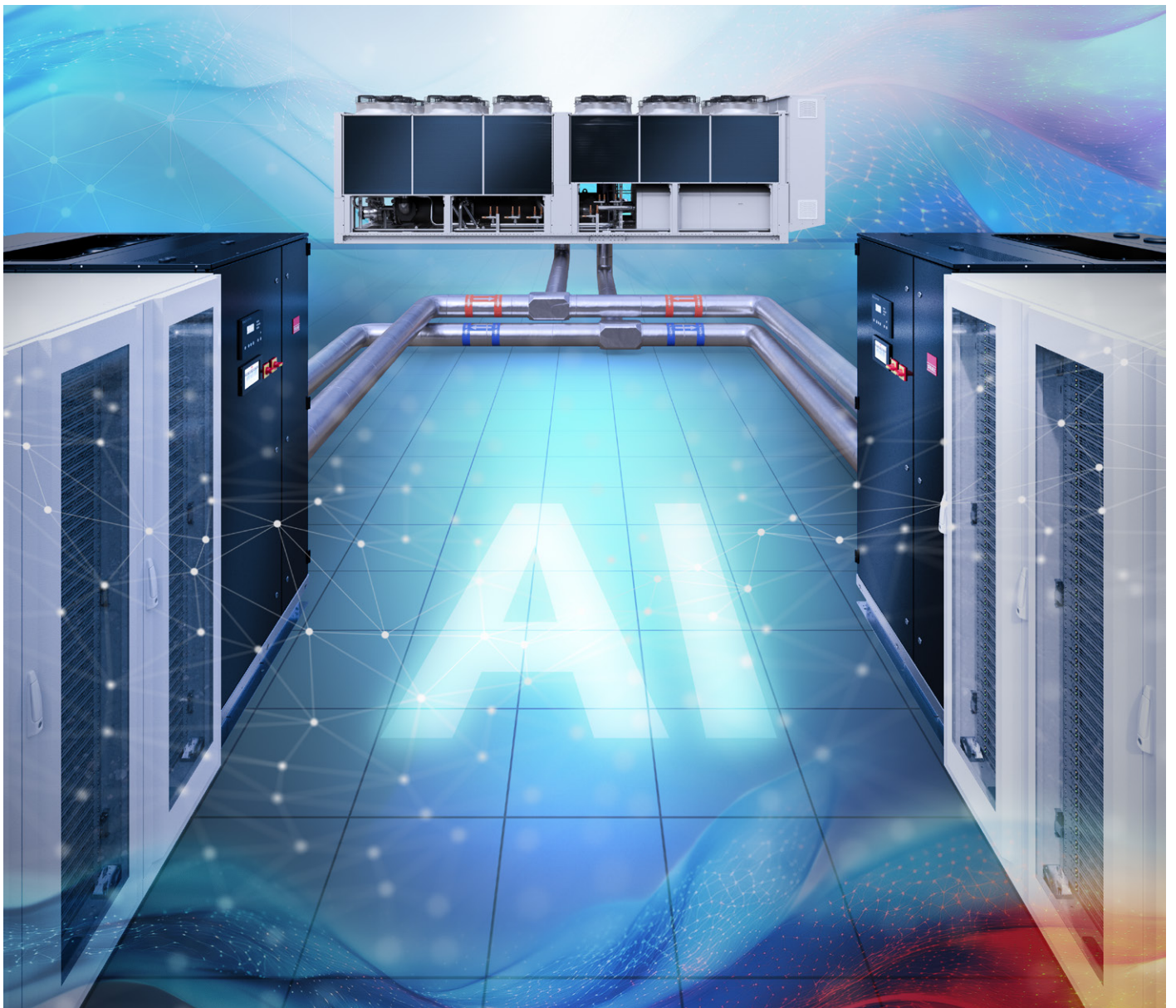


STULZ

CLIMATE.CUSTOMIZED.

End-to-End Cooling

The Key to Resilient AI Data Centers



Contents

- 1. Editorial 3
- 2. Liquid Cooling and Air Cooling are growing together 4
- 3. Key components of Liquid Cooling systems 5
- 4. Planning strategies for greenfield and retrofit architectures 10
- 5. Managing transient peaks in demand 15
- 6. Intelligent regulation and control architectures 20

1.

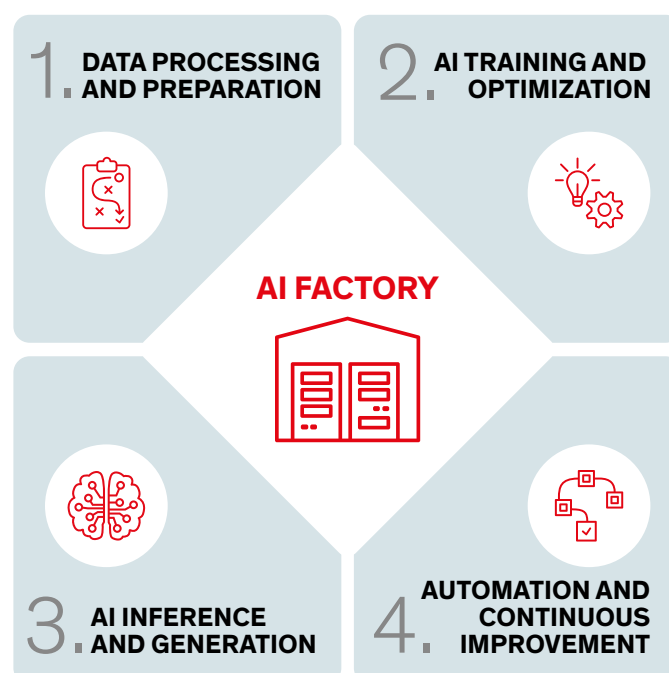
Editorial

Liquid Cooling as a system architecture: Planning, integration and control of high-density AI infrastructures

From a cooling perspective, the transformation of data centers into high-density AI factories represents a crucial turning point. Whereas the first wave of AI development was characterized primarily by centralized cloud infrastructures and rapidly scalable hyperscale clusters, the focus is now increasingly shifting to decentralized, highly specialized SLM (Small Language Model) applications. To minimize latency and achieve a greater level of operational sovereignty, they are being deployed seamlessly at the edge, either individually or in combination.

