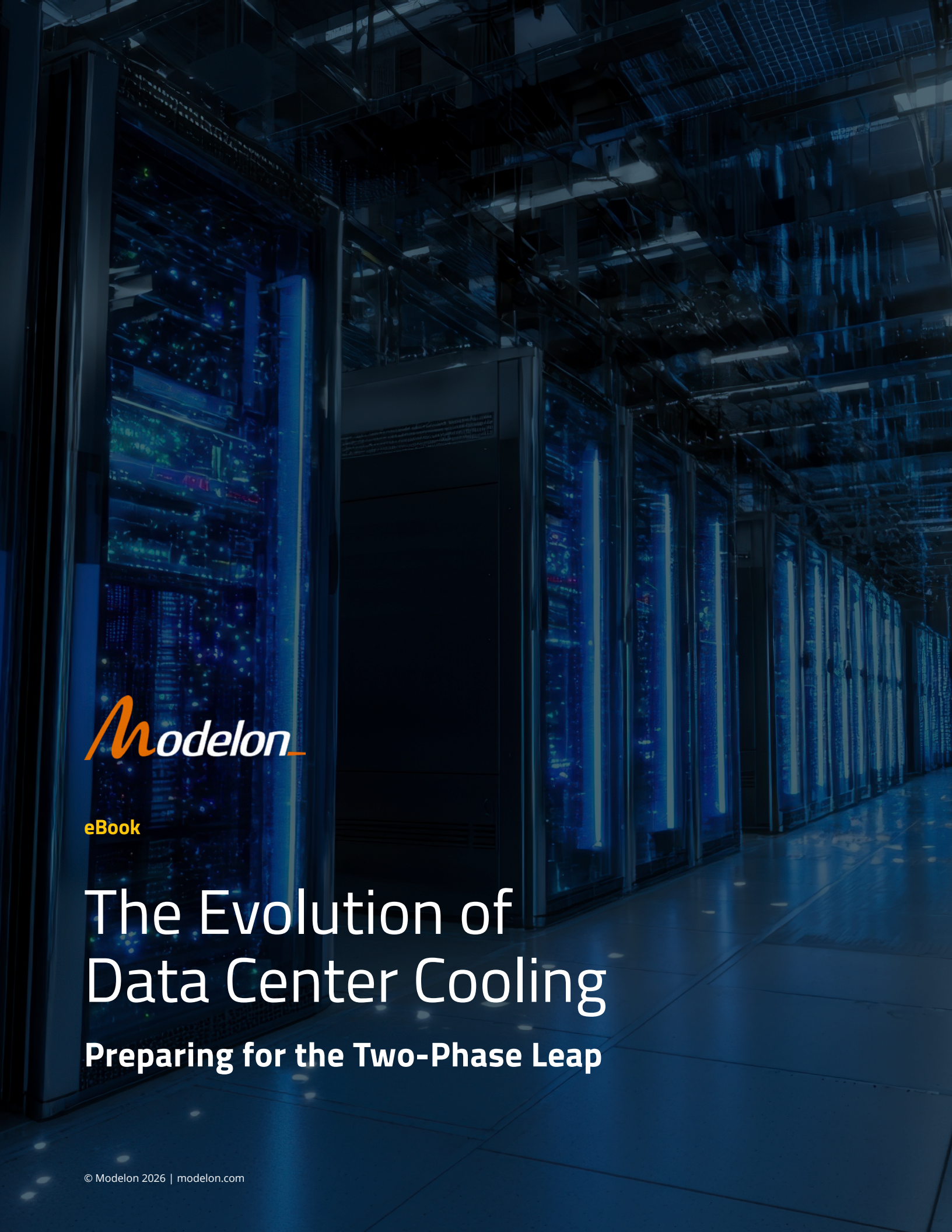


Modelon

eBook

The Evolution of Data Center Cooling

Preparing for the Two-Phase Leap

© Modelon 2026 | modelon.com

The Short Version

- AI and HPC workloads are rapidly pushing rack power densities beyond the practical limits of air and single-phase liquid cooling, making two-phase cooling increasingly important for next-generation data centers.
- Two-phase cooling can remove significantly more heat, but phase change introduces nonlinear behavior, fast transients, and risks such as pressure instability, refrigerant leakage, and dry-out.
- Dynamic system simulation helps engineers understand these complex interactions, right-size cooling equipment, and develop hardware and predictive control strategies together rather than relying on costly overprovisioning.
- Modelon Impact and its exclusive Data Center Library enable teams to model complete cooling systems, evaluate advanced two-phase architectures, and create digital twins that support design, validation, optimization, and real-time control.

