Industry

# Optimizing Data Center Cooling Efficiency During Transformational Growth

**AI, innovations, energy and water driving a move to liquid cooling**

Artificial intelligence and machine learning are rapidly changing the data center industry. To take advantage of new technology and meet its demands, companies are operating their centers at higher loads and building racks with significantly higher density. This change has led to a re-thinking of the existing cooling systems, and designing them with more efficient and water-saving solutions.

GRUNDFOS

Possibility in every drop

## Introduction

As companies, societies, and citizens, become increasingly dependent on digital services, fast and reliable dataflow has never been more important. Data centers are at the center of digital infrastructure, ensuring stable connectivity and data communication.

In 2023 alone, global data creation and replication used more than 120[1] zettabytes as individual and business use of digital technology skyrocketed. This is only expected to grow, driven by the continued adoption of artificial intelligence (AI) and machine learning (ML)[2]. AI has the potential to completely change our way of working – most recently seen with the launch of ChatGPT – and this is just the beginning. **A recent Accenture report notes that 98% of company leaders say that AI will play an important role in their strategies in the next three to five years, and McKinsey bluefield research finds that nearly 65% plan to increase their AI investment over the next three years[3 9]**.

To meet the increased demand for processing capabilities there has been a remarkable advancement in chips such as central processing units (CPUs) and graphics processing units (GPUs), resulting in powerful computing solutions. Consequently, high density racks (those greater than 30kW) have become prevalent, leading to challenges with thermal management and energy demand. This has put the data center industry in a paradigm shift where traditional air-cooling systems need to be replaced by thermal management system with higher cooling capacity.

The leading choice is liquid cooling which utilizes the higher cooling capacity of water compared to air. Recent advancement within liquid cooling has shown that this approach not only offers excellent thermal management for high density computation power but **liquid cooling also counterintuitively saves water**. Up to 15%[4] greater total usage effectiveness (TUE) compared to air cooling has been seen, leading to an improved power usage effectiveness (PUE) and more thoughtful use of the already pressured grid capacity.

Data centers are faced with public pressure and a significant social responsibility due to their enormous use of water in a time where around 2 billion people globally don't have access to clean and safe water. It has steered some governments to initiate the drafting of regulations to make it mandatory for data centers to report energy and water usage and ultimately reduce energy consumption by 11.7% between 2020 and 2030 in Europe[5]. The trend of designing more sustainable data centers is global, and hyperscale operators all over the world have clear ambitions and visible water and climate targets.

The pressure is real and underscores the importance for data centers to continue pushing the boundaries for what is possible in the transition to more effective and efficient cooling and use of resources.

> ***"We know people will eventually have to move to liquid, but we also hear feedback from customers that's challenging for them, their data centers aren't there, they need to build new data centers."***
>
> Charlie Boyle—Vice President & General Manager, DGX Systems, Nvidia

## Table of contents

## Data center of tomorrow

The power required to run a data center is already significant. Data centers in the United States account for about 3%[6] of the total power use and that's rapidly increasing. Driven by AI and cryptocurrency, the resulting energy consumption increase is between 20-40%[7] per year.

A newly released study by the Electric Power Research Institute estimates that data centers could consume 9% of the United States´ electricity generation by 2030. Hence, it becomes paramount to run the data centers as efficiently as possible to reduce this number[8].

### How do we lower power consumption?

Updating data center cooling systems is one of the most effective ways to decrease overall energy consumption, lower OPEX cost and save water. On average, 38% of data center energy usage can be attributed to cooling, which is around 80% of all the energy used for running supporting systems as showed in figure 1[9].

