB300 (liquid-cooled, 154-kilowatt rack) architectures as reference, the study compares cooling strategies across many dimensions. Important metrics used for comparison include secondary cooling to compute power ratio ($\text{SCCP}_{\text{ratio}}$), defined as the percentage of total rack power consumed by the secondary cooling loop, retrofit feasibility, compute density (C_{density}), and system complexity.

Single-mode cooling systems, such as air-cooled computer room air handler (CRAH) units, all-direct-to-chip (D2C) cold plates, and full immersion, vary widely in performance. Even the state-of-the-art CRAH systems exhibit high energy overhead ($\text{SCCP}_{\text{ratio}}$: 2.00 percent) and low rack density (C_{density}: 5.4 kilowatts per square meter). In contrast, all-D2C cold plates deliver superior efficiency ($\text{SCCP}_{\text{ratio}}$: 0.14 percent) and density (C_{density}: 82.7 kilowatts per square meter), though they pose significant retrofit challenges. Full-immersion cooling can achieve higher thermal performance than fully air-cooled solutions but is limited by requirements for lower facility water temperature, serviceability, and lack of standardization. For the facility water at 32°C, the $\text{SCCP}_{\text{ratio}}$ is 1.08 percent (about half of traditional air-cooling) and C_{density} is 16.3 kilowatts per square meter.

Hybrid cooling configurations offer more flexible pathways for retrofitting legacy facilities. Systems combining D2C with CRAH, rear-door heat exchangers (RDXs), immersion, or liquid-to-air coolant distribution units (CDUs) show varying trade-offs. D2C with CRAH setups are retrofit-friendly and efficient ($\text{SCCP}_{\text{ratio}}$: 0.65 percent; C_{density}: 18.9 kilowatts per square meter), while D2C with RDX units enhance air cooling but have little lower efficiency ($\text{SCCP}_{\text{ratio}}$: 0.89 percent). D2C with immersion enables ultra-high-density cooling (C_{density}: 63.4 kilowatts per square meter) but increases system complexity. Liquid-to-air CDUs, by contrast, are viable when liquid lines cannot be extended into the data center room, allowing selective integration of artificial intelligence or high-performance computing racks into traditional air-cooled halls. However, this approach carries a significant energy penalty ($\text{SCCP}_{\text{ratio}}$: 4.14 percent) and reduces the number of air-cooled racks that can be supported.

Key findings indicate that D2C cold plates are optimal for high thermal design power (TDP) chips above 1,500 watts, offering precise thermal control and reliability. Immersion cooling is best suited for lower TDP chips upto 1 kW, though hybrid immersion-D2C configurations extend its applicability to higher loads. Retrofit feasibility varies significantly: CRAH-based hybrids are the easiest to implement, RDX provides a strong option when auxiliary loads exceed CRAH capacity, and liquid-to-air CDUs serve as a fallback for facilities unable to extend liquid lines—albeit with efficiency trade-offs. Full immersion and two-phase systems require substantial infrastructure changes and remain limited by serviceability and standardization.

Fully liquid-cooled datacenters offer multiple advantages, including reduced secondary-side energy consumption and significantly increased compute density by eliminating air-based infrastructure. Additionally, they enable higher facility water temperatures, facilitate waste heat recovery when available, and reduce primary-side energy use by allowing the use of dry coolers instead of energy-intensive chillers.

From a commercial perspective, liquid cooling technologies are no longer experimental. D2C systems are widely available and increasingly standardized. CDUs from established vendors are sold as off-the-shelf products, while immersion systems are commercially available. Two-phase systems are technologically advanced but face barriers from fluid costs, environmental concerns (per- and polyfluoroalkyl substances and global warming potential), and limited field experience.

Cooling system market trends reflect three trajectories:

- D2C is the fastest-growing and most commercially mature.
- Immersion is technologically proven but slowed by cost and serviceability concerns.
- Hybrid systems are the most practical for retrofits, balancing efficiency with feasibility.

Emerging innovations include two-phase cold plates, modular rack-integrated CDUs, and new dielectric fluids with lower environmental impact. The industry trajectory is moving steadily from air-assisted liquid cooling toward fully liquid, high-density, and (eventually) two-phase systems for racks exceeding 200 kilowatts.

Datacenter market trends show that California is undergoing rapid transformation. Utilities project dozens of new facilities, ranging from 7-megawatt infill projects to 400+-megawatt campuses. Rack densities have risen from 10 kilowatts just a few years ago to 120 kilowatts today, with projections of 150 kilowatts or more in the near term. Hyperscale operators such as AWS, Microsoft, and Oracle are leading adoption of liquid cooling in purpose-built campuses, while colocation providers face retrofit constraints. Thermal management is becoming a competitive differentiator in site selection and market positioning.

California’s policy environment is evolving rapidly. The Small Power Plant Exemption has been leveraged by large datacenters to fast-track diesel backup generation, raising environmental and equity concerns. Legislative proposals—such as Senate Bill 1298, Assembly Bill 222, and Senate Bill 57—seek to increase transparency, require reporting of power usage effectiveness, and ensure that very large electricity users bear the costs of grid upgrades. Policymakers are also exploring incentives for liquid cooling adoption, R&D for serviceability and interoperability, and standards for retrofit-friendly designs. Utilities are considering how to integrate datacenters as flexible grid assets, balancing reliability with decarbonization goals.

Abbreviations and Acronyms

Acronym	Meaning
AI	Artificial intelligence
ARPA-E	Advanced Research Projects Agency–Energy
AUX	Auxiliary components in server
CDU	Coolant distribution unit
$C_{density}$	Computer power per unit area (kW/m^2)
CEC	California Energy Commission
CFM	Cubic Feet per Minute
CPU	Central processing unit
CRAC	Computer room air conditioner
CRAH	Computer room air handler
D2C	Direct-to-chip
DLC	Direct liquid cooling
FWS	Facility water supply
GPU	Graphics processing unit
GWP	Global warming potential
HDU	Heat dissipation unit
HBM	High bandwidth memory
HPC	High-performance computing
IT	Information technology
IOU	Investor-owned utility
kW	Kilowatt

Acronym	Meaning
kW/m^2	Kilowatt-compute per square meter
MW	Megawatt
OCP	Open Compute Project
Pa	Pascal
PFAS	Per- and polyfluoroalkyl substances
PG&E	Pacific Gas and Electric
PG25	25 percent concentration propylene glycol with water
PSU	Power supply unit
PUE	Power usage effectiveness
pPUE	Partial power usage effectiveness
RDX	Rear door heat exchanger
$SCCP_{ratio}$	Secondary Cooling to Compute Power Ratio (%)
SPPE	Small Power Plant Exemption
TDP	Thermal design power
2PF	Two-phase flow

Contents

Tables

Figures

Introduction

This report presents findings from a market characterization and energy comparison study on liquid cooling technologies for datacenters. The study was initiated in response to growing concerns about the energy impact of rapid datacenter expansion, particularly in California, where utilities anticipate significant load growth. Liquid cooling offers a promising alternative to conventional air-based systems, with the potential to reduce datacenter energy consumption by up to 20 percent and serve the rapidly increasing compute power density of datacenters.

The purpose of this report is to:

(1) Categorize liquid cooling technologies available or emerging in the commercial market.

(2) Evaluate their relative performance, energy savings potential, and ease of deployment.

(3) Summarize key market barriers and drivers.

(4) Present recommendations for energy efficiency programs, with a focus on technologies that merit further investigation.

The insights included here are intended for energy efficiency program administrators, utilities, technology developers, and stakeholders involved in emerging technologies and datacenter planning.

Background

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 data center 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 and 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 of air- lower thermal conductivity and lower heat capacitance.

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, unfamiliarity, 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 and 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.

At the same time, regulatory and market pressures are mounting. California's clean energy goals; grid constraints; and environmental, social, and governance reporting requirements are encouraging datacenter operators to pursue higher efficiency and lower-impact solutions. Modular datacenters and AI-optimized server designs further improve the viability of adopting liquid cooling at scale. Given this confluence of factors, a systematic market review is needed to inform policy and investment decisions.

Objectives

This study aims to characterize the current landscape of liquid cooling technologies for datacenters and identify the most promising options for future support through emerging technology programs.

Key Research Questions

- What liquid cooling technologies are currently available or nearing commercialization?
- How do they compare in terms of energy savings, reliability, cost, and retrofit feasibility?
- What are the primary barriers to market adoption?
- Which technologies should be prioritized for further validation and deployment?

Study Goals

- Categorize liquid cooling solutions and map them to specific use cases (e.g., retrofits, new construction, modular).
- Assess performance metrics using published data, lab findings, and manufacturer specifications.
- Identify non-energy benefits and potential risks (e.g., reliability, fouling, corrosion).
- Gather qualitative feedback from industry stakeholders across the supply chain.
- Provide actionable recommendations for program implementers and utility partners.

Deliverables and Tech Transfer

The expected deliverables of this study include a comprehensive report that summarizes the various categories of liquid cooling technologies, the current market landscape, and key challenges to adoption. The report will provide a curated list of high-potential technologies that are recommended for further study or demonstration, based on performance data, stakeholder feedback, and market readiness. In addition, it will offer guidance on how selected technologies can be integrated into custom or deemed energy efficiency programs led by investor-owned utilities (IOUs).

The technology transfer plan targets California IOUs, CalNEXT partners, and energy efficiency program managers as the primary audience for the study's findings. Dissemination will occur through multiple channels, including the Final Report, stakeholder briefings, webinars, and collaboration with emerging technology implementers. The intended impact of this transfer effort is to support informed program design, reduce the perceived risk of adopting new technologies, and accelerate the deployment of energy-efficient liquid cooling solutions across California datacenters.

Methodology and Approach

Datacenter Rack and Server Architecture

In this study, we adopt a consistent and comparative methodology to evaluate high-performance, high-density AI datacenter configurations, based on the latest commercially relevant architectures. We specifically select NVDIA's recent air-cooled and liquid-cooled offerings for AI factories, designs released at the GTC in March 2025. As the largest AI chip manufacturer, NVIDIA partners with a number of entities—including Supermicro, Dell, ASUS, ASRock Rack, GIGABYTE, Ingrasys, Inventec, Pegatron, Quanta Computer, Wistron, Wiwynn, and AIVRES—to build servers and racks based on NVIDIA reference cooling solutions for AI datacenter use cases. However, the cooling solution architectures and thermal models developed in this study are generic and are equally applicable to high-density AI racks from other providers.

Air-cooled rack reference design

For the air-cooled design, we consider a state-of-the-art server configuration using NVIDIA's Blackwell B200 GPUs, each with thermal design power (TDP) of 1,000 watts (W). A typical server integrates eight B200 GPUs along with supporting components, including two Intel CPUs and high-speed memory, yielding a total server-level TDP (Table 1) of approximately 10.3 kilowatts (kW). These servers are housed in a 10U chassis, and up to four such servers can be deployed within a standard 48U rack (as illustrated in Figure 1), resulting in a rack-level silicon TDP of approximately 41 kW (Table 2). Although an upper bound of 14 kW per 10U B200 server is published by NVIDIA, it has not yet been adopted by the partners, as there is ongoing work on optimal airflow management and containment strategies. As shown in Table 2, the total TDP of the rack (~44 kW) also includes heat dissipation from the power supply units (PSUs), which is approximately 7 percent of the rack TDP, i.e., 2.9 kW.

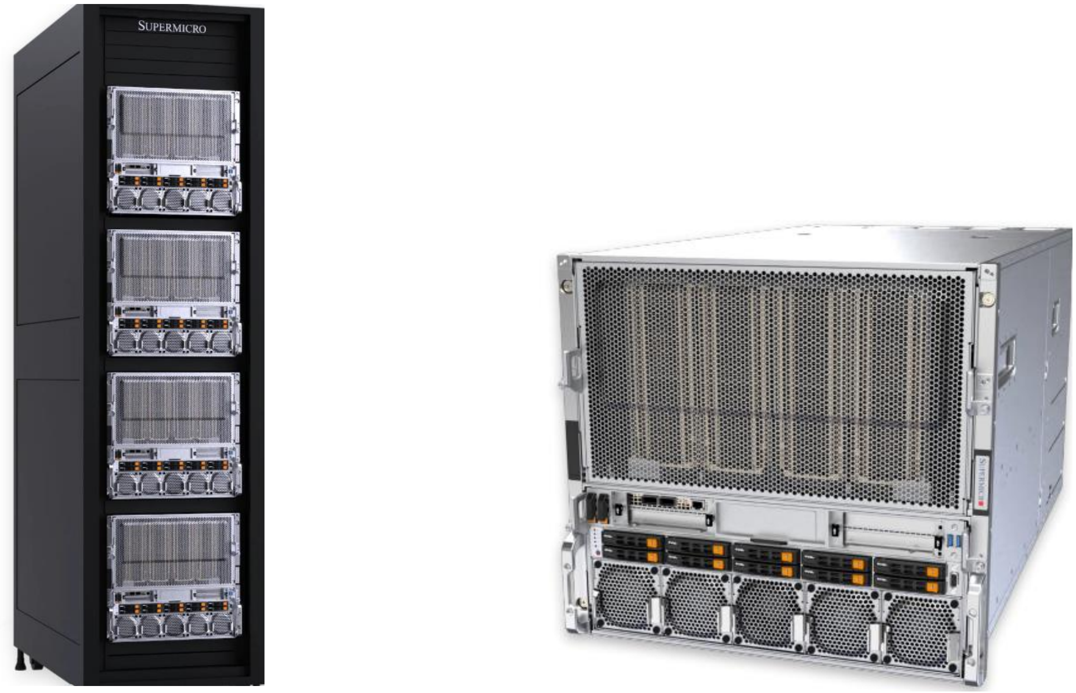


Figure 1: Latest air-cooled design: Supermicro HGX B200 rack (left) and 10U compute server (right).

Table 1: Air-cooled 10U server components and power density

Server components	TDP each (W)	Quantity (#)	Total power (kW)
Blackwell GPU/server	1000	8	8
CPU/server	300	2	0.6
NV link /server	35	8	0.28
DPU bluefield/server	75	2	0.15
SSD/server	15	36	0.54
DDR/server	15	24	0.36
Network switches/server	100	4	0.4
Total IT TDP/server			10.3

Table 2: Air-cooled 48U rack components and power density

Rack components	TDP each (W)	Quantity (#)	Total power (kW)
Servers /rack	10330	4	**41.3** **(Silicon Compute Power)**
PSU thermal dissipation/rack	2900	1	2.9
Total IT TDP/rack			**44**

Liquid-cooled rack reference design

The liquid-cooled configuration is modeled after NVIDIA's upcoming GB300 architecture and was unveiled at the GPU Technology Conference March 2025; it represents a fully liquid-cooled, high-density system. Each compute element—referred to as a Grace-Blackwell superchip—combines two Blackwell GPUs (both rated at 1,400 W TDP) with a single Grace CPU (350 W TDP). Two such superchips are integrated into a 1U compute tray, resulting in a total server tray-level TDP of approximately 7 kW (Table 3), shown in Figure 2. The full GB300 rack layout consists of 18 compute trays and 9 NVLink switch trays, arranged in an alternating compute-switch-compute pattern (Figure 2). The switch trays facilitate high-speed interconnects across the superchips and contribute additional thermal load. The combined silicon power dissipation for the entire rack totals approximately 126 kW, requiring direct-to-chip (D2C) liquid cooling to manage thermal loads effectively. As shown in Table 4, the total TDP of liquid-cooled rack (154 kW) also includes heat dissipation from the switch trays and PSUs.

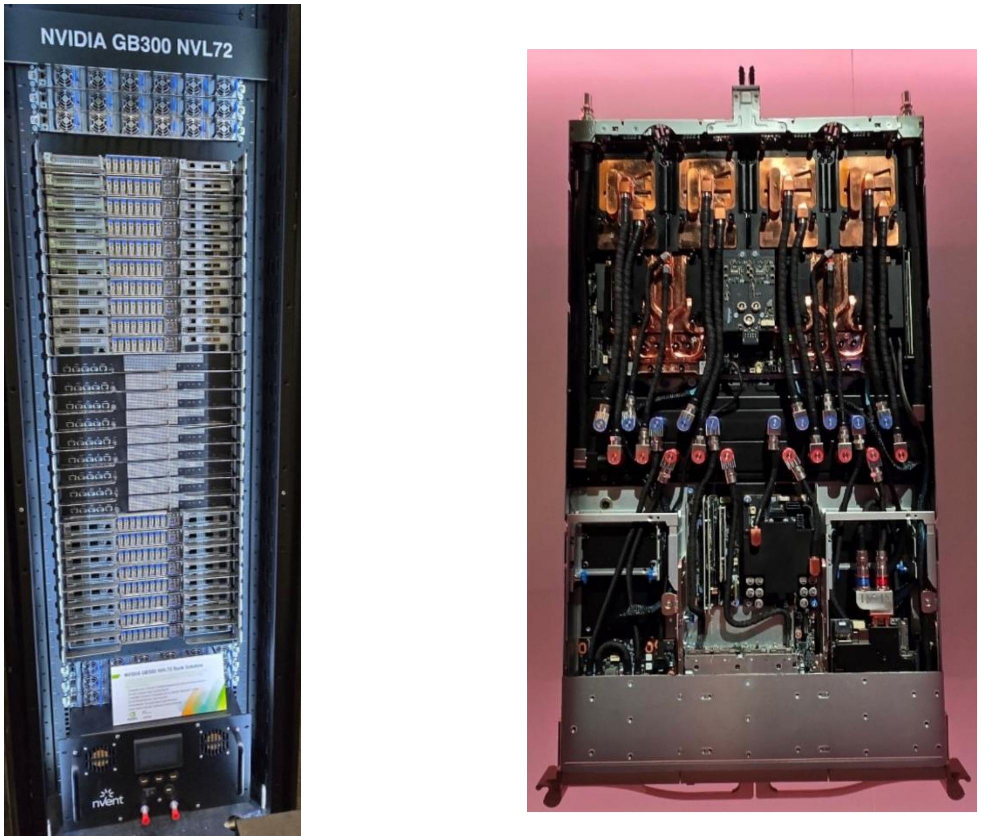


Figure 2: Latest liquid cooled design: NVIDIA GB300 rack layout (left) and compute/server tray (right).