Table of Contents

4 What has the Biggest Impact on PUE?

5 The Evolution of Cooling

- **5** Air Cooling
- **7** Single-Phase Liquid Cooling
- **10** Two-Phase Cooling: The Next Frontier

12 Solution Brief: Integrated CDU Design for AI/HPC Data Centers

13 Modelon Supports the Leap to Two-Phase

Introduction

Thermal management is proving to be a significant drag on AI progress. The heat intensity of AI workloads is increasing, as is public sensitivity to the amounts of energy and water consumed by data centers, and the need to show a clear path to profitability. Rather than continuing to slowly evolve, data center cooling technologies need to start progressing in leaps and bounds just to keep up.

While most existing data centers can still rely on air cooling, most new builds cannot—especially in resource-intensive AI and HPC (high-performance computing) facilities. Single-phase liquid cooling is adequately addressing current needs in AI/HPC data centers, but the physics of liquid cooling can only take you so far. It's increasingly hard to imagine a future that isn't dominated by two-phase cooling.

But the evolution of data center cooling isn't just one of pumps, pipes, and chillers. Control systems must evolve as well, moving away from reactive control loops and embracing real-time predictive controls that can anticipate how an entire system behaves over time, under changing load and environmental conditions.

At Modelon, we're helping industry leaders embrace dynamic simulation so they can use modeling to design data center cooling holistically, and then use those models to create the fast digital twins they need for predictive controls.



What Has the Biggest Impact on PUE?

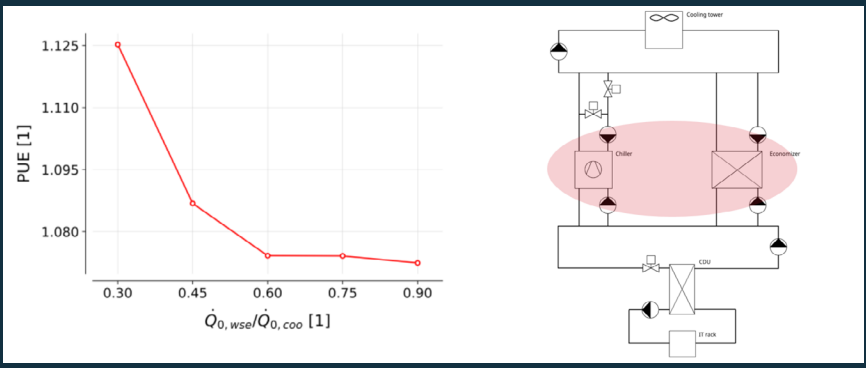
As data centers evolve, lowering operation costs hinges on determining what's driving cooling system energy consumption. This is a daunting challenge. There are non-linear, often counterintuitive, trade-offs at play, making it tricky to establish optimal control strategies and setpoints. In a presentation at SimBuild 2026, Dr. Michael Wetter of Lawrence Berkeley National Laboratory demonstrated how comprehensive system-level modeling can be used to make it significantly easier to explore these non-linear relationships and uncover new opportunities to improve power usage effectiveness (PUE).

Using **Modelon Impact** and the **Modelica Buildings Library**, Dr. Wetter modeled a liquid-cooled data center, outfitted with an economizer and a chiller, and operating at 45°C. He demonstrated how this model can be used to identify sources of energy consumption and explore sizing to improve PUE. Dr. Wetter chose to use Modelon Impact because this multi-domain platform excels at running simulations that fully couple control, pressure-driven flow distribution, and heat transfer. In Impact, designers can construct physics-based, data-driven models that scale from component to system level, and run both dynamic and steady-state simulations.

