uptime INTELLIGENCE

INTELLIGENCE UPDATE

AI and cooling: methods and capacities


Dr. Tomas Rahkonen

4 Feb 2025

AI and cooling: Part 1

Compute infrastructures for training large AI models are similar to high-performance computing (HPC) systems, which have long been used for demanding tasks in fields such as engineering, scientific research and finance. Both AI and HPC workloads require the processing of large datasets and complex computations. These demands are met by using specialized hardware such as high-performance servers with GPUs and high-speed interconnects, which tend to be much denser than typical IT equipment, with corresponding challenges in thermal management.

This Uptime Intelligence report is the first of three parts and introduces the cooling methods used for AI workloads and discusses their capacity ranges. Part two will look at AI cooling resiliency and the third part will focus on efficiency and heat reuse.

Note on AI rack densities

High-profile AI training infrastructure deployments, such as those used for training large language models, use many thousands of high-performance GPUs arranged in high-density racks. The drive to maximize training performance requires close coupling of chips, which in turn has led to jumps in rack loads — current generation systems can surpass 40 kW, while the 2025 generation will see some implementations that exceed 100 kW. Future leading-edge products will progress towards hundreds of kilowatts per rack.

By contrast, typical enterprise AI training setups are smaller and hosted either on-premises or in colocation data centers, with only hundreds or a few thousand GPUs. Rack loads in these facilities can be significantly lower, though still well above industry-average densities (the vast majority of racks are still well below 10 kW).

The primary driver behind increasing the density of AI compute racks is the close coupling of GPUs to enable low-latency, high-bandwidth data sharing. Notably, Nvidia’s rack-scale systems incorporate a copper interconnect to directly link up dozens of GPUs in a single mesh. If novel

model training methods ease the remote latency requirements between GPUs and compute nodes, the pressure on extreme densification would lessen accordingly, alleviating power and cooling challenges. For now, training leading-edge AI models favors tightly coupled, massively parallel system architectures.

IT cooling methods and typical capacities

Table 1 provides typical IT rack capacities for IT cooling methods commonly deployed in data centers. Immersion cooling is less frequently deployed than the other methods shown. These capacity numbers are provided as general guidelines for practical deployments. Higher rack capacities may be possible under favorable circumstances. The numbers assume the use of standard 19-inch 48U tall racks and large immersion vats, typically around 3 m x 1.5 m x 1.5 m (10 ft x 5 ft x 5 ft).

Table 1. Cooling methods and typical capacities

Cooling methods	Typical rack power
Air cooling	
Perimeter cooling *(CRAHs with containment system or fan walls)*	Up to 20 kW to 25 kW per rack
In-row	10 kW to 50 kW per rack
RDHx	20 kW to 50 kW per rack
Hybrid air and liquid cooling	
Sidecar (<70% DLC)	40 kW to 70 kW per rack
DLC (<70%) and perimeter cooling	50 kW to 70 kW per rack
DLC (<70%) and RDHx	70 kW to 150 kW per rack
DLC (>90%) and perimeter cooling or RDHx	150+ kW per rack
Total liquid cooling	
100% DLC coverage	150+ kW per rack
Single-phase immersion	50 kW to 150+ kW per vat
Two-phase immersion	50 kW to 150+ kW per vat

CRAH: computer room air handlers; DLC: direct liquid cooling; RDHx: rear-door heat exchanger

UPTIME INSTITUTE 2025

New IT cooling methods in data centers are usually adopted only when needed to support new IT types that result in higher rack power, have operating temperature restrictions of the IT components, or have other special requirements.

Perimeter cooling continues to be the preferred cooling method for traditional, low-density workloads (up to 20 kW to 25 kW of rack power) — with few signs of change in sight. Close-coupled cooling systems are also well proven and provide a relatively easy way to introduce higher rack power (typically up to 50 kW) in a few rack rows (or a dedicated room) within a perimeter-cooled data center.