### Is liquid cooling hot?

Traditional cooling systems are based on air cooling, which are built on refrigeration infrastructure that includes chillers, pumps, humidity control, fans, thermal managed aisle, cooling towers, and filtration systems. The system works by heat rejection when cold air is blown over the servers transferring the hot air to the hot aisle and removing it from the server room. The system has no direct use of water in the server room thereby removing the risk of leakage and limiting overall maintenance needs. However, a significant amount of water is consumed if the system uses evaporative cooling to dispatch the captured heat from the hot aisle to the atmosphere[10]. The air-cooled system is and will still be the preferred solution for many data centers if the cooling capacity for a single rack is low enough and the climate in the location allows it. Air cooling systems have continually evolved to address higher densities with greater efficiency, but there is a point at which air does not have the thermal transfer properties required to provide sufficient cooling to high-density racks in an

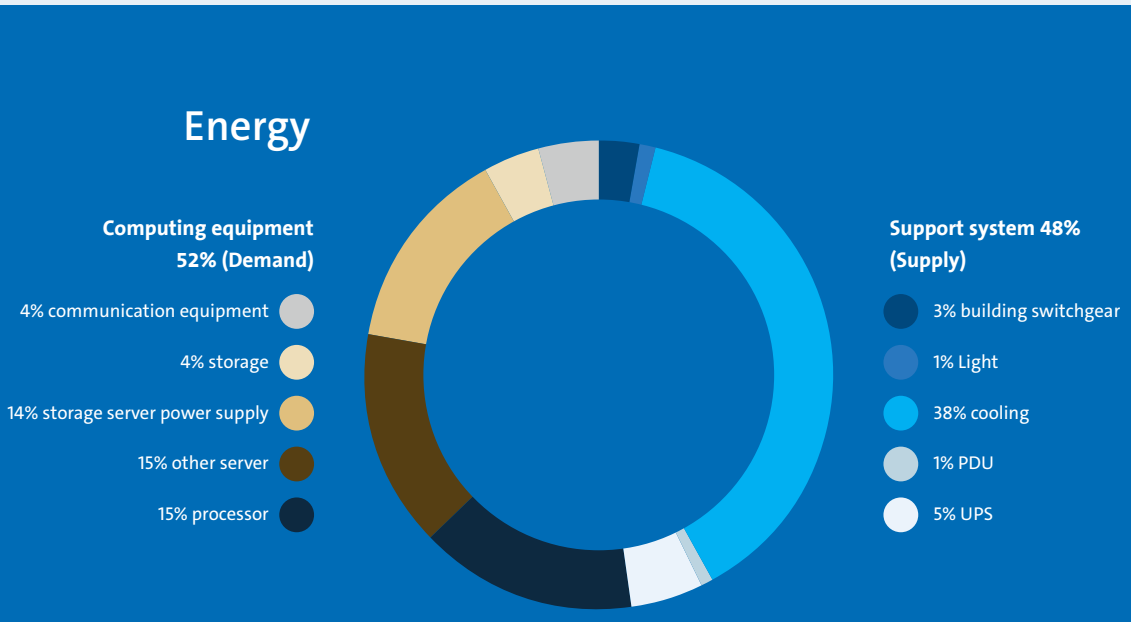



***Figure 1:*** *Energy distribution in data centers*

economically or sustainably efficient manner. In fact, as more high-power racks are deployed, air cooling no longer becomes physically possible.

At the heart of increasing rack densities are the new generation of CPUs and GPUs, which have thermal power density much higher than previous generations. An example is a newly released GPU by Nvidia (DGX H100). Its maximum power consumption is 160% more than that of the company´s previous generation chip[2]. Simultaneously, server manufacturing is packing more CPUs and GPUs inside each rack to avoid high latency by physical distance which otherwise would exist.

Fortunately, extensive work has been done to develop liquid cooling technologies that leverage the higher thermal transfer properties of water or other fluids. Multiple solutions are available today, including rear-door heat exchange, direct-to-chip cold plate, and the emerging technology of immersion cooling. Figure 2 shows a schematic overview of cooling systems and their preference operation area regarding cooling capacity.

### Rear-door heat exchanger

Rear-door heat exchangers can manage rack densities above 20 kW, and don't bring liquid directly in contact with the servers but utilize the thermal properties of liquid. The cooling solution works by installing a heat exchanger on the backside of the rack, extracting the heat from the air when it passes the liquid-filled coils. Systems can be either active or passive, where active heat exchangers include fans to pull air through the coils and remove the heat, enabling higher rack densities compared to passive.