Conventional wisdom holds that upsizing the economizer is a good way to improve PUE. But after running annual simulations with variable time steps, Dr. Wetter found that this was not necessarily the case. Upsizing the economizer to handle more than 60% of the data center's peak cooling load yields diminishing benefits, in terms of PUE. Instead the model and simulations revealed that, unexpectedly, it was pumping coolant that had a major impact on PUE under these conditions. Pumping accounted for nearly 79% of their cooling-related energy consumption. The model likewise exposed previously under-appreciated potential to improve PUE through part-load control improvements during periods of decreased GPU utilization.

PUE VERSUS ECONOMIZER CAPACITY AS A SHARE OF PEAK COOLING LOAD

Beyond 60%, upsizing the economizer yields diminishing PUE benefits.



Source: Dr. Michael Wetter, Lawrence Berkeley National Laboratory, SimBuild 2026.

The Evolution of Cooling

TYPE OF COOLING	MAX POWER DENSITY	MAX HEAT FLUX	APPLICATION
Air	**Up to 35 kW**	**5–25 W/cm²** Up to 50 W/cm² with in-row cooling and rear-door heat exchangers	Personal computers, mobile devices, enterprise data centers
Liquid	**Up to 200 kW**	**50–100 W/cm²** ~250+ W/cm² with microchannels	Conventional AI/HPC data centers
Two-phase	**Currently up to 250 kW** Theoretical limit of 1,000 kW	**300–500 W/cm²** Experimental systems reach 800–1,500+ W/cm²	Next-gen AI/HPC data centers

Air Cooling

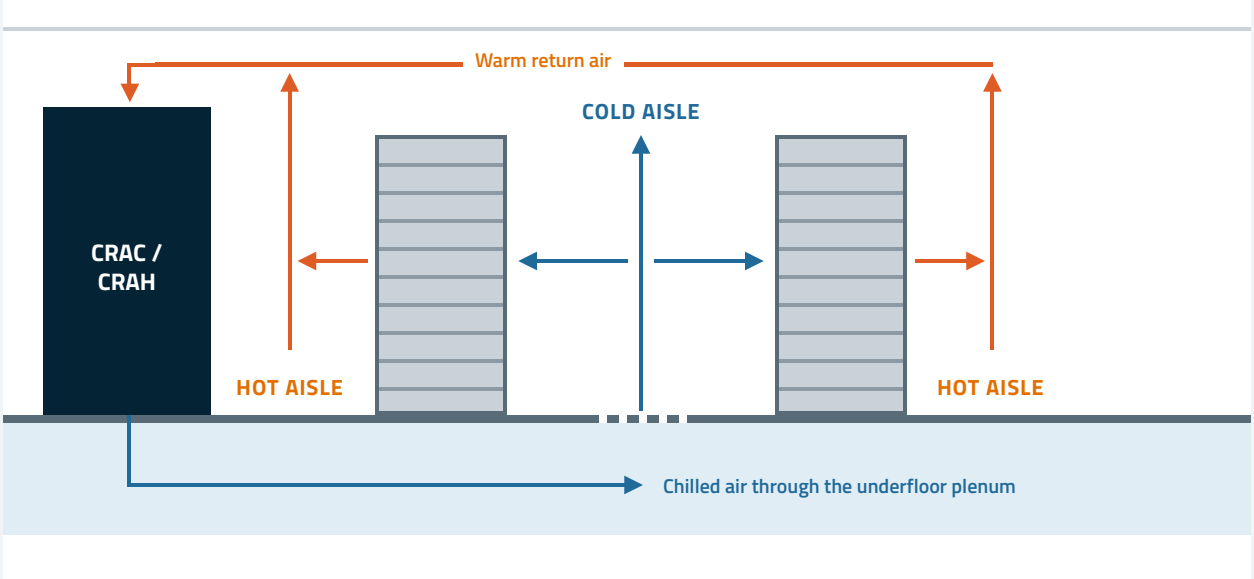
Air cooling is the traditional standard used in everything from laptops to enterprise data centers for banking, web hosting, and e-commerce. The active air-cooling systems in data centers are, in essence, heavy-duty HVAC systems. They pump cold air up through the floor to create “cold aisles” in front of each server rack, then draw away the hot air expelled from the backs of the racks into the “hot aisles.” This hot air is cycled back through the CRAC or CRAH (computer room air conditioner or computer room air handler), cooled, and pumped back under the server room’s false floor to replenish the cold aisle. The CRAC or CRAH is most often managed by a simple control loop: The system has a desired setpoint, knows the current temperature, tracks how long the temperature has deviated from the setpoint, calculates a trend, and modulates the supply air temperature and pressure to bring setpoint and reality closer together.

