Green and resilient now:

Report on fluid systems enabling water–energy coordination for computing infrastructure in the AI era

September 2026

格兰富 GRUNDFOS | CAICT 中国信通院

Preface

AI is devouring power and water at an unprecedented rate. As computing density soars, infrastructure is no longer just scaling up - it is being forced to reinvent itself. The central question is no longer how much compute can be delivered, but how sustainably. High-density AI clusters are pushing power and water systems to their limits, while liquid cooling is fundamentally reshaping cooling architecture. The imperative is clear: boost resource efficiency and operational resilience, without sacrificing compute capacity. This is the new frontier for green infrastructure.

Across the world, the boom in AI is turning data centres into a test of resource capacity as much as technological ambition. In America, Goldman Sachs and others expect data centre power demand to rise from 41 GW in 2026 to 66 GW in 2027, while additional water-supply capacity of 697-1,450 million gallons a day (MGD) may be needed by 2030. Europe faces the same squeeze: its data centres are estimated to have consumed about 76 billion litres of water in 2025, and indicators such as water-usage effectiveness (WUE) are becoming part of sustainability assessments. China, meanwhile, is expanding fast: by the end of June 2026 its intelligent-computing capacity had reached 2,185 EFLOPS (FP16), with national computing facilities operating at an overall utilisation rate of 71.4%. The figures differ by market, but the underlying challenge is global: AI infrastructure is running up against tighter constraints in energy, water, land and grid capacity.

This report argues that the next priority for AI-era computing infrastructure has therefore shifted from the scale of computing capacity alone, to the ability to manage the resources that make such capacity possible. Supporting greater growth in computing capacity with lower resource consumption has become a shared challenge for the industry. As rack power density rises and liquid cooling spreads, thermal management, water use, energy efficiency and system reliability can no longer be treated as separate engineering issues. A growing number of countries and regions are incorporating coordinated resource use into sustainability assessments for computing centres.

Against this background, the importance of fluid systems is being reassessed. In the past, fluid systems primarily ensured heat removal. In the AI era, however, they are becoming an important hub connecting computing equipment, cooling systems, energy systems and water-resource systems. How heat is collected, transported, regulated, used and recovered affects not only cooling efficiency, but also energy use, water consumption and long-term system operation. Fluid systems are gradually evolving into key connecting infrastructure for green computing.

The implication therefore is a change in mindset. Water–energy coordination cannot be achieved by improving one product at a time. It requires the industry to look beyond standalone equipment and manage the fluid system as an integrated chain - from cooling architecture and distribution networks to pumps, CDUs, controls and digital platforms. This report sets out a three-layer framework across architecture, equipment and platform and identifies seven critical technologies that can turn that approach into practice: distributed demand-driven fluid distribution, leak-risk management, waste-heat recovery, intelligent pump coordination, multi-layer resilience, data-driven closed-loop control and lifecycle fluid management. Together, they offer a practical route to higher efficiency, stronger operational resilience and more intelligent use of resources.

Overall, this report seeks to encourage greater industry attention to the foundational role of fluid systems in green computing and to move computing infrastructure from isolated energy-saving measures to system-level optimisation, from resource consumption to coordinated resource use and from experience-based maintenance to intelligence-powered resilient operations. Through collaborative innovation across the industry value chain, we look forward to exploring more efficient, reliable and sustainable pathways for computing-infrastructure development and supporting the high-quality development of green computing in the AI era.

The AI era is still being written. As computing infrastructure and fluid-system technologies continue to evolve, so will the ideas and practices that support sustainable growth. We welcome feedback and perspectives from industry experts, partners and readers and look forward to advancing this dialogue as we collectively shape the future of green computing.

Contents

List of figures

List of tables

Chapter 1

The water–energy challenge for AI-era computing infrastructure

01

1.1 Resource constraints are reshaping global computing infrastructure

U.S.

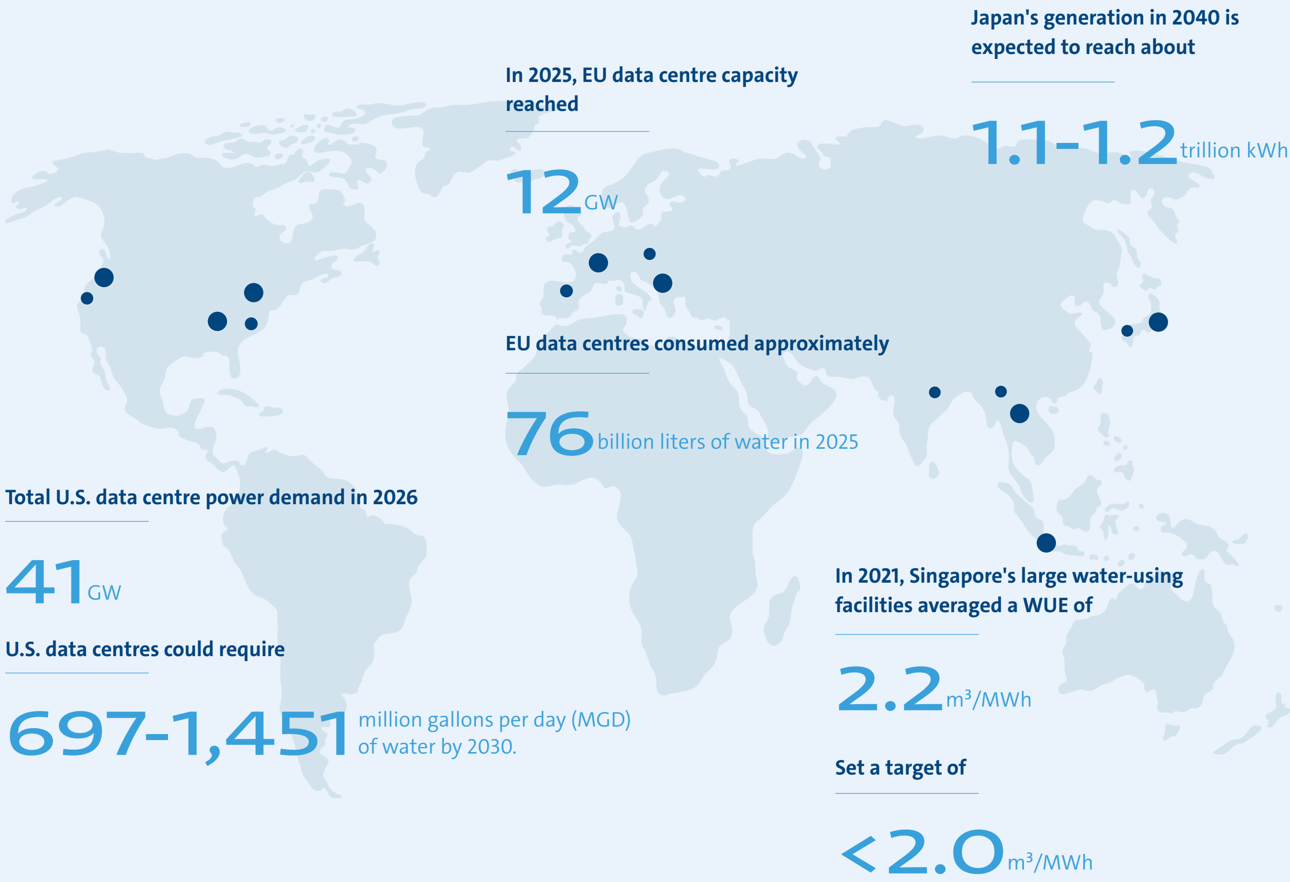
The rapid expansion of U.S. AI data centres is intensifying pressure on regional power systems, water supply and resource governance. Goldman Sachs estimates that total U.S. data centre power demand will rise from 31 GW in 2025 to 41 GW in 2026 and then surge to 66 GW in 2027, an increase of approximately 112% from 2025. Pressure is therefore rising across grid expansion, power delivery and energy-efficiency management. Alongside this rapid increase in electricity load, the cooling demand of high-density AI computing facilities is creating severe pressure on water consumption, making water–energy coupling constraints increasingly prominent. The 2026 study *Small Bottle, Big Pipe: Quantifying and Addressing the Impact of Data Centres on Public Water Systems,* produced by researchers from the California Institute of Technology (Caltech), the University of California, Riverside (UCR) and other institutions, concludes that if 2024 water-use intensity persists, U.S. data centres could require an additional 697-1,451 MGD of water-supply capacity by 2030 to meet growing cooling demand. Large data centres using evaporative cooling may require substantial amounts of water during hot periods: peak daily use at some facilities exceeds one million gallons, while some planned large facilities could require as much as eight MGD. In industry practice, leading cloud-computing companies such as Microsoft and AWS have established joint PUE/WUE operational monitoring, substituted reclaimed water for municipal freshwater, adopted innovations such as direct-to-chip and microfluidic cooling and set Water Positive targets for 2030. These companies increasingly treat water-use performance and energy efficiency as parallel core KPIs, embedding water–energy coordination into long-term data centre operations.

Europe

In Europe, the rapid expansion of computing capacity is forcing energy and water efficiency into the same sustainability framework. Data centres underpin Europe's digital services and are enabling rapid advances in computing and AI, yet their substantial energy use could slow Europe's clean-energy transition and increase electricity costs by straining the grid. EU data centre capacity is expected to more than double over the next few years, from 12 GW in 2025 to 28 GW in 2030. Data centres currently account for approximately 3% of EU electricity demand, a share that could rise to 7%–9% by 2030 and place greater strain on power systems. In terms of water consumption, expansion of the computing industry is steadily increasing cooling-water demand. The European Commission’s 2026 *Impact Assessment Report for the Cloud and AI Development Act* estimates that data centres across the EU consumed approximately 76 billion litres of water in 2025 Cooling systems at large data centres may require several million litres of freshwater per day, with actual use determined directly by cooling method, climate and regional water conditions. The European Commission's 2026 Country Report - Ireland shows that large energy users, primarily data centres, increased their share of Ireland's metered electricity consumption from 14.8% in 2015 to 31.0% in 2024, with demand concentrated in the Dublin area. Ireland also faces water-infrastructure constraints, including insufficient wastewater-treatment capacity and high leakage rates in water-supply networks, further limiting room for computing expansion. To address fragmented water and energy management, the EU continues to improve its overarching institutional framework. *Under the Energy Efficiency Directive (Directive (EU) 2023/1791),* a 2025 assessment report from the European Commission's Directorate-General for Energy further refined a unified framework for data centre sustainability assessment and disclosure. The framework places WUE alongside traditional metrics such as PUE in routine public disclosure and requires systematic consideration of grid capacity, regional water endowment, equipment efficiency and other key factors throughout data centre planning, design and operation, promoting coordinated optimisation of efficient energy use and water conservation in computing infrastructure.

Japan and Southeast Asia

In Japan and Southeast Asia, data centre development is increasingly shaped by local resource endowments, infrastructure capacity and green data centre requirements. Japan's Agency for Natural Resources and Energy stated in the 2025 *Seventh Strategic Energy Plan* that digital and green transformation will continue to increase electricity demand; national generation in 2040 is expected to reach about 1.1–1.2 trillion kWh. The plan calls for more renewable energy and stronger energy-efficiency measures to meet this demand.[1] Land and water-constrained Singapore has made resource efficiency a core requirement for green data centres. Its Infocomm Media Development Authority reported in the 2024 Green Data Centre Roadmap that large water-using facilities averaged WUE of 2.2 m^3/MWh in 2021 and set a target of 2.0 m^3/MWh or less. The roadmap also identifies cooling towers as the source of about 97% of data centre water use, making cooling systems optimisation a priority.[2] Malaysia's *2024 Guidelines for Sustainable Development of Data Centres* similarly require higher water efficiency in design and operation and encourage sustainable cooling solutions.[3]



1. "The 7th strategic energy plan", agency of natural resources and energy, 2025
2. "Green data centre roadmap", IMDA, 2024
3. "Guideline for sustainable development of data centre", MIDA, 2024

1.2 Status of green development of China's computing infrastructure

China continues to expand its computing infrastructure and resource base. By the end of June 2026, China’s intelligent computing capacity had reached 2,185 EFLOPS (FP16), while the nationwide rack deployment rate for computing facilities stood at 71.4%.