Liquid cooling, whether hybrid or total, is typically used for high rack power (more than 50 kW) or high-performance IT with specialised cooling demands. A complicating factor is that liquid cooling can limit IT hardware options. Unlike air cooling, which is compatible with almost all IT hardware, support for cold plates or immersion cooling must be confirmed for each type of IT hardware. This is less problematic for large AI training data centers with standardized hardware but can pose challenges for smaller colocation or mixed-use facilities.

Direct liquid cooling (DLC) further impacts facility infrastructure and operations through additional piping (tertiary cooling rings to IT cold plates) in the white space, redundant coolant distribution units (CDUs) that are typically powered by UPS, and blurred boundaries for service level agreements between facilities and IT teams (or colocation providers and tenants). For rack powers approaching 150 kW or more, total (or close to total) liquid cooling solutions will be required.

Perimeter air cooling

Most data centers use perimeter air cooling, where computer room air handlers (CRAHs) or fan wall units supply cold air to IT equipment. Typically, cooling units are strategically placed to allow for maintenance without entering the IT area. CRAHs release heat to the facility's cooling system, usually a chilled water loop. Fan wall units release heat either to a chilled water loop or directly to the outside air. Hot or cold aisle containment is often used to manage airflow, improve cooling efficiency, and prevent hot spots.

In systems using direct expansion cooling, computer room air conditioning units are used instead of CRAHs.

Perimeter cooling is generally effective for rack loads up to 20 kW to 25 kW when combined with optimized airflow (or 10 kW to 15 kW in older systems). At higher rack loads, these systems may struggle to provide adequate airflow, leading to hot spots, throttling or equipment failures — especially at high utilization or during maintenance. Key limitations include the cooling capacity of the coils and the airflow capacity of the data center and its containment systems.

Close-coupled air cooling

Close-coupled cooling uses air to cool IT equipment but places cooling units and piping close to the IT racks. This setup typically supports higher rack densities, up to 50 kW per rack, but requires facilities teams to access the IT area for maintenance, which can complicate the interaction between facilities and IT teams.

In-row air cooling

In-row data center cooling is a well-established close-coupled cooling method in which cooling units are placed between server racks. These units deliver cool air directly to heat sources and typically use (efficient) prefabricated air containment systems. In-row cooling systems that release heat to a facility's water-cooling system can provide a straightforward solution for

establishing high-density zones within a data center.

Rear-door heat exchangers

Another close-coupled cooling method commonly used to cool high-density racks is rear-door heat exchangers (RDHx). A door with a depth of 10 cm to 30 cm (4 in to 12 in) is attached to the back of each IT rack and removes heat from the exhaust air using a heat exchanger. For high-density racks, air-to-liquid heat exchangers are used.

Multiple RDHx doors are typically connected via piping to a secondary facility chilled water loop. In some deployments, a cooling distribution unit (CDU) circulates coolant at the required flow rate and temperature, ensuring that heat is effectively dissipated from high-density IT equipment while reducing the impact on facility-wide cooling systems.

RDHx systems and CRAHs can work together, where the RDHx removes most of the heat directly at the rack level, while the perimeter cooling handles the residual heat and maintains ambient room conditions.

Direct liquid cooling

The main advantage of DLC is its high thermal conductivity, which is far superior to air cooling. Water and alternative engineered fluids can transfer heat orders of magnitude more effectively than air, allowing liquid cooling systems to maintain stable temperatures for high-power components like CPUs and GPUs, even under heavy workloads. In immersion cooling systems, the coolant makes direct contact with IT hardware, whereas cold plate systems remove heat indirectly by circulating liquid to cold plates attached to high-power components.

The threshold at which liquid cooling becomes necessary for servers depends on several factors, such as the thermal design power (TDP) and operating temperature restrictions of the server components, the density of the deployment, and the data center’s cooling performance. Liquid cooling is typically considered when TDP exceeds 300 W per processor. Although DLC has so far been mainly used for compute servers, it can also be used for network switches and storage systems.

Most liquid cooling systems use cooling distribution units (CDUs) with pumps to circulate the cooling fluid (either water or dielectric) through heat exchangers and cooling feeds to the IT equipment. CDUs supply a regulated flow of cooled fluid to the system and include electrical components to power the pumps, as well as controls for system operation (such as flow, pressure and temperature). They also provide monitoring and alarms.