Air cooling of this sort is simple and effective, provided your racks are usually drawing between 15 to 20 kW with few transients, and rarely exceeding 35 kW. Since standard enterprise data center server racks typically consume 3.5 kW to 10 kW, and that power density is fairly evenly distributed across the rack, air cooling is a great solution.

Air-Cooling System Design: "Macro-level" Modeling

An enterprise data center is essentially a slow-moving microclimate in a warehouse. To design an air-handling system to manage that microclimate, you need a "macro-level" view of the data center's airflow. While a comprehensive multi-domain system-level platform like Modelon Impact can prove extremely helpful in modeling the cooling equipment itself—chillers, air-handlers, water loops, CRAH/CRACs, etc.—it's the air in the data hall, which serves as the direct cooling medium, that most concerns engineers. Controlling this huge, slow-moving, slowly changing mass of air is best accomplished using 3D Computational Fluid Dynamics (CFD).

True, CFD cannot capture all of the potential nuances and air-flow complexities of what is happening in a data center, but enterprise computing is fairly forgiving: Engineers can design their air-handling system based on a CFD model that captures the data hall airflows, and work out the details for the control loop once everything is installed.



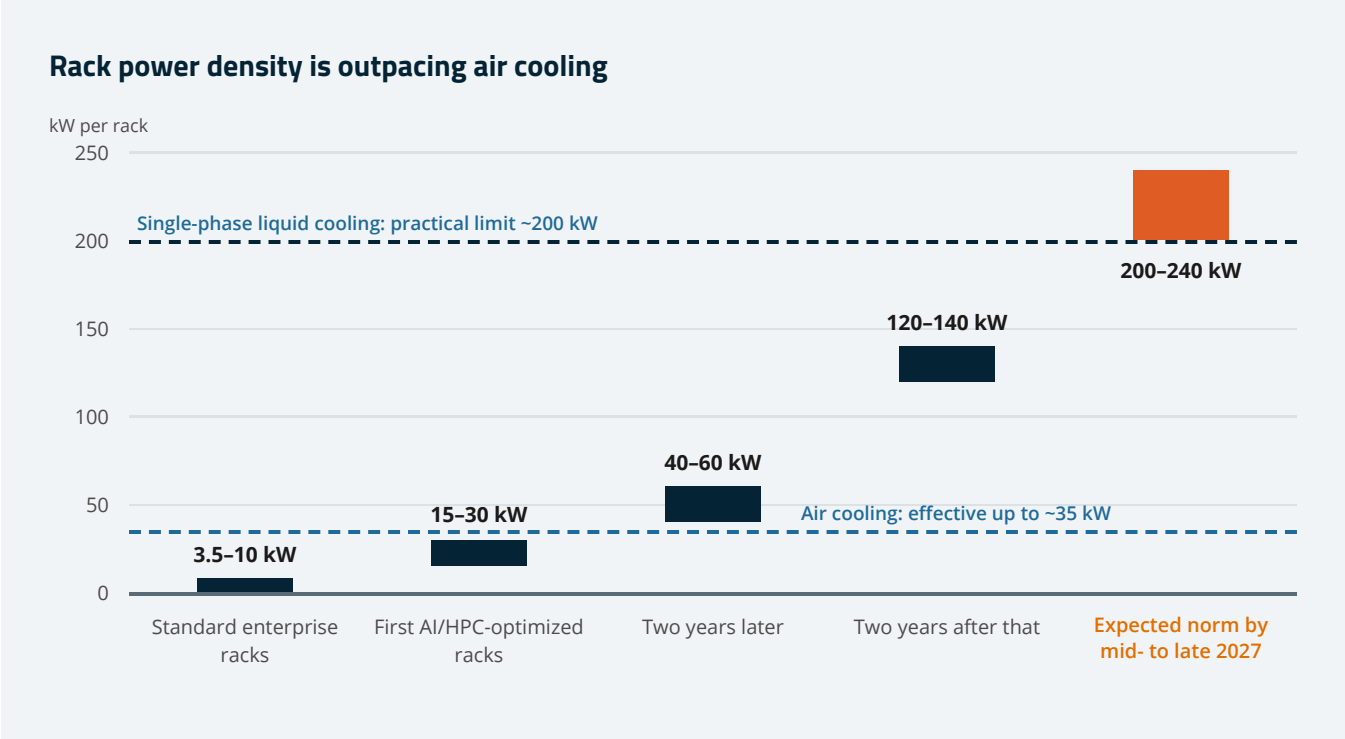
Raised-floor air cooling: the CRAC/CRAH pushes chilled air through the underfloor plenum and up into the cold aisle. Servers exhaust warm air into the hot aisles, and it returns overhead to the CRAC/CRAH to be cooled again.