China’s intelligent computing capacity had reached

2,185 EFLOPS(FP16)

Nationwide rack deployment rate for computing facilities stood at

71.4%

By the end of June 2026

China has introduced a series of policies that make green and low-carbon development a core requirement for high-quality computing infrastructure. Before 2020, policy focused mainly on data centre planning, siting and layout. *The Guiding Opinions on Strengthening Green Data Centre Development and the Three-Year Action Plan for New Data Centres (2021–2023)* subsequently called for higher standards in new facilities, greener operation and retrofits of existing sites and faster innovation and adoption of green technologies. The national carbon-peaking and carbon-neutrality strategy, together with the Eastern Data, Western Computing initiative, then broadened the policy foundation for green development across the computing value chain. Later measures - including the *Action Plan for High-Quality Development of Computing Infrastructure and the Special Action Plan for Green and Low-Carbon Development of Data Centres* - set requirements for energy and carbon efficiency, computing–power coordination and green electricity use at national hub nodes. In 2026, *the Government Work Report, the Action Plan for Bidirectional Empowerment of AI and Energy and the 15th Five-Year Plan Carbon-Peaking Action Plan* identified AI, large AI-computing clusters and computing–power coordination as new infrastructure priorities and strengthened requirements for green electricity consumption and carbon accounting.

Table 1 Policies related to green computing in recent years

Publication date	Policy document	Key provisions
October 2023	Action Plan for High-Quality Development of Computing Infrastructure	Calls for stronger computing supply, transport, storage, industry enablement, green development and security. By 2025, total computing capacity is to exceed 300 EFLOPS, with AI computing accounting for 35%.
December 2023	Implementation Opinions on the Eastern Data, Western Computing Project and the National Integrated Computing Network	Coordinates general, intelligent and supercomputing capacity across eastern, central and western China and integrates computing with green electricity. By the end of 2025, a bidirectional computing–power coordination mechanism is to be established, with green electricity exceeding 80% at new data centres in national hub nodes.
July 2024	Special Action Plan for Green and Low-Carbon Development of Data Centres	Targets an overall rack utilisation rate of at least 60%, average PUE below 1.5 and 10% annual growth in renewable-energy use by the end of 2025. By 2030, national average PUE and energy and carbon efficiency per unit of computing are to reach advanced international levels.
August 2024	Action Plan to Accelerate Construction of a New Power System (2024–2027)	Promotes computing–power coordination projects, joint planning of computing and electricity infrastructure, integration of generation, load and storage and aggregated green-electricity supply through nearby generation, trading and on-site consumption.
March 2025	Opinions on High-Quality Development of the Renewable-Energy Green Electricity Certificate Market	Advances mandatory green-certificate consumption and accounting and raises the share of green electricity used by major energy consumers, including data centres. New facilities at national hub nodes are to exceed the existing 80% requirement.
August 2025	Opinions on Deepening the AI Plus Initiative	Optimises national AI-computing resources and the integrated computing network; strengthens coordination among data, computing, electricity and networks; and promotes accessible, cost-effective, green and secure AI-computing supply.
March 2026	Government Work Report	Expands AI Plus, develops new infrastructure such as hyperscale AI-computing clusters and computing–power coordination, strengthens national monitoring and dispatch and supports public-cloud development.
April 2026	Action Plan for Bidirectional Empowerment of AI and Energy	Coordinates energy, computing, applications, data and models; strengthens energy support for AI; and uses AI to accelerate an efficient, reliable, green, low-carbon, open and integrated energy system.
July 2026	15th Five-Year Plan Carbon-Peaking Action Plan	Coordinates computing facilities with renewable energy, supports direct green-electricity supply and storage and promotes non-fossil electricity for new facilities. It also strengthens efficiency and carbon standards, upgrades noncompliant sites, phases out inefficient technology and encourages green and zero-carbon computing facilities.

China's green-computing industry is developing a coordinated value chain spanning facilities, equipment, platforms and applications.

At the facility level, national computing hubs increasingly follow a "computing follows green electricity" model, using wind and solar resources in integrated source-grid-load-storage projects and applying free cooling and waste-heat recovery to reduce energy use and emissions.

At the equipment level, high-voltage DC power supplies, energy storage, smart pumps, high-density efficient servers and low-loss network equipment continue to improve through materials, products and process innovation.

At the platform level, computing virtualisation, pooled scheduling, AI-assisted optimisation and end-to-end energy monitoring enable dynamic allocation and reduce redundant capacity.

At the application level, green computing supports industrial decarbonisation, smart cities and scientific research. Together, these developments are creating an ecosystem that extends from hardware and platform software to algorithms and industry applications.

The xUE framework for computing infrastructure is becoming more comprehensive, and both energy and water performance are receiving greater attention. Power Usage Effectiveness (PUE), introduced in 2007, remains the principal energy-efficiency metric; the closer PUE is to 1.0, the more efficiently a data centre utilises energy. Water Usage Effectiveness (WUE), Carbon Usage Effectiveness (CUE), Infrastructure Usage Effectiveness (IUE) and Space Usage Effectiveness (SUE) broaden the assessment to water, carbon, infrastructure and space. As green computing develops, the industry is moving beyond electricity savings alone towards integrated water–energy performance. Indirect evaporative cooling, liquid cooling and water reuse can lower PUE while controlling water demand. Leading green data centres in China report PUE below 1.2 and WUE below 0.7 L/kWh. Routine monitoring of both metrics, combined with intelligent operations and process optimisation, can help manage the trade-off between energy and water use and support a more integrated low-carbon operating model.

China's leading green data centres deliver strong performance on both PUE and WUE.

PUE Management	WUE reduced to
<1.2	<0.7 L/kWh

1.3 Reconfiguring cooling systems for high-density computing

High-density AI loads are turning cooling capacity into a binding constraint on computing deployment, while accelerating the shift from air cooling to single-phase and two-phase liquid cooling. A 2025 global survey found an average rack density of about 9 kW, and 82% of respondents reported maximum rack densities below 30 kW. Above 60 kW, however, HPC and AI become the dominant workloads, and some racks exceed 100 kW, revealing a widening density gap between conventional facilities and leading AI clusters[4]. Rack-scale systems such as NVIDIA NVL72 integrate up to 72 accelerators, with individual accelerator thermal design power reaching 1,000 W. Two-phase direct-liquid-cooling trials have demonstrated 2,200 W per socket, heat flux near 250 W/cm² and support for 1U[5] servers above 10 kW. Hybrid air–liquid cooling is therefore well suited as a transitional solution for retrofitting existing data centres, while single-phase direct-to-chip liquid cooling is expected to remain the dominant cooling technology for AI racks over the coming years. Two-phase direct-to-chip liquid cooling, meanwhile, is positioned to support chips with even higher heat flux densities.

Cooling-system transformation requires more than replacing terminal equipment; it requires system-level thermal-fluid management across chips, racks, coolant distribution units (CDUs), piping and outdoor cooling sources. PNNL, ASHRAE and NEMA recommend coordinating energy and water efficiency throughout planning, design, commissioning, operation and retrofit, with technology selected for the climate and load density[6]. Reliability must also extend beyond cold plates to piping cleanliness, water quality and commissioning. Single-phase cold-plate channels can be about 100 μm wide and vulnerable to construction debris and corrosion products. OCP therefore requires pressure testing, cleaning and flushing before operation, with at least eight system volumes circulated and final turbidity below 5 NTU[7].

The next stage of high-density cooling will emphasise load sensing, dynamic demand-driven coolant delivery and coordinated system control rather than nominal cooling capacity alone. A digital-twin study using 47,186 annual operating records from the Frontier supercomputer reported cooling energy reductions of 20.4% from flow optimisation alone and 30.1% from joint optimisation of flow and supply temperature. With practical limits on pump and actuator ramp rates, modelled savings still remained 27.8% overall and 33%[8] in summer. Liquid cooling is therefore not inherently efficient; efficiency depends on pumps, CDUs, heat exchangers and heat-rejection equipment operating as a coordinated closed loop. NVIDIA's roadmap for IT racks of 1 MW and above in 2027 likewise points to integrated, upfront design of cooling, power and building systems.[9]

In 2025, single-rack density has surpassed

100 kW

Frontier's digital-twin study: optimizing flow and supply temperature cuts energy use by

30.1%

4. 《Uptime institute global data centre survey 2025》, july 2025
5. 《Pumped two-phase refrigerant-based direct liquid cooling technology for next generation AI clusters with high TDP accelerators》, february 14, 2025
6. 《PNNL/ASHRAE/NEMA AI data centre energy performance framework》, 2026
7. 《Guidelines for pre-commission preparation of technology cooling system row manifolds in liquid cooled data centres》, january 27, 2025
8. 《Digital twin-based cooling system optimisation for data centre》, march 3, 2026
9. 《NVIDIA 800 VDC architecture will power the next generation of AI factories》, may 20, 2025

1.4 Current industry challenges

As AI clusters scale from kilowatts to hundreds of megawatts, energy efficiency is becoming a system-level bottleneck.

Training workloads may involve thousands of GPUs for weeks or months, and some hyperscale AI centres consume billions of kilowatt-hours annually. Continued growth in AI, cloud computing, big data and the industrial internet will place sustained pressure on regional energy supplies and grids. Demand and utilisation are concentrated in eastern China, where energy quotas and grid capacity are tightening, while western regions have abundant clean energy but face utilisation and latency constraints. By June 2025, the national average PUE for computing infrastructure had fallen to 1.42; some liquid-cooled high-density facilities report project-level PUE below 1.2. The marginal benefit of conventional measures—variable-speed cooling equipment, hot/cold aisle containment and free cooling—is narrowing. Further gains require deeper changes in cooling systems, IT efficiency and workload scheduling, coordinated across chips, servers, materials and controls. Diverse climates, building conditions and load profiles make standardised, scalable solutions difficult in the near term.

Green retrofits involve high capital costs and long payback periods.

Liquid cooling, energy storage, waste-heat recovery, efficient power systems and intelligent monitoring usually require higher upfront investment than conventional air-cooling and power infrastructure. They also place greater demands on facility design and operations, including airflow, floor loading, piping, leak protection and technical skills. Older facilities may also face limits in ceiling height, structural capacity, layout, electrical capacity and fire safety codes. Retrofitting can require extensive building and MEP work, equipment relocation, new pumping and coolant systems and planned downtime, increasing both direct cost and operational risk. Although green computing offers environmental, social and reputational value, stable green premiums and differentiated pricing remain limited. Budget-constrained operators may prioritise capacity expansion, while smaller facilities face constrained financing, uneven policy support, technology uncertainty and commercial risk.

National average PUE for computing infrastructure	Some liquid-cooled high-density facilities report project-level PUE	Marginal improvement in PUE from conventional measures is
↓ 1.42	<1.2	Narrowing

*By the end of June 2026

Lifecycle management remains inadequate.

Industry attention is still concentrated on operational PUE, while emissions and environmental impacts from equipment manufacture, transport, installation, retirement and recycling receive less scrutiny. Advanced chips, servers, storage, network equipment, semiconductor materials and critical metal extraction are energy- and emissions-intensive, yet embodied carbon is not fully incorporated into mainstream accounting.

Water management is also uneven. Medium- and large-sized data centres that employ evaporative cooling and water-cooling technologies can intensify pressure in water-stressed regions, while many sites still lack robust reuse, filtration, purification and WUE management. Scaling, low wastewater recovery and evaporative losses reduce efficiency. Because water and electricity are coupled, hot weather can increase both intake and discharge, further widening local supply-demand gaps as computing capacity expands.