However, this trend is triggering a domino effect that should not be underestimated, because any increase in local computing power then results in a significantly higher demand for air conditioning at different levels. This starts with the space in the data center, includes the rack within micro data centers, and extends throughout the entire cooling system path. The resulting heat loads can no longer be physically absorbed by conventional Air Cooling concepts.

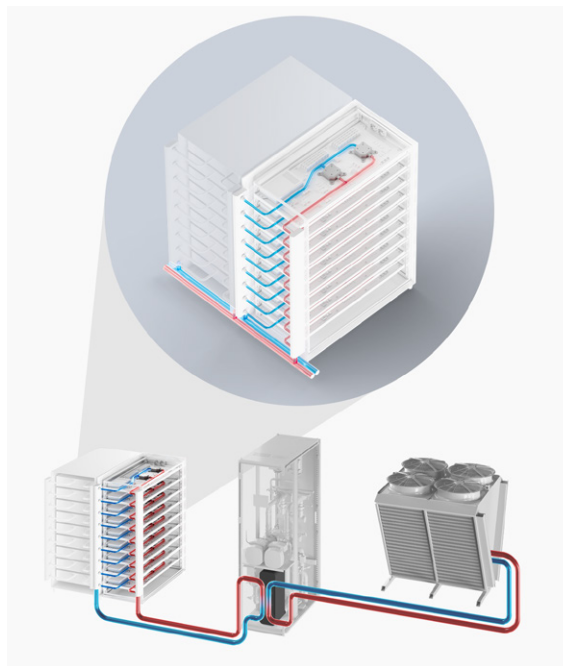
This is why data center operators and planners that want to achieve a reliable return on investment over the long term and a high level of energy efficiency should regard the cooling and system architecture as a key business-critical factor from the very beginning.



2. Liquid Cooling and Air Cooling are growing together

In modern data center planning, the debate has long since moved on from a basic discussion of the benefits of Air Cooling or Liquid Cooling. The reality of operating high-density AI infrastructures clearly demonstrates that future-proof data centers can only remain competitive if they embrace hybrid cooling concepts. While Direct Chip Liquid Cooling (DCLC) captures the extreme hotspots at the processor level (CPUs, GPUs), a significant proportion of the thermal power loss that is produced by voltage converters, memory modules, optical transceivers and the general server electronics remains in the room. But this heat input still needs to be dissipated reliably via closed-circuit precision air-conditioning solutions.

The challenge facing specialist planners and manufacturers is no longer just to create a technical design for a single cooling component, but to synchronously control two air-conditioning methods that are thermodynamically and hydraulically completely different in one common cooling system path in continuous and partial load mode. In hybrid scenarios like this, uncoordinated island solutions at individual interfaces will quickly result in efficiency losses, hydraulic vibrations or, in extreme cases, may even cause temperature-induced reductions in performance for processors and GPUs. Consequently, the cooling infrastructure should not be viewed as a collection of purchased components that exist in isolation. It is the business-critical overall system that needs to link together seamlessly from the cold plate in the server to the Coolant Distribution Unit (CDU) and the recirculating air units to the chiller. Experience gained in recent years clearly shows that successful implementation of precision air-conditioning systems is ultimately always dependent on cooperation between planners, operators and an experienced system partner that need to work together to handle the complex nature of modern data centers. In modern cooling systems, precision air cooling works together closely with the Liquid Cooling infrastructure to create an overarching integrated concept that covers all the requirements for delivering efficient cooling and a high level of system availability.



3. Key components of Liquid Cooling systems

Liquid Cooling in modern AI data centers needs to be conceived of as a complete system that is continuously interlinked, not merely as a series of individual components in isolation. Each level, from the cold plate on the silicon die to the external dry cooling, has an impact on the thermal performance, hydraulic stability and the controllability of the system as a whole. Accordingly, it is absolutely vital to coordinate all these subsystems. Scaling of the pump capacities, reliable bleeding of the risers and continuous filtration in the secondary circuit are crucial for guaranteeing thermal stability under full load and for permanently securing demanding AI workloads and training processes against risks of failure.

To ensure that the system is actually integrated smoothly in practice, it is essential to have a basic understanding of the individual components in the Liquid Cooling system together with their physical interfaces. Each component, from the quick-disconnect coupling on the server slide-in unit to the plate heat exchanger, forms a crucial component for heat transfer. For ease of orientation and to enable critical interfaces to be analyzed in precise detail, the following section divides the end-to-end topology into six coordinated levels:

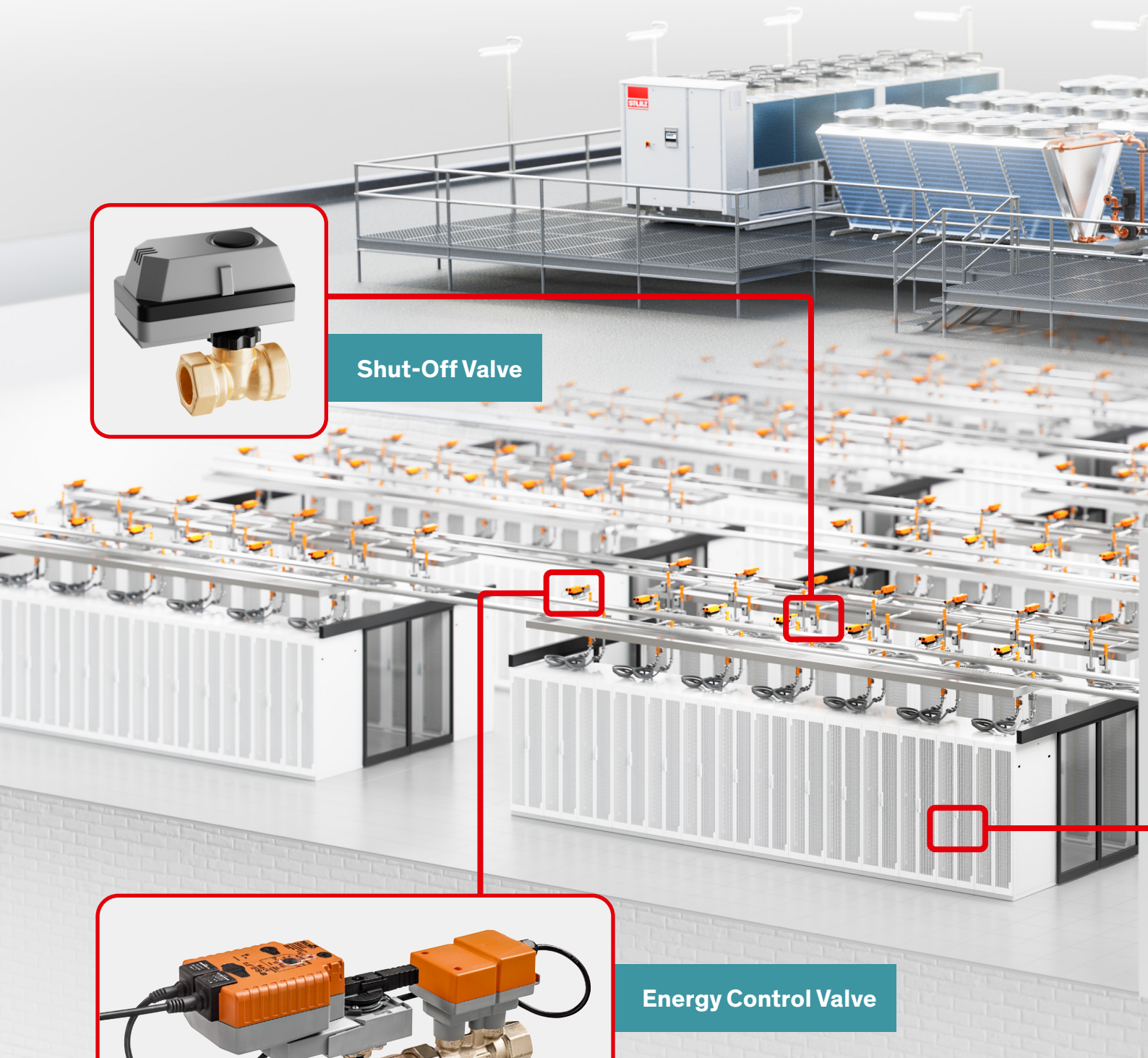
Level 1: Server level

The closed Liquid Cooling circuit begins at the point where the heat is generated directly on the chip. Special cold plates dissipate the thermal energy directly from the CPUs, GPUs, memory modules and voltage converters. The quality of the Thermal Interface Material (TIM) used plays a crucial role in determining how well this dissipation works. To enable server blades to be installed and removed during ongoing operation without any liquid escaping, internal hose systems with process-reliable plug-in couplings (known as quick disconnects) ensure a clean transfer. Additional leak detection sensors in the bottom of the housing also sound the alarm instantly whenever any moisture escapes.

Level 2: Rack subdistribution

The rack level is where the server lines converge – supplied and controlled via in-rack manifolds for supply and return. Isolation valves allow individual slide-in units to be disconnected in a targeted manner, while flow meters, temperature and pressure sensors, and leak trays or leak detection ropes permanently monitor the condition. Depending on the architecture, in-rack CDUs or Rear Door Heat Exchangers (RDHs) are also used to absorb load peaks directly on site.