### Direct-to-chip liquid

Direct-to-chip leads liquid directly to the server's high heat generating components (CPUs, GPUs, memory modules), efficiently removing heat through single-phase cold plates or two-phase evaporation units. Direct-to-chip offers better cooling capacity compared to rear-door heat exchanger, but the solution can typically only remove 70-75% of the generated heat, due to difficulties applying cold plates to all components, so a hybrid solution with air cooling is often used.

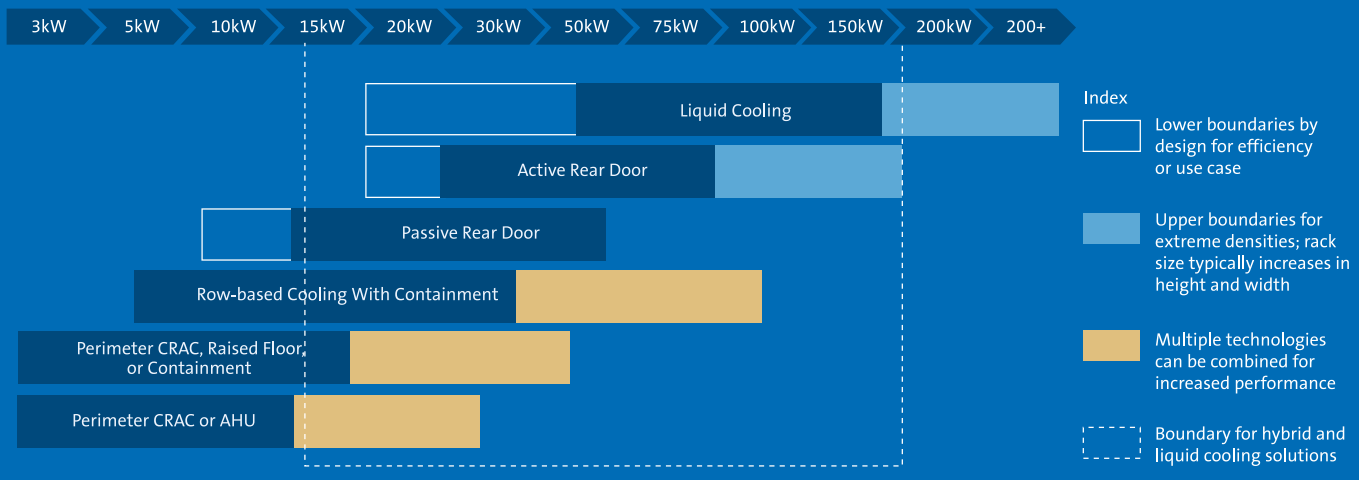
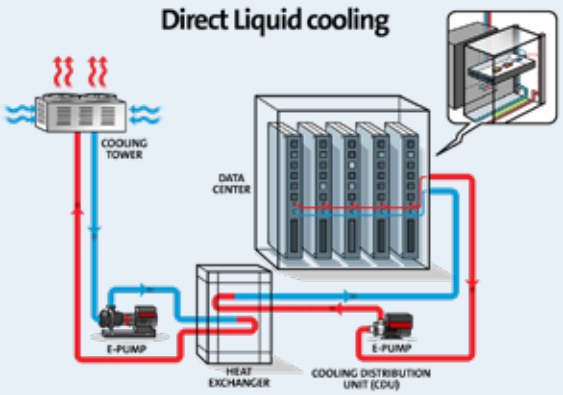
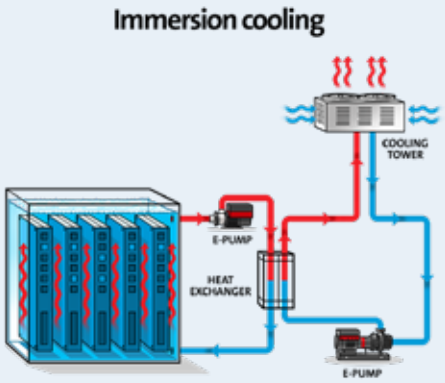



***Figure 2:*** *Air-based cooling is losing its efficiency when a rack density exceeds 20 kW where liquid points become the better solution*[11].
VERTIV whitepaper: Understanding data center liquid cooling options and infrastructure requirements

## Liquid cooling




### Immersion cooling

Immersion cooling offers the highest cooling capacity. The system works by submerging servers directly in a dielectric liquid, which could be natural or synthetic oils, hydrocarbon or liquids based on fluorochemicals. This approach maximizes the thermal transfer of the liquid and is the most energy efficient cooling solution. Like direct-to-chip, immersion cooling systems can be designed as single or two-phased. The two-phased systems use a special formulated dielectric fluid with a low boiling point, allowing the liquid to boil and change phase by the heat generated by the servers. The generated vapor is then condensed by coils positioned in the top of the tank, and gravity pulls the liquid back into the tank for the continuous cycle.