Chapter 2

How fluid systems enable green computing infrastructure

02

Like the vascular network of the human body, fluid systems deliver cooling where it is needed across computing infrastructure. As rack densities rise and liquid cooling expands, this network is becoming central to how data centres manage heat, water, energy and operational risk.

In a data centre, the fluid system collects, transports, exchanges and rejects heat through coordinated piping, pumps, heat exchangers, treatment equipment and controls. It manages water, refrigerants, coolants and airflows while regulating temperature, humidity, cleanliness, fluid quality and water use. Typical subsystems include cooling sources, hydronic distribution, terminal heat exchangers, thermal storage, water treatment, heat recovery, monitoring and controls. Together they form the delivery backbone of thermal management and a primary target for PUE improvement.

2.1 Fluid systems connect thermal management and water stewardship

Fluid systems, particularly liquid cooling technologies, are reshaping data centre thermal management by bringing coolant into the server or close to chips and GPUs. Liquids carry far more heat than air at a smaller temperature difference, shortening the heat-transfer path and reducing fan power. Cold-plate liquid cooling has already been deployed at scale in 30–50 kW rack environments, while immersion cooling is gaining traction in higher-density computing clusters. Some new AI data centres report PUE below 1.15. These systems support rack-density increases from 10 kW to more than 100 kW while reducing cooling energy per unit of computing. Phase-change cooling, heat pipes and microchannel cold plates extend the range of chip, server and rack-level solutions.

Fluid systems are becoming a critical enabler of water stewardship in computing infrastructure, improving water efficiency through closed-loop circulation, advanced water-quality management and reclaimed-water utilisation. To address the high evaporative losses of open cooling towers, the industry is promoting closed-loop cooling systems, hybrid dry-wet cooling and dry coolers. These technologies reduce direct air-water contact and increase cycles of concentration, substantially lowering both makeup-water demand and wastewater discharge. Hybrid dry-wet cooling, for example, relies primarily on dry coolers during dry periods and activates evaporative cooling only under high-temperature, high-humidity conditions, reducing evaporation by more than 50% without compromising cooling performance. Water softening, filtration, scale control, disinfection and online monitoring extend equipment life and preserve heat-transfer efficiency, avoiding the increase in energy use caused by scaling and corrosion. In water-scarce regions, some data centres are piloting the use of deeply treated reclaimed water for cooling-tower makeup and landscape irrigation, reducing dependence on municipal water supply. At the system level, technology choices should also account for water use embedded in electricity generation: a configuration that lowers direct water use may increase energy demand and shift part of the water footprint upstream.

Table 2 Comparison of energy and water characteristics of typical heat-rejection methods

Heat-rejection method	Energy use	Water use	Primary constraint	Water–energy role
Open cooling tower	Low (latent-heat transfer; close approach)	High (evaporation and blowdown)	Wet-bulb temperature	Baseline option where energy use has priority
Closed-circuit cooling tower	Relatively low	Medium (spray evaporation; closed internal loop)	Wet-bulb temperature; larger temperature approach	Compromise where water quality is critical
Dry cooler	High (fan power and sensible-heat transfer)	≈0 on site	Dry-bulb temperature; reduced capacity in hot weather	Baseline option in water-stressed regions
Adiabatic dry cooler	Medium to high (spray assist in hot weather)	Low (water used only during hot periods)	Improved hot-weather capacity but some makeup water required	Peak cooling with low annual water use
Hybrid dry–wet system	Dispatchable (switches between dry and evaporative modes)	Dispatchable	Greater integration and control complexity	Control lever for the water–energy trade-off

WUE is emerging as a critical design and operational metric, bringing water-use considerations into decisions on data centre siting, cooling strategies and infrastructure architecture. Measures such as cooling-water leak detection, optimisation of cooling-tower drift eliminators and spray-cooling retrofits can further reduce water consumption without altering overall system architecture. By optimising the water side of fluid systems, data centres can maintain cooling performance while reducing water demand, helping to balance the rapid expansion of computing capacity with regional water-resource constraints.

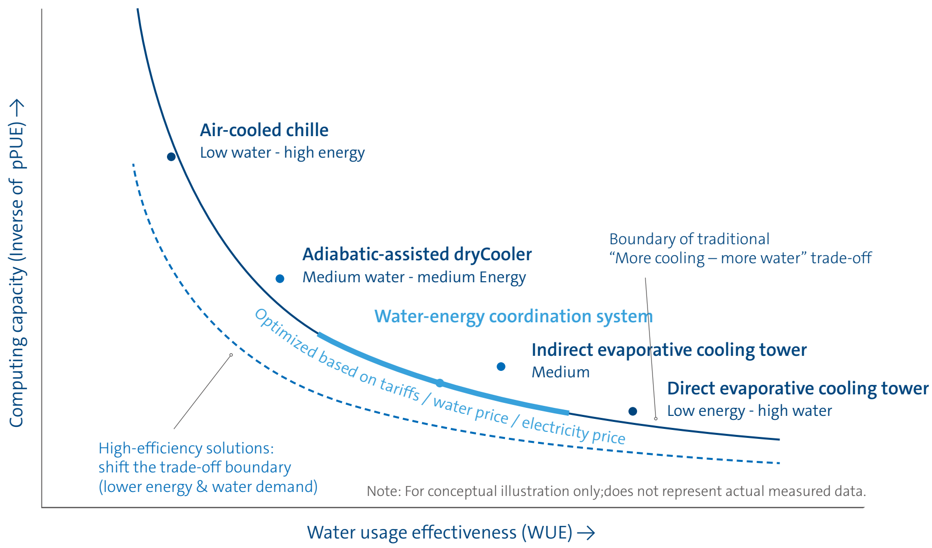


Figure 1 Energy–water trade-offs across heat-rejection methods

AI and digital twins are shifting fluid-system management from experience-based operation to data-driven optimisation. Sensors for temperature, pressure, flow, liquid level and energy use can create real-time models of chillers, pumps, cooling towers, piping and terminals. These models help operators identify hydraulic imbalance, overcooling, hot spots and inefficient control strategies.

Where sensing, data quality and control capability are sufficient, coordinated optimisation may reduce cooling energy by roughly 10%–20%, subject to project-specific validation. Anomaly detection can also identify pump vibration, leaks, heat-exchanger fouling and valve faults before they cause disruption. When combined with weather, electricity-market and workload data, these systems can optimise supply and return temperatures on an hourly basis and thermal-storage schedules on a daily basis, enabling more proactive thermal management.

2.2 Fluid systems improve energy efficiency and flexibility in high-density computing

Fluid systems improve energy efficiency and flexibility by matching cooling delivery to real-time IT demand. Conventional cooling plants are typically designed for peak load and often operate with excess capacity. By contrast, variable-speed pumps, modulating valves, terminal units, CDUs and variable-speed cooling towers can continuously adjust flow, differential pressure and supply temperature in response to real-time heat load. When AI training workloads ramp up, the system can proactively raise CDU flow and operating margin; when demand subsides, pump speeds and heat-rejection capacity can be reduced accordingly. This turns cooling from a fixed-capacity utility into a dispatchable resource: cooling delivery follows IT demand, overprovisioning and idle operation decline and low PUE can be maintained under highly dynamic workloads.

Optimising the heat-transfer chain allows more computing capacity within the same footprint. Coordinated planning of rack layouts, floor loading, piping, CDU capacity, leak protection, power distribution and network architecture brings cooling closer to heat sources at the chip, server and rack levels. High-density liquid-cooled racks can be deployed alongside modular data halls, prefabricated power systems and dense network architectures, accelerating construction and enabling phased expansion. The result is greater computing output from the same footprint, requiring fewer racks and less supporting infrastructure for a given workload. This can reduce civil work, cable losses and auxiliary energy consumption while supporting larger AI clusters within power and space constraints.

2.3 Fluid systems reconfigure computing infrastructure for water–energy coordination and operational resilience

Fluid systems integrate four interconnected functions—cooling source, distribution, thermal storage and control—to optimise the architecture of computing infrastructure. On the source side, cooling strategies can be aligned with local climate, energy mix, policy environment and water availability. Technologies such as dry coolers, plate heat exchangers, indirect evaporative cooling and adiabatic humidification can harness ambient air, groundwater, lake water and seawater as free-cooling resources, extending natural cooling hours and reducing mechanical refrigeration demand. On the distribution side, ring-main networks, zonal architectures, staged pumping systems, variable-speed pumps and dynamic balancing valves improve hydraulic performance and enable demand-based cooling delivery. In large AI campuses, distribution networks must also coordinate piping systems, utility corridors, CDU deployment and hot-aisle/cold-aisle designs to deliver cooling efficiently and reliably to every computing node. On the storage side, chilled-water storage, ice storage, phase-change materials and large-temperature-differential thermal storage can support peak shaving and load shifting. Future integration with advanced electrical energy storage may further improve renewable-energy utilisation and operational flexibility. On the control side, fluid systems are evolving from equipment-level control to integrated orchestration of computing and energy resources. This requires a sensing and actuation layer, a data and analytics layer and an optimisation and decision-making layer, together enabling coordinated management of cooling, energy and computing resources across the facility.

As computing chips push thermal limits higher, fluid systems are extending from facility-level cooling infrastructure to chip-level thermal management. Spanning the entire cooling chain, they operate at the intersection of water and energy, where heat extraction, distribution efficiency and heat-rejection strategies directly influence computing density, energy intensity and water consumption. As such, fluid systems have become a critical enabler of water–energy coordination in modern computing infrastructure.

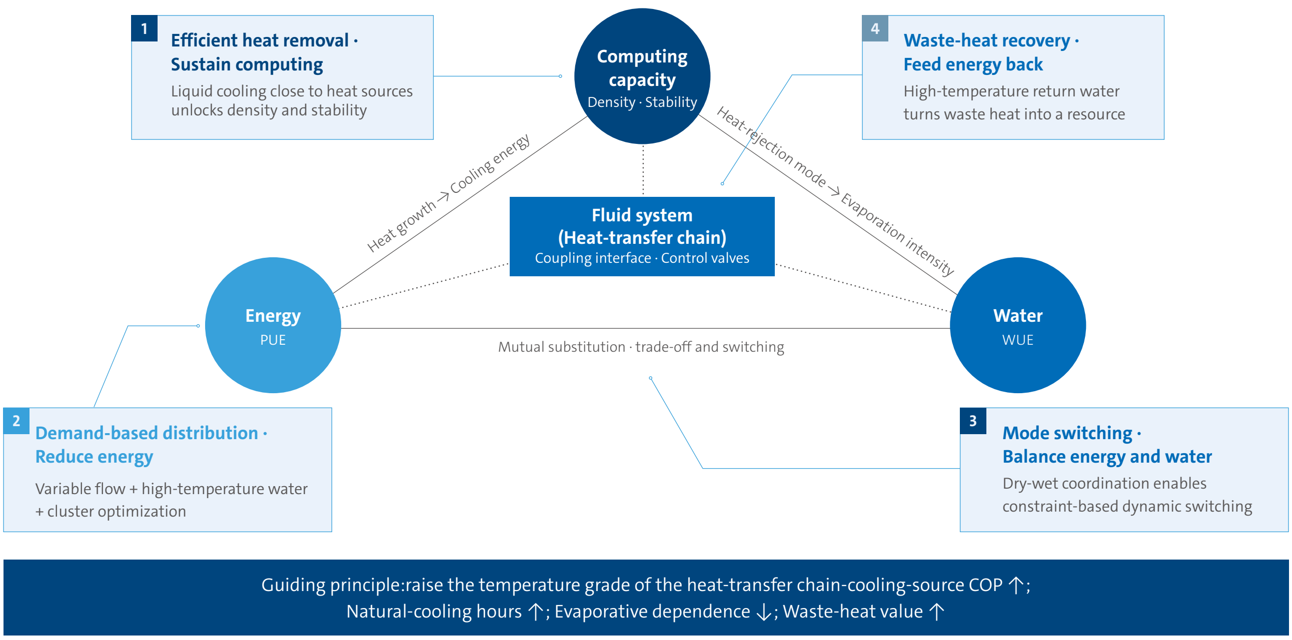


Figure 2 Computing capacity and water–energy coordination

As AI data centre architectures evolve, fluid systems are continuously improving heat-transfer efficiency and therefore reducing the cooling energy required per unit of computing capacity. As a result, the marginal energy required for additional computing capacity continues to decline, helping computing growth decouple from resource consumption. At the same time, advances in high-temperature liquid cooling, heat recovery and water-management systems are driving the coordinated optimisation of PUE, WUE and energy reuse efficiency (ERE). This reduces the dependence of computing growth on additional water resources, allowing water productivity and computing capacity to improve in parallel.