There are two types of CDUs based on their heat exchange method:

- **Liquid-to-liquid CDUs** transfer heat to the facility cooling system, which is typically a chilled water system.
- **Liquid-to-air CDUs** transfer heat to air, typically directed back to CRAH units via a hot aisle.

Sidecars

A recent DLC approach that makes use of cold plates and liquid-to-air CDUs is commonly referred to as sidecars. These are typically used for IT rack loads between 40 kW and 70 kW and provide a practical approach for introducing some DLC IT into an air-cooled data center. Sidecars are self-contained units placed within regular rack rows and do not require connection to external piping systems. Compared with in-rack CDUs, sidecars also help free up space in the IT rack for more compute, however, at the same time, they take up rack positions in the data hall.

Each sidecar includes one liquid-to-air CDU that typically supports a single IT rack and releases heat as warm air into a hot aisle. The total number of sidecars that can be supported depends on available CRAH cooling capacity and floor space, among other things. Some sidecars (liquid-to-air CDUs) use a refrigeration cycle to increase cooling capacity.

Cold plates (with liquid-to-liquid CDU)

Cold plate systems designed for high-density racks typically use liquid-to-liquid CDUs, which release heat to the facility cooling system. Each CDU (or pair of CDUs) serves multiple IT racks. A cold plate system may have a dedicated facility cooling feed or share it with perimeter cooling systems.

While cold plate systems primarily remove IT heat via liquid cooling, they still depend on air cooling to remove 5% to 30% (and sometimes up to 50%) of the heat. The air-cooling load can be significant. For example, in a 70 kW IT rack where cold plates handle 70% of the required cooling, 21 kW still requires perimeter air cooling (70 kW x 30% = 21 kW).

Looking at an Nvidia GPU rack of 130 kW with 70% of heat handled by cold plates, 39 kW is left for air cooling, which will require the use of RDHx.

Cold plate systems can absorb more than 90% of heat by extending DLC to additional components like memory and storage. This supports rack loads over 150 kW and can reduce or eliminate the need for IT equipment fans. While such designs were previously used in supercomputers, the rise of high-power GPU racks (150 kW and above) for AI training is likely to make them more common at high-end AI sites. Some cold plate systems, such as the HPE Cray Olympus, can absorb 100% of IT heat.

Immersion cooling

Immersion cooling systems submerge IT equipment in vats (tanks) filled with dielectric coolant. In single-phase systems, CDUs cool and circulate the coolant through a heat exchanger. Two-phase systems rely on the coolant changing phase from liquid to gas to circulate naturally, eliminating the need for pumps.

Both single-phase and two-phase immersion systems release heat to the facility cooling system. Single-phase vats may use either built-in or external CDUs, while two-phase systems use condenser coils. Some two-phase immersion systems make use of a CDU to circulate a non-

water coolant to the coil, avoiding the presence of water (and the associated risks) within the vat.

The commercial availability of two-phase immersion systems significantly declined due to 3M's decision to stop producing per- and polyfluoroalkyl substances (PFAS), including two-phase coolants, by 2025. Some companies, such as Chemours, are developing new alternative two-phase coolants.

Immersion cooling can support vat loads above 150 kW and does not depend on support from air cooling. While immersion cooling is used for high-density crypto mining deployments, it has yet to see uptake in AI compute workloads.

The Uptime Intelligence View

High-profile AI training deployments use DLC and, in the future, may adopt immersion cooling to manage their extreme heat loads. However, edge colocation inference and on-premises AI deployments will likely continue using close-coupled cooling systems, such as RDHx or in-row cooling, for the foreseeable future. These cooling methods can be sufficient for managing the lower heat loads typically associated with inference and less demanding enterprise AI training workloads.

ABOUT THE AUTHOR

Dr. Tomas Rahkonen

Dr. Rahkonen is the Research Director Sustainability, Europe at Uptime Institute. Rahkonen has spent the last 25 years in positions within the telecommunications, mobile communications, and data center sectors globally, and most recently served as the CTO of Flexenclosure, where he managed the design and delivery of prefab data centers across four continents.

trahkonen@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.