### How does it all work?

Liquid cooling systems have some obvious differences depending on whether direct-to-chip, rear-door heat exchanger, or immersion cooling is chosen. However, there are also many similarities between them. All liquid systems need a dedicated infrastructure to create a fluid cooling loop, enabling the servers to reject heat and a place for the liquid to dispatch the heat.

A central part of all liquid cooling systems is the cooling distribution unit (CDU) which provides controlled cooling liquid for rear-door heat exchangers, direct-to-chip and immersion cooling. Liquid cooling systems typically consist of two loops connected by a heat exchanger in the CDU. The primary loop absorbs the heat from the secondary loop and dispatches the heat either by a cooling towers or fans. The secondary loop is the cycle in contact with the servers, where the flow is controlled by a CDU aiming to balance the flow of the liquid to meet the specific cooling demand.

CDUs come in a broad range of sizes and configurations depending on the overall design, both in-row and in-rack are available.

It is important that the entire cooling system is well configured to optimize the full potential for a given data center. The system should be designed to not only handle the average cooling load, but to also respond to sudden load increases and adapt to increased demand in the future.

> *"Facilities deploying high-density racks (>30 kW) will have to use liquid cooling, as air cooling cannot provide the needed heat removal to maintain IT system reliability. This applies to both edge and core data centers."*
>
> VERTIV whitepaper: Understanding data center liquid cooling options and infrastructure requirements

## The perfect balance between PUE and WUE

Energy efficiency is a critical factor in the operation of data centers, as they consume approximately 2% of the world's energy. For a long time it has been the focus point for data centers, determined to lower the PUE to as close to 1 as possible and there is no doubt, that it's still a target for data centers. It is perhaps also more important now than ever! In many areas of the world, the grid hasn't been designed for the rapid increase in electrification, which can be a serious bottle neck for the continuous industrial development.

However, it is also important not to forget another parameter – water usage effectiveness (WUE). It has for a long time been known that data centers are using tremendous amounts of water, but there has still not been much focus on reporting the actual usage. A data center can require as much as 19 million liters of drinking water per day either directly for cooling or indirectly for power generation[12].

Understanding water use for a given data center is an important part of determining where a data center can be located, as there needs to be a sufficient volume of source water today and in the future. This balancing of supply and demand often puts data centers at conflict with players competing for shrinking available volumes of water. Depending on the location, data centers are in competition with municipalities, farmers, and other industrial users for affordable and reliable water sources.

There are an increasing number of examples of data centers that pump large volumes of water directly from rivers, lakes, or seas. In Toronto, both Cologix and Equinix have attached their facilities to the deep lake water cooling system, which draws on the chilly waters of Lake Ontario. While the upper layer is drawn and used as potable water, the water underneath, which remains at 4°C all year round, is drawn via three 5-kilometer of intake pipes, filtered, and used for cooling.

Another example of alternative water source usage is Google which use seawater at its data center in Hamina, Finland as the primary means of cooling. This data center was an innovative conversion of an existing decommissioned paper mill that had an existing sea water tunnel into the building that was previously used for cooling in the paper manufacturing process. The data center has been such a success that Google is investing $670M to build a second data center in Hamina.

Microsoft announced water sustainability targets in September 2020, pledging to make its operations net water positive by 2030[13]. More so, companies like Google and Amazon are already using a certain level of reclaimed water for cooling their data centers. Reclaimed water can be tertiary effluent wastewater, greywater, black water, and rainwater.