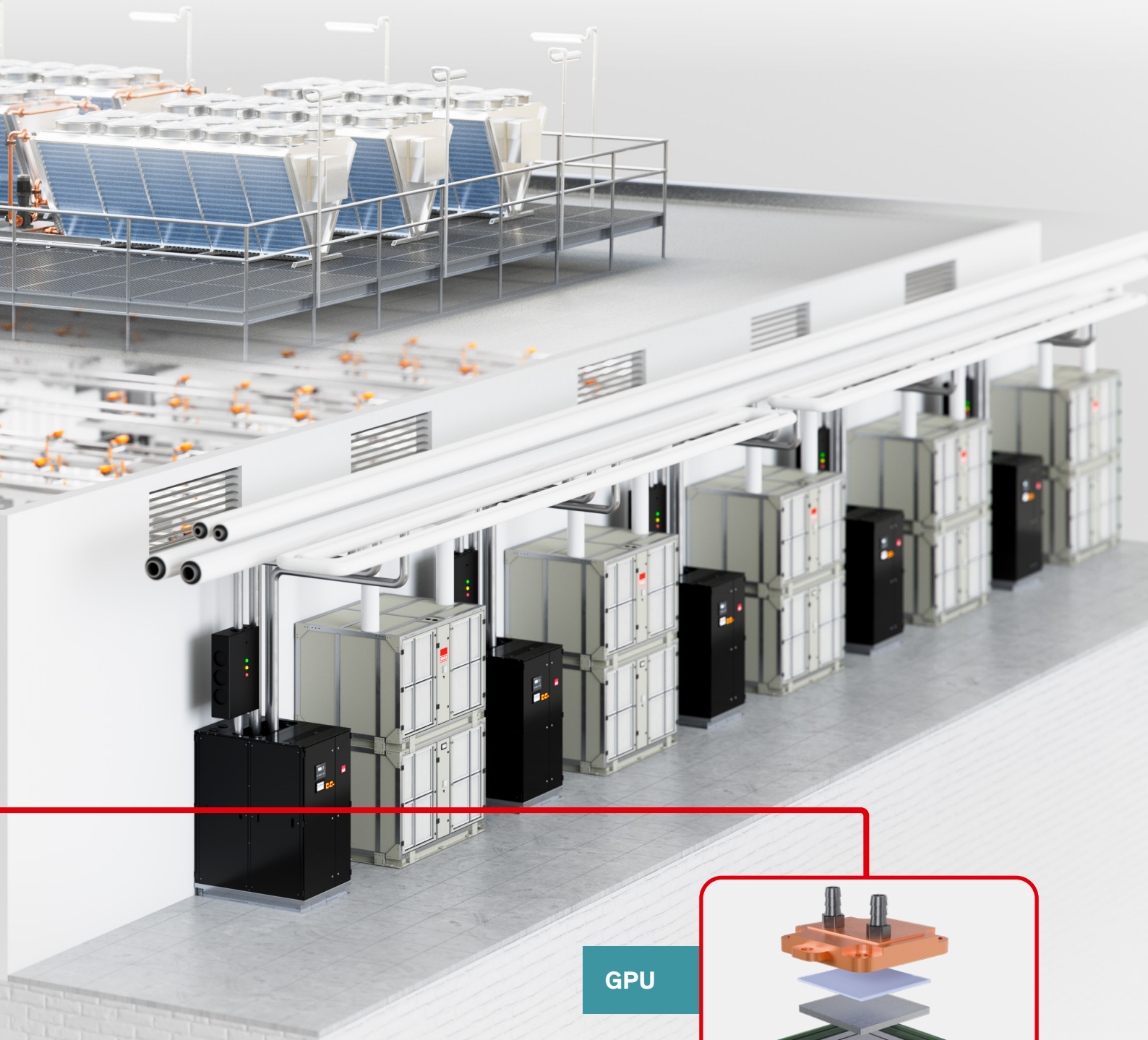
DATA CENTER LIQUID COOLING TOPOLOGY - END TO END VIEW



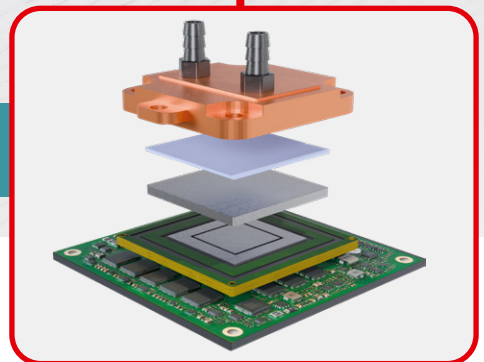
Shut-Off Valve



Energy Control Valve



GPU



Level 3: Row and zone distribution

Within the zone distribution, the racks are hydraulically connected to the higher-level building infrastructure. Row manifolds as well as balancing valves and check valves ensure that the flow rate is distributed evenly. Pressure-independent control valves (PICV / DP controlled) or intelligent energy valves control the mass flow dynamically according to the level of demand. They reliably prevent hydraulic unbalanced loads, while drain and fill valves and flexible connections compensate for any thermal expansion, making it easier to perform subsequent maintenance and service work. It is also crucially important to create technical levels of redundancy at level 3, for example through dual A/B line routing (dual feed) with isolation valves in pairs, which allows individual line sections to be serviced during ongoing operation without interrupting the supply to the remaining racks.

Level 4: Building services management system

At the building level, all supply lines converge in the main supply and return headers. Central flow rate measurements and temperature and differential pressure sensors provide an overview of the entire system. Safety valves, a central maintenance bypass and dedicated flushing connection and commissioning points provide the basis for smooth commissioning and stable operating conditions over a long period of time. At the same time, the treatment and monitoring of water quality needs to be firmly embedded as a continuous operating process here. Constant monitoring of the pH value, electrical conductivity and inhibitor concentration level prevents fouling and corrosion and ensures efficient heat transfer throughout the entire building network.

Level 5: Central Coolant Distribution

The Coolant Distribution Unit (CDU) connects the primary cold water supply to the secondary rack circuit as a central hydraulic separating layer. Internal plate heat exchangers ensure strict separation of media. Circulation in the secondary circuit is ensured by controlled pumps, with reliability being provided by coordinated group control with N+1 redundancy at the equipment level. An integrated PLC coordinates all control loops. The CDU is supplemented by dirt traps, sampling ports for water analysis, expansion vessels, electronic pressure maintenance, degassing units for microbubbles, and chemical dosing units.

Level 6: Dry Coolers

The chain ends with the dissipation of the heat into the environment. Dry Coolers, adiabatic dry coolers, chillers, optional cooling towers and Free Cooling or economizer coils work hand in hand here. Heat rejection control valves and precise outdoor temperature detection also ensure that, depending on the time of year, the optimum operating point between mechanical chilling and Free Cooling is maintained.

Engineering checklist: Overarching planning principles for Liquid Cooling systems

- **Consistently avoid single points of failure:** Critical systems and units – from central CDUs and dry coolers to control systems and current paths – need to be designed with some redundancy (e.g. N+1 or 2N). Redundancy at the equipment and system level ensures that the failure or maintenance of one unit does not adversely affect cooling in the white space.
- **Enable maintenance work to be done during ongoing operation:** Strategically placed isolation valves and bypasses allow components, filters or manifolds to be replaced without having to reduce the IT load in the white space.
- **Modularity and scalability:** Think in modular and scalable terms right from the outset: The main piping and distribution headers should be designed to be sufficiently large that future generations of chips with a higher power density can be integrated without having to make any structural interventions.
- **Quality assurance:** Efficient, reliable operation of the system as a whole can only be guaranteed at all times through careful hydraulic planning and coordination right across the system.

4. Planning strategies for green-field and retrofit architectures

While the first generation of AI clusters was often integrated into existing data center environments designed primarily for air cooling, new generations of GPUs with power densities of 100 kilowatts per rack and well beyond require a more in-depth planning approach. The central question that now needs to be asked is no longer whether Liquid Cooling is deployed, but how the overall system comprising a chilled water connection, Coolant Distribution Unit (CDU) and Liquid Cooling subdistribution can be adapted to suit different structural realities and customer expectations in an error-free, energy-efficient and scalable manner.

The wide variety of geographical, climatic and architectural factors mean that the requirements for comprehensive implementation of AI infrastructures sometimes have very different impacts. Planning engineers are faced with two fundamental scenarios to start with: creating a new design on a greenfield site or the challenge of modernizing an existing site (retrofit/brownfield).