Taken together, fluid systems create value across four dimensions: more precise thermal management to support higher computing density; greater energy efficiency to help decouple computing growth from energy consumption; more productive water use through closed-loop circulation, water-quality management and reclaimed-water utilisation; and stronger operational resilience through redundancy, fault tolerance and predictive maintenance. These objectives must be optimised together. Higher energy efficiency should not come at the expense of water efficiency, nor should greater density compromise resilience. This is the essence of water–energy coordination in computing infrastructure.

Table 3 Key technology pathways and expected outcomes for fluid systems

Area	Technical pathway	Core measure	Reported outcome
Energy use	Efficient distribution	Replace fixed-speed pumps with smart variable-speed pumps	25%–40% lower pumping energy
	Demand-based control	Variable-flow and variable differential-pressure control	Flow deviation <5%; idle power reduced by 40%
	Distributed architecture	Distributed pumping and pumps in place of control valves	20%–35% lower distribution energy
	Dynamic control	AI algorithms and dual-drive AI control	40% lower cooling-system energy
Water use	Closed-loop cooling	Replace evaporative cooling with liquid cooling	31%–82% water savings
	Water-quality management	Filtration, treatment, and higher cycles of concentration	Increasing cycles of concentration from 3 to 6 saves 30% water
	Reclaimed-water use	Reclaimed water and rainwater harvesting	97% lower potable-water use
Operating risk	Real-time sensing	Comprehensive multiparameter sensor network	98% fault-prediction accuracy
	Redundant design	N+1/2N pump redundancy	System remains operational after a single-pump failure
	Condition monitoring	Trend analysis of vibration, differential pressure, and temperature	40% lower failure rate
	Duty/standby transfer	Automatic staged transfer to standby pumps	Cooling interruption <10 seconds

Note: Data are based on theoretical calculations and project-based estimates

Chapter 3

Fluid-system technologies across the architecture, equipment and platform layers

03

AI computing and liquid cooling are pushing fluid systems beyond component-level efficiency. Modern fluid systems must coordinate heat generation, transport, regulation and reuse across the facility. As systems scale, risks also multiply. Hydraulic imbalance, slower control response, coolant degradation and mismatches between waste-heat supply and demand. Future systems must manage these risks through dynamic control, intelligent operation and water–energy coordination—without compromising cooling continuity.

Delivering the four value dimensions outlined above—thermal management, energy efficiency, water stewardship and operational resilience—requires a three-layer fluid-system capability framework spanning architecture, equipment and platform. The architecture layer sets the performance foundation by optimising heat-transfer pathways, defining safety boundaries and enabling waste-heat recovery. The equipment layer translates design intent into operating capability through pumps, controllers, sensors and other critical components that regulate flow, monitor conditions and sustain cooling continuity. The platform layer closes the optimisation loop by turning operational data into insights, control strategies and continuous improvement, enabling fluid systems to move from monitoring towards intelligent optimisation.

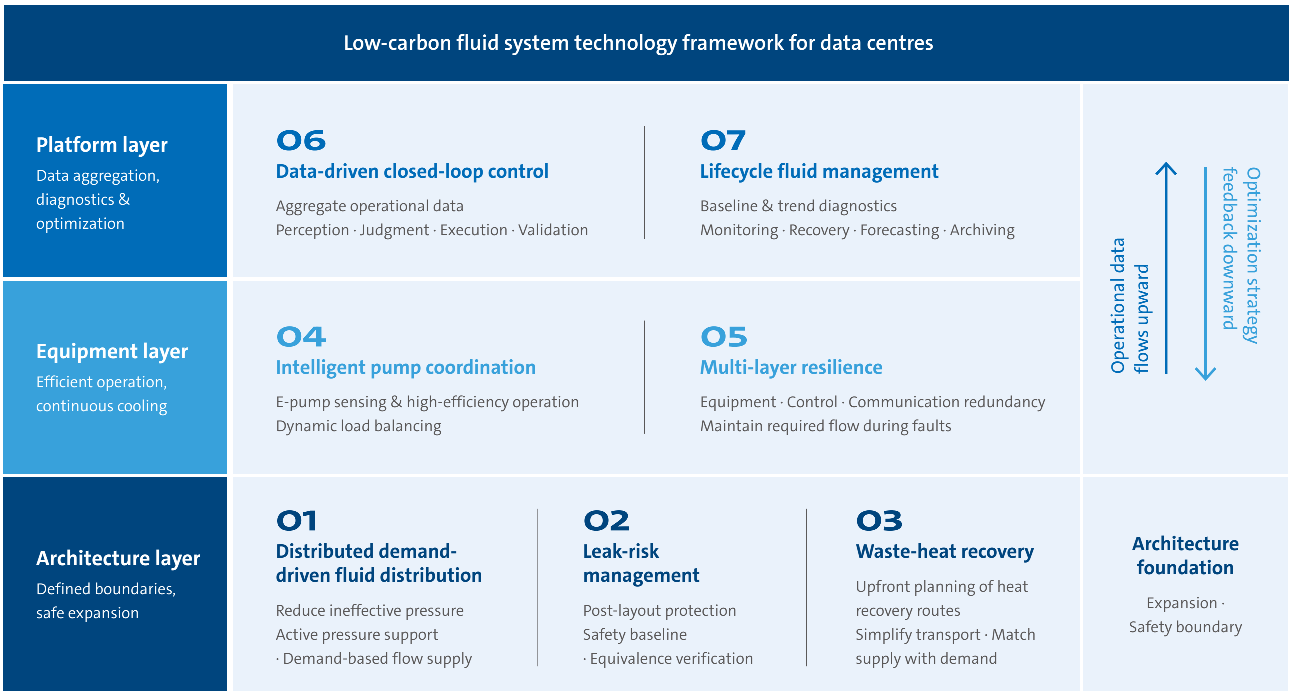


Figure 3 Seven key technologies in the three-layer framework

Together, these three layers define seven key technology domains: distributed demand-driven fluid distribution, leak-risk management, waste-heat recovery, intelligent pump coordination, multi-layer resilience, data-driven closed-loop control and lifecycle fluid management.

Table 4 Mapping seven key technologies to four value dimensions

Layer	Key technology	Thermal management	Energy efficiency	Water stewardship	Operational resilience
Architecture	① Distributed demand-driven fluid distribution	●	●	○	◐
Architecture	② Leak-risk management	◐	○	○	●
Architecture	③ Waste-heat recovery	●	●	◐	○
Equipment	④ Intelligent pump coordination	●	●	◐	◐
Equipment	⑤ Multi-layer resilience	◐	○	○	●
Platform	⑥ Data-driven closed-loop control	◐	●	◐	●
Platform	⑦ Lifecycle fluid management	◐	◐	●	◐

● = core support (the technology is a primary pathway for realising the value dimension)
◐ = collaborative support (a coordinating or safeguarding role)
○ = relatively weak association

3.1 Architecture-layer fluid-system technologies

01 Distributed demand-driven fluid distribution

Distributed demand-driven fluid distribution improves hydraulic efficiency by matching fluid supply to actual cooling demand in high-density liquid-cooling systems. As AI training and inference loads and rack power density rise, cold-plate microchannels, filters, heat exchangers, valves and piping create increasingly complex resistance. Conventional centralised systems typically maintain differential pressure for the most demanding circuit, even when branch-level thermal loads differ significantly. The resulting excess flow and valve throttling waste pumping energy.

The architecture moves regulation closer to thermal loads and adjusts operating conditions as temperature, pressure, flow and IT load change. This reduces distribution energy use, improves response and scalability and aligns cooling more closely with high-density liquid-cooling demand. The transition requires more than additional pumps or controls; it changes the architecture and control strategy so cooling capacity can follow computing load.

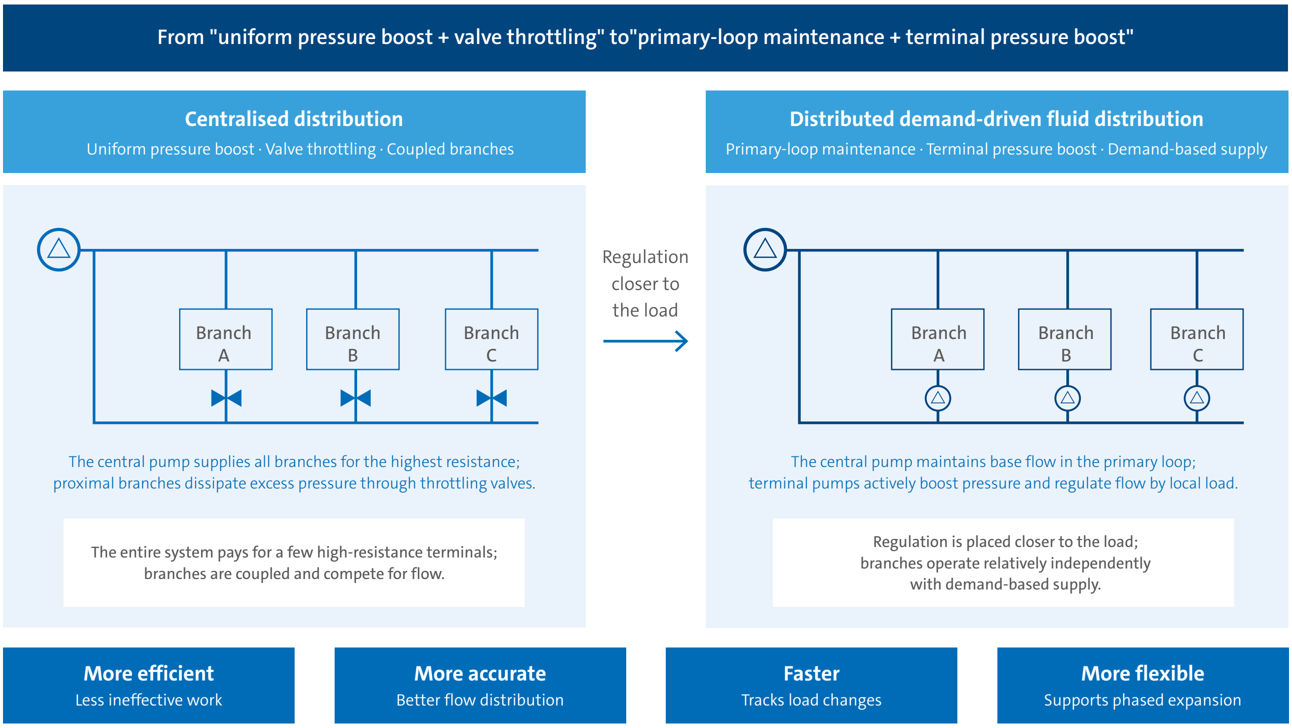


Figure 4 Centralised vs. distributed demand-driven fluid distribution

In practice, distributed distribution can be implemented in three main ways. Distributed demand-driven fluid distribution allocates cooling capacity according to computing-load demand, while distributed pumping moves variable-frequency control to the zone, CDU or terminal. A third approach replaces throttling valves with pumps, using active pressure boosting and flow regulation to reduce persistent throttling losses. Isolation, check and pressure-relief protection must remain in place.