Large data centers are looking to justify their social license to operate and attempt to integrate recycled water and or rainwater for direct use. These emerging alternative sources of water will create new business opportunities for water treatment companies, as well as innovation opportunities for manufacturers of cooling and refrigeration equipment to build their products from materials that can function effectively by using alternative water sources[14].

| Company | Location | Amount of reclaimed water daily (Liters) |
|---|---|---|
| Microsoft | Washington | 11 million |
| Google | Arizona | 4-15 million |
| Amazon | Red Oak Texas | 15 million |
| Google | South Carolina | 19 million |

## Data centers want to be sustainable

Microsoft announced water sustainability targets in September 2020, pledging to make its operations net water positive by 2030. More so, companies like Google, Amazon and Meta are already using a certain level of reclaimed water for cooling their data centers.

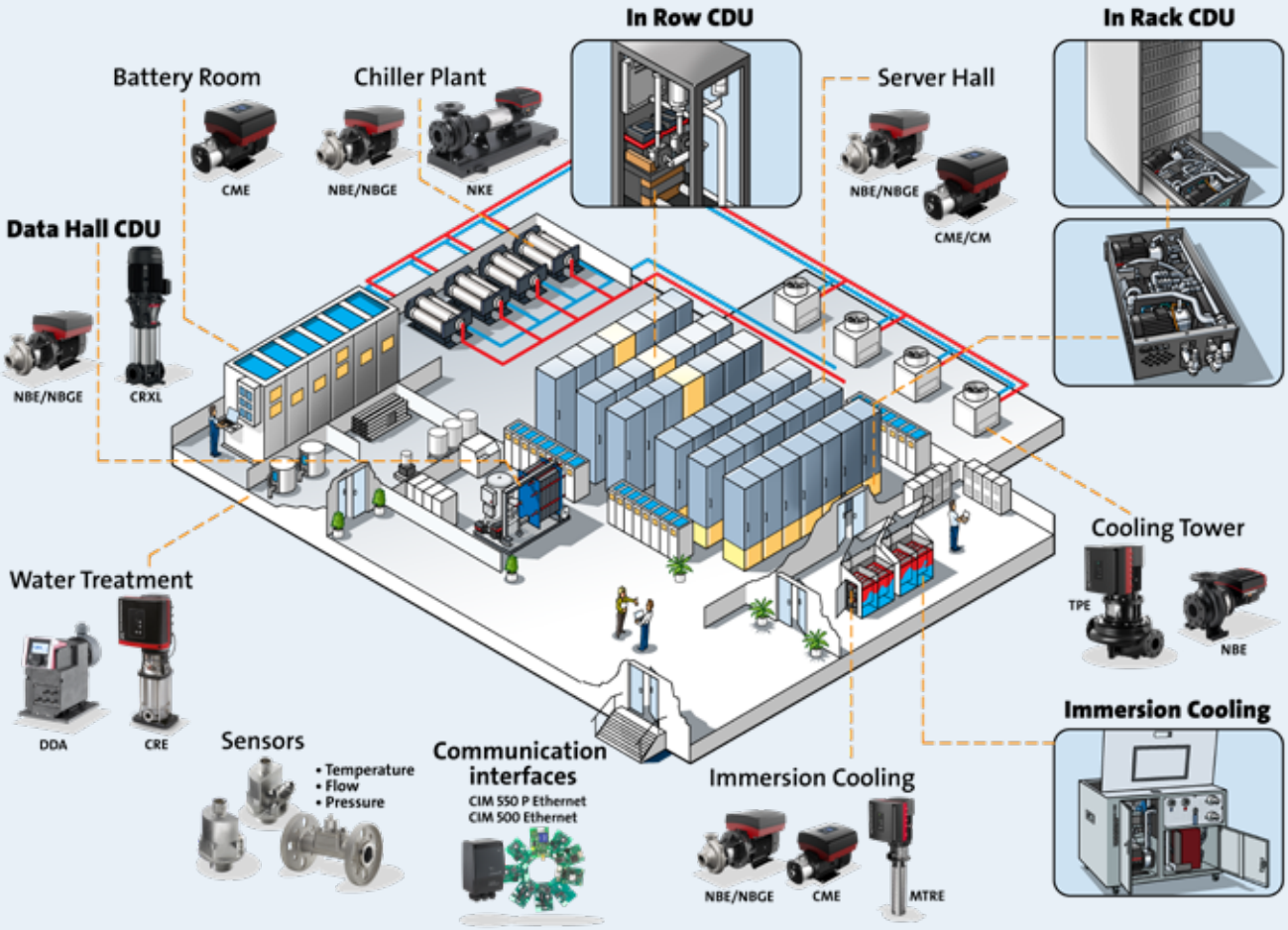



***Figure 3:*** *Example of an overview of a data center architecture*

## Uptime is crucial

Data centers are important for our digital economy. They act as the physical infrastructure, supporting services we rely on every day. It is important that data centers guarantee high reliability and never fail. Standards describing the requirements for data centers are categorized in four tiers defining allowable downtime, system architecture, cooling, and power redundancy.