4.1 The thermodynamic foundation: System separation and temperature spread

Regardless of the structural context, the Coolant Distribution Unit forms the crucial link between the primary cold water supply for the data center and the secondary, closed circuit of the LC system. As central chillers – indoors or placed on the roof – are designed primarily for the moderate temperature levels that are found in conventional air cooling areas, the design of high-performance CDUs requires strict hydraulic system separation. It is only this system decoupling that will ensure the primary circuit remains untouched while the CDU on the secondary side provides the precise flow rates, pressures and liquid qualities that modern AI clusters require.

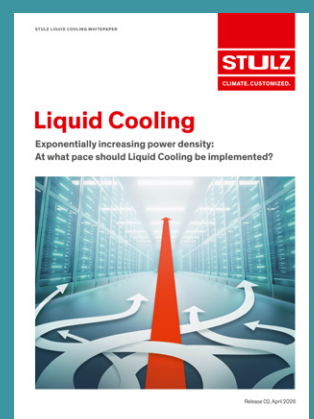
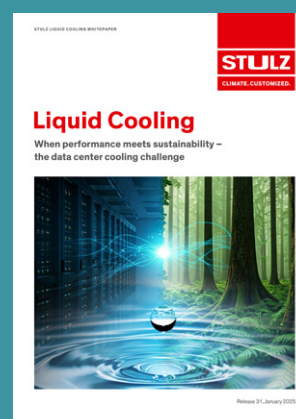
A key lever that can be used to increase the overall level of energy efficiency is precise optimization of the approach temperature at the CDU's internal heat exchangers. If it proves possible to reduce the difference in temperature between the outgoing secondary medium and the incoming primary medium from a previously conservative 4 kelvin to an ambitious 2 kelvin, this fundamentally changes the entire thermal balance model for the data center.

+ Take AI-ready infrastructure to the next level

Explore our Liquid Cooling whitepapers for further insights into modern cooling strategies.

>> [Liquid Cooling: When performance meets sustainability](#)

>> [Liquid Cooling: Exponentially increasing power density](#)



APT comparison			
	4K APT	2K APT	2K APT (two units)
Cooling Capacity	1,000 kW	1,000 kW	500 kW
ΔT TCS	10 K	10 K	10 K
ΔT FWS	10 K	6 K	8 K
APT	4 K	2 K	2 K
TCS Flow / Capacity	1.5 LPM/kW	1.5 LPM/kW	1.5 LPM/kW
TCS ESP	250 kPa	250 kPa	250 kPa
TCS Pump Power Consumption	13.7 kW	13.7 kW	5.9 kW



This physical correlation has a direct impact on the facility water system (FWS): If the inlet temperatures on the chiller and dry cooler side are raised accordingly, the chillers operate with a significantly higher coefficient of performance (COP). However, if the primary chilled water temperatures and flow rates need to be strictly maintained for structural or operational reasons, the system compensates for the increased thermal load solely by adjusting the surface areas of the plate heat exchangers in the CDU. This is why early on in the design phase engineers need to carefully weigh up whether the primary network is capable of being adapted flexibly or whether additional heat exchanger capacities need to be provided in the secondary line.

Power Consumption			
Temp in/out	Modes	%	Power Consumption (kWh/a)
31/21	Dx	0.50	235539
	Mix	83.10	
	FC	16.40	
29/23	Dx	1.80	252496
	Mix	82.40	
	FC	15.80	
31/23	Dx	0.50	213627
	Mix	79.60	
	FC	19.90	



Engineering checklist: Thermal principles and system separation

- **Approach temperature:** Lowering the temperature difference from 4 K to 2 K will maximize efficiency, but requires precise coordination of the primary and secondary circuits.
- **FWS link:** Unchanged chilled water temperatures in the existing network at the building will require correspondingly larger plate heat exchanger surface areas in the CDU.
- **Hydraulic** decoupling: Strict separation protects the system as a whole and guarantees that specific rack flow rates can be provided precisely.

Greenfield architectures: Maximum scalability for new designs

The fundamental design of a new data center presents a rare opportunity to model the entire cooling and air-conditioning infrastructure from scratch to reflect the requirements of Liquid Cooling. Nevertheless, even a new design comes with its own planning challenges, especially with regard to the long-term evolution of AI loads.

Modular zones and scalable room concepts

Rather than immediately equipping the building with high-pressure piping throughout, modular zoning has proven to be a good option. The infrastructure can then be expanded as required alongside the rollout of new chip generations (ranging from Blackwell with 125 kW to future architectures that will exceed 500 kW). CDUs and distributor skids are positioned locally within the rack rows to minimize the pipe lengths and the associated pressure drops in the secondary circuit.

Dimensions of the main piping

As the thermal loads for each rack are growing at an ever faster rate, the main piping in particular needs to be provided with strategic power reserves right from the outset. Primary and secondary lines need to be designed so that subsequent adjustments to the flow rates can still be made without needing to interfere with the building's structural fabric. Openly accessible pipe routes will prevent any concealed bottlenecks and make it easier to expand the infrastructure while the system is operating in the future.