The choice of the GB300 architecture allows us to normalize load density and layout, ensuring that we can accurately compare energy efficiency and compute density of each liquid cooling and hybrid strategy. Compared to the fully air-cooled configuration (Table 2), the liquid-cooled GB300 architecture (Table 4) offers a 3.5 times higher rack-level TDP, enabling significantly greater compute throughput within the same physical footprint. This design illustrates the scalability benefits of liquid cooling as GPU power demands exceed the thermal limits of traditional air-cooled infrastructure.

Table 3: Liquid-cooled 1U compute/server tray components and power density

Server Tray Components	TDP (W)	Quantity (#)	Total power (kW)
Blackwell GPU/server	1400	4	5.6
CPU/server	350	2	0.7
NV link/server	35	4	0.14
DPU bluefield/server	75	1	0.075

Server Tray Components	TDP (W)	Quantity (#)	Total power (kW)
SSD/server	15	16	0.24
DDR/server	15	16	0.24
Total IT TDP/server			7

Table 4: Liquid-cooled 48U rack components and power density

Rack Components	TDP (W)	Quantity (#)	Total power (kW)
Servers (1U)/rack	6995	18	**125.9 (Silicon compute power)**
NV link Switch trays (1U)/rack	2000	9	18
PSU thermal dissipation/rack	1680	6	10.1
Total IT TDP/rack			**154**

Metrics of Evaluation

Several metrics exist for datacenter cooling energy characterization and are attributable to the boundary of the analysis and datacenter designs. A datacenter cooling plant can be broken into two main loops: 1) a primary or facility cooling loop that rejects heat to the outdoor ambient, and 2) the secondary or technology system cooling loop that interfaces and collects heat from the IT equipment, as shown in Figure 3. Heat generated by server components, such as chips, is first transferred to the secondary cooling loop. This loop circulates either air (in traditional systems), liquid (in liquid-cooled systems), or a combination of both (in hybrid systems) to remove heat from the IT equipment. The secondary loop then transfers this heat to the primary cooling loop via a heat exchanger: CRAH (in traditional air-cooled systems) or a coolant distribution unit (CDU) (in liquid-cooled systems). The primary cooling loop rejects heat to the ambient environment, typically using one or more systems such as chillers, cooling towers, dry coolers, or adiabatic coolers.

The shift from traditional air cooling to modern liquid cooling primarily affects the secondary loop. Since the performance of the primary loop depends on external factors like location and weather—typically designed to maintain a Facility Water Supply (FWS) temperature of 30°C regardless of the cooling strategy (OCP 2024) —this study focuses on evaluating the performance of the secondary cooling loop, which varies significantly across different cooling technologies.

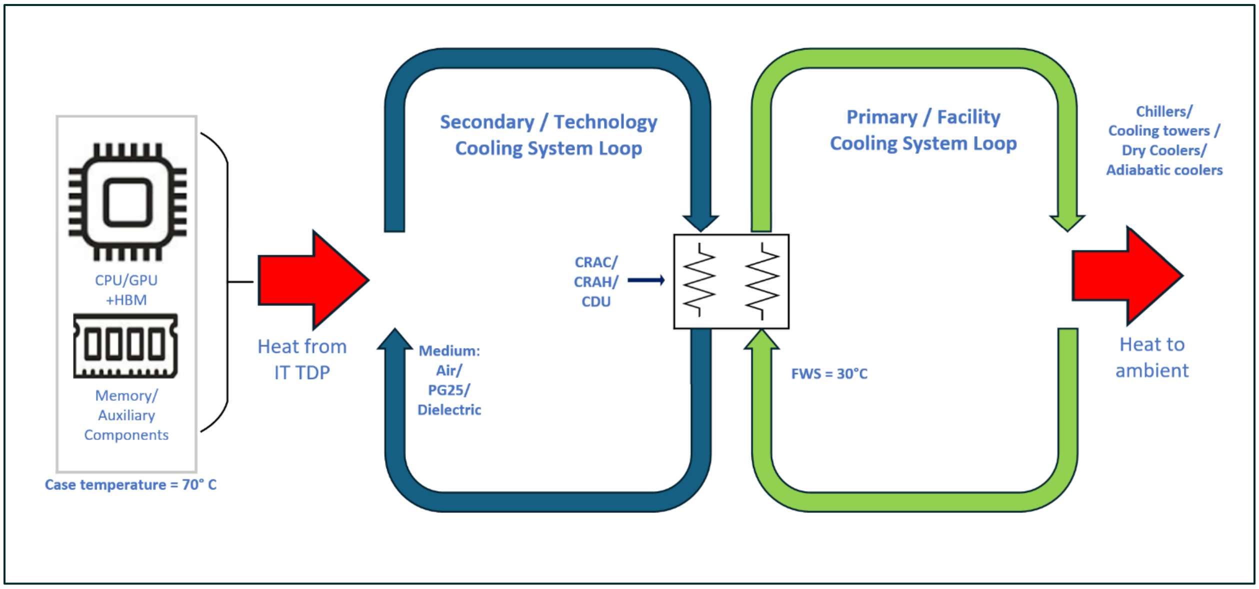


Figure 3: Datacenter cooling plant primary and secondary loops.

Our methodology benchmarks each liquid cooling configuration against a state-of-the-art air-cooled datacenter as the baseline. The comparison focuses on the following key criteria:

- **Secondary loop cooling energy consumption:** Quantified as the electrical energy required in the secondary loop to maintain a maximum chip case temperature at or below 70°C. This includes the power used by fans, pumps, and other components involved in transporting heat from the servers to the primary or facility cooling loop.

- A metric defined for rack level secondary energy use is the percentage of secondary cooling to compute power ratio ($SCCP_{ratio}$). Total IT compute TDP includes all thermal dissipation components in the racks (CPU+GPU + memory + storage + switches + PSU).

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

- **Compute density:** Compute power density ($C_{density}$) (kW/m^2) is a metric that measures compute power per unit floor area occupied by the secondary cooling loop system. AI factories would need cooling systems that can support high computing power density. This metric evaluates how much compute power can be supported per unit area and will help analyze the compactness of the AI datacenter design that is enabled by the selected cooling technology.

- $C_{density} = \frac{IT\ compute\ power\ supported}{Total\ floor\ area\ of\ secondary\ loop\ cooling\ system}$

- **Retrofit feasibility:** An assessment of the practical challenges and infrastructure changes required to upgrade from traditional air-cooled systems to liquid-cooled solutions.

- **System complexity, failure modes, and cost considerations:** A qualitative evaluation of the hardware components, plumbing requirements, and relative cost drivers associated with each cooling strategy.

- The above comparison approach provides both technical rigor and practical relevance, supporting informed decisions by datacenter operators, policymakers, and efficiency program managers. The findings are intended to guide future design best practices and potential updates to code requirements and utility programs.

Datacenter Facility and Layout

For our energy use comparison, we assume a FWS temperature of 30°C (OCP 2024), which was agreed upon by implementers and manufacturers to balance chip reliability and energy efficiency. A datacenter facility is modeled and used as a reference for comparing various cooling systems. For datacenters with air-cooled capacity above approximately 35 kW/rack, a fan wall design of CRAHs is recommended instead of a raised floor design. The fan wall design was modeled using Coolsim, an industry-leading air-cooled datacenter modeling software (Applied Math Modeling 2025), for a datacenter room with 26 by 25 meter dimensions, a room height of 4.85 meters, and a ceiling plenum height of 1.3 meters, as shown in Figure 4. This facility design provides a realistic and standardized basis for evaluating thermal performance across cooling strategies.

For the baseline fully air-cooled datacenter design, 80 racks each with total TDP of 44 kW were placed in the room on both sides of the hot aisle containment. 16 CRAHs each with design capacity of 228.6 kW (65 Refrigerant tons), with chilled water entering at 30ºC and a leaving air temperature of 35ºC.

For the liquid with air cooling hybrid datacenter design, the racks have air side heat dissipation of approximately 40.5 kW generated by the auxiliary components and the PSUs. Similar to fully air-cooled design, four double rows of racks with hot aisle containment are placed in the middle of the room, with total 16 CRAHs supplying air from both sides of the room. Additionally, for this hybrid cooling system (liquid and air), the datacenter room also has 16 heat dissipation units (HDU), which simulate the parasitic heat losses from an in-row liquid-to-liquid CDU. Each HDU has an estimated parasitic heat loss of 1,875 W. The total air-side IT cooling load for the hybrid datacenter room with 80 racks and 16 HDUs is approximately 3.27 megawatts (MW). Each of the CRAH have a design cooling capacity appropriately sized for the datacenter room of 211 kW (60 Refrigerant tons), with chilled water entering at 30ºC and a leaving air temperature of 35ºC. Thus, the total design cooling capacity of the CRAHs is 3.37 MW.

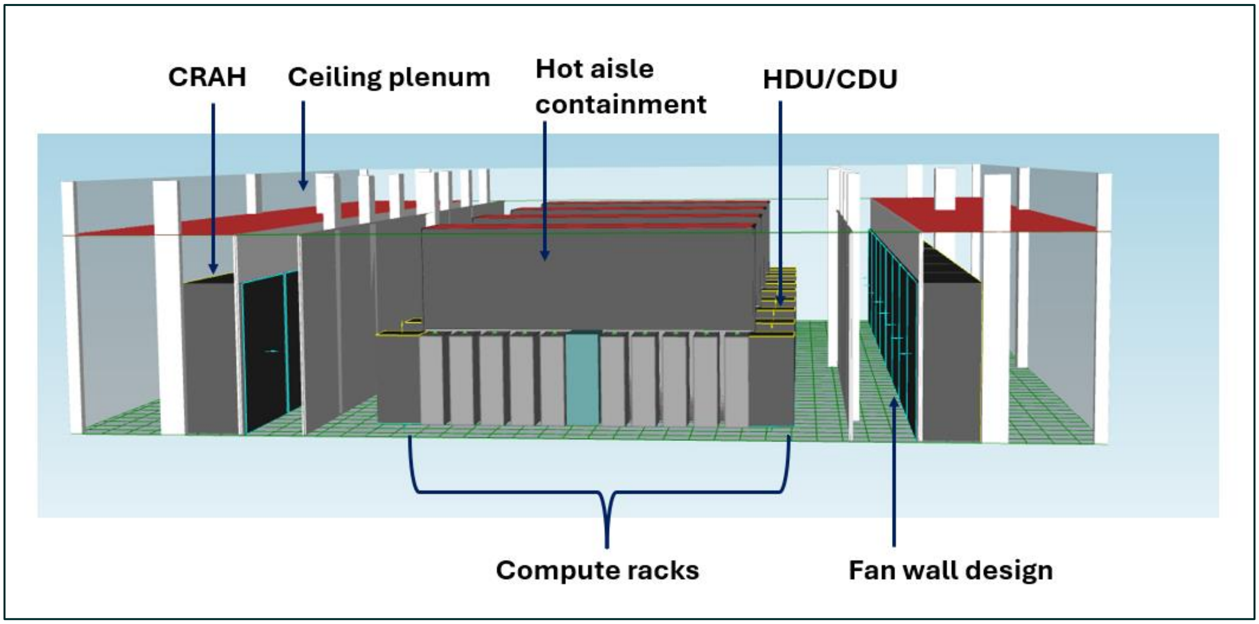


Figure 4: Air and Liquid hybrid cooled datacenter room layout in Coolsim for modeling secondary cooling loop energy consumption and compute power density

Findings

Datacenter Cooling Strategies

Our review of datacenter cooling strategies identified both single-mode and hybrid approaches currently in use, each with distinct advantages and limitations. Single-mode refers to only one method of cooling the compute components (both CPU + GPU and auxiliary). Hybrid approaches use a combination of two modes to cool the compute components. Among single-mode strategies, air-cooling with CRACs/CRAHs remains the most common approach, but it struggles with efficiency at higher rack densities due to limitations in heat removal capacity. Furthermore, this approach cannot dissipate heat from the latest high TDP GPUs, leading to throttling of the performance.

In contrast, an all-cold plate solution uses direct-to-chip (D2C) cooling combined with liquid cooling of auxiliary components. In this scheme, cold plates on CPU, GPU, and auxiliary components are connected to liquid CDUs. Higher cooling efficiency is attained by targeting heat sources directly and eliminating or greatly reducing the use of fans. Full-immersion cooling, where servers are submerged in dielectric fluid, offers significant potential for high-density deployments but presents challenges in serviceability, component standardization, and ease of retrofits.

Hybrid cooling strategies are being increasingly explored as datacenter densities continue to rise. One hybrid strategy combines D2C with air cooling, with D2C handling loads from high heat dissipation components, while residual heat is removed through air systems. A second hybrid strategy consists of D2C combined with rear-door heat exchangers (RDX). The rear door heat exchangers extract heat dissipated by the auxiliary components from air into the liquid, thereby reducing or eliminating the load on the CRAHs in the datacenter hall. A third hybrid strategy combines D2C with immersion cooling of the auxiliary components. Each of these approaches balances energy efficiency, technical feasibility, cost, and operational complexity in different ways. Additionally, all the liquid cooling approaches mentioned above can be modified to two phase based on refrigerants instead of single-phase liquid. Single-phase liquid options use predominantly a 25 percent concentration of propylene glycol with water (PG25) for D2C and dielectric oils for immersion. A summary of the cooling strategies is described in Table 3.

Table 5: List of cooling categories and approaches.

Cooling category	Type of cooling	Options
Single mode	**Traditional CRAC/CRAH air-cooling**	• Without Rear Door HX • With Rear Door HX
Single mode	**D2C with cold plates**	• Single phase liquid • Two phase refrigerants
Single mode	**Full-immersion cooling**	• Single phase liquid • Two phase refrigerants
Hybrid mode	**D2C cold plate + traditional air-cooling**	• Single phase liquid • Two phase refrigerants
Hybrid mode	**D2C cold plate + liquid-to-air CDU for traditional air-cooling**	• Single phase liquid • Two phase refrigerants
Hybrid mode	**D2C cold plate + Rear door HX for traditional air-cooling**	• Single phase liquid • Two phase refrigerants
Hybrid mode	**D2C cold plate + immersion cooling**	• Single phase liquid • Two phase refrigerants

Single-Mode Cooling Approaches

TRADITIONAL CRAH AIR-COOLING

Traditional air-cooling, as shown in Figure 5, primarily utilizing CRAH units with chilled water, has been the standard for datacenters for decades and remains prevalent for many general-purpose server environments Its main advantages include a generally lower initial capital investment due to simpler infrastructure and a widely understood maintenance process, as datacenter personnel are typically familiar with managing air-based systems, fans, and ductwork. However, air cooling is increasingly inefficient for modern high-density workloads, consuming a significant portion of a datacenter's total electrical energy—typically 33 percent to 40 percent (Offutt and Zhu 2025). This method struggles to maintain optimal temperatures for high-TDP GPUs, often leading to throttling of the processors to maintain safe temperature limits. It also requires a larger physical footprint for cooling infrastructure, generates high noise levels from numerous fans. On the primary side, traditional air-cooled datacenters can consume substantial amounts of water when relying on water-intensive evaporative cooling towers for the chiller, or when using indirect evaporative cooling in place of mechanical chillers.

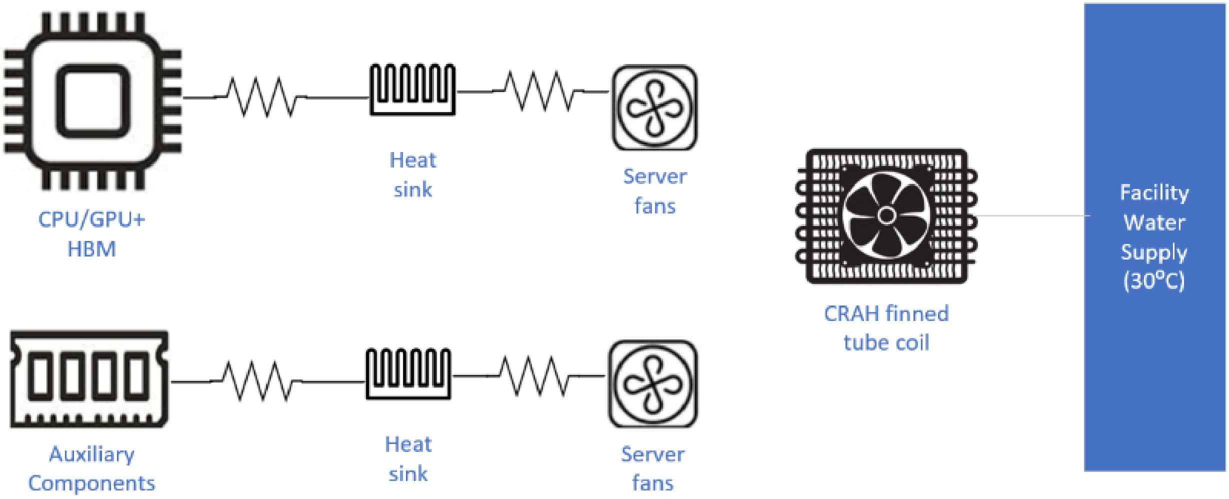


Figure 5: Traditional air-cooling with CRAH

DIRECT-TO-CHIP (D2C) WITH COLD PLATES

D2C cooling with cold plates is gaining significant traction, especially in environments with high-performance computing (HPC) and AI workloads, where it targets heat-intensive components like CPUs and GPUs. In recent implementations, such as the upcoming NVIDIA GB300 design (Figure 2), D2C systems are designed to cool all components within a server or rack using cold plates, effectively eliminating the need for supplemental air cooling within the IT equipment itself.