Tier 4 is the most reliable data center rating, allowing a maximum yearly downtime of 26.3 minutes and ensuring high redundancy, see table 1[15]. When selecting cooling equipment or upgrading to liquid cooling, the reliability of the equipment, as well as sensors to provide the intelligence and detect issues before they happen, become critical parts of the selection process.

| Data center tier rating | Tier 1 | Tier 2 | Tier 3 | Tier 4 |
|---|---|---|---|---|
| Uptime guarantee | 99.671% | 99.741% | 99.982% | 99.995% |
| Downtime per year | <28.8 hours | <22 hours | <1.6 hours | <26.3 minutes |
| Component redundancy | None | Partial N+1 | Full N+1 | 2N or 2N+1 |
| Concurrently maintainable | No | No | Partially | Yes |

## Pumps as the core

Pumps are an essential part of modern data centers and are found in a variety of applications: temperature control, water treatment, wastewater handling and boosting, as illustrated in figure 3. They are a vital part of the journey toward lower PUE and WUE for data centers, as new pump solutions, enabled by intelligent controls, can optimize cooling cycles resulting in up to 70% lower energy consumption.

## Intelligent pumps

Electronically controlled pumps – known as E-pumps – offer a unique potential for data centers to control their cooling system in a better and more efficient way. In traditional systems, balancing valves are used to throttle the flow, leading to back pressure in the system, energy loss and heat generation, all negatively impacting PUE as you can see in figure 4.

**Figure 4: Traditional system**

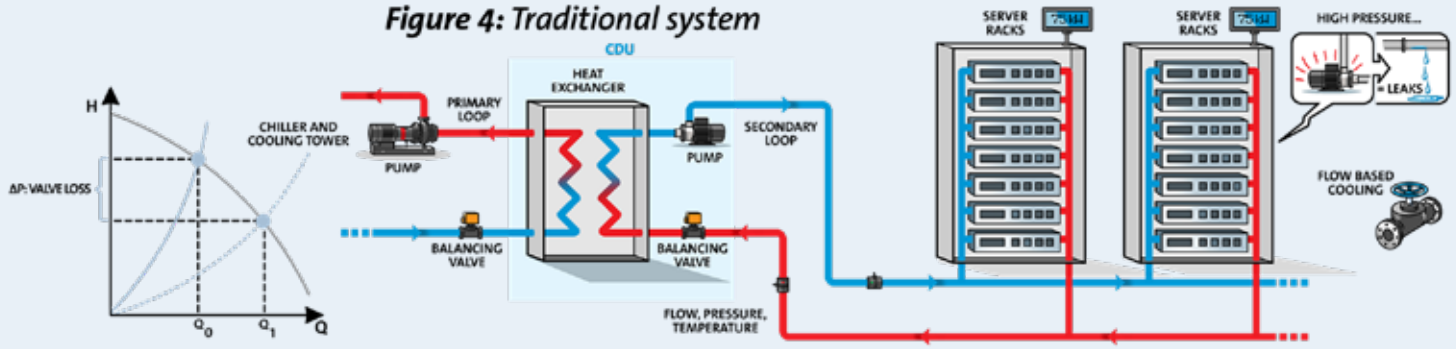
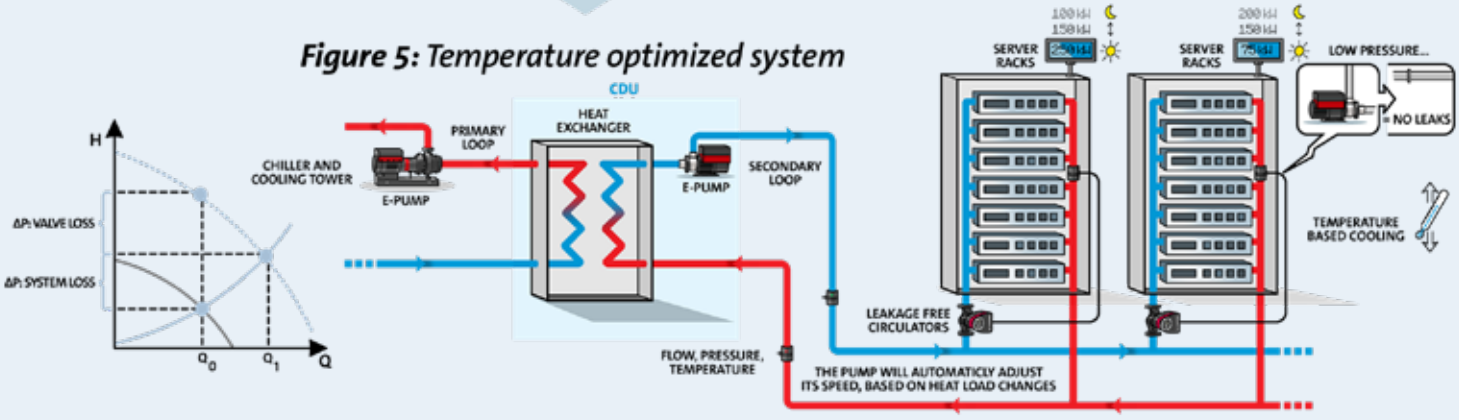