An enterprise data center is essentially a slow-moving microclimate in a warehouse.

Single-Phase Liquid Cooling

Advancements in chip design and growth in AI compute demand has rapidly pushed chip power densities beyond what air cooling can reliably handle. The first racks optimized for AI/HPC were in the 15 kW to 30 kW range, near the maximum for effective air cooling. Just two years later, new AI racks were consuming around 40 to 60 kW (twice the power density of the average server rack of the preceding decade). Two years after that, new racks were consuming 120 to 140 kW. Now, the latest AI racks have power densities four times greater than what air cooling can effectively handle. With heat transfer coefficients around 3,000 times those of open air cooling systems, liquid cooling systems are effective up to 100 to 200 kW, making them a good fit for the current generation of AI/HPC data centers.

But the cooling challenge posed by modern AI/HPC workloads isn't just one of higher power densities. AI/HPC racks both draw more power, and do so irregularly across the rack—and even across individual chips. By delivering coolant (usually either water or glycol) directly to the chip itself, liquid cooling can much more effectively address these thermal gradients across servers.



~3,000×

Heat transfer coefficients of liquid cooling, compared with open air cooling systems

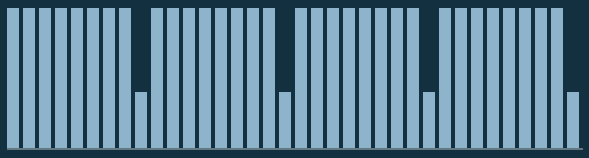
100–200 kW

Rack power densities where single-phase liquid cooling remains effective

AI Workloads are Fundamentally Different

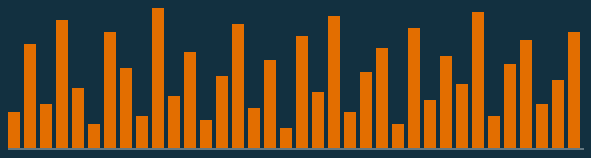
Traditional enterprise data center workloads are well understood, often quite predictable, and fairly steady. AI workloads are the opposite. During training, the entire system experiences a massive power draw for weeks or months at a time, with large transients during regular checkpointing and data synchronization. Then, during inference (when users are querying the model), the power draw becomes "spikey" and asynchronous, creating uneven power demands across a rack. In such a situation, the cooling system needs to both correctly anticipate which specific servers or chips will need cooling, and precisely deliver it within milliseconds.

TRAINING



Sustained, massive power draw for weeks or months, with transients at checkpoints

INFERENCE



"Spikey," asynchronous power draw that varies across the rack

"Hotspots" and Thermal Gradients

Squeezing more power into the same silicon square means that AI chips have higher heat flux (measured in watts/cm^2) than their predecessors. By working these chips unevenly, inference creates wider thermal gradients, not just across the rack, but across the surface of an individual chip. If not managed properly, the speed and magnitude at which this heat varies will create wide thermal gradients in small areas, damaging the chips and physically straining surrounding components. Poorly managed thermal cycling, hotspots, and thermal gradients accelerate fatigue, drive down reliability, and result in early failure of expensive components.