As illustrated in Figure 6, this method involves circulating a liquid coolant through cold plates attached directly to processors—like CPU/GPU + high bandwidth memory (HBM)—and other auxiliary heat-generating components, such as memory, storage, network, and power components. The coolant absorbs heat and is then pumped to a coolant distribution unit (CDU), which transfers the heat to the FWS, before the coolant is recirculated. The advantages of this comprehensive D2C method include superior heat dissipation directly at the source, leading to high energy efficiency, significantly lower noise levels due to no reliance on fans, and the ability to support high component densities.

Many D2C systems are closed loop and offer substantial potential for waste heat recovery due to heat extraction at high temperatures (>50 °C) from the chip, as well as operating at higher FWS temperature. However, D2C systems typically involve a higher initial investment in specialized equipment and complex plumbing due to limited space and many cold plates. Although modern D2C systems incorporate safeguards and have become very reliable, they also carry a potential risk of corrosion, erosion, fouling, and leaks. Liquid cooled racks come with specialized coolant distribution manifolds in the rear of the racks and blind-mate quick disconnects to assist with ease of assembly and serviceability.

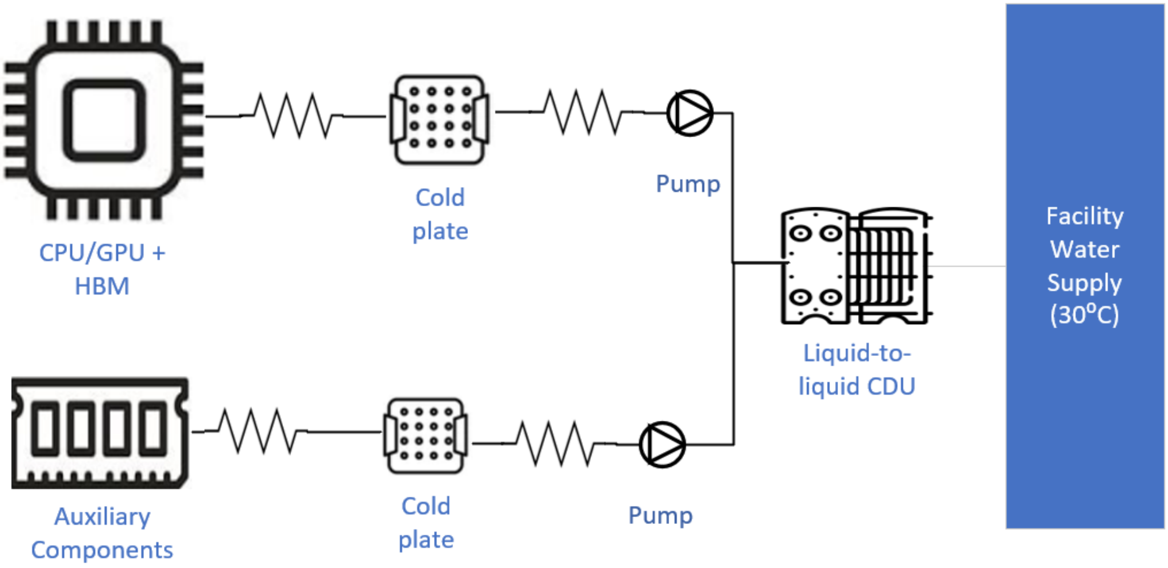


Figure 6: D2C liquid cooling with cold plates for all components.

FULL-IMMERSION COOLING

Full-Immersion cooling, which involves submerging entire server components or racks in a non-conductive dielectric liquid, is growing in prevalence, especially for hyperscale, AI, HPC, and edge datacenters (OCP 2023). This method—which uses components like immersion tanks, dielectric fluid, heat exchangers, and pumps—offers high heat removal capabilities, shown in Figure 7. Immersion cooling delivers significant energy savings with a partial power use effectiveness (pPUE) of 1.03 reported by several manufacturers. This pPUE number likely includes both primary and secondary cooling loop electrical energy consumption (Liquidstack 2022, Shell and Asperitas 2025, Submer 2025). Other benefits are the high potential for waste heat recovery, reduced noise levels, and increased reliability by eliminating dust and vibration. Vendors offer immersion cooling in both single- and two-phase versions.

In single-phase immersion cooling, the immersion tank is coupled with a liquid-to-liquid CDU. A dielectric fluid, such as mineral oil, is pumped through the tank in which the server boards are submerged. The warm fluid exchanges heat in the CDU heat exchanger with the primary loop coolant. In order to cool higher-heat dissipating components like GPUs and CPUs effectively, flow is preferentially directed at the heat sinks located over CPUs and GPUs.

In two phase immersion cooling, a dielectric coolant is contained within the immersion tank and boils at a temperature lower than the limit of the electronic components at atmospheric pressure—typically between 50°C to 60°C. The fluid extracts heat from the components and changes its phase to vapor. Primary facility water is directed through a condenser coil at the upper part of the tank, which is then used to convert the dielectric coolant back to liquid. Typically, there is no forced circulation of the dielectric coolant in the tank due to the high heat transfer coefficient associated with phase change. Thus, two phase immersion cooling can be more efficient when compared with single-phase immersion cooling.

The typical upper limit for thermal power dissipation of chips that can be cooled by single-phase immersion is roughly 1 kW. Higher power chips will need dedicated D2C cold plates through which the cooler fluid is directed to provide targeted removal of heat— a hybrid scheme discussed further in the next section on hybrid cooling strategies. The heat dissipation from a single-phase immersion tank can vary widely depending on the temperature of the primary cooling loop. As an example, for a 42U immersion tank, if the facility water is at 32°C, the heat dissipation is only 100 kW, whereas it increases significantly to 184 kW for chilled facility water at 13°C (Green Revolution Cooling 2025). Note that the pPUE of 1.03, including the primary and secondary cooling loop energy use, can be achieved only in the absence of a chiller.

Full-immersion cooling presents multiple challenges, several of which are being solved through active research and standardization of practices. Both single-phase and two-phase immersion systems typically require a higher initial capital investment compared to traditional air cooling. Retrofitting existing datacenters can be complex if using existing servers, as there are concerns regarding component-fluid compatibility, necessitating comprehensive material compatibility testing (Shah, et al. 2019, Dymyd 2020)

The nascent market also suffers from a lack of clear industry standards for immersion cooling management, although significant progress has been made recently to requirements and standards through the OCP (OCP 2023). Other barriers to adoption include the prevalence of proprietary systems, which can lead to vendor lock-in, and complex warranty coverage for immersion-cooled servers. For single-phase systems, coolant selection is critical, often involving hydrocarbon oils—e.g., PAO oil—which require continuous monitoring and periodic replacement. Two-phase immersion, while highly efficient, faces additional health and environmental concerns, particularly related to per- and polyfluoroalkyl substances (PFAS) and their global warming potential (GWP).

Alternatives with little-to-no PFAS-related impact and low global warming potential, such as hydrofluoroolefins, have recently been introduced, e.g., COVEN FP4000 by Gluditec and Opteon SF33 by Chemours. The coolants used in two-phase systems are also significantly more expensive—often seven times or more than single-phase coolants—and pose concerns like coolant vapor loss and toxicity, which necessitates additional infrastructure for sealing and filtration (Kelly, et al. 2025). Maintenance for both types requires specialized training for datacenter staff to safely handle submerged equipment and fluids, and to manage the intricate plumbing and coolant systems.

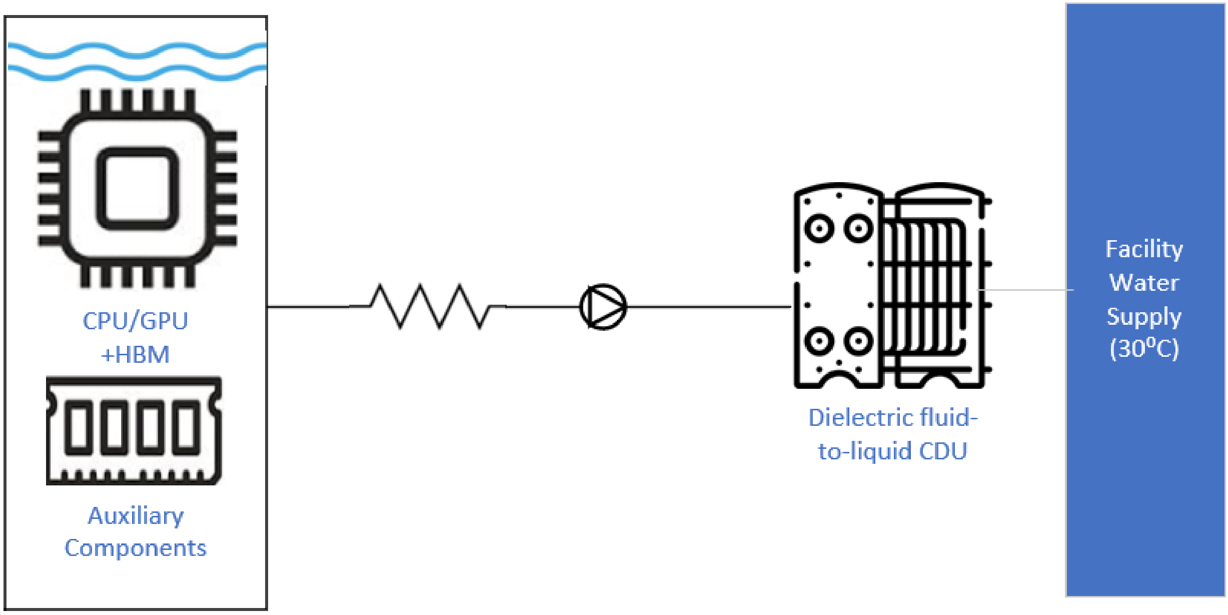


Figure 7: Immersion cooling with tub for all datacenter components.

Later in this section, Table 6 shows the detailed comparison of the single-mode cooling technologies in terms of energy use, ease of retrofitting, rack energy density, non-energy impacts, failure mechanisms and reliability, cost levels, and components required.

Hybrid Cooling Approaches

Hybrid cooling strategies combine elements of liquid cooling with traditional air cooling or next-generation immersion cooling, aiming to leverage the strengths of both technologies for optimized performance, energy efficiency, and scalability.

D2C COLD PLATE WITH TRADITIONAL AIR-COOLING

This hybrid cooling strategy combines elements of liquid cooling with traditional air cooling. The approach involves D2C cold plates for high-heat-generating components such as CPUs and GPUs, complemented by traditional air-cooling using server fans for other server components. The CRAH unit in the data hall is sized to remove the heat from the auxiliary components.

As previously illustrated in Figure 8, this system dedicates liquid cooling via cold plates, pumps, and a CDU to critical components like CPU/GPU + HBM, which generate the most intense heat. Simultaneously, less heat-intensive components, such as memory and auxiliary components, continue to be cooled by conventional air-cooling methods, using heat sinks, server fans, and CRAH units that interact with the facility's air circulation. This hybrid model offers a practical transition path for air-cooled datacenters, allowing them to manage increasing thermal loads from advanced processors while leveraging their existing air infrastructure.

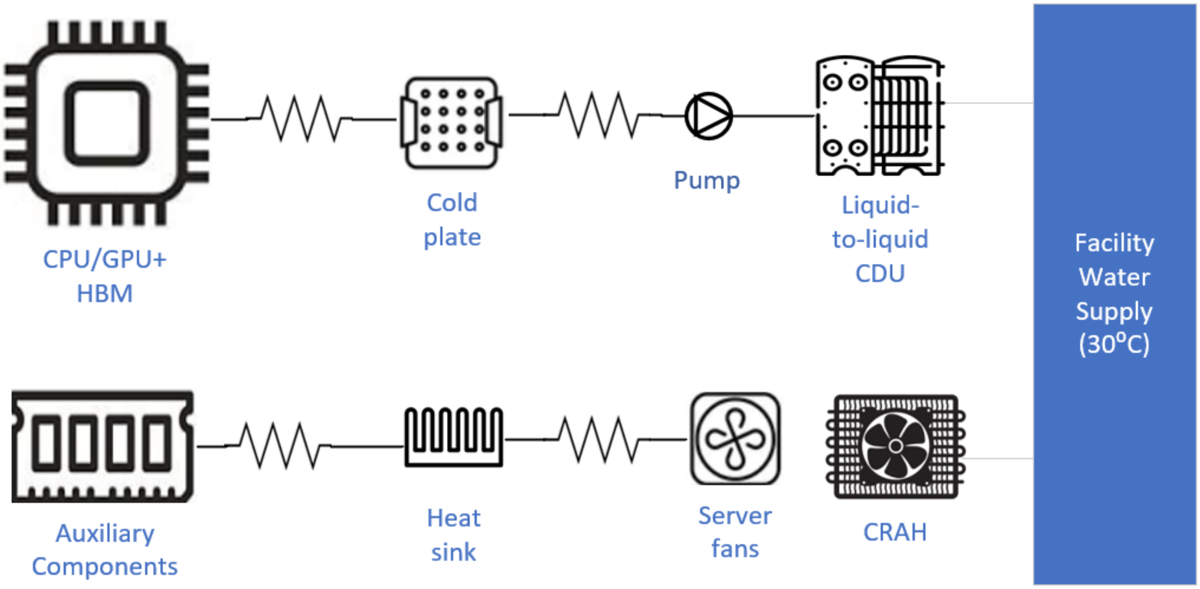


Figure 8: Hybrid D2C cold plate and traditional air-cooling

D2C COLD PLATE WITH LIQUID-TO-AIR CDU AND TRADITIONAL AIR-COOLING

Another specific hybrid approach for retrofitting involves D2C cold plates for high-heat-generating components, complemented by traditional CRAH air-cooling. As depicted in Figure 9, liquid cooling is dedicated to critical components like CPU/GPU + HBM, where a cold plate absorbs heat, which is then pumped to a liquid-to-air CDU. This CDU rejects the heat directly to the ambient air to be picked by the CRAH, making it a suitable choice when a facility-wide cooling water loop is unavailable or difficult to implement for retrofits.

Simultaneously, less heat-intensive components, such as memory/auxiliary components, continue to be cooled by conventional air-cooling methods, using heat sinks, server fans, and CRAH units that interact with the facility's air circulation. This hybrid model allows datacenters to manage increasing thermal loads from advanced processors while leveraging their existing air infrastructure. However, using a water-air CDU can negate some of the efficiency gains typically associated with D2C liquid cooling, as it still relies on air for heat rejection from the liquid loop, and additional energy is expended to pump the coolant.

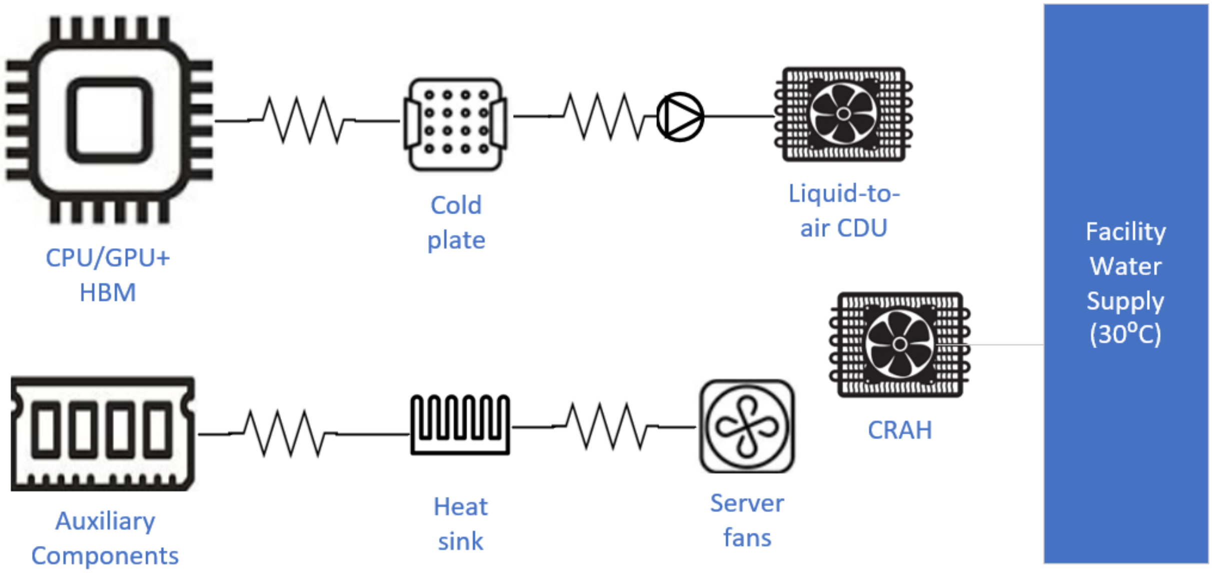


Figure 9: Hybrid D2C cold plate with liquid-to-air CDU and traditional air-cooling.

D2C COLD PLATE WITH REAR-DOOR HEAT EXCHANGER (RDX) AND TRADITIONAL CRAH AIR-COOLING

Another relevant hybrid approach integrates D2C cold plates with rear-door heat exchangers (RDX) and traditional CRAH air-cooling. As depicted in Figure 10, liquid cooling is dedicated to high-heat-generating components like CPU and GPU, where a cold plate absorbs heat, which is then pumped to a heat exchanger (CDU) connected to the FWS. Simultaneously, less heat-intensive auxiliary components are cooled by air, using heat sinks and server fans. Instead of the CRAHs picking up the heat from auxiliary components in traditional methods, a significant portion of heat —approximately 90 to 100 percent—is then collected by the RDX. The RDX uses a fan to pull the air across a liquid-cooled coil, transferring heat directly to the facility water supply.

The CRAH unit is also present, handling any remaining air-cooling requirements for the data hall or uncaptured heat. This hybrid model offers a flexible solution, providing targeted liquid cooling for the hottest components while enhancing the efficiency of air-cooled racks through the RDX, thereby reducing the load on the CRAH and improving overall thermal management within the datacenter. It is an especially useful retrofit in a scenario where the CRAH is undersized relative to the auxiliary component load. Heat recovery into the liquid from auxiliary components enhances waste heat reuse capacity.