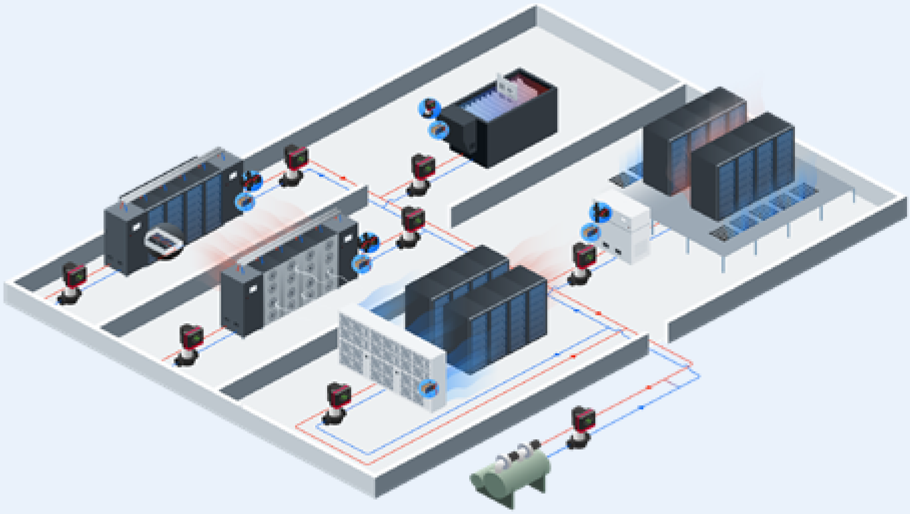


Figure 5 Distributed demand-driven fluid distribution architecture

At its core, distributed demand-driven fluid distribution shifts flow control closer to thermal loads, matching cooling delivery to actual demand while maintaining safe boundaries for flow, pressure and response time. High ΔT and low flow are not objectives in themselves, but means of improving efficiency within the constraints of chip thermal safety, system stability and operational reliability.

Balancing efficiency and resilience therefore remain a primary design principle. Different cooling architectures establish different hydraulic and coolant operating boundaries. Table 5 summarises the key technical considerations associated with each terminal cooling configuration.

Table 5 Fluid-system considerations across cooling configurations

Cooling configuration	Typical location	Fluid-system considerations	Focus in this chapter
Rear-door heat exchanger (RDHx)	Rear of rack	Supply/return conditions, pressure drop, condensation, and air–liquid coordination	Common loop requirements
Cold-plate liquid cooling	Server	Minimum flow, CDU control, filter pressure drop, and temperature response	Design limits
Immersion cooling	Immersion tank	Fluid compatibility, pump selection, materials, and seals	Coolant compatibility
Hybrid air–liquid cooling	Existing-facility retrofit	Air–liquid coordination, local hot spots, and phased retrofit	Control relationships

02 Leak-risk management

As liquid-cooling deployments scale, safety management must shift from equipment-level protection to zone-based risk management. Compared with conventional air cooling, liquid cooling introduces additional interfaces among coolant, IT equipment, racks and piping, making leak control a critical design requirement. Because the potential impact of a failure varies by location, protection measures should be tailored to different risk zones. Within servers, the priority is to prevent coolant leakage from affecting IT operation. In white-space areas, such as CDUs and rack manifolds located close to IT equipment, the focus is on leak detection, isolation and impact containment. In dedicated mechanical and electrical areas, including heat-exchange stations, CDU rooms and pump rooms, greater emphasis is placed on maintenance accessibility, fault detection and recovery capability. A zone-based risk-management approach provides a common framework for designing and evaluating different liquid-cooling architectures and equipment configurations. Rather than prescribing a specific technology pathway, it enables diverse solutions to meet consistent safety objectives while improving both system safety and design flexibility.

Table 6 Leak-risk management objectives for liquid-cooling systems

Risk zone	Typical location	Primary consequence	Control objective and acceptable technical approach
High-risk zone	Cold plates, internal server piping, connections near GPUs	Coolant may contact energised IT equipment	Prevent coolant ingress into IT equipment; rigorously verify connections, materials, pressure, thermal cycling and detection
IT-adjacent zone	White-space CDUs, rack-side manifolds, RDHx	Adjacent servers may be affected or the isolation area may expand	Provide higher-grade detection, containment and local isolation; select the approach for distance, drainage and residual cooling
Controlled MEP zone	Dedicated CDU rooms, heat-exchange stations, pump rooms	Local equipment damage or reduced cooling capacity	Verify detection, drainage, isolation, backup cooling and recovery time; different pump types may demonstrate equivalent outcomes

Liquid-cooling safety should classify risk by consequence rather than equipment architecture or pump type.

A system may use sealless, mechanically sealed, magnetic-drive or canned-motor pumps, but every design must still meet the relevant detection and isolation requirements. It must also provide drainage, residual cooling and recovery. Acceptance should be based on recorded metrics, including detection time, isolation scope, minimum guaranteed flow and recovery time. The objective is to define acceptable consequences for each area, not mandate a single equipment type. Converting these boundaries into verifiable and maintainable requirements allows high-density fluid systems to coordinate thermal performance and resource use without weakening reliability.

03 Waste-heat recovery

Waste-heat recovery is where fluid systems connect data centres to wider energy systems: most electricity consumed by a data centre ultimately becomes heat. Conventional cooling systems reject this heat, while liquid-cooling systems can collect and transport it more efficiently for reuse. High-temperature liquid cooling may supply heat to buildings, industrial processes and other applications. Recovered heat creates value only when it matches actual demand. Source temperature, transport distance, demand stability and the energy used by additional equipment all affect performance. Fluid-system architecture should therefore plan recovery pathways in advance and coordinate heat sources, transport conditions and end uses with external energy systems.

Cooling IT equipment must take priority over waste-heat recovery. External heating networks should be isolated so their fluctuations cannot disrupt server cooling. Data centres should also retain independent backup heat-rejection capacity for periods when recovered heat cannot be used or external demand is interrupted. Assessments must consider recovered heat, user demand, transport conditions and the energy consumed by heat pumps, circulation pumps and other equipment. PUE, WUE, CUE, ERE and operational reliability should therefore be assessed together. Otherwise, improvement in one metric may mask higher resource use, weaker carbon performance or greater operational risk.

Overall, the significance of waste-heat recovery lies not merely in recovering more heat, but in shifting data centres from one-way resource consumption towards circular energy use. By strengthening coordination with regional energy systems while preserving cooling safety, fluid systems can become an important link among computing capacity, energy and thermal-energy utilisation, supporting more circular, resource-efficient computing infrastructure.

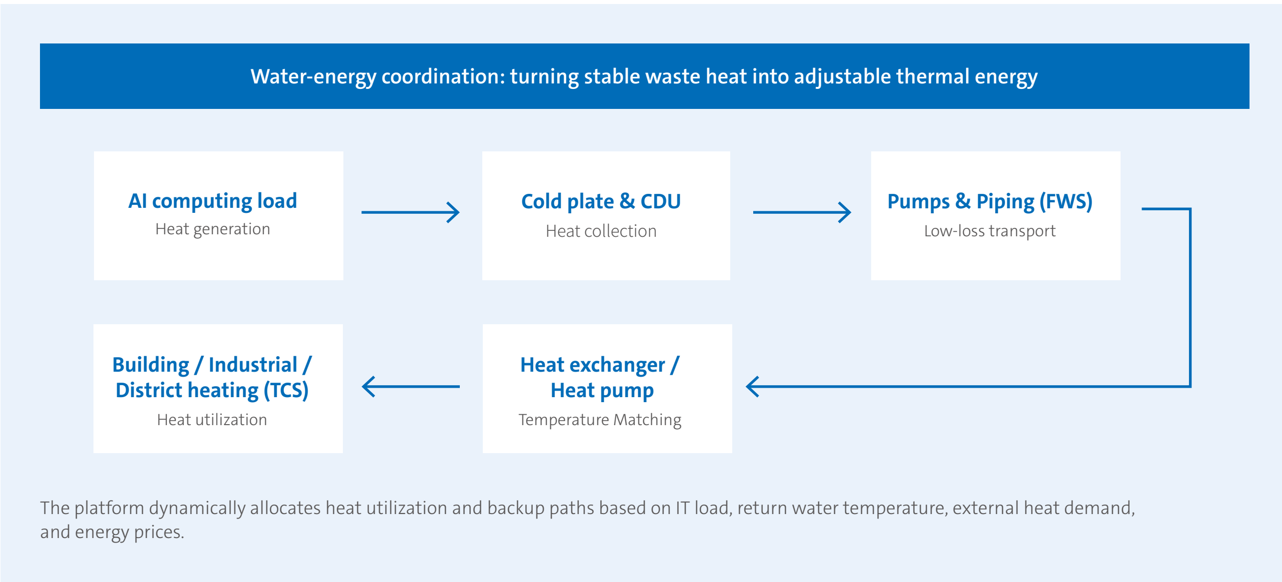


Figure 6 Water–energy coordination through waste-heat recovery

Case study 1 | Microsoft–VEKS waste-heat recovery project, Denmark

Microsoft is developing a new data centre in Denmark with VEKS, a nonprofit district-heating company owned by 12 municipalities west of Copenhagen. The project combines waste-heat collection, heat-pump temperature lift and district heating, with high-efficiency pumping supporting stable circulation and lower distribution energy. It illustrates how a data centre can become a low-carbon urban heat source.

Microsoft is responsible for heat collection inside the data centre and piping within the site boundary. VEKS is building the external network, heat-exchange stations and large heat pumps that upgrade low-grade heat for users in 12 municipalities. A 33,000 MWh underground thermal store balances supply and demand. Efficient pumping and precise hydraulic control support stable circulation and lower distribution energy. The data centre operates on 100% renewable electricity.

The system is scheduled to enter service in the 2025–2026 heating season. As the data centre expands, it is expected to supply heat to nearly 6,000 homes at a competitive cost compared with conventional VEKS heat sources, reducing primary-energy use and regional emissions.

Underground thermal store with a capacity of

3.3 MWh

Renewable electricity ratio of data centre operates

100%

As the data centre expands, it can supply heat to nearly

6,000 homes

3.2 Equipment-layer fluid-system technologies

04 Intelligent pump coordination

Intelligent pump coordination is foundational to distributed demand-driven fluid distribution and efficient fluid-system operation. Its purpose is not to add functions for their own sake, but to integrate sensing, decision-making and execution within pump equipment so operating conditions can shift from fixed operation to demand-based control. In a liquid-cooling system, pump sets not only transport coolant; they also act as key actuators linking fluid-state sensing with system regulation. Conventional pump sets are typically assembled from pumps, drives and external controllers. Intelligent pumps integrate high-efficiency motors, variable-frequency drives, controllers and sensor interfaces to acquire real-time pressure, flow, temperature, power and operating status. They then adjust speed, differential pressure or flow-control mode so that cooling capacity remains aligned with changes in thermal load.

The value of an intelligent pump lies not in any single feature—variable-speed operation, communications or a digital interface—but in its ability to select and switch control strategies according to the operating goals of the fluid loop. Different systems may target constant differential pressure, constant temperature difference, constant flow or load-following operation. It adapts its control logic to actual conditions and falls back to a predefined safe state when abnormalities occur. This helps the system respond to changing server loads and cooling demand, reducing unnecessary distribution energy and improving operational adaptability.

Intelligent pump coordination is a system capability, not a feature of one device. Its performance still depends on equipment selection, sensor configuration, controller tuning and compatibility with the wider fluid system. The control framework must reflect hydraulic behaviour, thermal-management requirements and failure boundaries. When these conditions are in place, intelligent pump coordination can reduce energy use, accelerate response and support high-density liquid cooling. Moving control closer to the equipment layer helps liquid-cooling systems shift from static operation to dynamic optimisation.