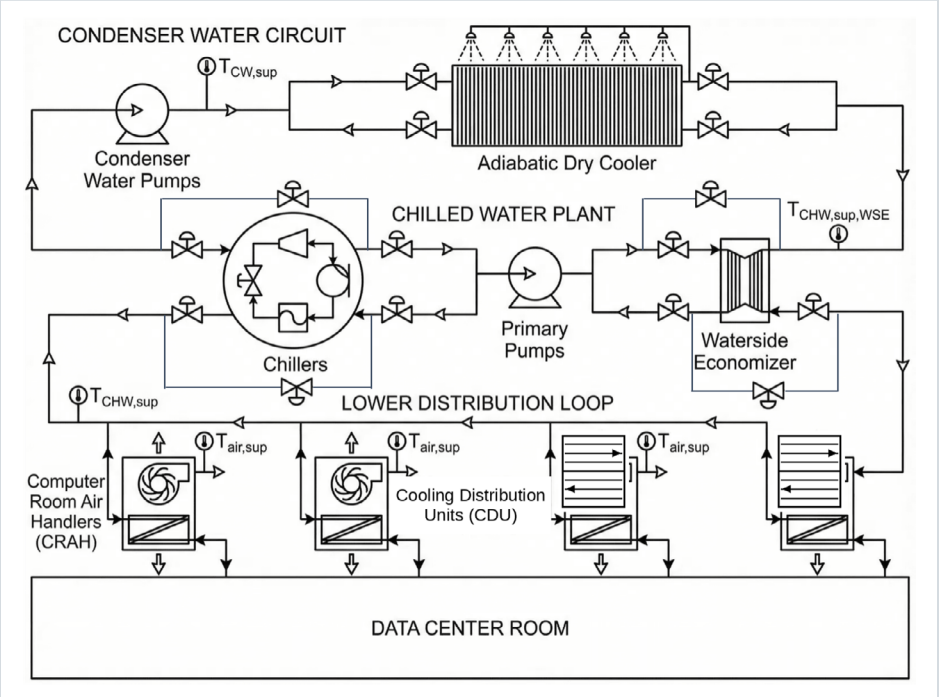
Liquid-Cooled System Design and Control: Fast System-Level Modeling

Liquid cooling can be fairly easy to design, build, and maintain. With dripless quick-connect (QD) couplings, maintenance and repair can be almost as easy as swapping batteries in a power tool: Unclip a server, pull it, and plug a new one in.

At first glance, modeling a single-phase cooling system likewise seems *less* challenging than modeling for air cooling. In a liquid-cooled system, you are pumping chilled water to cold plates: You control precisely where the water goes, its pressure, and its temperature. This is a discrete 1D system and can be modeled at a fraction of the computational cost of a full CFD solution for an air-cooled data center.

But while modeling and maintaining liquid cooling systems is arguably easier, the stakes are inarguably higher: Thermal tolerances in AI/HPC data centers are tighter, transients faster, and mistakes more expensive. You must perfectly model pump pressures, flow rates, and heat exchanger efficiency to ensure you can deliver cooling anywhere on the rack quickly enough to catch transients before they cause damage.

Since your thermal gradients are greater and more unpredictable in these data centers, simple control loops alone are insufficient. You need a control solution that can also predict which specific chip—and even which portion of which chip—will be in a hotspot. This requires a shift in tooling to more robust dynamic system simulation. Platforms like Modelon Impact are designed to accurately model entire interconnected systems in real time. The same model used to create the system can then serve as the digital twin underlying a proactive control system.



Thermal tolerances in AI/HPC data centers are tighter, transients faster, and mistakes more expensive.

Two-Phase Cooling: The Next Frontier

Practically speaking, single-phase liquid cooling is limited to power densities below 200 kW—a threshold we're likely to exceed soon. Industry experts agree that 200 to 240 kW racks will become the norm in AI/HPC data centers by mid- to late 2027.