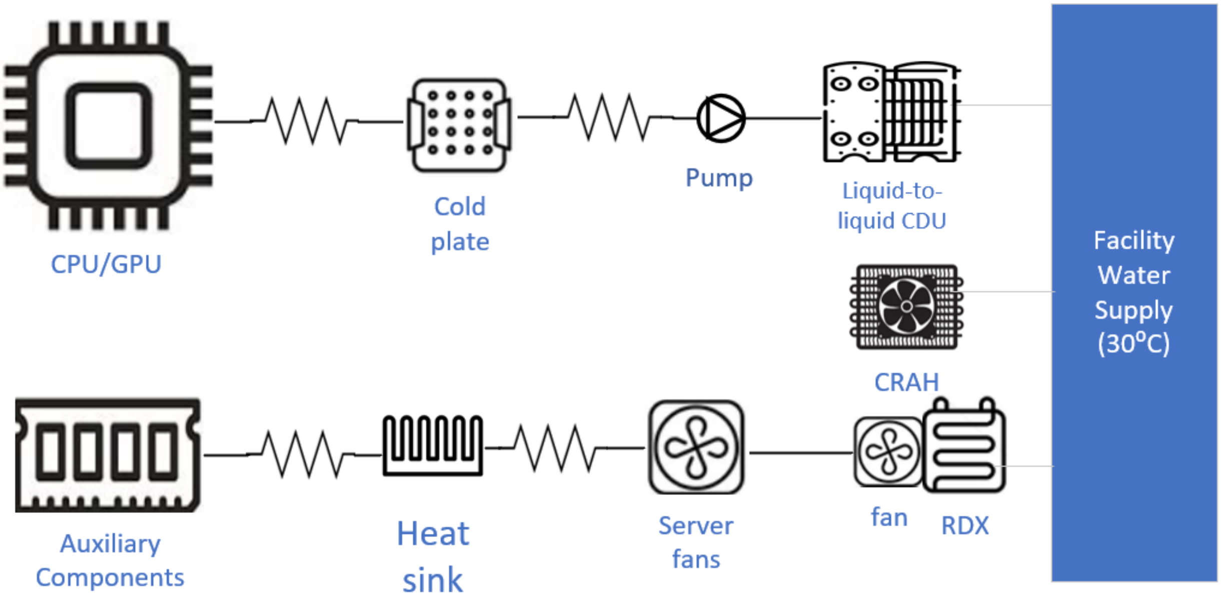


Figure 10: Hybrid D2C cold plate and air-cooling with RDX.

D2C COLD PLATE WITH IMMERSION COOLING

An advanced hybrid approach involves combining D2C cold plates with immersion cooling within the same system. As illustrated in Figure 11, high-heat-generating components like CPU and GPU are directly cooled by a liquid circulating through cold plates, which is then pumped to a CDU connected to the facility water supply. Simultaneously, other components, such as memory/auxiliary components, are fully submerged in an immersion cooling fluid within a tank. This immersion fluid absorbs heat from these components and is then circulated by a separate pump to its own CDU, also connected to the facility water supply.

It is also possible to configure this system in series, where the liquid first cools the cold plate for the CPU or GPU and then flows to fill the tank or server for the immersion cooling of the auxiliary components, before proceeding to the CDU for heat rejection. This hybrid model allows for targeted liquid cooling of the most demanding chips while providing comprehensive, efficient cooling for all other components through immersion, enabling ultra-high-density computing environments.

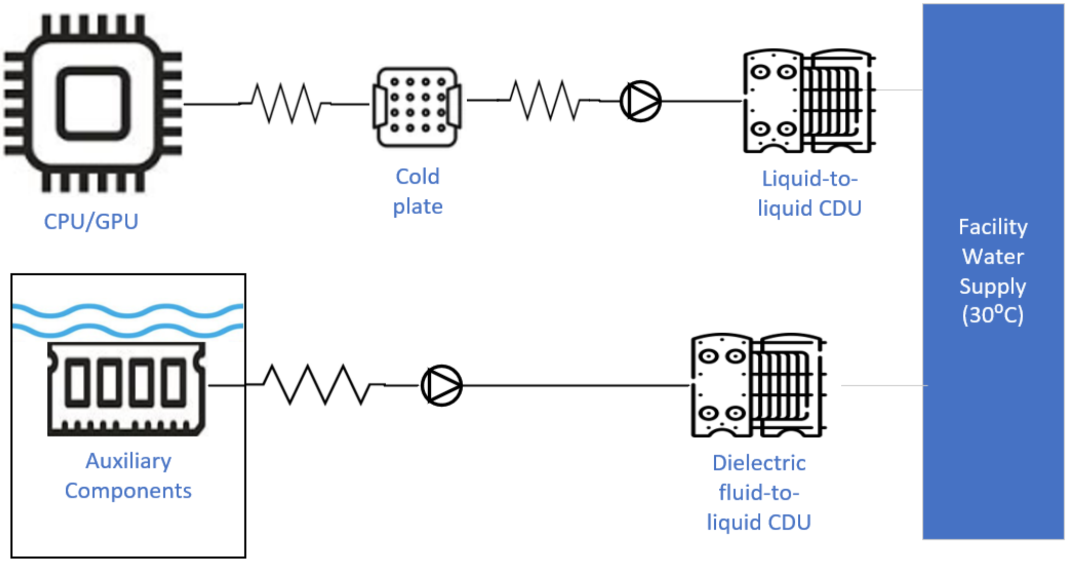


Figure 11: Hybrid D2C cold plate and immersion cooling.

Later in this section, Table 7 shows the detailed comparison of the hybrid mode cooling technologies in terms of energy use, ease of retrofitting, rack energy density, non-energy impacts, failure mechanisms and reliability, cost levels, and components required.

Single-Mode Cooling Technology Detailed Comparison

Table 6: Detailed list of single mode cooling technologies with salient features.

Type	0. Full air-cooled		1. Full D2C Cold plate		2. Full Immersion	
Options	a	b	a	b	a	b
Description	Air-cooled with CRAH	Air-cooled with CRAH and Rear Door HX (RDX)	Liquid Cooling cold plate	2-phase flow (2PF) in cold plate	Immersion all liquid cooling	Immersion all 2PF cooling
Energy Use	high (30-40% of compute energy)	moderate to high	low	very low	low	very low
Rack power density	low (<50 kW/rack)	low (<50 kW/rack)	high (150+ kW/ rack)	very high (200+ kW/ rack)	moderate (100+ kW/ tank)	high (150+ kW/ tank)
Ease of retrofitting	- (baseline)	moderate; needs CDU and liquid lines	Hard; needs server design to permit all-liquid cooled components; tubing management within server; rack level fluid distribution and CDU	very hard; in addition to 1a, need to account for vapor management and two-phase flow distribution	moderate to hard: moderate for immersion baths with server/rack form factor, hard for change out of racks to immersion baths and CDU	moderate to hard; similar to 2a with reduced number of pumps, but needs condenser in immersion bath
Retrofitting-relative comparison	need to upsize CRAH for higher density racks; need hot and cold aisle containment	existing CRAH can be used; but CDU and RDX needed	fans replaced by pumps; custom designs and increased cost	condenser needed in CDU; flow distribution needs careful design; potential use of PFAS and GWP refrigerants; higher cost than strategy 1a	best for new construction—requires different architecture than racks; potential to get locked into a specific vendor	new construction-requires different architecture than racks; potential to get locked in with vendor; PFAS and GWP; significantly higher cost than strategy 2a

Type	0. Full air-cooled		1. Full D2C Cold plate		2. Full Immersion	
Waste heat recovery potential	Hard	Moderate (limited recovery capacity from high density racks)	Easy	Easy (recovery with less pump power)	Easy	Easy (recovery with less pump power)
Failure mechanisms	• Fan failure (in Servers, CRAH units or RDX fans). • Compressor failure in CRAH systems • Refrigerant leaks or charge loss • Condensate drain blockage leading to water damage • Dust and particulate buildup on coils and filters • Airflow imbalance across racks or rows • Rear door HX fouling or coolant loop failure (if liquid-cooled RX) • Hot aisle/cold aisle containment breach		• Coolant leaks, • Pump failure or degraded flow rate • Cold plate clogging due to particulates or biofouling • Thermal interface material (TIM) degradation • TIM Pump-out: migration of TIM away from the center • Air entrapment in the loop reducing heat transfer • Corrosion in mixed-metal systems	• Dry-out or flow instabilities in evaporators • Pressure spikes due to vapor expansion • Non-condensable gas accumulation • Pump cavitation or vapor lock • Leakage of working fluid, which may be more volatile or regulated	• Fluid degradation (oxidation, contamination, moisture ingress) • Material compatibility issues (TIM, plastics, seals, PCB coatings) • Pump or fluid circulation failure • Thermal stratification in poorly mixed tanks • Component corrosion if fluid is not fully inert	• Boiling instability or dry-out at high heat flux • Vapor management issues (condensation inefficiency) • Fluid loss due to evaporation or leaks • Pressure control failure in sealed systems • Contamination affecting boiling characteristics
Reliability considerations (pros/cons)	• Mature and well-understood; easy to maintain • Redundancy (e.g., N+1 CRAHs, dual RDX fans) improves uptime • Less efficient at high rack densities • Susceptible to thermal hotspots if airflow is not well managed		• High thermal efficiency and predictable performance • Easier to control and monitor than two-phase systems • Requires robust leak detection and maintenance protocols	• Very high TDP chips can be cooled • More complex control and safety requirements • Sensitive to orientation and flow balance • Requires precise charge management and sealed systems	• Excellent thermal uniformity and simplicity • Reduced moving parts (no fans on servers) • Requires fluid monitoring and periodic replacement • Maintenance can be more involved (cleaning, fluid handling)	• Very high TDP chips can be cooled • Complex system design (condenser, vapor containment) • Requires precise thermal and fluid management • Limited field experience compared to other methods

Type	0. Full air-cooled		1. Full D2C Cold plate		2. Full Immersion	
Environmental Impacts	Low; may have higher land use		Low; traditional PG25 fluid is used	Medium; small refrigerant volume, GWP and PFAS concerns	Medium; uses dielectric oils which require proper recycling	High; limited refrigerant options and large refrigerant volume, GWP and PFAS concerns
Cost (relative to baseline), major cost components	(baseline)	$ RDX and CDU	$$$$ cold plates; fluid hose and fittings; CDU	$$$$$ cold plates; fluid hose and fittings; CDU; fluid	$$$$ Immersion bath; CDU	$$$$ Immersion bath; CDU; fluid
Components						
Cold plates	-	-	✓	✓	-	-
Rear Door HX	-	✓	-	-	-	-
CDU	-	-	✓	✓	✓	✓
CRAH	✓	✓	-	-	-	-
Immersion tub	-	-	-	-	✓	✓
2 Phase condenser	-	-	-	✓	-	✓

Hybrid Mode Cooling Technology Detailed List

Table 7: Detailed list of hybrid cooling technologies with salient features.

Type	3. Cold plate CPU/GPU + CRAH cooled AUX		4. Cold plate CPU/GPU with liquid-to-air CDU + CRAH cooled AUX	5. Cold plate CPU/GPU + CRAH with RDX for AUX		6. Cold plate CPU/GPU + Immersion cooling AUX	
Options	a	b	a	a	b	a	b
Description	Direct Liquid Cooling (DLC) in cold plate	2-phase flow (2PF) in cold plate	AUX Air-cooled with CRAH	Direct Liquid Cooling (DLC) in cold plate	2-phase flow (2PF) in cold plate	Direct Liquid Cooling (DLC) in cold plate	2-phase flow (2PF) in cold plate
Energy Use	moderate	moderate-low	high	moderate	moderate-low	low	very low
Rack power density	moderate (100+ kW/rack)	high (150+ kW/rack)	moderate (100+ kW/rack)	high (150+ kW/rack)	very high (200+ kW/rack)	moderate (100+ kW/rack)	high (150+ kW/rack)
Ease of retrofitting	moderate	hard	easy	moderate-hard	hard	hard	very hard
Retrofitting-relative comparison	CPU/GPU are DLC while rest are air cooled with fans; existing CRAH can be used; easier retrofit than strategy 1a	Condenser needed in CDU; flow distribution needs careful design; potential use of PFAS and GWP refrigerants; higher cost than strategy 1a	need to upsize CRAH; more expensive than strategy 0a; however, rack densities can be increased and there is no need for liquid lines into the datacenter	promising retrofit scenario—existing CRAH can be used; but CDU, cold plates and RDX needed	condenser needed in CDU; flow distribution needs careful design; PFAS and GWP of refrigerants; needs separate CDU for RDX; higher cost than strategy 5a	can be fitted into a rack architecture, however serviceability low; potential to get locked into a specific cooling solution and provider	rack architecture, serviceability low; potential to get locked into a specific vendor; PFAS and GWP refrigerants; significantly higher cost than strategy 6a, needs two CDUs

Type	3. Cold plate CPU/GPU + CRAH cooled AUX		4. Cold plate CPU/GPU with liquid-to-air CDU + CRAH cooled AUX		5. Cold plate CPU/GPU + CRAH with RDX for AUX		6. Cold plate CPU/GPU + Immersion cooling AUX	
Waste heat recovery potential	Easy (heat recovery possible from CPUs and GPUs)	Easy (heat recovery possible from CPUs and GPUs, recovery with less pump power)	Hard		Easy (heat recovery possible from all components)	Easy (heat recovery possible from all components, recovery with less pump power)	Easy (heat recovery possible from all components)	Easy (heat recovery possible from all components, recovery with less pump power)
Failure mechanisms	• Leak or pump failure in cold plate loop • CRAH compressor or fan failure • Airflow imbalance between cold plate-cooled and air-cooled zones		• CDU pump or heat exchanger failure • Leak in cold plate loop or CDU plumbing • Air cooling inefficiency for high-power auxiliary components		• Rear door HX coolant loop failure or fouling • CRAH unit failure or refrigerant leak • Cold plate loop leak or TIM degradation		• Fluid contamination or degradation in immersion tank • Leak or pump failure in cold plate loop • Material compatibility issues between immersion fluid and components	
Reliability considerations (pros/cons)	• Uses traditional air-cooling components with higher level of understanding • Requires careful thermal zoning and airflow management • Dual maintenance protocols (liquid and air systems)		• Liquid is confined in a small loop, easy management • Requires robust leak detection and flow assurance • Hybrid airflow and liquid loop coordination is critical		• Rear door HX improves rack-level cooling redundancy but multiple cooling paths increase complexity • Requires coordinated control across three subsystems		• Immersion cooling offers uniform thermal management but fluid monitoring and maintenance are essential • Integration of two liquid systems requires careful isolation and control	
Environmental Impacts	Low; traditional PG25 fluid is used	Medium; small refrigerant volume, GWP and PFAS concerns	Low, may have higher land use		Low; traditional PG25 fluid is used	Medium; small refrigerant volume, GWP and PFAS concerns	Medium; uses dielectric oils which require proper recycling	High; limited refrigerant options and large refrigerant volume, GWP and PFAS concerns

Type	3. Cold plate CPU/GPU + CRAH cooled AUX		4. Cold plate CPU/GPU with liquid-to-air CDU + CRAH cooled AUX	5. Cold plate CPU/GPU + CRAH with RDX for AUX		6. Cold plate CPU/GPU + Immersion cooling AUX	
Cost	$$ Cold plates on only CPU/GPU; liquid-to-liquid CDU	$$ Same as 3a	$$$ Cold plates on CPU/GPU; liquid-air CDU	$$$$ Cold plates on only CPU/GPU; RDX; liquid-to-liquid CDU	$$$$ Same as 5a	$$$$$ Costs of 3a + cold plates on CPU/GPU	$$$$$$ In addition to costs in 6a, two different fluids needed
Components							
Cold plates	✓	✓	✓	✓	✓	✓	✓
Rear Door HX	-	-	-	✓	✓	-	-
CDU	✓	✓	✓	✓	✓	✓	✓
CRAH	✓	✓	✓	✓	✓	-	-
Immersion tub/server	-	-	-	-	-	✓	✓
2 Phase condenser	-	✓	-	-	✓	-	✓

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 or simulate any specific product from any individual manufacturer. 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. As can be seen from the “Components” section in Table 5, modeling of the above components allows for energy modeling of the secondary side of all technology scenarios with the exception of full-immersion cooling. 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 since two-phase technologies are emerging and require further characterization relative to refrigerants used and specifics of the system design that are not readily available.

The $SCCP_{ratio}$ and $C_{density}$ are calculated for each cooling strategy (Table 8) using the component modeling methodology discussed in the following section. For more information on the components used in each strategy, refer to schematics of single-mode cooling (Figure 5, Figure 6, Figure 7) and hybrid mode cooling (Figure 8, Figure 9, Figure 10, Figure 11) strategies.

Air-Cooled room and CRAH Simulation

For the fully air-cooled solution, 80 racks of approximately 44 kW TDP each were placed in the middle of the datacenter room and simulated using Coolsim (Applied Math Modeling 2025). As shown in Figure 12 (a), the racks are supplied with 35°C air by the CRAHs from both sides of the room. The hot air containment aisle picks up the hot air (45.5°C on average) leaving the racks and is circulated back as return air to the CRAH through the ceiling plenum Figure 12 (b). The total pressure drop measured in the room that each of the fans need to overcome is 36 pascals (Pa). Additionally, to determine the air side pressure drop in the CRAH coil, a regular finned-tube coil is sized to provide 228 kW of cooling through a flow area of 3 x 3 m^2. The CRAH coils are sized with 0.625 in outer diameter of copper tubes and 14 fins per inch of 0.006 in thick aluminum fins. The pressure drop through the coil is calculated to be about 56.4 Pa. Thus, the total air-side pressure drop for the CRAH fans is calculated to be 92.4 Pa for the given room layout and CRAH design. The same methodology was used to calculate secondary side energy consumption of air-based hybrid cooling strategies, where the air-side heat duty of each rack is 40.5 kW and each CRAH has a design cooling capacity of 211 kW.