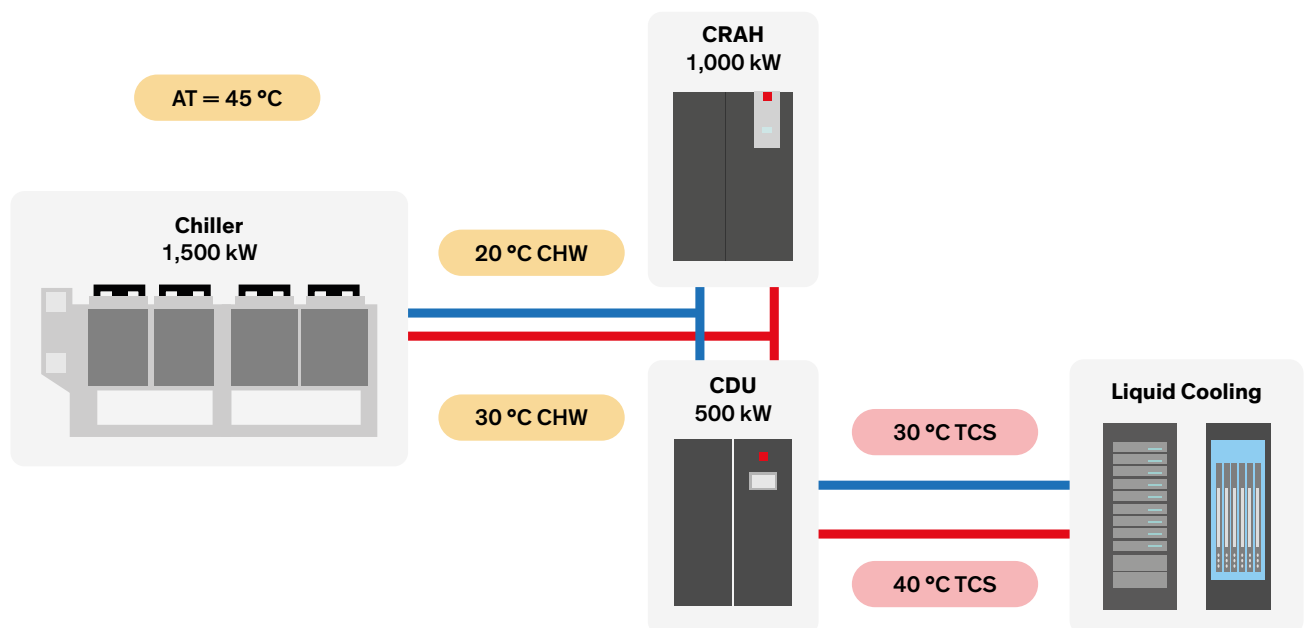
Engineering checklist: Greenfield practice

- **Modular zoning:** Investments are equalized because the infrastructure is scaled to reflect demand as new chip generations emerge.
- **Power reserves:** Generously dimensioned primary and secondary lines prevent the need for expensive modifications as rack densities increase.
- **Accessible piping:** Open pipe routes safeguard the long-term capacity for the system to be maintained and expanded with the system still in operation.

4.3 Retrofit and brownfield architectures: The art of integration in an existing facility

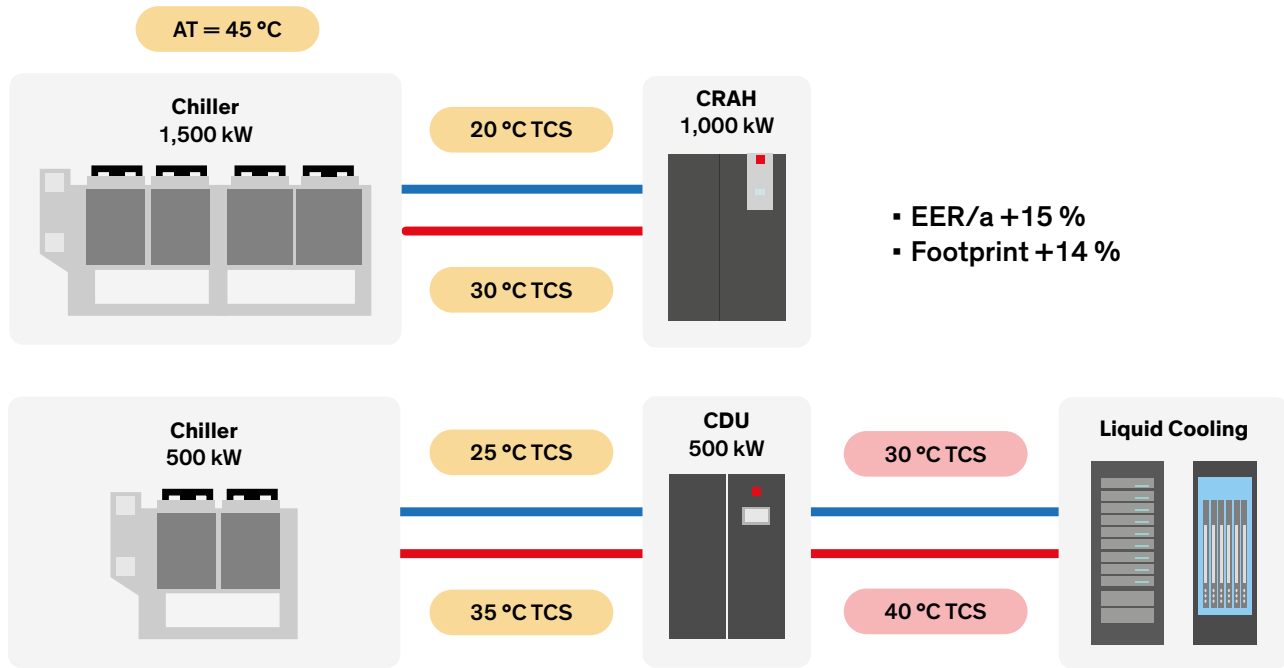
Retrofitting within existing data centers that are designed primarily for air cooling presents a far greater challenge. This is where a lack of space, static loads and hydraulic restrictions all collide. In existing data centers, new chilling components have to compete directly with the very limited amount of space available in the raised floor or need to be capable of being integrated in the rack rows alongside existing side coolers. To save valuable space for IT equipment, compact in-rack CDUs or space-saving row solutions will be used, depending on the density and room concept. For high total loads, it is advisable to use central industrial CDUs or modular skid solutions outside the pure white space to enable the cooling capacity to be provided without losing server footprint.

Thermodynamic exploitation of the temperature levels can offer a key approach for optimizing the outcome in retrofit projects. Retrofitting high-density AI racks quickly pushes existing chillers and dry coolers to their performance limits. However, instead of replacing all the mechanical chilling at great cost, the load on the overall system can also be eased via the secondary side. Operating with increased coolant temperatures will protect the existing infrastructure. However, this requires the tolerances of the server cold plates to be approved for the new temperature window. This workaround ensures that Liquid Cooling systems can be used reliably and increases the cost-effectiveness of the existing facility at the same time. Last but not least, a retrofit always requires hydraulic integration. Liquid Cooling is generally operated in parallel with the legacy air cooling systems. To rule out any instability here, strict hydraulic separation of the circuits via valves is absolutely essential. Additional professional bleeding concepts and integrated fine filters can also be implemented to prevent cavitation in the pumps and deposits in the microchannels of the server cold plates as well.