This is why two-phase cooling systems have become a hot topic. Two-phase cooling uses a refrigerant that changes phase (from liquid to gas) as it passes the chip. Harnessing the latent heat on the surface of the GPU/TPU to boil the refrigerant allows the coolant to absorb an enormous amount of energy with a much smaller amount of fluid. This approach also effectively addresses higher levels of heat flux, more precisely moving more heat away from hotspots more quickly. Theoretically, two-phase cooling architecture could deliver heat transfer coefficients 10 to 20 times those of single-phase liquid cooling, which in turn is 50,000 times greater than that of air cooling. Current two-phase cooling systems handle up to 250 kW, providing 2 to 3 times the total heat removal capacity of single-phase liquid systems.

Given the extremely high theoretical heat transfer coefficient of two-phase cooling, it's reasonable to assume there is considerable room to grow. But there are also considerable design challenges. Two-phase behavior is non-linear: In a two-phase system a 10% increase in heat doesn't result in a 10% increase in temperature—it results in a phase change. When that liquid turns to vapor, its volume, velocity, and pressure dynamics all change drastically and instantly, more than doubling refrigerant fluid acceleration. Two-phase cooling loops are also much more sensitive to any fluid or gas leakage, eliminating the option of convenient quick-connect hoses.

With heightened capacity also comes heightened risk. If you're designing a system where you intend to vaporize the coolant, then you run the risk of "dry-out": intense transient heat spikes can vaporize your refrigerant before the chip has properly cooled, and potentially trap that vapor against the surface you are attempting to cool. This creates an insulating blanket that hobbles thermal management.

10–20×

Theoretical heat transfer coefficients of two-phase cooling, compared with single-phase liquid cooling

250 kW

Rack power densities handled by current two-phase cooling systems

2–3×

Total heat removal capacity of single-phase liquid systems

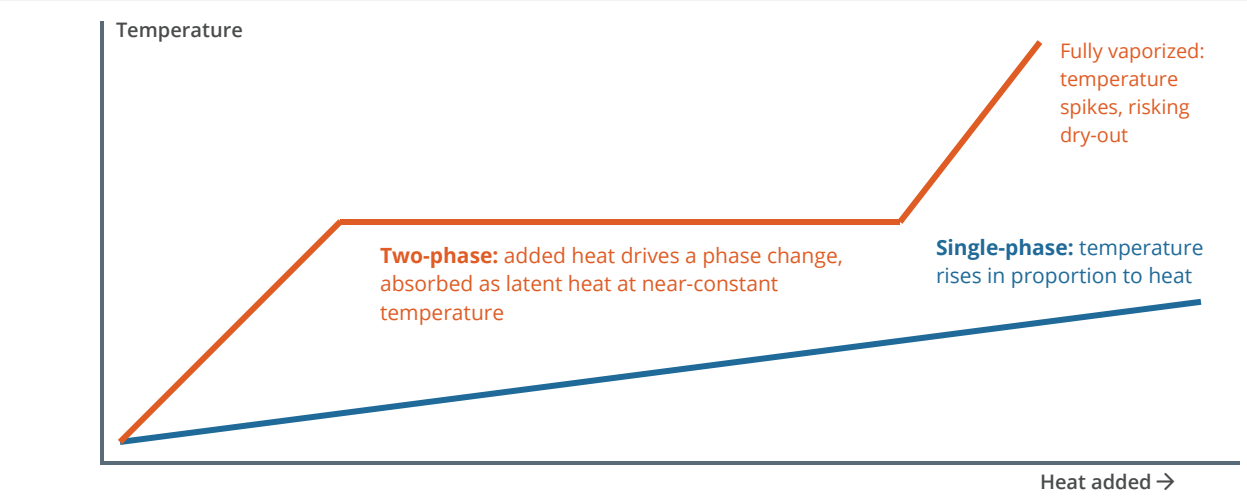
Two-Phase Cooling System Design: Transient System Modeling

When it comes to modeling, there is a fundamental similarity between air-cooled and single-phase liquid-cooled systems: In both cases, you are working to capture behaviors in a system where temperature spikes and fluid flow are proportional to the heat absorbed. In such systems, if you leave some margin of error to absorb occasional transients, all will be well.