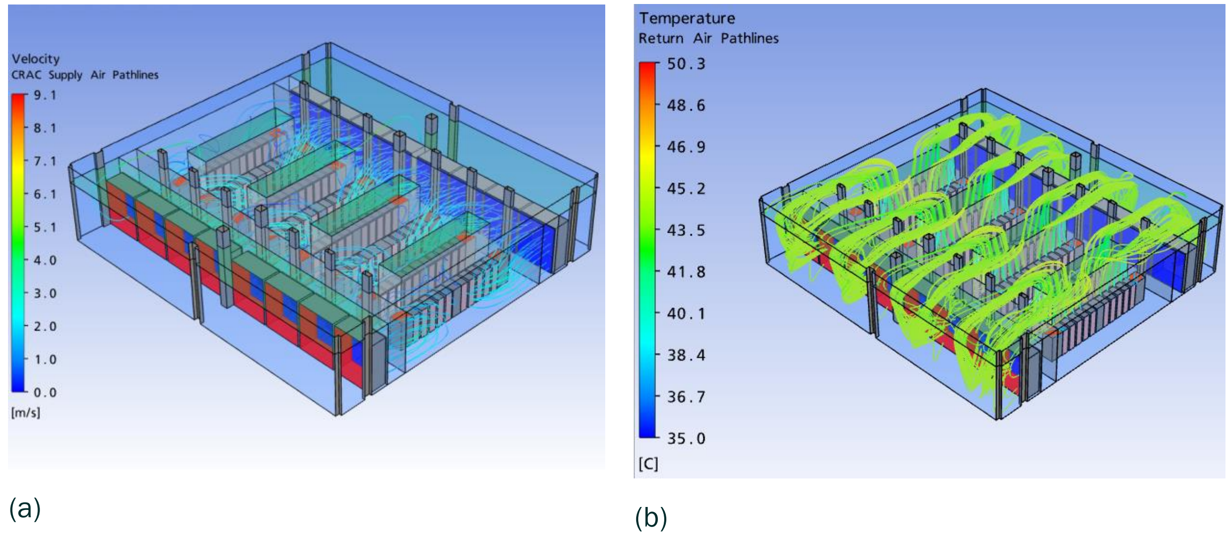


Figure 12: Fully air-cooled AI density datacenter room simulation pathlines from CRAH to racks (a) and from racks to CRAH (b).

Server Fan and Heat Sink Analysis

A simplified analysis was adopted to model server fan energy for air-cooling to dissipate heat from a fully air-cooled case and the auxiliary components of the air-based hybrid cooling strategies. The individual trays in the rack have server fans that provide adequate air supply through the heat sinks that sit on top of the heat dissipating components. The fan, chip, and heat sink are modeled in Ansys Icepak to determine the air flow required and the static pressure of the fans (Ansys by Synopsis 2025). As shown in Figure 13, a 40 x 40 mm^2 fan is modeled to supply air to two chips each dissipating 35 W of heat (based on heat dissipation of NV link chip in Table 3) in a 1 U server. Two heat sink designs, one with tight fitting of the heat sink (a) and another with gaps around the heat sink (b), are modeled as shown in Figure 14. A fan curve, obtained from Sunon fans, was used to determine the operating point and subsequently, the flow rate and static pressure needed to maintain the chip at a case temperature of 70 °C for an inlet air temperature of 35 °C (SUNON 2025).

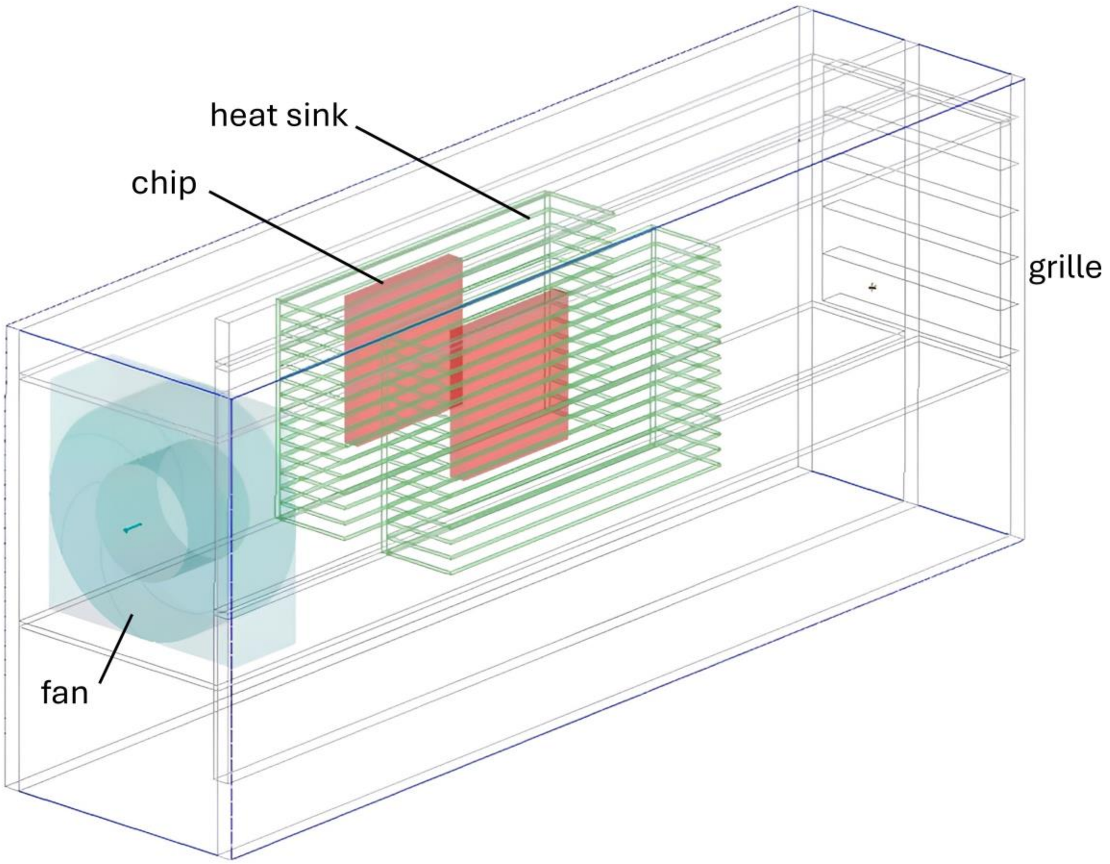


Figure 13: Server fan simulation with chips and heatsinks in Ansys Icepak.

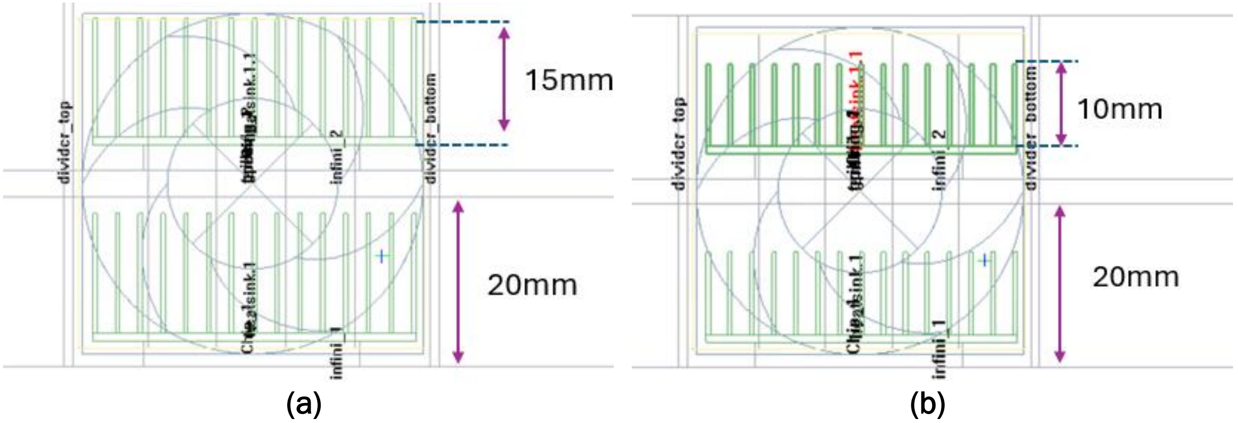


Figure 14: Heat sink designs modeled for server fan energy use simulation.

Based on simulations in Ansys Icepak, to keep the chip case temperature below 70°C, heat sink design (a) requires 172 cubic feet per minute (CFM) of air and heat sink design (b) requires 247 CFM of air for every kW of thermal TDP . Assuming a total of 35 percent combined motor and server fan efficiency, this equates to roughly 0.005 kW for heat sink design (a) and 0.01 kW of fan electrical power, for every KW of thermal TDP for the heat sink design (b). In a study comparing air-cooled and liquid-cooled datacenter designs, 100 CFM per kW TDP was noted for heat sinks with fans (Curtis, Shedd and and Clark 2023). Thus, heat sink design (a) was chosen for the metrics calculated in Table 9 for air-cooled auxiliary components.

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 2 on right.

In this study, cold plates were specifically designed to handle 1,400 W—the highest thermal design power (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 15. 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.

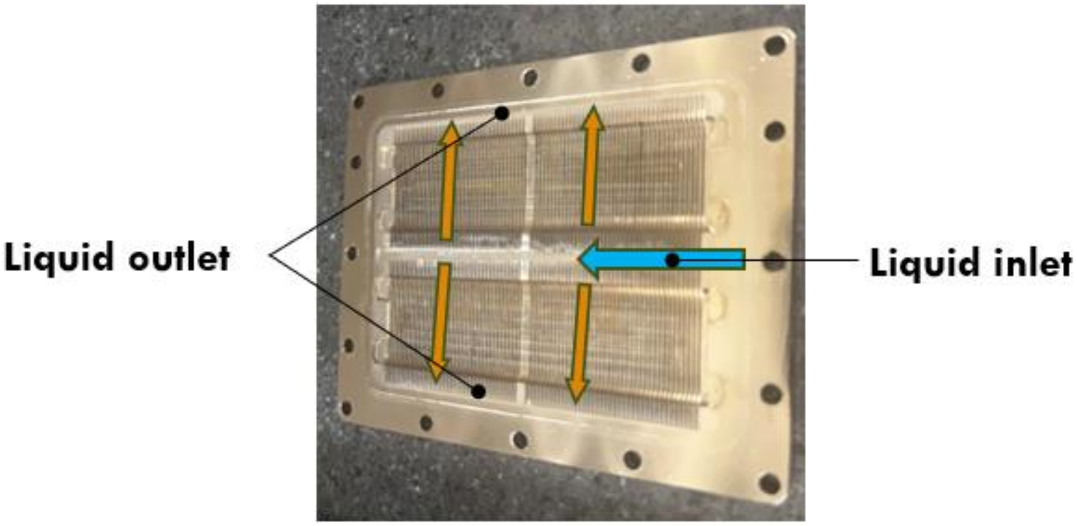


Figure 15: Example of Skived-fin cold plate with liquid flow directions. (Alphacool 2025)

Figure 16 illustrates the relationship between volumetric flow rate, GPU case temperature, and pumping power for an advanced heat sink cooling a 1,400 W GPU with a 35°C inlet of PG25. As the volumetric flow rate of the PG25 coolant increases, the GPU case temperature decreases significantly, particularly at lower flow rates, before leveling off. Conversely, the pumping power required to circulate the fluid rises exponentially with increasing flow. It is crucial to find a balance between achieving a low GPU case temperature, which is vital for optimal component performance and longevity, and maintaining low pump power, which directly impacts energy consumption and operational costs. For this optimal design, 2.25 liters per minute (LPM) flow rate ensures 70°C max case temperature, thereby maximizing the overall efficiency and cost-effectiveness of the liquid cooling system.

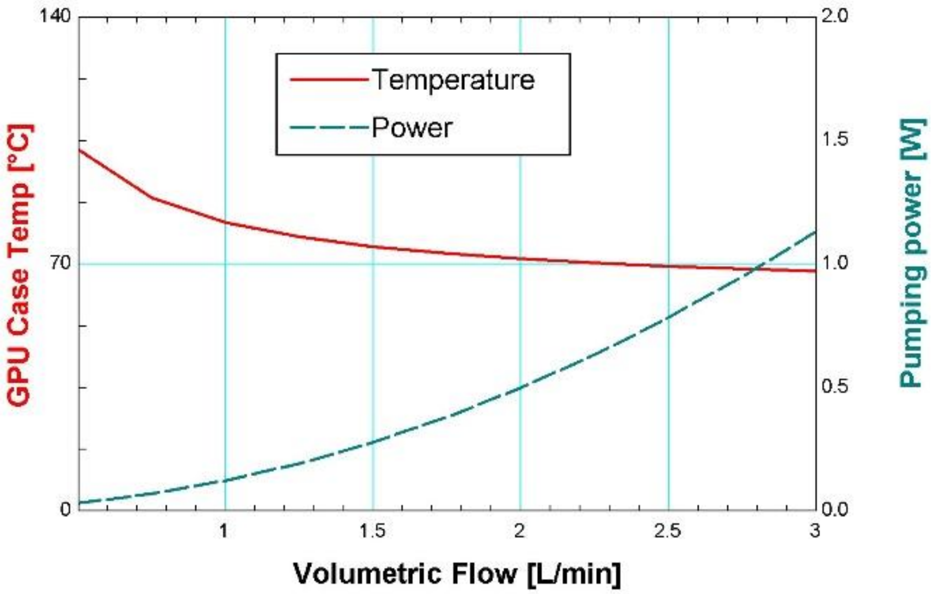


Figure 16: Skived cold plate simulation of case temperature and pumping power for 1400 W GPU

Liquid-to-Liquid CDU Performance

Liquid-to-liquid CDUs are essential components in modern data centers utilizing any liquid cooling technology, acting as the centralized system that manages the circulation, temperature regulation, and pressure control of the fluid delivered to high-density IT equipment. The CDU safely isolates the facility's water loop (the primary loop) from the server-facing coolant loop (the secondary loop). The facility water side interfaces with the heat rejection plant, while the secondary side circulates fluid to the server components. A typical liquid-to-liquid CDU contains vital components such as pumps to maintain coolant flow, a plate-and-frame heat exchanger for thermal transfer, a filtration system to ensure fluid purity, and a pressurization/expansion tank. For this analysis, a simplified in-row CDU model of 2300 kW maximum capacity serving multiple adjacent racks was analyzed informed by designs of several liquid-to-liquid CDUs (CoolIT Systems 2025, Vertiv Group Corp. 2025). A 2300 kW capacity CDU is capable of providing sufficient cooling for the 20 modeled hybrid racks positioned on both sides of the hot aisle containment. Flow rate requirements for the rack derived from D2C cold plate modeling were used to estimate the pressure drop of the secondary loop for each liquid cooled technology. Secondary coolant modeled for D2C cold plate is PG25 and for immersion hybrid is dielectric oil. A plate heat exchanger sizing software was used to estimate the secondary coolant pressure drop in the heat exchanger transferring heat to facility water side. Additionally, pressure drop from the filtration system and valves in the liquid line was estimated to be twice of the pressure drop in the cold plate of the GPU based on recent datacenter liquid-cooling loop modeling performed by the team for Coolerchips project (Milan 2023, ARPA-E 2021). The total pressure drop of the system was used to calculate the liquid-to-liquid CDU pump power consumption for respective fully liquid cooled and hybrid strategies.

Liquid-to-Air CDU Performance

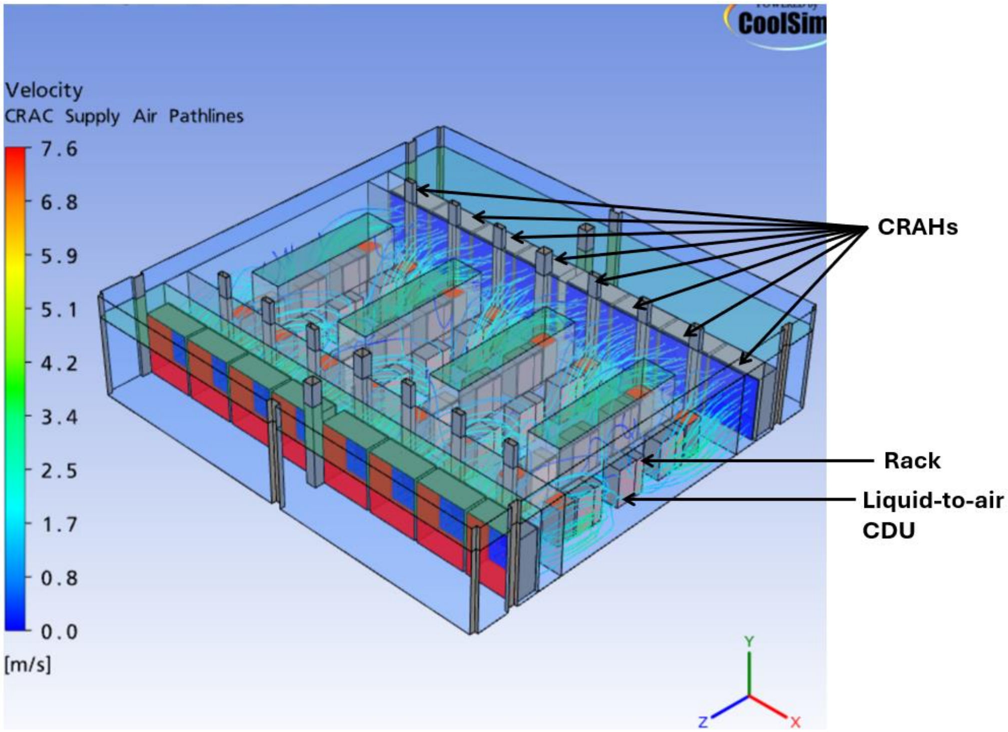
There are many models of liquid-to-air CDUs brought by several manufacturers of varying capacity (CoolIT Systems 2025, Envicool 2025, Data Center Knowledge 2025) (Figure 17). The liquid-to-air CDU design needed to dissipate the heat from CPUs and GPUs (example shown in Figure 9) was modeled using CoilDesigner (Optimized Thermal Systems, Inc 2025). The heat from the secondary loop must be passed from the liquid to the air using the CDU's air-coils, and then to the room and CRAHs. This extra step of heat transfer increases the chips' secondary loop liquid supply temperature to 43°C (approach of 8°C) for the CRAH supply air temperature of 35°C. This higher liquid supply temperature (43°C) necessitates a higher flow rate of the PG25 coolant through the D2C cold plate to maintain GPU temperatures below the 70°C threshold. Based on the optimized cold plate modeling, even with an increased flow rate of 4 LPM, the GPU could not operate at its full 1,400 W TDP and had to be throttled down to 1,300 W to stay within the temperature limit.