Engineering checklist: Retrofit practice

- **Space optimization:** Use of compact, wall-integrated units or external skids to overcome shortages of space in an existing building.
- **Temperature flexibility:** Allowing higher coolant temperatures can prevent the need to fully replace existing chillers.
- **Operational reliability:** Strict separation of circuits and filtration equipment will protect both existing and new systems during operation.



4.4 Summary of the key planning principles

Whether it is a greenfield or retrofit site, successful delivery of future-proof Liquid Cooling is ultimately based on four irrefutable engineering principles:

- **Consistently avoid single points of failure:** Critical systems and units need to have a redundant design (e. g. N+1 at the equipment level) to protect them against partial failures.
- **Live maintainability and uninterruptible service:** At the rack and server level, drip-free quick disconnect couplings (QDs) allow components to be replaced during operation without the need to use tools. At the overarching distribution and unit level, strategically placed isolation valves and consistent unit redundancy ensure continuous cooling. This means that any maintenance work or replacement of components can be done without interruption or load shedding in the white space.
- **Modular scalability:** The infrastructure needs to be flexible enough to accommodate changes in chip power densities and changing coolant properties.
- **IT infrastructure coordination at an early stage:** The exact thermal load profiles and pressure drops of the specific server OEMs need to be reflected in the hydraulic planning.

5. Managing transient peaks in demand

Appropriate implementation of Liquid Cooling systems in modern AI data centers is not defined solely by achieving the maximum thermal design load, but primarily by the extreme temporal and spatial dynamics of the load states. While classic cloud workloads and traditional high-performance environments display relatively sluggish and predictable workload curves, training and operating large language models results in sudden load changes. For the system as a whole, this means that the air conditioning, as a flexible control system, needs to react intelligently to transient power peaks within the cooling system parameters.

5.1 The transient problem: From static TDP to dynamic load spikes

Modern AI accelerators and GPU clusters often operate at the physical load limit of the semiconductor material. When a training job starts or large batches of inference are processed, the power consumption of the processors can jump from minimum stand-by or idle loads to the maximum thermal design power (TDP) load in the space of a few seconds.

Two fundamentally different operating modes produce specific load profiles in each case:

- **Training:** This is a computationally intensive phase in which a model learns from large amounts of data and there are very high load peaks, which are often synchronized across many GPUs.
- **Inference:** The operating phase after training in which the finished model processes new data and infers decisions, which tends to result in a more consistent load profile. However, this may be at a constantly high level.

The main reason why traditional air cooling designs are insufficient for handling AI workloads is down to a combination of extreme power density and significant differences in the reaction times, because conventional air cooling systems respond within minutes, whereas within liquid-cooled AI applications peak loads occur in a matter of seconds.

In this context, professionals in the industry often raise the question of whether the thermal inertia of the water can act as a thermal buffer here. As water has a much higher heat capacity than air, a liquid-cooled circuit is in principle better able to absorb spikes in power. However, the physics of the closed secondary circuit will dictate whether the thermal mass of the fluid is sufficient to absorb AI load spikes in the megawatt range or whether the buffer will be used up within a few seconds. As the cooling medium is usually admixed with a glycol fraction (for example, PG25 or 25% propylene glycol), the specific heat capacity drops to around 3.8 kJ/(kg·K) compared to 4,186 kJ/(kg·K) for pure water.

$$t_{Buffer} = \frac{\rho \cdot V \cdot c_p \cdot \Delta P}{\Delta P}$$

Figure 5.1: Formula for estimating the theoretical bridging time of the secondary circuit

The buffer time t_{Buffer} describes the time window until the maximum permissible temperature increase is reached (ΔT). In this formula, ρ represents the density, V represents the effective fluid volume and c_p represents the specific heat capacity of the cooling medium. The parameter ΔP defines the thermal load of the load spike that is temporarily uncovered during the valve transit time and reaction time.

The thermal buffering effect of the secondary circuit can be estimated via the heat energy that can be stored in the fluid. The crucial factor here is not the overall IT workload, but the proportion of a load spike that the active cooling does not dissipate during its reaction time. With the conservative assumption that this additional heat output must initially be absorbed completely by the fluid, a volume of 500 liters of a water-glycol mixture with a specific heat capacity of around 3.8 kJ / (kg·K) and an average temperature increase of 10 kelvin is able to absorb roughly 19 megajoules of heat. This corresponds to a theoretical bridging time of around 38 seconds with an uncovered power output of 500 kilowatts or around 10 seconds for 2 megawatts. In the case of 300 liters, the bridging time is reduced to around 23 or 6 seconds, respectively.

This means that the fluid content can effectively dampen rapid load spikes and so provide additional time for the active cooling to respond. However, it is not a substitute for sufficiently dimensioned dissipation of heat. The buffer time that is actually available will depend on the effective volume of liquid in each case, the material values of the cooling medium, the permissible temperature increase and in particular on how quickly the CDUs and the higher-level cooling system are able to take on the additional heat output.

If the overall system responds with a delay, the processors and GPUs will automatically switch to heat-induced throttling, which will have a noticeably adverse impact on training times and computing efficiency. In addition, time-averaged values should not be used as the basis for a design because they do not adequately reflect short peaks in demand; the decisive factors here are the peak load and the rate of increase.

Modern server cold plates operate with ultra-fine fin structures so they can dissipate peak loads on the chip surface without delay. To ensure that the heat exchangers can always work reliably even under extreme conditions, the chemical stability of the cooling medium needs to be guaranteed in addition to homogeneous mechanical contact pressure on the chip surface. In high-density direct-to-chip systems, the fluid performs many more tasks than merely acting as a transport medium. This is why the term fluid integrity has become

established in modern data center planning. This term encompasses much more than the classic definition of water quality. It is only if the fluid has permanent chemical and physical stability that the heat transfer in the cold plates can be kept constant and the formation of local hotspots can be prevented if any dynamic load spikes occur. Today, integrated fluid management is used as a protective architectural component right from the outset and is no longer considered to be a purely downstream maintenance task.