Table 7 Intelligent pump coordination modes and engineering boundaries

Control mode	Application	Primary variables	Safety and tuning limits
Constant differential pressure	Loops with relatively stable terminal differential pressure	Differential pressure; pump speed	Set a minimum differential pressure; verify flow, temperature and pump power at the index circuit
Proportional differential pressure	Friction-dominated variable-flow loops	Differential pressure; flow	Reduce setpoint with flow while preventing terminal underflow, overshoot and oscillation
Constant flow/ temperature/ΔT	Stable-flow systems or defined thermal targets	Flow; supply and return temperatures	Maintain cold-plate minimum flow and temperature limits; fall back safely on sensor failure
IT feedforward plus thermal feedback	Rapidly changing AI loads with thermal lag	IT power; temperature; differential pressure/flow	Limit ramp rate and overshoot; feedforward does not replace thermal protection
Multipump staging	Parallel pumps over a wide load range	Total flow; differential pressure; efficiency	Stage and balance pumps under control; capacity staging is not the same as fault transfer

Case study 2 | Smart pumps for leading immersion cooling

Immersion cooling places specific demands on pumps: reliable circulation, embedded monitoring and compatibility with dielectric fluids. In the latest platform of a leading Euro immersion cooling, smart variable-speed pumps were integrated into the cooling loop to simplify the hydronic system and provide operating data. Embedded sensing measures pressure and temperature, while differential pressure across the pump can be used to estimate flow and reduce the need for external flow meters. The data helps identify efficient operating points and detect abnormal trends before they affect uptime. Because dielectric fluids can affect elastomers and pump seals differently, fluid samples and sealing materials were tested for compatibility. The case shows that pump selection for immersion cooling must integrate control capability, operational visibility and material compatibility.

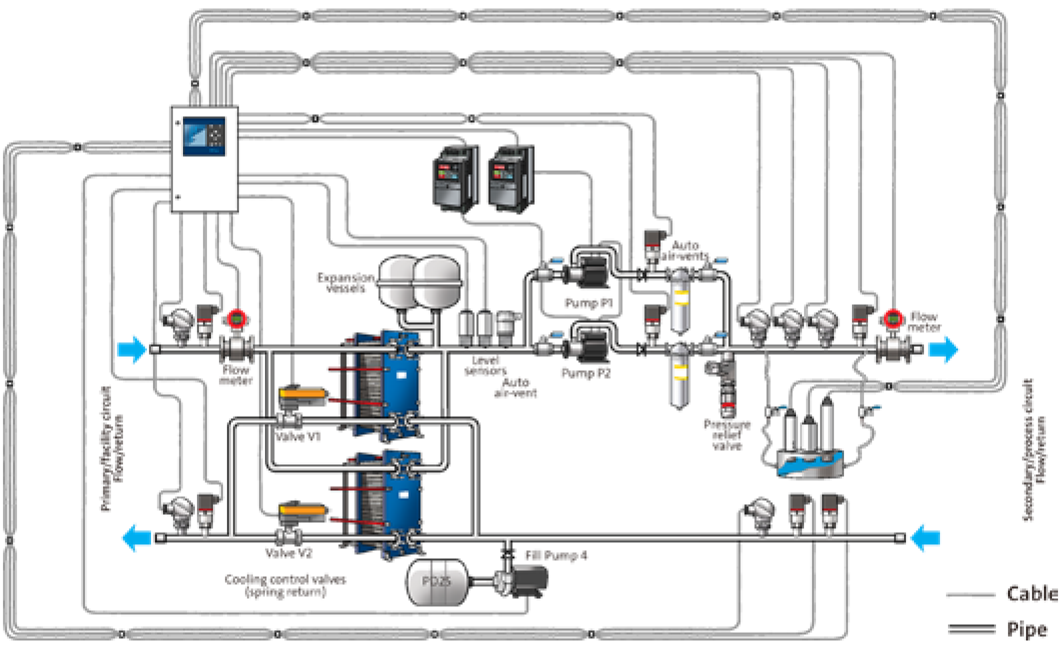


Figure 7 CDU control under traditional centralised architecture

05 Multi-layer resilience

Multi-layer resilience safeguards continuous cooling in high-density computing environments by keeping a single failure from breaking the cooling chain. Conventional duty/standby arrangements are often insufficient because operating and standby equipment may still share power supplies, controllers, communication links or critical components. Multi-layer resilience therefore spans equipment, control and communications. Multiple-pump coordination, reserve capacity and isolatable branches limit the impact of equipment-level failures. Control design must prevent any single node from becoming a bottleneck and preserve essential local control strategies, while communications design must preserve essential operation when higher-level information is lost. The goal is cooling continuity at system level, not reliability at device level alone.

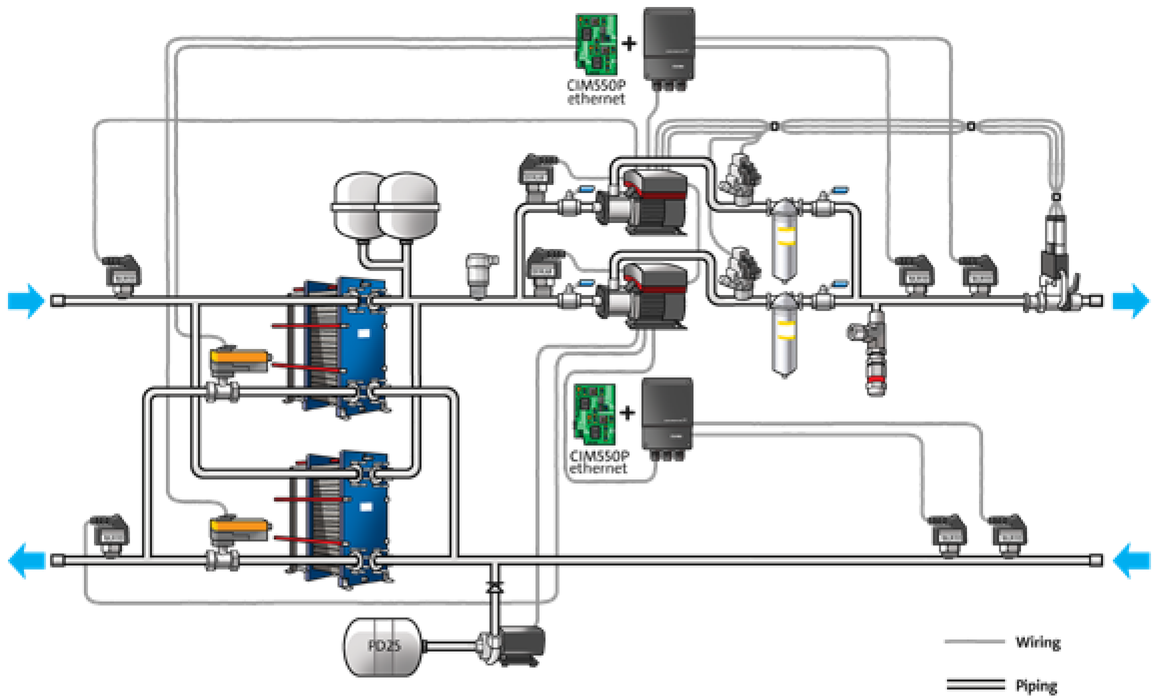


Figure 8 CDU control relationships with Syspro local-control redundancy

Redundancy should be judged by whether required cooling continues during a failure, not by the number of spare devices. Figures 7 and 8 show why standby capacity alone is insufficient. The system must maintain cooling, temperature limits and controlled recovery if a pump, controller or communication link fails. Integrated programmable pump control is one option, but it must be backed by independent power, hardwired interlocks and full-load system validation. When communications return, supervisory control should resume at a controlled ramp rate to avoid abrupt changes in differential pressure or flow.

Table 8 Typical failure scenarios and acceptance requirements

Failure scenario	Minimum action and protection	Recorded acceptance criteria
Single-pump cut-out	Start standby pump; remaining pumps maintain flow and prevent reverse flow	Minimum flow, temperature excursion, transfer time
Controller failure	Local or standby controller maintains safe setpoints	Degraded state, takeover time, recovery time
Communications loss	Local control maintains flow and temperature and rejects high-risk remote commands	Outage duration, alarm, recovery ramp rate
Sensor drift	Validate and switch to a redundant signal, substitute value or conservative fallback	Deviation threshold, fallback trigger, temperature/flow limits
Filter blockage	Alarm on differential pressure; isolate or switch filters; protect pump inlet	Differential pressure, minimum flow, cavitation margin, response time
Maintenance or power transfer	Isolate equipment; independent power or remaining pumps maintain capacity	Available cooling, interruption duration, recovery time

3.3 Platform-layer fluid-system technologies

06 Data-driven closed-loop control

Data-driven closed-loop control is the platform-layer capability that turns fluid-system data into continuous operational improvement. As liquid-cooling systems become more complex, manual inspections and isolated parameter monitoring are no longer sufficient to understand system behaviour or manage operating risk.

Fluid systems generate large volumes of operational data, including flow, pressure, temperature, power consumption, equipment status, coolant condition, water quality and water-efficiency metrics. By integrating these data through a unified model, operators can identify deviations, efficiency losses and emerging risks across devices, loops and system layers. The value of the platform is therefore not visibility alone, but the ability to translate insight into recommendations, control actions and validated improvement.

A true closed loop requires four connected stages: sensing, decision-making, execution and validation. The platform must first confirm that data sources are trustworthy, identify abnormalities and root causes and determine whether an operational adjustment is needed. It then executes controlled changes and validates the results by tracking temperature, flow, pump power, redundancy status and other key indicators. Data from different devices must share a common time base and use consistent equipment identifiers, units of measure and quality labels. These controls reduce errors caused by unsynchronised data or anomalous values.

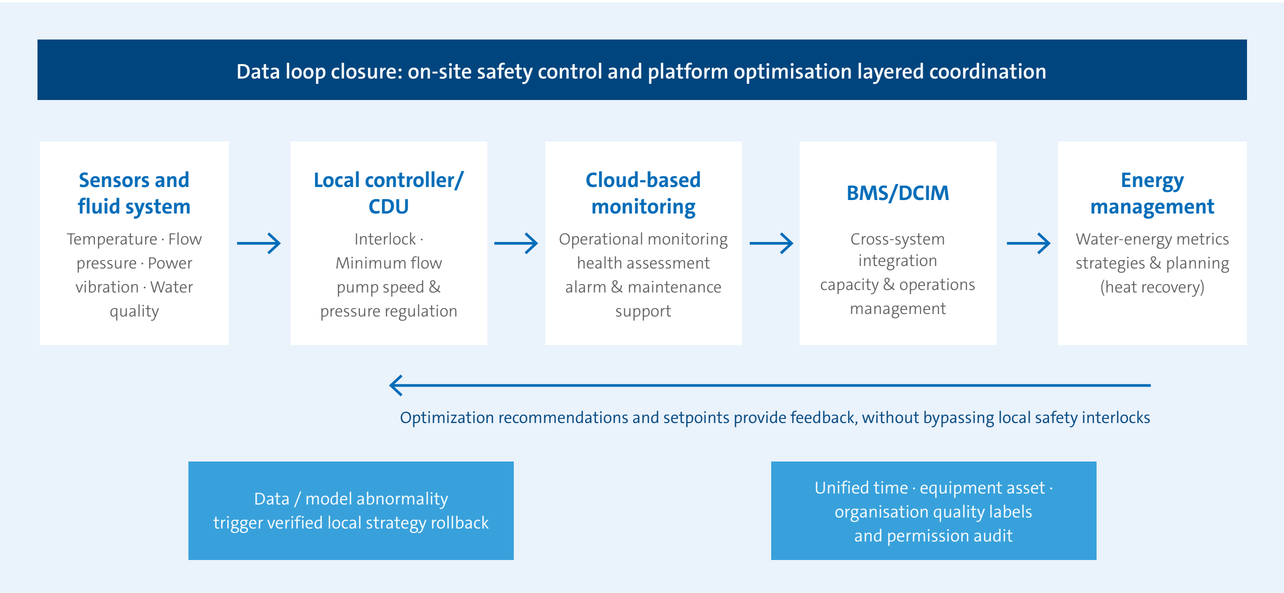


Figure 9 Data-driven closed-loop control architecture

Data-driven closed-loop control does not mean that all control should be centralised at the platform. Equipment-level controls must continue to provide safety interlocks, minimum-flow protection, temperature protection and other basic safeguards. CDUs or zone controllers should handle regulation on time scales of seconds to minutes, while the platform layer focuses on operational optimisation, capacity forecasting and maintenance decisions. When data is abnormal or communications are interrupted, the system should automatically revert to a validated local-control strategy so that cooling safety remains local, deterministic and independent of the upper-level platform.