Liquid-to-air CDU heat rejection coil was modeled in CoilDesigner as an A-frame micro-channel heat exchanger with four rows and an air flow rate of 15,600 CFM. The total air-side pressure head required by the liquid-to-air CDU fans was estimated to be 94.5 Pa. This equates to electrical power consumption of 1.39 kW for the fans with a combined motor and fan efficiency of 50 percent (Ziehl-Abegg AG 2021).

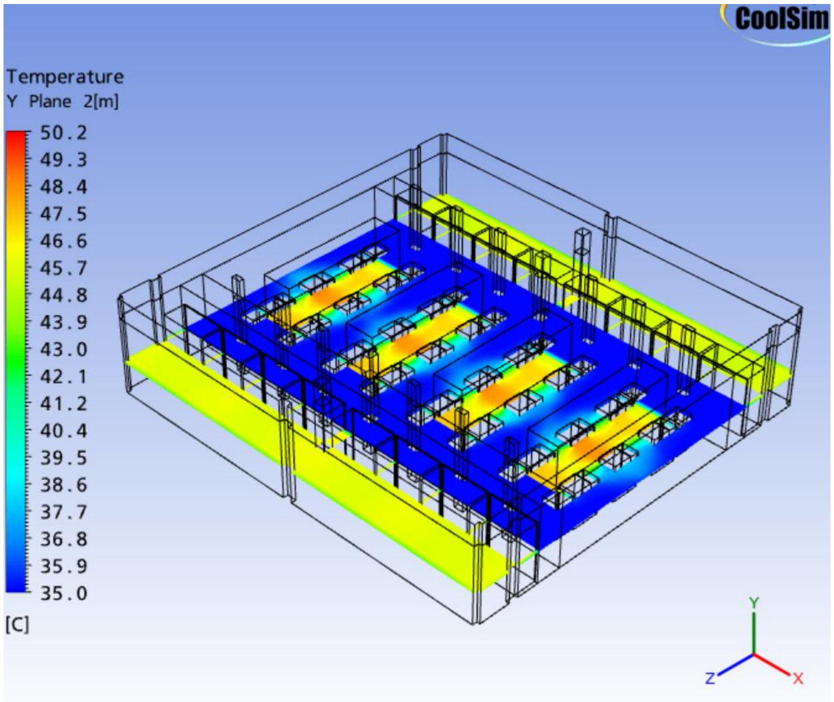
Figure 17: Examples of liquid-to-air CDUs: AHX180 by CoolIT systems of 180kW capacity (left) and Cool inside

Cabinet by Envicool of 120 kW capacity (right)

An air-side model was created to determine the achievable compute density by fitting as many racks as possible into the same datacenter room with the existing infrastructure as the fully air-cooled configuration. In this configuration (Figure 9), the racks were configured to reject heat from auxiliary components into the air directly, while the liquid-to-air CDUs handled the heat from the GPUs and CPUs through the cold plate. The analysis showed that only 24 racks and liquid-to-air CDU pairs could be placed in the same room as the 80 racks of a fully air-cooled layout, all while staying within the cooling capacity of the 16 CRAHs of 228 kW cooling capacity each. The temperature and velocity distribution, as shown in Figure 18, confirmed that both the racks and CDUs received adequate air supply, and that the leaving air temperatures in the hot aisle containment did not exceed thresholds.



(a)

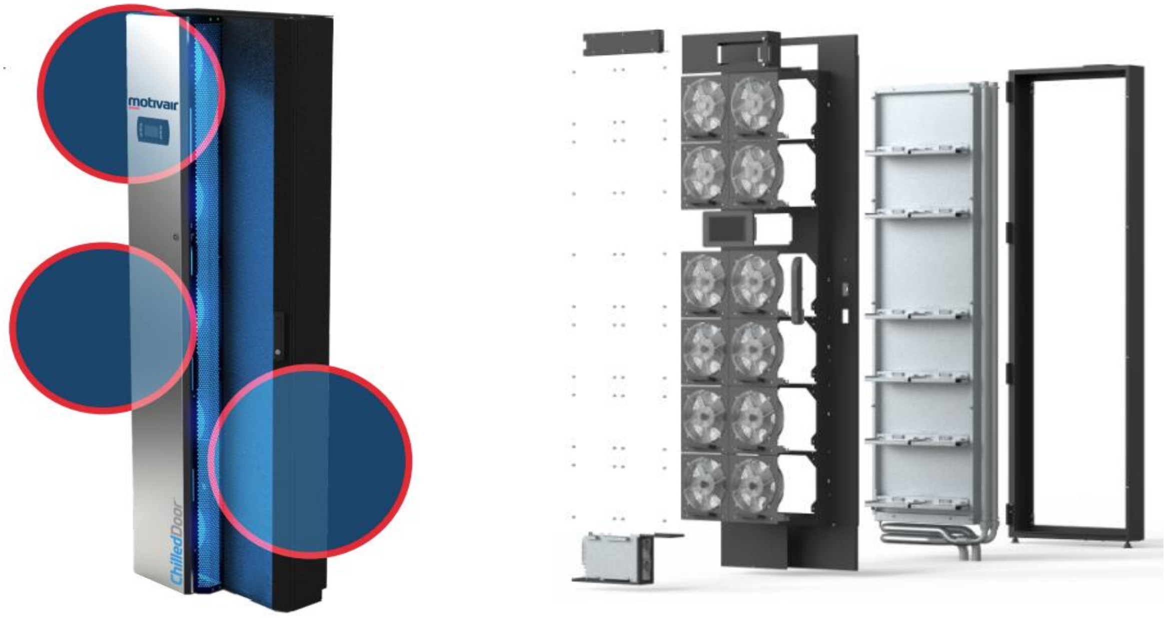


(b)

Figure 18: Coolsim modeling of liquid-to-air CDU of (a) supply air path lines, and (b) temperature distribution

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 in hybrid liquid-cooled design, the team sized and modeled the RDX 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). Examples of RDX from two manufacturers is shown in Figure 19 (a) (Motivair 2022) and Figure 19 (b) (Nvent data-solutions 2025).



(a)

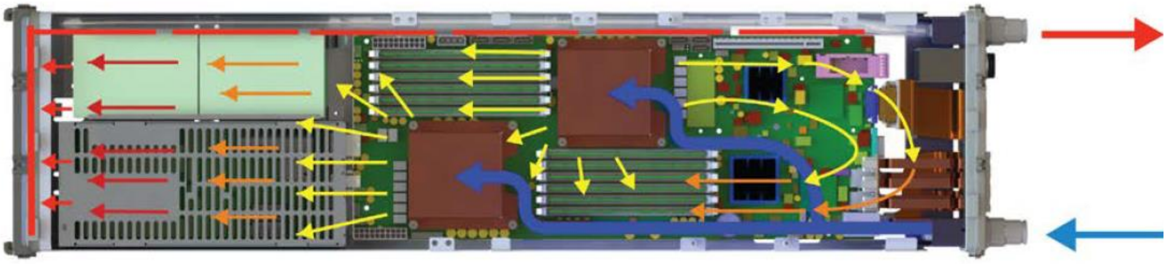
(b)

Figure 19: Examples of rear-door heat exchangers: Motivair Chilled door (a) and by nvent RDHx Pro (b).

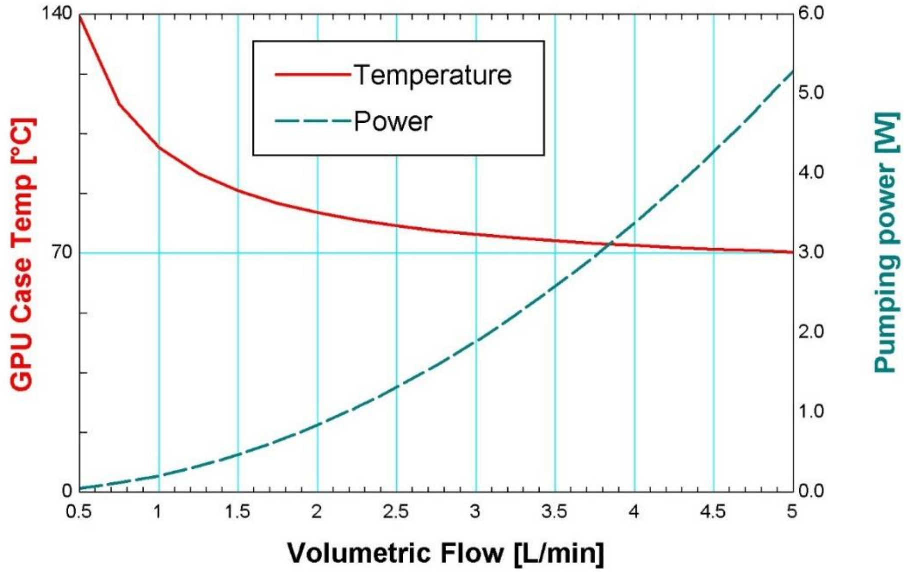
Hybrid Cold Plate and Immersion Analysis

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 20 (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 20 (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 auxiliary components 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 20: Hybrid immersion modeling and simulation. (a) Dielectric coolant flow pattern and (b) cold plate modeling of the case temperature.

Cooling strategy performance comparison

The component models were used together to simulate overall cooling technologies to compare across strategies. Table 8 shows a summary of the different component models (described above) that were used to simulate secondary cooling energy consumption and compute density for each datacenter cooling strategy.

Table 8: 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	Not modeled. Manufacturer and published research data used for comparison with other technologies in analogous conditions			
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

Cooling Technology	Model 1	Model 2	Model 3	Model 4
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 simulation for GPU and CPU with dielectric fluid	Dielectric oil flow through AUX with heat sinks	Liquid-to-liquid CDU	

The outcome of the energy modeling analysis is presented in Table 9 in terms of $\text{SCCP}_{\text{ratio}}$(%) and $\text{C}_{\text{density}}$ (kW/m^2) for each datacenter cooling strategy. The single model full-air cooling with CRAH utilizes the most cooling energy with a $\text{SCCP}_{\text{ratio}}$ of 2.00%. Note from the air-cooled rack (Table 2), that the TDP is throttled from 1400 W to 1000 W. Moreover, the rack density is only 44 kW, which translates to a low compute density 5.4 kW/m^2 for the datacenter room geometry shown in Figure 4.

The All-D2C scenario has a rack power density of 154kW (Table 4) and results in the lowest cooling power consumption on the secondary side, with an $\text{SCCP}_{\text{ratio}}$ of 0.14% and the highest compute density of 82.7 kW/m^2. The All-D2C cooling scenario can support the highest density compute because it can avoid all the large air-cooling infrastructure such as the CRAHs, hot aisle containment, and ceiling plenums with return air passages. The full-immersion cooling $\text{SCCP}_{\text{ratio}}$ of 1.08% (calculated based on D2C+immersion hybrid model) is high because of high pump power needs of the dielectric oil compared to PG25. However, this number can vary a lot based on manufacturer design and dielectric fluid type. Using manufacturer and published research data, $\text{C}_{\text{density}}$ for full-immersion was calculated to be 16.3 for a FWS of 32°C as compared to 30°C FWS modeled for all other liquid cooled scenarios. Higher densities are noted by the manufacturers for lower coolant temperature of 15°C, but that will result in significant increase in the primary loop cooling energy consumption.

Among the hybrid cooling technologies, D2C for CPU/GPU with traditional air cooling and RDX for auxiliary components yields the lowest $\text{SCCP}_{\text{ratio}}$ of 0.65 and 0.89, respectively. Since the RDX uses additional fans, the $\text{SCCP}_{\text{ratio}}$ is a bit higher at the back of the racks when compared to that of the D2C with CRAH. However, if the auxiliary component heat dissipation cannot be handled by the existing CRAH during retrofit, an RDX hybrid might be a good option. The $\text{C}_{\text{density}}$ of the hybrid D2C +RDX is similar to the hybrid D2C with CRAH option at 18.9 kW/m^2.

The hybrid immersion uses a dielectric with poorer thermal properties compared to PG25 that is used in D2C, the GPU power had to be throttled to 1100 W from 1400 W for the same maximum case temperature of 70°C and leads to $\text{SCCP}_{\text{ratio}}$ of 1.08%. Despite lower secondary side energy efficiency, the compute density for D2C+immersion is the highest at 63.4 kW/m^2 among all the hybrid strategies by avoiding air as the media.

The D2C with liquid-to-air CDU cooling strategy requires the highest amount of secondary cooling energy ($SCCP_{ratio}$ =4.14%). In this mode, heat extracted from the high-density liquid-cooled racks is rejected by the liquid-to-air CDU into the data center room air. This heat then places an additional load on CRAHs to transfer to the loop for dissipation. Consequently, the energy consumed by the CDU fans and pumps, compounded by the operational energy of the CRAHs handling the rejected heat, significantly increases the total cooling energy use for this strategy. Furthermore, the maximum number of racks supported is constrained by the fixed capacity of the CRAHs. Since this comparative study maintains a constant CRAH capacity (like the full-air cooled scenario), the achievable compute density ($C_{density}$ = 5.4 kW/m^2) is ultimately limited by the heat rejection capability of the room's air-side infrastructure. The primary advantage of this cooling mode, however, is its ease of integration: it allows high-density liquid-cooled racks to be deployed into existing data centers without requiring the installation of new facility liquid lines within the data center room itself.

Table 9: Energy and compute density comparison across different datacenter cooling strategies.[1]

Cooling mode	$SCCP_{ratio}$ (%)	$C_{density}$ (kW/m^2)
Traditional CRAH air-cooling (baseline)	2.00	5.4
D2C with cold plates	0.14	82.7
Full Immersion tank	1.08*	16.3**
D2C cold plate + traditional air-cooling	0.65	18.9
D2C cold plate and liquid-to-air CDU + traditional air-cooling	4.14***	5.4
D2C cold plate + RDX for traditional air-cooling	0.89	18.9
D2C cold plate + immersion cooling	1.08*	63.4

*Based on CDU pump power with secondary side dielectric oil flow and GPU throttled to 1100 W.
** Calculated from product brochure (Green Revolution Cooling 2025) using FWS of 32°C
*** GPU needed to be throttled to 1300 W to maintain case temperature of 70°C

Two-phase liquid performance

For all the D2C cold plate liquid cooling strategies, another option would be to use two-phase (2PF) refrigerants instead of regular single-phase fluids (Table 5). These two-phase options are an area of active research and are in need for further characterization relative to fluids used, system design, and cost. A recently shown study at the OCP summit 2025 comparing 2PF D2C to PG25 D2C showed similar capital expenditures but lower operational costs due to higher energy efficiency leading to about 12% saving on the total cost of ownership for 2PF solution (Accelsius 2025).

The fundamental advantage of 2PF liquid cooling solutions lies in the latent heat of vaporization during the phase change process, which provides superior heat transfer and exceptional temperature uniformity across the chip surface, thereby enhancing component reliability. This uniform heat removal allows for the handling of higher heat fluxes while requiring significantly lower pumping power and flow rates compared to single-phase liquids. However, 2PF technology for datacenter cooling is still emerging and presents challenges, including the use of refrigerants that may contain PFAS and have a high GWP. Furthermore, the systems inherently possess higher complexity, initial cost, and maintenance costs compared to simpler single-phase loops. Among the 2PF technologies, D2C cold plates can use a wider variety of refrigerants and operate over a range of pressures and use lower volume of refrigerant. Design considerations include sizing the lines to consider high vapor velocities and potential for non-condensable gas removal if operating under sub-atmospheric conditions. Two-phase immersion cooling requires larger volumes of refrigerant and is more limited in choice of refrigerants, and could be harder to service, compared to 2PF D2C cold plates.

Components and Commercial Availability

Each cooling approach depends on a distinct set of components, with varying levels of commercial maturity and vendor availability. For air-cooling systems, core components include CRAC and CRAH units, cold/hot aisle containment structures, and, in some cases, raised floors or overhead ducts. These systems are widely available from vendors such as Vertiv, AIRSYS, and Stulz, who have a long history of deployment in datacenters of all sizes. Traditionally, large commercial HVAC manufacturers such as Daikin, Carrier, Trane, Munters, Danfoss, Johnson Controls, etc. are also suppliers of various components of the traditional air-cooled solution.

Liquid cooled D2C setups mount cold plates directly on processors and route coolant via manifolds to coolant distribution units. Single phase cold plate vendors are: Vertiv, CoolIT Systems, JetCool Technologies, Motivair, Coolermaster, LiteOn, Mikros, Boyd Corporation, Chilldyne, AIRSYS, Fabric8Labs, Wieland, etc. These single-phase cold plates are all commercially available and continuous improvements are being made to increase energy efficiency.

The liquid-to-liquid CDU suppliers supporting the D2C cold plates are: Advanced Cooling Technologies, Boyd Corporation, Chilldyne, Vertiv, CoolIT, Rittal, Asetek, Stulz, Scheider Electric, nVent, LiquidStack, Opticool Technologies, etc.

Liquid-to-air CDUs are seen as a temporary solution for retrofitting existing air-cooled datacenters with high density liquid cooled racks with TDP of 100+ kW. The vendors that we can find offering 100+ kW of liquid-to-air CDU are: CoolIT, nVent, Boyd, Delta, and Stulz.