**Figure 5: Temperature optimized system**




*"We need to move away from traditional flow-based control to temperature control where we can vary the flow for each rack based on the demanded cooling load, thereby ensuring that only the flow needed is supplied resulting in energy efficient coolant distribution."*

Jens Mielke – Chief Key Account Manager, Grundfos

Instead, the operators can use variable frequency drive (VFD) motors, as illustrated in figure 5, which are integrated directly into the pump and motor as a single component, increasing intelligence and connectivity, and offering system insights, all while reducing the footprint and number of components required. Motors with an integrated VFD also have an integrated proportional integral derivative (PID) controller that eliminates the need for an external controller.

This allows operators to adjust the speed and flow of the pump to meet the actual cooling load during average operation, but it also enables the system to meet peak loads, ensuring optimal operating conditions. Using VFD and control MPC on both the primary and secondary loop offers the possibility for optimized balance of entire cooling systems bringing down energy consumption and saving more water compared to fixed speed pumps.

Grundfos high efficiency motors with VFDs can take in signals from temperature sensors, pressure sensors or flow meters, for example, and adjust the pump speed to maintain a certain amount of pressure, flow or temperature. With Grundfos direct sensor technology CDUs are able to operate at a constant temperature setting and react to any environmental changes in real time.

“Grundfos offers advanced CIM/CIU modules, integrated with SCADA systems, for centralized monitoring and control of pumps in liquid cooling systems. Combined with Integrated VFD motor monitoring functions, we provide enhanced safety and performance. With duty/standby functionality, our backup pumps ensure uninterrupted operation and maintaining optimal temperature and pressure,” says Peter Ørsted, Application Manager, Grundfos Industry.

Furthermore, E-pumps enable easier commissioning compared to traditional liquid cooling systems because they eliminate tuning and calibration of system valves by using self-adjusting pumps, which can optimize the temperature control based on real-time conditions. By switching from a fixed speed motor to an E-motor with VFD, there are some significant energy savings by balancing the primary and secondary cooling loop thereby optimizing the energy usage.

Additionally, Grundfos E-pumps are rated IE5, or ultra-premium efficiency, one of the highest

international efficiencies (IE) on the market. These IE codes serve as a reference for governments who specify the efficiency levels for their minimum energy performance standards for motors in their respective countries. While the market standards or requirements are at IE3 level, Grundfos offers superior efficiency with IE5 level motors, which exceeds the IE5 efficiency requirements by as much as 2%.