The purpose of data-driven closed-loop control is to build a continuous optimisation capability, not another layer of passive monitoring.

By accumulating long-term operating data, fluid systems can manage not only equipment condition but also coolant condition, operating efficiency and overall system health. This supports a transition from detecting problems after they occur to identifying risks in advance, from scheduled to predictive maintenance and from experience-based to data-driven management, providing a foundation for the long-term stable operation of high-density computing infrastructure.

Case study 3 | Pump control and remote operations platform

A layered pump-control platform connects on-site self-learning regulation, supervisory coordination and remote operations into a continuous optimisation loop. In a large data centre in Sichuan, high-efficiency pump sets, dedicated pump controllers, supervisory interfaces and a remote monitoring platform were deployed for high-density computing loads.

On-site controllers used pressure, flow, operating frequency, energy consumption and other data to refine pump operating curves and determine the number and frequency of operating pumps in real time. The supervisory system aggregated operating status and coordinated setpoints without overriding site safety interlocks. The remote platform recorded energy use, alarms and equipment-health information for early warning and maintenance planning, while retaining local operating strategies for communication failures.

The project achieved stable round-the-clock operation while reducing cooling-distribution energy use. It demonstrates that the platform layer should connect on-site execution, cross-equipment analysis and lifecycle operations into a verifiable closed loop, rather than replace equipment-level control.

07 Lifecycle fluid management

Lifecycle fluid management extends platform-layer control from equipment operation to fluid condition, material compatibility and long-term reliability.

Water quality, particulates, coolant ageing, erosion, blockage and cavitation require continuous monitoring and intervention throughout the system lifecycle. The objective is to establish baselines, track deviations, verify corrective actions and update operating strategies so that coolant condition becomes part of the same closed-loop management system as equipment performance.

Table 9 Coolant risks, monitoring and verification points

Risk dimension	Monitoring focus	Verification and response
Flow and hydraulics	Flow/velocity, filter differential pressure, inlet pressure, cavitation	Verify maximum velocity and NPSH; alarm on differential pressure and clean filters
Material compatibility	Metal corrosion, elastomer swelling, polymer ageing, particle shedding	Test immersion, thermal cycling and seals for the fluid, material, temperature and service life
Coolant condition	Viscosity, acid number, particles, precipitates, water quality, dielectric properties	Set baselines and limits; use trends to determine purification, dosing or replacement
Maintenance loop	Differential pressure, heat transfer, flow and coolant indicators before and after intervention	Retest and archive results; update operating baselines and maintenance strategy

Coolant management should cover the full lifecycle: baseline establishment, continuous monitoring, condition assessment, purification, chemical treatment or replacement, revalidation and record archiving. Each sampling event, coolant replenishment, filter replacement or material change should be documented with pre- and post-treatment results, allowing operating baselines and maintenance strategies to remain verifiable and traceable.

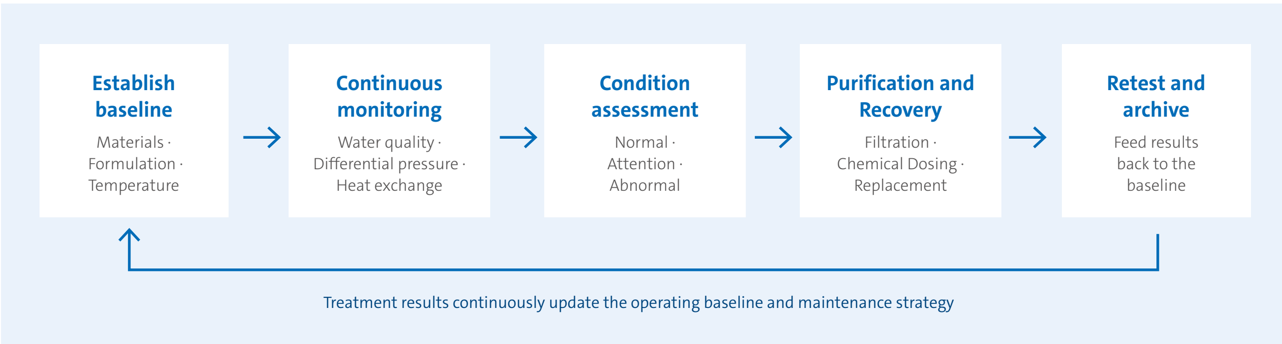


Figure 10 Lifecycle fluid management loop

Facility-level water stewardship and technical coolant management are distinct but interdependent parts of the fluid-system lifecycle. Water stewardship covers intake, treatment, makeup, circulation, discharge and reuse; coolant management focuses on heat-transfer performance, material compatibility and reliability within the cooling loop. Projects should select water sources for local conditions and configure treatment to meet water-quality requirements. During operation, teams should monitor the makeup-blowdown balance, WUE and water quality while controlling microbial growth, scaling and corrosion. The two must be coordinated, and the table below summarises the distinction.

Table 10 Hierarchical differences between water stewardship and coolant management

Level	Managed object	Key issue
Facility	Water sources and makeup	Municipal, recycled, rainwater or reclaimed supplies; makeup-water quality
System	Facility water system (FWS) and technology cooling system (TCS) loops	Filtration, softening/reverse osmosis, chemical treatment, pressure drop and reliability
Coolant	Cooling fluid	Compatibility, corrosion, particles, viscosity and ageing
Environment	Discharge and waste heat	WUE, ERE, blowdown and waste-heat recovery

Looking ahead, fluid-system management will evolve in three directions: from asset-level optimisation to system-wide orchestration; from thermal assurance to coordinated management of computing, energy and water resources; and from experience-based operations to data-driven lifecycle management. Future data centre competitiveness will increasingly depend on the ability to scale efficiently through intelligent control, resource coordination and optimised distribution. In this context, fluid systems will extend beyond reliable cooling to become a foundational capability for sustainable, resource-efficient computing infrastructure.

Chapter 4

Representative case studies of fluid systems in green computing infrastructure

04

The following cases illustrate smart pumping, liquid-cooling optimisation, water reuse, heat recovery, high-reliability operation and retrofit of existing facilities, showing the shift from component support to system-level optimisation.

4.1 Intelligent pumps enable efficient liquid cooling for stable, energy-efficient high-density computing

NorthC data centre: intelligent flow control for high-power direct-to-chip cooling

As AI and HPC increase chip power, efficiency and capacity limits of conventional air cooling become more pronounced. NorthC, a leading Dutch data centre operator, is deploying efficient liquid cooling across more than ten facilities and aims to achieve carbon-neutral operations by 2030.

At its Eindhoven High Tech Campus facility, NorthC uses direct-to-chip (D2C) liquid cooling in a closed water loop to remove heat at the server. Intelligent temperature and flow control adjusts coolant flow and supply/return temperatures to real-time server load. NorthC reports about four times the cooling capacity of conventional air cooling, supporting rack deployments of 300–500 kW.

Figure 11 NorthC data centre intelligent flow-control installation

Zhenshitong AI computing centre: pump-driven two-phase liquid cooling for backplane-level heat removal

Rising GPU power and rack density are increasing the limits of air cooling and single-phase liquid cooling. To support a high-density AI facility, the Zhenshitong AI Computing Centre adopted an integrated chip- and backplane-level, pump-driven two-phase liquid-cooling system.

The system uses liquid–vapor phase change in microchannels, with a reported heat-transfer coefficient of 65,000 W/(m²·K)—20 times that of single-phase liquid cooling and 1,000 times that of air cooling—and supports ultra-high thermal design loads of up to 1,500 W per chip. Reported coolant flow is 10% of that required for single-phase liquid cooling and 1% of the equivalent air-cooling flow, reducing distribution energy. The design PUE is below 1.20, with an optimum of 1.10 with chip-level cooling solution in new builds. The system adopts a water-free cooling design, achieving a WUE of 0, waste-heat recovery of 100% at temperatures above 45 °C, low maintenance and noise, and a two-year payback period.

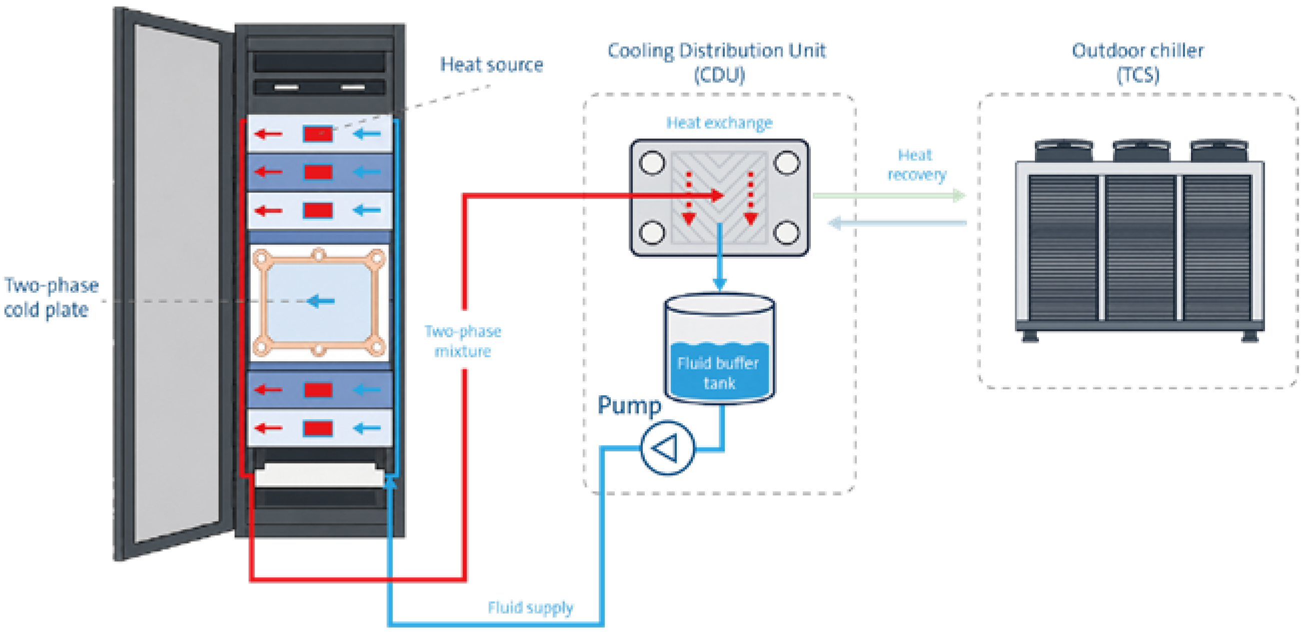


Figure 12 Chip-level pump-driven two-phase liquid-cooling cold-plate heat-dissipation system

With an optimum PUE of

1.10

Payback period

2 years

4.2 Fluid-system architecture optimisation enables data centre-wide energy-efficiency upgrades

Izmir data centre, Türkiye: fluid-system optimisation with distributed pumping

As computing facilities expand, centralised cooling can incur high throttling losses, difficult hydraulic balancing and slow response. During construction, a Tier III data centre in Izmir, Türkiye, redesigned its cooling-distribution architecture to improve energy efficiency and operating stability.

The 11,838 m^2 facility has 4 MW of IT capacity and about 1,500 m^2 of white-space and holds Uptime Institute Tier III and LEED Gold certifications. It replaced centralised pumps and dynamic balancing valves with distributed pumping, assigning independent variable-speed pumps to dozens of fan-wall cooling units. Active flow control matches cooling to real-time heat load, reducing piping pressure losses and improving hydraulic control.