RDX has been used with traditional air-cooling and is provided by several vendors: nVent, USystems, Motivair, OptiCool Technologies, CoolIT Systems, Attom Technology, Legrand, Panduit, Vertiv, Delta Power Solutions, Envicool, Coolcentric, etc.

Full-immersion tanks are also provided by several vendors: Green Revolution Cooling (GRC), Iceotope, Submer, LiquidStack, Hypertec, DCX Liquid Cooling Systems.

The hybrid D2C with immersion technology is only provided by a few vendors: Liquid Cool Solutions, Aceotope, and LiquidStack.

Two-phase cold plates are an emerging technology and are provided by vendors such as: ERG, Zutacore, Accelsius, Fabric8Labs, Seguente, Wieland. The specialized CDUs associated with these 2PF systems are soon-to-be-available technology.

New emerging liquid cooling solutions emphasize transformative improvements in energy efficiency, thermal performance, and reliability to meet the demands of high-density workloads such as AI and HPC. One innovation is the helical turbulator cold plate technology (Chilldyne 2022). Complementing this is the negative-pressure cooling distribution unit (CDU), which uses vacuum-driven coolant circulation to eliminate leaks and reduce complexity, offering scalable, energy-efficient cooling for next-generation AI data centers (Chilldyne 2025).

Another cutting-edge approach is the confined direct two-phase jet impingement cooling system developed with topology optimization, additive manufacturing, and advanced surface engineering to maximize thermal performance while minimizing pumping power (Purdue Univeristy; Binghamton University; Seguente Inc. 2023).

Additionally, a deep decarbonization project led by the University of Florida aims to revolutionize data center sustainability by integrating state-of-the-art phase-change heat transfer and optimization technologies under ARPA-E's COOLERCHIPS program. Together, these solutions promise significant reductions in energy consumption, carbon footprint, and operational risks for modern data centers (University of Florida 2023).

California Data Center Market Trends

California’s datacenter landscape is expanding rapidly, shaped largely by demand from AI, cloud computing, and enterprise services (CalCCA. 2025). Pacific Gas and Electric (PG&E) anticipates spending $73 billion to support 10 gigawatts of new datacenter load in the next 10 years (Skidmore 2025). Recent filings and announcements indicate dozens of new facilities under development, ranging from 7-MW infill projects to planned campuses of 400+ MW. These projects are geographically clustered: The Bay Area and Santa Clara Valley continue to dominate activity, with additional developments in Stockton, Gilroy, and Hayward, along with emerging regions of Southern California, such as Vernon and Imperial County’s “Lithium Valley.” Facilities are at varied stages of progress, from early land acquisition (e.g., AWS in Santa Clara) to full operation (e.g., Prime Data Centers’ 33-MW Vernon site, pre-leased to an AI company) (Mordor Intelligence.2025).

Interviews with datacenter operators and technical staff reinforced how rapidly California’s market is shifting toward ultra-dense, AI-driven compute environments. Rack densities that were once 10 kW to 15 kW have escalated up to the range of 90 kW to 100 kW, with projected growth to 150 kW in the near term. They also noted that hyperscalers are better positioned than co-location providers to manage this transition, as their new campuses can integrate liquid cooling, advanced containment, and very large coolant distribution units at scale. Meanwhile, retrofits in older, multitenant (co-location) facilities are often uneconomic or technically constrained by specific server requirements required by each client.

These perspectives highlight a divide in infrastructure readiness: Operators who have purpose-built campuses are beginning to normalize liquid and two-phase cooling for AI racks, while legacy facilities remain reliant on hybrid or incremental upgrades. As such, thermal management is emerging not only as a technical challenge, but also as a differentiating factor in market competitiveness and site selection.

Ownership and use patterns also illustrate a mixed model of self-build and co-location. Hyperscale operators like Microsoft and AWS are expanding their self-built campuses—some exceeding 90 MW—often while navigating streamlined state permitting via Small Power Plant Exemptions (SPPE) (Esram et al. 2024, Shehabi et al. 2024). At the same time, developers such as Equinix, Prime Data Centers, and Centersquare are constructing multitenant facilities where space is leased to enterprise or AI-specific customers.

Operators are also taking novel approaches—for example, Nautilus Data Technologies' floating 7-MW Stockton datacenter uses water-based cooling, and CalEthos recently announced a 420-MW campus in Southern California aimed at leveraging abundant land and renewable geothermal energy (Matthew, G. 2024). Collectively, these trends suggest California is entering a new phase where large-scale hyperscale campuses coexist with specialized colocation hubs, each serving different segments of the market but together contributing to substantial new electrical load growth across the state.

Cooling Technology Market Trends

The commercial availability of liquid cooling technologies has expanded considerably over the past several years, with a broad spectrum of vendors now supplying D2C cold plates, CDUs, and immersion systems. Direct-to-chip solutions are the most widely adopted, particularly in GPU- and CPU-intensive environments, where component-level targeting provides superior efficiency and scalability. Products from manufacturers such as CoolIT, JetCool, and ZutaCore are already being integrated into high-performance computing clusters, often in combination with rack-level CDUs provided by companies like Vertiv, Chilldyne, and DCX. These offerings demonstrate that liquid cooling is no longer a speculative technology, but one that is available off-the-shelf for both new builds and, with some difficulty, retrofits.

Immersion cooling, while technologically mature, remains less widely adopted in California. Companies like Green Revolution Cooling and DCX have deployed single-phase immersion systems in select environments. Wider adoption is still limited, as challenges remain in areas like serviceability, compatibility between fluids and components, and the absence of standardized design guidelines. Two-phase immersion systems offer stronger thermal performance than air cooling, but they encounter additional barriers due to the high cost of fluids and unresolved questions about environmental and safety impacts. In this context, hybrid approaches that pair cold plates with conventional CRAH units or RDX are often regarded as the most practical retrofit option (Madeiros McEnroe, 2023). The market is segmented, with D2C systems growing fastest, immersion limited to specialized uses, and hybrids serving as a bridge between air and full liquid cooling.

Emerging technologies currently under development indicate a broader adoption of liquid cooling and increased support for high-density applications. Trends include two-phase cold plates using refrigerant evaporation for higher heat transfer, immersion systems exploring new dielectric fluids, and hybrids that pair D2C cooling with immersion cooling for auxiliary components. Several ARPA-E COOLERCHIPS projects emphasize modular, rack-integrated CDUs and non-water refrigerants designed to operate at warmer temperatures, reducing dependence on facility chillers. Development is moving from hybrid toward fully liquid systems above 200 kW per rack, with emphasis on reducing fan energy, supporting waste heat recovery, and lowering water use. In short, the industry trajectory appears to be moving from air-assisted liquid cooling toward all-liquid, high-density, and potentially two-phase designs, though widespread adoption will depend on resolving issues of reliability, standardization, and retrofit feasibility.

California Data Center Policy Trends

U.S. data center electricity demand is accelerating to unprecedented levels, driven by AI workloads that require far denser server configurations than utilities have historically planned for. In 2023, data centers consumed an estimated 176 TWh, accounting for approximately 4.4 percent of total U.S. electricity, and projections indicate that this will increase to 325–580 TWh by 2028, representing up to 12 percent of national demand (Shehabi, Newkirk and Smith 2024). In California, data centers make up 60% of Silicon Valley Power's commercial load, and demand is set to double by 2035 (Esram and Elliott 2024). AI racks now draw 25–35 kW each, with new systems exceeding 100 kW, significantly higher than the 7.5 kW typical of older facilities. A single AI rack now consumes roughly 39 times the energy of a typical U.S. home, and a large AI campus can rival the electricity use of a city (Sunkara and Narukulla. 2025). For policymakers and program administrators, these dynamics underscore that technical findings, such as the relative efficiency of liquid versus air cooling, cannot be separated from their system-level consequences. Grid constraints, transmission expansion, and capital recovery will shape ratepayer impacts, while cooling choices will determine energy, water, and emissions outcomes. Understanding this broader context is essential for translating technology characterization into feasible, incentive-ready measures that maintain affordability, reliability, and sustainability as AI-era loads reshape California's electric system.

Water consumption linked to data centers is an emerging concern as AI deployment accelerates. National water withdrawals were approximately 322 billion gallons per day in 2015, with thermoelectric power accounting for 41 percent, underscoring that a significant portion of computing's water burden is embedded upstream in electricity generation (Li, et al. 2025). In addition to the embedded water use, there is also direct water use in datacenters that use indirect evaporative cooling as a means to reduce mechanical chiller energy use.

AI data center growth is rapidly outpacing California's energy, water, and air quality frameworks. According to interviews with various utilities, many still plan for air-cooled racks under 15 kW, even as industry stakeholders state that AI systems already exceed 100 kW and are projected to surpass 200 kW.

This section addresses four intersecting challenges that require a coordinated policy response: surging electricity demand that could raise data centers' share of U.S. load to 12 percent by 2028, rising direct and embodied water use that current metrics like Water Usage Efficiency (WUE) fail to capture, growing reliance on diesel backup generation through the SPPE that threatens local air quality and the state's decarbonization goals, and speculative infrastructure buildouts that risk stranded assets and carbon lock-in. The throughline is the need for integrated planning that aligns data center permitting, utility forecasting, and environmental standards with California's climate and equity goals, including those laid out in Title 24.

While direct data center water use accounted for less than one-half of one percent of national withdrawals in 2018, on-site consumption likely doubled from 2018 to 2023, reaching 66 billion liters, and could increase to 150–280 billion liters by 2028, straining supplies in hubs such as California and Virginia. High-density AI unquestionably increases cooling needs, but water outcomes depend on technology choices. Dry or closed-loop liquid cooling can nearly eliminate on-site water use but would shift the burden to off-site electricity generation, thereby increasing Scope 2 water. Because WUE captures only site-level withdrawals, it overlooks larger embodied impacts from electricity and manufacturing, with indirect consumption from power generation alone at nearly 800 billion liters in 2023, significantly exceeding the 66 billion liters used directly for cooling. Semiconductor fabrication amplifies this; Apple reports that 99 percent of its water footprint is in the supply chain. Even when liquid-cooled AI systems consume less than five percent of the IT load for cooling energy, the true water cost remains substantial once upstream and embodied impacts are factored in. A comprehensive accounting framework that integrates Scope 1 (direct), 2 (indirect/energy-related), and 3 (supply chain) is therefore necessary to assess energy and water tradeoffs among air, liquid, and hybrid systems and to guide sustainable AI growth.

California's data center buildout is creating an air quality risk as hyperscale facilities utilize the SPPE to install large fleets of diesel backup generators, often forty or more per site, each with a capacity of roughly 2–3 MW. In Santa Clara, SPPE-approved diesel capacity now exceeds 759 MW, surpassing Silicon Valley Power's 2024 peak generation of 713.5 MW (CEC 2025, Pineda 2025). Intended for emergencies, these generators emit 200–600 times more NO_x and fine particulates than natural gas turbines, pollutants tied to asthma, cardiovascular disease, and premature mortality. Proposals to utilize these fleets for grid support could shift rarely used standby units into routine emitters, mirroring patterns in northern Virginia, where diesel emissions are linked to an estimated $220–$300 million in annual health damages (Han, et al. 2024). While batteries are cleaner, limitations on operation time and costs hinder deployment at the 100 MW-plus scale required by AI facilities. Consequently, SPPE-enabled diesel is expanding more rapidly than regulatory controls or mitigations, raising concerns they are undermining decarbonization goals and air quality standards.

The pace of AI infrastructure development has also become a speculative driver of grid expansion and fossil fuel-based power generation. Although projections indicate data center demand will grow from about 4 percent of U.S. electricity today to 11–12 percent by 2030, assumptions are uncertain, and many facilities are being built ahead of verified compute needs. Utilities say operators often request 50 MW connections but use only 30–35 MW, leaving costly idle capacity. With large renewable generation projects stalled for years, many regions turn to quick gas or diesel builds, locking in emissions. Modeling suggests that if renewable deployment lags, data center growth alone could raise U.S. carbon emissions by 5.5 percent by 2030. One approach would be to link growth to proven demand, phase in connections, and favor renewables and flexible resources over fossil fuel stopgaps, preventing AI hype from derailing decarbonization. Building code provides a potential mechanism for aligning data center growth with broader state goals.

Title 24, California's statewide vehicle for building energy performance, provides the current compliance infrastructure for data centers through prescriptive and performance paths. Energy Budgets are defined in terms of "Time Dependent Valuation site energy (LSC)" and "Source Energy," not water or local emissions. As AI-era loads accelerate and densities diverge sharply from legacy "computer room" assumptions, several gaps emerge. First, while Title 24 sets mandatory and prescriptive requirements specific to computer rooms, those provisions do not scale by data hall size or rack power density, limiting their effectiveness for very high-density AI deployments. Segmenting requirements by total IT load and kW/rack - and pairing them with liquid-cooling readiness, acceptance tests (e.g., CDUs, leak detection, containment), and updated modeling of liquid, hybrid, and refrigerant-economizer systems - would make the code more responsive to observed thermal loads and field practices. Second, because the performance method budgets only energy, a companion compliance element could integrate site and upstream water accounting as well. This approach would better reflect technology trade-offs between air, liquid, and hybrid cooling, aligning with our objective to expand data center modeling and reporting rather than relying on PUE alone. Third, to mitigate air-quality externalities associated with large backup fleets, Title 24 could establish data-center-specific limits on generator standby energy (e.g., annual kWh per kW of installed capacity), offer prescriptive/credit pathways for non-diesel alternatives, and recognize heat-recovery designs as compliance credits, improving alignment between reliability planning and decarbonization goals. Together, these updates would keep the energy code's core structure intact while extending its reach to the intertwined energy, water, and air-quality impacts surfaced in this section.

Utility Responses to Data Center Growth

This section summarizes publicly available information and interview findings on California utility offerings relevant to data center efficiency and near-term readiness amid rapid load growth. The materials indicate outdated guidance, few sector-specific programs, and limited state coordination. Each paragraph profiles one utility, what exists or does not, how recent the materials are, and how prepared they appear for AI-driven densities. The discussion sets up the need for coordinated support so deployment aligns with affordability, reliability, decarbonization, and clean air.

PG&E remains the most visible IOU in this space, yet the core public-facing guidance on data center efficiency is dated. The utility's Data Center Best Practices Guide was published in 2012, and the companion Energy Efficiency Baselines for Data Centers document was issued in 2013, both of which reflect earlier industry practices and IT load profiles (PG&E 2013, PG&E 2012). In recent years, program implementation has been routed through the Advanced Energy Program, which PG&E administers while third-party implementer Resource Innovations manages day-to-day delivery, a structure that can ensure knowledgeable oversight but can also diffuse accountability for measure updates and sector-specific refreshes as technologies evolve. The result is a portfolio with historical credibility but limited evidence of current alignment with rapid increases in rack density, liquid-cooling adoption, and measured-savings approaches now common in the sector.

SDG&E's Market Access Program (MAP), its primary energy-efficiency vehicle, relies on NMEC-based savings but provides no data-center-specific technical guidance. The only available reference, the Tools and Tips for Estimating Energy Efficiency guide (2015), directs users to PG&E's Data Center Best Practices Guide (2012) and Energy Efficiency Baselines for Data Centers (2013), both of which predate modern high-density and liquid-cooled systems. The guide also lists modeling tools such as eQUEST®, EnergyPro®, and EnergyPlus®, but the document notes that none of these tools include a data-center building type, requiring users to approximate with office or hospital templates. SDG&E's publicly available resources continue to mirror PG&E's outdated prescriptive requirements, covering economizers, reheat, fan power, and humidification under Title 24, leaving aggregators and facility operators without current methods or validated models for today's high-density, hybrid-cooled facilities.

Southern California Edison's (SCE) current approach to data center efficiency reflects a cautious and evolving posture. While publicly available materials still emphasize mid-2010s measures, such as variable-speed drives, economizers, and airflow management for conventional air-cooled rooms, staff at Energy Solutions, which oversees SCE's programs in this area, now recognize that these frameworks no longer align with the realities of high-density, AI-driven facilities. Previous data center programs were discontinued when practices such as server virtualization became Industry Standards (SCE 2015, SCE 2025). Still, in interviews SCE explicitly identified the need to re-engage this sector with updated technical assumptions. Staff expressed a deliberate interest in approaching future efforts with care, including integrating power, water, and non-energy benefits into program design to ensure a more holistic and effective response. SCE also noted that within its service territory, hyperscale or AI-focused data centers are not expected to represent major near-term growth, with most new activity anticipated from traditional enterprise and colocation facilities. While the new industry-focused program SCE is preparing to launch will rely primarily on deemed measures, staff anticipate that data center-specific implementations will require custom pathways, consistent with Silicon Valley Power's approach of aligning incentives with verified performance. Taken together, these developments suggest a measured, adaptive strategy, one that balances caution with readiness as data center technologies and grid needs continue to evolve.

Silicon Valley Power stands out as the only utility in this set with data-center-specific offerings, having separate rebate applications for retrofit and new construction that pay for verified kWh savings over multiple performance periods, use site-level or measure-specific M&V plans, and reference PUE to validate results, when appropriate. The materials set minimum eligibility thresholds - for example, IT loads at or above several hundred kilowatts, and define baseline assumptions that include economizer requirements and cap incentives per project and per program year, which signals a more mature administrative framework than is visible at the IOUs (SVP 2025-2026, SVP 2025). The significant expansion within SVP's territory is already documented elsewhere in this report. Yet interviews indicated continued emphasis on air-side economizers and guidance that presume sub-25 kW racks, suggesting a gap between the program architecture and the near-term shift toward 150 kW+ liquid-enabled, AI-oriented builds. The program design is therefore more specific and measurable than peers, while still reflecting legacy assumptions that may understate the scale and pace of load growth now arriving.