### The right solution for any system

As data centers increase in rack densities, the demands to the cooling system are also changing. Grundfos' broad product range ensures that the right pump can be selected for each individual cooling system, minimize the risk of oversizing and waste of energy.

Furthermore, Grundfos E-pumps enable the motors to operate at higher speed, enabling a smaller and more compact pump to be chosen while maintaining the needed flow and head, this could be a great solution for applications where space is an issue.

The new developments in cooling systems have led to the introduction of new engineering fluids for immersion cooling, which offers unique possibilities. These liquids can be natural/ synthetic ester oils, hydrocarbon fluids or fluorochemical based liquids. The compatibility between the fluid and the components in contact with it are extremely important to ensure minimal risk of unexpected maintenance need. Parameters such as boiling point, vapor pressure, viscosity, density, and material suitability are all important when selecting the right pump and configuring the size, selecting shaft seals and O-ring material.

## Conclusion

The rapid increase seen within energy consumption for data centers, driven by AI advancement and crypto mining has created a pressing need for more efficient cooling solutions. Traditional air-cooling methods are becoming less effective at managing the intense heat generated by high-density racks making the shift to liquid cooling essential.

Liquid cooling not only offers better efficiency leading to less power consumption, but also addresses crucial concerns related to water scarcity. Traditional cooling methods, counterintuitively, often rely heavily on water, exacerbating global water shortages. By adopting liquid cooling, data centers can significantly reduce their water usage, as these systems are designed to be more water efficient by employing a closed loop and recirculating the liquid in the system.

Large advancements have been made within liquid cooling systems over the last couple of years, but there are still improvements to be made. Optimizing cooling systems to operate based on real-time temperature data rather than the traditional flow-based approach is a vital next step. Temperature-based cooling employing state of the art sensor and integrated VFD technology, applying cooling only where and when its needed. This method conserves both energy and water, leading to low operational costs and a more sustainable use of resources.

The transition to liquid cooling and the adoption of temperature-based optimization are crucial steps for data centers to manage rising power and cooling demands. These advancements will enable data centers to operate more efficiently and conserve valuable resources like water thereby contributing to a more sustainable future in and increasingly resource-constrained world.

## References


1. Amount of data created daily (2024) [Link]
2. Forbes | Al, Liquid cooling and the data center of the future [Link]
3. Accenture technology vision 2023 [Link]
4. Vertiv | Liquid cooling options for data centers [Link]
5. CIO | EU moves towards regulating data center energy and water use [Link]
6. Presentation at GBI sustainable data centers seminar from Garrett Helmer, CRO at Cohesion
7. Data Centres and Data Transmission Networks [Link]
8. US data center electricity demand could double by 2030, driven by artificial intelligence: EPRI [Link]
9. McKinsey & Company bluefield research
10. Lenovo StoryHub | The world's AI generators: rethinking water usage in data centers to build a more sustainable future [Link]
11. Vertiv whitepaper: Vertiv liquid cooling [Link]
12. Data centers are draining resources in water-stressed communities [Link]
13. Sustainable by design: Transforming datacenter water efficiency [Link]
14. BlueTech research: Water & data centers
15. PhoenixNAP | Data center tiers explained [Link]


## Table of abbreviations and acronyms

| | |
|---|---|
| OPEX | Operating expense |
| DC | Data center |
| EU | European union |
| PUE | Power usage effectiveness |
| WUE | Water usage effectiveness |
| CPU | Central processing units |
| GPU | Graphical processing units |
| CDU | Cooling distribution unit |
| CRAC | Computer room air conditioning |
| AHU | Air handling unit |
| PDU | Power distribution unit |
| UPS | Uninterruptible power supply |
| VFD | Variable frequency drive |
| IE | International efficiencies |
| PID | Proportional integral derivative |
| TUE | Total usage effectiveness |

**Grundfos Holding A/S**
Poul Due Jensens Vej 7
DK-8850 Bjerringbro
Tel: +45 87 50 14 00
www.grundfos.com

GRUNDFOS