But with two-phase cooling, the properties of your fluid change dramatically as you heat or cool. This has a correspondingly dramatic impact on overall system behaviour and, crucially, cooling capacity. By intentionally forcing your refrigerant to change states, you are making transient management a core part of your control design. This requires dynamic system modeling that brings together complex thermodynamic equations with the physical realities of the system as a whole.

Under these conditions, control system tuning becomes much more difficult, with traditional reactive control logic more likely to fail. Ideally, a two-phase cooling system will rely on predictive control systems, where a digital twin accurately modeling what is about to happen in your data center allows the control system to anticipate and meet cooling demand.

Why two-phase behavior is non-linear



Illustrative only, not to scale. In a two-phase system, a 10% increase in heat doesn't result in a 10% increase in temperature—it results in a phase change.

By intentionally forcing your refrigerant to change states, you are making transient management a core part of your control design.

SOLUTION BRIEF

Integrated CDU Design for AI/HPC Data Centers

When the dynamic behavior of the cooling system cannot be predicted, engineers tend to compensate with margin. That means investing in oversized CDUs (Coolant Distribution Units) and establishing conservatively low control setpoints. This in turn creates redundant capacity that will rarely (if ever) be used. Every kilowatt of overprovisioned cooling is capital spent and efficiency lost. At the scale of today's AI data centers, that's an enormous wasted investment.

This is an inconvenience in single-phase cooling systems, but becomes critical as data centers begin to adopt more complex two-phase systems. In a single-phase cooling system, the CDU is a fairly simple thermal "heart," combining a pump and a heat exchanger. In a two-phase system, the CDU is more like a complex thermodynamic "brain," actively managing every aspect of the refrigeration cycle. It balances pressure, flow rate, and condensation to regulate the quality of the refrigerant (the ratio of liquid to gas) and ensure reliability.

Because two-phase CDUs are in their infancy, with many fundamental questions about two-phase flows and heat transfer as-of-yet unanswered, it's easier to see how using modeling and simulation to develop control and hardware together makes sense. But modeling and simulation have just as much to offer with mature technologies, like single-phase cooling: When you can rely on a validated model throughout the design process, you can right size your entire system, reducing costs while improving reliability.

With Modelon's tools, you can create a digital twin of your CDU at a component level and test control code against the physics before you ever build the hardware. The result is a system that provides more hours of cooling at no additional cost, uses less energy, and reaps measurably better PUE (power usage effectiveness). That same validated model can then become the foundation of the cooling architecture's control system, solving the time derivatives of pressure and temperature in near real-time to anticipate transient heat spikes before they can cause trouble.



SINGLE-PHASE CDU

The thermal "heart"

A fairly simple unit combining a pump and a heat exchanger.

TWO-PHASE CDU

The thermodynamic "brain"

Actively manages the refrigeration cycle, balancing pressure, flow rate, and condensation to regulate refrigerant quality and ensure reliability.

Modelon Supports the Leap to Two-Phase

Many of the fundamental questions underlying two-phase physics, as it applies to data center cooling, remain unanswered. Nonetheless, a two-phase system is likely the only way to adequately cool AI/HPC data centers. The ultimate goal isn't just to design better cooling system hardware—it's to design hardware systems and control strategies hand in hand, so those systems can adapt and evolve as well.

Modelon has an established track record of helping data center builders, hyperscalers, and component suppliers embrace system-level simulation for their projects and supporting control design by modeling volatile transient behaviours in critical infrastructure. Our cloud-based system simulation platform, Modelon Impact, supports intelligent collaboration across teams, accelerating project turnaround while lowering costs.

Over the last 20 years, we've distinguished ourselves in thermal management, reliably and accurately modeling fluid flow and heat transfer at scale over different flow regimes in the automotive, aerospace, and HVAC industries. We've excelled at capturing tightly coupled systems, creating a platform for understanding the system-wide interactions of individual components under varying operating conditions. Modelon is currently collaborating with engineers at a leading hyperscaler and researchers at University of Maryland, working together to translate HVAC phase-change physics into real-world data center cooling solutions, apply our proven libraries to new problems, and share the resulting tools with component manufacturers.

View Data Center Solutions

© Modelon 2026 | modelon.com