Project tests report about 15% lower total cooling-system energy use than conventional control, together with faster load response and better hydraulic balance. The case shows how distribution architecture and control strategy can extend savings from individual equipment to the whole system.

Could lower total cooling energy consumption by

15%

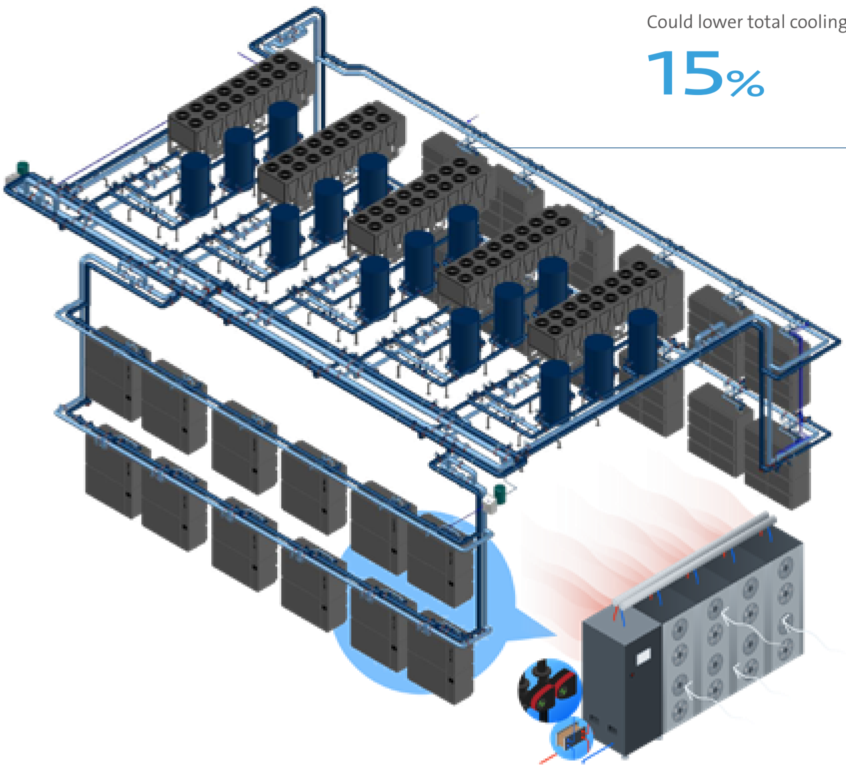


Figure 13 System architecture of an Izmir data centre in Türkiye

4.3 Coordinated water–energy use builds a circular system for green computing

JUPITER supercomputing centre, Germany: warm-water direct liquid cooling with heat recovery

JUPITER is a new-generation European exascale-class supercomputing platform with IT load of up to 27 MW. Its heat-removal demand and low-carbon objectives cannot be met reliably over the long term with conventional air cooling alone.

The project combines 25 MW of warm-water direct liquid cooling with 2 MW of air cooling, using prefabricated modules, waste-heat recovery and renewable electricity. High-capacity heat pumps upgrade low-grade server heat for regional use, while distributed smart-pump control and real-time monitoring regulate coolant temperature, differential pressure and flow.

A low-temperature district-heating network operates at about 40 °C and uses supercomputer waste heat directly for buildings; heat pumps can raise the temperature to about 70 °C where required. Project tests report PUE of 1.05–1.08 and annual carbon-dioxide savings of about 5,000 metric tons. Heat recovery currently uses 20% of the available waste heat, with a target of 50%.

PUE can reach	Annual CO_2 reduction	Waste-heat utilization rate
1.05-1.08	5,000 tonnes	20% → 50%

Figure 14 Warm-water direct liquid-cooling system at the JUPITER supercomputing centre

4.4 Fluid-system optimisation and retrofit unlock energy-saving potential in existing data centres

QNB Finansbank data centre: energy-efficiency retrofit of an existing facility

After 13 years of stable operation, QNB Finansbank's data centre in Türkiye launched efficiency retrofit without interrupting business operations. The project targeted cooling energy and the remaining savings potential of the existing data facility.

The original refrigerant-based cooling system was replaced by a hybrid solution combining dry coolers, precision air conditioning and free cooling. Fifteen IE5 high-efficiency twin-head variable-speed pumps were installed, and the system now switches automatically to free cooling when outdoor conditions permit. The upgrade addresses both cooling architecture and hydronic distribution.

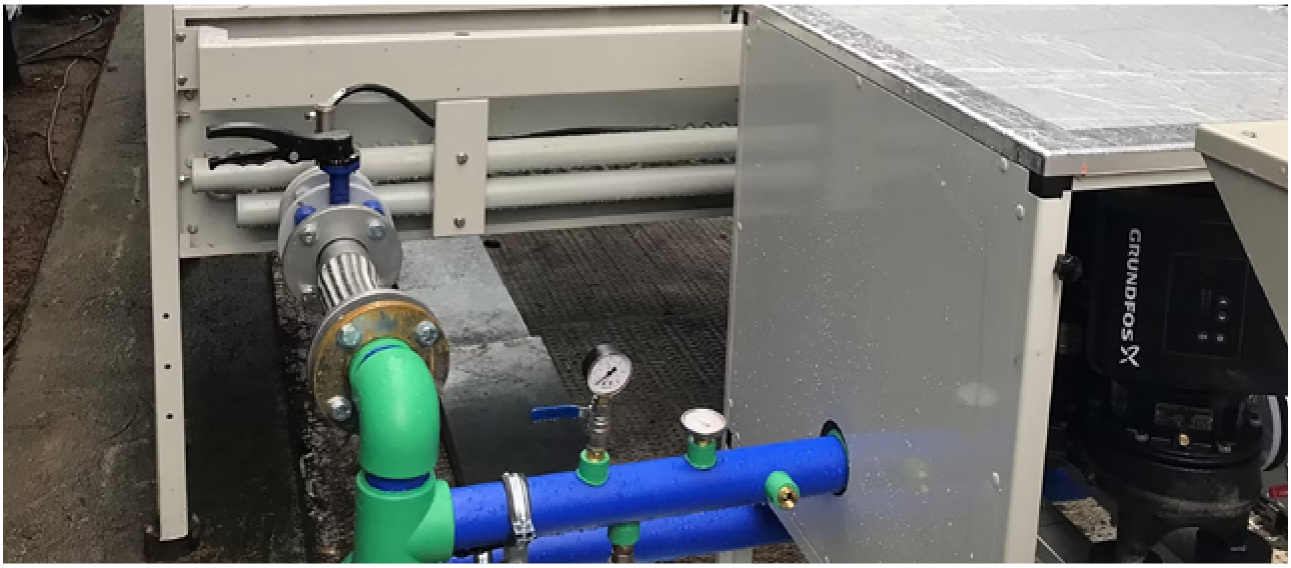


Figure 15 Retrofit Installation at the QNB Finansbank Data Centre

After the retrofit, cooling energy consumption fell by more than 50% and PUE improved from 1.52 to 1.21, while cooling capacity, carbon performance and reliability also improved. The case demonstrates a cost-effective pathway for upgrading existing data centres through coordinated changes to cooling architecture and fluid distribution.

Cooling energy consumption fell by more than

50%

PUE improved from

1.52 → 1.21

Chapter 5

Future trends and recommendations

05

5.1 Future trends

Coordinated xUE management will deepen as high-density computing and liquid cooling expand. Green-computing governance is moving from energy control alone towards integrated management of energy and water performance.

Leading operators — including Huawei, Alibaba, Microsoft and AWS — are combining intelligent fluid optimisation, free cooling, non-potable-water reuse and campus-wide digital monitoring and dispatch to reduce cooling energy while improving water circulation. The broader ecosystem is also becoming more collaborative: operators, engineering designers, equipment suppliers and research organisations are working towards common specifications, data interfaces and evaluation methods for liquid-cooling fluid systems. Future progress will depend on consistent green-computing framework standardisation, full-process monitoring and reporting and pilot projects that connect policy, technology and deployment. Together, these measures can create a water–energy ecosystem with clearer standards, stronger industry coordination and more predictable implementation.

Pumps and hydronic distribution are becoming foundational to computing infrastructure efficiency and reliability.

As AI workloads and liquid-cooling deployment expand, distribution design, control and fault tolerance have a growing effect on cooling energy, operating stability and controllability. Distributed smart-pump systems should therefore be integrated into CDU development and cooling-network planning, together with intelligent sensors and standardised data interfaces. Variable-speed pumping and precise flow and differential-pressure control can then track real-time load rather than operating at fixed design conditions. Safety must advance at the same level. Pump redundancy, coolant and material compatibility, leak protection, independent power and controls, equipment-health monitoring and water-quality management should be designed as one system. The objective is not simply efficient pumping, but a distribution system that can reduce energy use while maintaining stable, recoverable operation under high-density computing loads.

High-standard new projects and systematic retrofits will proceed in parallel, extending water–energy optimisation across planning, construction, operation, maintenance and renovation.

New AI and supercomputing hubs can combine hybrid air–liquid cooling, smart pumping and integrated monitoring in replicable designs suited to high-density loads. These projects should coordinate cooling architecture, hydronic distribution, controls and operational practices from the outset. Existing facilities present a different challenge: many have ageing cooling equipment, insufficient fluid-control capability and high pumping and distribution losses. Targeted upgrades to efficient pumps, piping networks, valves, controls and site-wide energy and water monitoring can unlock capacity while reducing operational energy consumption. Applying high standards to new projects and focused, measurable improvements to existing assets can lower lifecycle resource use and improve green and standardised operation of computing infrastructure.

5.2 Recommendations

At industry level, China should strengthen fluid-system standards and collaboration mechanisms.

PUE, WUE and other xUE metrics need consistent system boundaries, calculation methods and evaluation criteria so that energy and water performance can be compared meaningfully. Standardised collection of pump-system flow, temperature, differential pressure, power and equipment-status data would support reliable operating records and more transparent assessment. CDU manufacturers, fluid-equipment suppliers, data centre operators, design institutes and research organisations should coordinate technical requirements, data interfaces and verification methods through a shared industry platform. Pilot projects can then convert effective technologies and operating practices into repeatable guidance, while regular exchange and joint review can build consensus and connect policy, technology and implementation.

At the technical level, fluid-system upgrades should combine efficiency improvement with end-to-end risk control.

Distributed smart pumping and optimised hydronic distribution should integrate sensors and standard interfaces into CDU and cooling-system design, allowing flow and differential pressure to follow high-density computing loads. For large liquid-cooling systems, reliability provisions should combine pump redundancy, coolant and material compatibility, leak detection and containment, local safety functions, online monitoring and fault warning. Refined water-quality management and equipment-health monitoring should prevent scaling, corrosion, abnormal energy use and unplanned shutdowns. These measures should be verified at system level rather than assessed as isolated component features. Coordinated optimisation can then improve energy efficiency and operating reliability without trading one objective against the other.

Implementation should address new projects and existing facilities in parallel.

New projects should demonstrate liquid cooling, smart pumping and system-level energy optimisation through standardised, repeatable design, commissioning and operating practices. Their results should be measured and documented so that effective approaches can be transferred to other sites. Existing data centres should target high cooling energy, low water efficiency, weak controls and ageing equipment through focused upgrades to hydronic distribution, heat-rejection equipment, monitoring and control systems. Retrofit plans must protect business continuity and prioritise measures that can be verified under actual load. Together, high-standard delivery and carefully staged retrofits can reduce energy and water use, recover value from existing assets and raise the overall level of green, low-carbon computing infrastructure in China.

GRUNDFOS (China) Holding Co., Ltd.
Unit 903, 9F, Dinghao DH3 Tower A,No. 3
Haidian Street, Haidian District, Beijing, China
Post Code: 100080
www.grundfos.cn

格兰富 GRUNDFOS | CAICT 中国信通院