There is an urgent need to improve clarity, consistency, and coordination across utilities to support policy that is informed by current realities. Within CalNEXT's remit under the Process Loads TPM, near-term funding should emphasize work that moves measures into the portfolio, performance validation, and market analysis to establish credible baselines and use cases, measure development to translate findings into custom packages, and program development to pilot sector-specific pathways for data centers. Priorities include sector-specific technical guidance aligned to current and anticipated rack densities and liquid or hybrid cooling, updated measurement and verification that pays for verified load and emissions outcomes, and shared data infrastructure that improves observability and trialability. Non-cost barriers deserve equal attention, including alignment with interconnection and permitting timelines, program rules that recognize demand flexibility and waste-heat reuse, and workforce training that moves beyond air-only playbooks. In parallel, the CPUC can place guardrails on cost recovery by limiting rate increases tied to data center growth until utilization and demand persistence are verified, using milestone-based approvals and prudency reviews to reduce stranded-asset risk and avoid unnecessary investment in higher-emitting infrastructure. Framed this way, state and utility partners can support rapid build-out while protecting affordability, reliability, decarbonization, and clean air.

Further Stakeholder Insights

The research team has consulted and collaborated with key stakeholders across the datacenter cooling ecosystem. Our strategy focused on engaging representatives from each critical segment of the supply chain, from silicon providers to utilities and end users, ensuring we captured comprehensive market perspectives. To keep the analysis cohesive, stakeholder insights are woven throughout the report in the sections where they are most relevant, though highlights from one-on-one interviews with major industry players are also synthesized here.

Silicon providers and server integrators: Silicon providers and server integrators (AMD, Intel, NVIDIA, Dell, Supermicro, and Oracle) consistently noted the rapid increase in server and rack power densities, with projections exceeding 10 kW per 1U and reaching up to 200 kW per rack by the year 2030. While D2C liquid cooling for all components offers high thermal performance, providers emphasized it is often cost-prohibitive and requires custom engineering. A more practical approach involves hybrid cooling schemes—using cold plates for high-heat-flux components such as CPUs and GPUs, while relying on air cooling for lower-power elements. Cold plate specifications are tightly controlled by silicon providers, and the increasing weight of liquid-cooled servers and racks is a growing concern for mechanical design and serviceability.

Cooling solution providers: Cooling vendors (CoolIT, Submer, MyHeatSinks, and Liquid Cool Solutions) highlighted the potential of hybrid immersion cooling systems, particularly for high-flux applications. These systems may require separate fluids—one for cold plate cooling and another dielectric fluid for immersion of remaining components—to optimize thermal performance and maintain serviceability. The providers broadly agreed that higher facility water temperatures can be accommodated with advanced cooling technologies such as D2C and immersion, improving overall system efficiency. While some stakeholders expressed concern about the challenges of cooling HBM, this was not universally shared. Overall, the feedback emphasized the need for flexible, scalable cooling architectures that balance thermal performance, cost, and operational complexity.

Data center design firms: Feedback from couple of datacenter design firms highlighted the challenges and tradeoffs associated with achieving Tier III datacenter reliability, particularly in the context of cooling system design. One design firm emphasized a modular approach to power and cooling redundancy that allows for concurrent maintenance and fault isolation, minimizing disruption while avoiding the inefficiencies (in terms of cost and space) of overly complex redundancy schemes. The other design firm reinforced that Tier III certification often requires highly intricate cooling architectures—such as multi-path chilled water systems or redundant liquid cooling loops—which can be prohibitively expensive and operationally complex for many end users. They noted that many customers are frequently willing to tolerate limited downtime in portions of the datacenter, such as individual racks or servers, rather than implementing Tier III-level redundancy across all cooling and power components. Ultimately, both firms agreed that the nature of the IT workload and customer expectations should guide the level of cooling system redundancy, allowing for tailored solutions that balance reliability, cost, and maintainability.

Energy providers and utilities: The team engaged Chevron, Pacific Gas and Electric (PG&E) through their 2050 Partners Code Readiness program, Southern California Edison, Valley Clean Energy, San Diego Gas & Electric, and Silicon Valley Power to assess utility perspectives on datacenter load growth and efficiency opportunities. The IOU responses to data-center growth, as reflected in publicly available materials and recent interviews, show that while utilities are currently not fully prepared for rapid load increases, there is an emerging consensus that custom measures will be required to meet the expectations of future operators. PG&E remains the most visible IOU in this space. Yet, its guidance on data-center efficiency dates back to 2012, with only a few sector-specific programs and limited coordination across state agencies. SDGE's MAP program relies on NMEC-based savings. Still, it provides no data-center-specific technical guidance, leaving operators without current methods or validated models for today's high-density, hybrid-cooled facilities. In interviews, SCE recognized that their frameworks still emphasize mid-2010s measures such as variable-speed drives, economizers, and airflow management for conventional air-cooled rooms, and they anticipate a need to develop custom measures that align incentives with verified performance. Silicon Valley Power stands out as the only utility in this set with data-center-specific offerings, having separate rebate applications for retrofit and new construction that pay for verified kWh savings over multiple performance periods. Interviews indicate continued emphasis on air-side economizers and guidance that presume sub-25 kW racks, suggesting a gap between the program architecture and the near-term shift toward 150 kW+ liquid-enabled, AI-oriented builds. Taken together, these developments mean a measured, adaptive strategy that balances caution with readiness as data center technologies and grid needs continue to evolve.

Communities: Integrating datacenters with rural electric co-ops would be a great opportunity to bring broadband/fiber access to rural communities. However, local communities must be engaged in the decision process. There also seems to be an increased interest in locating datacenters in tribal lands.

Beyond individual interviews, we leveraged several informal discussion venues from late 2024 to early 2025 to gather broader market insights. These included participation in the 2024 Open Compute Summit, NVIDIAs 2025 GPU Technology Conference, the 2025 ARPA-E Summit, and the 2025 UC Davis Energy Efficiency Institute Board Meeting. These events provided opportunities for conversations and panel discussions that supplemented our more formal interview process.

Recommendations

To support the thermal demands of HPC and AI workloads, particularly those using chips with TDP approaching 1,500 W, D2C liquid cooling emerges as the most effective solution. D2C systems provide superior energy efficiency and precise thermal control, ensuring reliable operation of high-TDP processors while minimizing cooling overhead. For new datacenter builds targeting dense compute environments, D2C should be considered a foundational cooling strategy.

In retrofit scenarios, where existing facilities rely on aircooled CRAH/CRAC infrastructure, hybrid integration of D2C cooling requires careful evaluation of auxiliary heat loads. If the CRAH system can absorb the residual (auxiliary) TDP—typically from memory, power delivery, and other boardlevel components—then a hybrid approach is feasible with D2C cooling for CPUs and GPUs, with CRAHs handling the remaining heat. However, if CRAHs cannot meet the auxiliary TDP, RDX provides a more capable retrofit path, capturing residual heat at the rack level and reducing the burden on central air systems.

In cases where liquid lines cannot be extended into the datacenter room due to layout or operational constraints, a liquid to air CDU becomes a viable alternative. This approach allows selective integration of AI or HPC racks into a traditional air-cooled hall, though it requires either increasing CRAH capacity or reducing the number of air cooled racks to maintain thermal balance. Importantly, this configuration carries an energy penalty due to the inefficiencies of double heat exchange. For lower TDP chips, single-phase immersion cooling provides a compact and efficient solution, particularly in edge or modular deployments. Hybrid immersion D2C configurations can extend immersion cooling to higher TDP workloads by combining submersion with targeted liquid cooling for the hottest components.

From a market perspective, liquid cooling technologies are commercially available and increasingly deployed in hyperscale environments. D2C cold plates and rack-level CDUs are the most mature and widely adopted, with multiple vendors offering standardized solutions. Immersion cooling systems are available but remain niche due to serviceability and interoperability challenges, while two-phase systems are still emerging and face barriers related to fluid costs and environmental concerns. Market trends show a clear trajectory: hyperscale operators are normalizing liquid cooling for AI workloads exceeding 200 kW per rack, while colocation providers are adopting hybrid retrofits to incrementally increase density. This divide underscores the importance of flexible deployment pathways that accommodate both new builds and legacy facilities.

On the policy side, California's regulatory environment is evolving rapidly in response to datacenter growth. The use of Small Power Plant Exemptions for diesel backup generation has raised environmental and equity concerns, while new legislative proposals (e.g., SB 1298, AB 222, SB 57) seek to increase transparency, require reporting of efficiency metrics such as PUE, and ensure that very large electricity users bear the costs of grid upgrades. Policymakers are also considering incentives for adoption of liquid cooling, R&D support for serviceability and interoperability, and the development of retrofit-friendly standards. Meanwhile, utilities are exploring how to integrate datacenters as flexible grid assets, balancing reliability with decarbonization goals.

Taken together, these findings suggest a set of integrated recommendations:

- **For new builds**, prioritize all-D2C liquid cooling as the baseline strategy for AI and HPC workloads.
- **For retrofits**, adopt hybrid approaches tailored to facility constraints: CRAH-assisted D2C where auxiliary loads are manageable, RDX where CRAHs are insufficient, and liquid-to-air CDUs where liquid distribution is infeasible.
- **For lower-density or modular deployments**, consider immersion cooling, with hybrid immersion-D2C for higher TDP chips.

- **For utilities and policymakers,** align incentive programs with technologies that deliver low $SCCP_{ratio}$, higher facility water supply temperature, and higher rack density, and establish standards that reduce the risk of integrating two-phase cooling technologies.
- **For the market,** encourage vendor collaboration on interoperability, expand service models to address operational concerns, and accelerate the transition from hybrid to fully liquid systems in high-density environments.
- **For future research work,** laboratory characterization of efficiency and reliability of most promising hybrid cooling technologies (D2C + air and D2C+RDX)**.** Also, two-phase liquid cooling systems (cold plates and CDUs) are emerging technologies and will need modeling, testing, and field validation to support future datacenter rack power density above 200+ kW.

References

Accelsius. 2025. "Optimizing Data Center Cooling: TCO and Design Comparisons Between Single-Phase and Two-Phase Direct-to-Chip Facilities." *OCP Summit.* San Jose. https://drive.google.com/file/d/1olzTvM1EGs0g-cTnxVc1JoZTxj7H4uEe/view.

Alphacool. 2025. *Alphacool ES Jet LGA 4677 2U CPU Cooler.* 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.* https://www.ansys.com/products/electronics/ansys-icepak.

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

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

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

Buffington, Brent. 2025. *SCE Data Center Tracking and Forecasting.* Presentation at DAWG Workshop, Sacramento: California Energy Commission, 25-IEPR-03.

n.d. *CalNEXT.* Accessed 10 18, 2023. https://calnext.com/.

CEC. 2025. "Data Center Applications: Power Plants - Backup Generating Systems." https://www.energy.ca.gov/programs-and-topics/topics/power-plants.

Chilldyne. 2025. "Chilldyne Negative Pressure Cooling Distribution Unit (CDU 300)." https://chilldyne.com/cooling-distribution-unit-cdu-cf-cdu300/.

Chilldyne. 2022. "Helical Turbulator Robust Nucleate Boiling Cold Plate." https://arpa-e.energy.gov/programs-and-initiatives/search-all-projects/helical-turbulator-robust-nucleate-boiling-cold-plate.

Coles, Henry C., and Steve Greenberg. 2014. *Direct Liquid Cooling for Electronic Equipment.* Lawrence Berkeley National Laboratory. https://eta.lbl.gov/publications/direct-liquid-cooling-electronic.

CoolIT Systems. 2025. *Cooland Distribution Units AHx180.* https://www.coolitsystems.com/cdu-product/ahx180/.

—. 2025. *Coolant Distribution Unit - CHx2000.* https://www.coolitsystems.com/cdu-product/chx2000/.

Curtis, R., T. Shedd, and E. B. and Clark. 2023. "Performance Comparison of Five Data Center Thermal Management Technologies." *39th Semi-Therm Symposium.* San Jose, CA. doi:https://doi.org/10.23919/SEMI-THERM59981.2023.10267908.

Data Center Knowledge. 2025. *Exploring the Benefits of Liquid-to-Air Coolant Distribution Units (CDUs).* March 04. https://www.datacenterknowledge.com/cooling/exploring-the-benefits-of-liquid-to-air-coolant-distribution-units-cdus-.

Dymyd, L., Ciubotaru, L., Helezen, M., Shah, J. M., Pai, L., Brink, R., Payne, R., Gullbrand, J., and Gore, N.,. 2020. *Design Guidelines for Immersion-cooled IT Equipment.* Open Compute Project. https://www.opencompute.org/documents/design-guidelines-for-immersion-cooled-it-equipment-revision-1-01-pdf.

Envicool. 2025. *Coolinside Full Chain Liquid Cooling Solution.* https://en.envicool.com/upload/download/Coolinside%20Full%20Chain%20Liquid%20Cooling%20Solution.pdf.

Esram, Nora Wang, and and Neal Elliott. 2024. "Turning Data Centers into Grid and Regional Assets: Considerations and Recommendations for the Federal Government, State Policymakers, and Utility Regulators."

Green Revolution Cooling. 2025. *ICEraQ: Powering High-Performance Data Centers with Next-Gen Cooling.* https://www.grcooling.com/iceraq/#.

Han, Yuelin, Zhifeng Wu, Pengfei Li, Adam Wierman, and and Shaolei Ren. 2024. *The Unpaid Toll: Quantifying the Public Health Impact of AI.* arXiv. doi:https://doi.org/10.48550/arXiv.2412.06288.

Kelly, J, Y Wang, E. Leka, and J. Coe. 2025. *Compact Thermal Management Solutions for High-Density AI Data Centers.* Open Compute Project. https://www.opencompute.org/summit/ocp-educational-webinar-program/on-demand-webinars.

Kozubal, Eric. 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).

Li, Pengfei, Yang Jianyi, A. Islam Mohammad, and and Shaolei Ren. 2025. "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models." *arXiv Preprint.* doi:https://doi.org/10.48550/arXiv.2304.03271.

Liquidstack. 2022. "DataTank 4U: Slashing Energy Use in Edge & Micro Data Centers." https://liquidstack.com/content/uploads/2022/06/DataTank4U-Data-Sheet.pdf.

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

Motivair. 2022. *The ChilledDoor: Datacenter and IT cooling.* Schneider Electric. https://www.motivaircorp.com/products/chilleddoor/.

Nvent data-solutions. 2025. *RDHx Pro the game changing solutions for high density racks.* https://www.nvent.com/en-us/data-solutions/rear-door-heat-exchangers/.

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

OCP. 2023. *Immersion Cooling Requirements Rev 2.0.* Open Compute Project. https://www.opencompute.org/documents/ocp-acs-immersion-requirements-rev-2-1-pdf.

Offutt, Martin C., and Ling Zhu. 2025. *Data Centers and Their Energy Consumption.* Nonpartisan report for Members of congress, Washington D.C: Congressional Research Service. https://www.congress.gov/crs_external_products/R/PDF/R48646/R48646.3.pdf.

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

PG&E. 2013. "Data Center Best Practices Guide Energy Efficiency Solutions for High-Performance Data Centers."

PG&E. 2012. "Energy Efficiency Baselines for Data Centers."

Pineda, Manuel. 2025. *Powering Silicon Valley's AI & Data Center Growth.* CMUA Water and Power Conference. https://www.cmua.org/files/Manuel%20Pineda%20CMUAFinal.pdf.

Purdue Univeristy; Binghamton University; Seguente Inc. 2023. "Confined Direct Two-phase

Jet Impingement Cooling with Topology Optimized Surface Engineering and Phase Separation Using Additive Manufacturing." *ARPA-E.* https://arpa-e.energy.gov/programs-and-initiatives/search-all-projects/confined-direct-two-phase-jet-impingement-cooling-topology-optimized-surface-engineering-and-phase-separation-using-additive-manufacturing.

SCE. 2015. "Computing the Benefits: Solutions for Creating Energy-Efficient Data Centers."

SCE. 2025. "Data Center Energy Savings & Efficiency." https://www.sce.com/business/save-costs-energy/savings-by-business-type/data-centers.

Shah, J. M., R. Eiland, P. Rajmane, A. Siddarth, D. Agonafer, and V. Mulay. 2019. "Reliability Considerations for Oil Immersion-Cooled Data Centers." *Journal of Electronic Packaging.* doi:DOI: 10.1115/1.4042979.

Shehabi, Arman, Alex Newkirk, and Sarah Smith. 2024. "United States Data Center Energy Usage Report." doi:https://doi.org/10.71468/P1WC7Q.

Shell and Asperitas. 2025. *Integrated Immersion Cooling Solution.* https://www.asperitas.com/resources/downloads.

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

Submer. 2025. *SmartPoD EXO: Ensuring extreme density and climate-resilient datacenters.* https://submer.com/smartpod/exo/.

Sunkara, Krishna Chaitanya, and and Krishnaiah Narukulla. 2025. "Power Consumption and Heat Dissipation in AI Data Centers: A Comparative Analysis." *International Journal of Innovative Research in Science, Engineering and Technology.* doi:https://doi.org/10.15680/ijirset.2025.1402015.

SUNON. 2025. "DC Brushless fan & blower." Ltd. SUNONwealth Electric Machine Industry Co. 08 25. 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.

SVP. 2025-2026. "2025-2026 New Construction Data Center Rebate Application." https://www.siliconvalleypower.com/home/showpublisheddocument/87487/638881819076400000.

SVP. 2025. "Data Center Rebate: Maximize Your Data Center Efficiency." https://www.siliconvalleypower.com/home/showpublisheddocument/84943/638935341760800000.

University of Florida. 2023. "Hyperefficient Data Centers for Deep Decarbonization of Large-scale Computing." *ARPA-E.* https://arpa-e.energy.gov/programs-and-initiatives/search-all-projects/hyperefficient-data-centers-deep-decarbonization-large-scale-computing.

Vertiv Group Corp. 2025. *Vertiv™ CoolChip CDU – 2300kW Data Sheet.* https://www.vertiv.com/4a667f/globalassets/shared/vertiv-coolchip-cdu-2300kw-data-sheet-sl-80005.pdf.

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