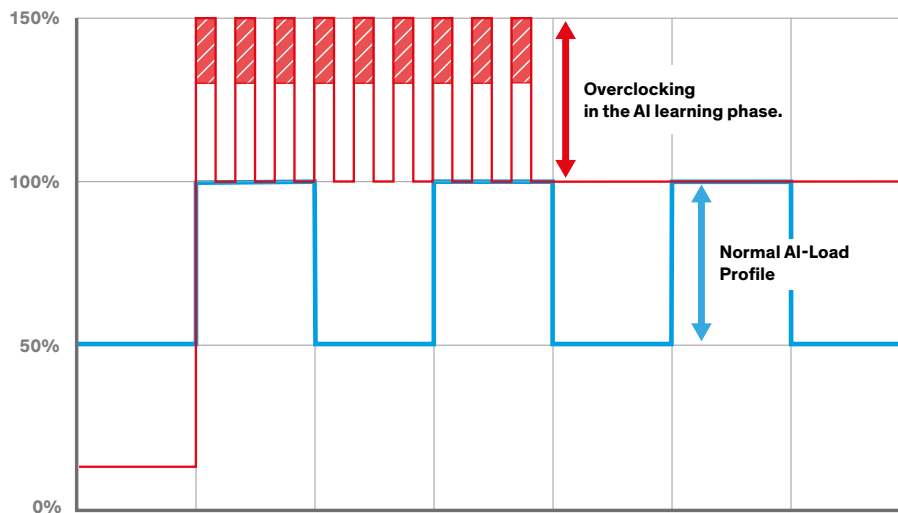
Even in closed secondary circuits, the medium is subject to natural interactions with the materials that are used and changing thermal loads during the years in which the system operates. To maintain the integrity of the system throughout the life cycle of the facility, operators are advised to make the transition from selective spot checks to continuous monitoring of the system. If parameters such as the pH value, electrical conductivity, turbidity and differential pressure trends are recorded constantly via the filter, any gradual changes in the fluid can be identified reliably at an early stage.

To operate a data center reliably without any disruption, it has proven to be a good idea in practice to link this fluid management to a multi-stage maintenance plan. A model like this starts with regular trend analysis to identify minor parameter shifts and, if necessary, evolves into defined warning levels for facility management. This approach makes it possible to implement prompt corrective measures such as controlled redosing of stabilizers or increasing the level of filtration in the bypass flow as required. Such a structured approach gives operators the level of certainty they need to act and ensures reliable protection of the cold plates, the server hardware and all the cooling infrastructure.

5.2 System strategies for managing dynamic loads

With the rapid increases in the performance of modern AI processors, purely reactive readjustment in the refrigerant circuit often encounters physical limits. A changed mass flow simply cannot be transported to the semiconductors in the distribution network without delay. To manage load peaks, it is therefore necessary to implement forward-looking structural and technical system-related measures directly in the hydraulic network.

Hydraulic buffering via generously sized pipe cross sections in the supply lines plays a key role here. This measure reduces the pressure drops during normal operation and, in combination with deliberately placed fluid buffers, provides a sufficient volume of inertia. This reservoir absorbs the first thermal and hydraulic shock for the crucial initial seconds before the higher-level cooling system kicks in.



Management of dynamic load spikes is also based on coordinated unit redundancy in which multiple CDUs work together in accordance with the N+1 principle. If a unit fails or a load peak kicks in, the additional units spread the demand for cooling evenly among each other. This reliably prevents any extreme operating states or uncontrolled hydraulic pressure surges in the system as a whole.

Engineering checklist for managing dynamic AI loads

- **Generously sized pipe cross sections and deliberate fluid buffers in the distribution network** continuously reduce the pressure drops and absorb sudden thermal load surges for several seconds.
- **The design with unit redundancy rather than component redundancy** guarantees the required level of operational stability without any thermal bottlenecks through coordinated operation of a number of active units.
- **The thermal dimensioning** is consistently focused on transient load peaks and their rate of rise and its design is not based on time-averaged measured values.

5.3 The CDU as an integral part of hybrid air-conditioning systems

A significant change is taking place when it comes to the placement and integration of Coolant Distribution Units (CDUs) as traditional data centers have previously operated the cooling largely decoupled from the actual IT load. However, with high-density AI workloads, there is an increasing focus on IT-side parameters such as the component temperatures and the instantaneous power consumption of the processors.

The integration of control technology is based on robust interfaces between the building automation system and programmable logic controllers. The distribution unit no longer acts as an isolated individual unit, but continuously adapts its hydraulic output to the thermal requirements of the hardware to counteract any fluctuations in temperature at an early stage.

Engineering checklist for hybrid air-conditioning systems

- **Deliberately dimensioned buffer volumes** in the TCS circuit in combination with fast-reacting control valves absorb transient load changes for the first few crucial seconds.
- **Careful hydraulic calculation of the pressure conditions** in the pipe network prevents any pressure surges and ensures stable rates of flow at the consumers.
- **The thermodynamic design** takes account of the reduced heat capacity of water-glycol mixtures and is consistently based on real peak loads rather than time-averaged measured values.

Conclusion

Managing highly dynamic AI workloads demands a fundamental departure from isolated approaches. In modern data centers, a Coolant Distribution Unit can no longer be operated as an autonomous building services unit, but must instead be synchronized directly with the thermal load states of the IT infrastructure via programmable logic controllers and automated interfaces.

Neither a passive fluid volume nor oversized individual components are capable of compensating for transient megawatt surges on their own. Uninterrupted operation can only be guaranteed by ensuring smooth interaction between buffers placed deliberately in the distribution network, precisely calculated pressure conditions and interlinked unit redundancy. If these interfaces are already taken into account holistically within the hydraulic design, heat-induced throttling of the semiconductors can be permanently averted and high-density server structures can be operated safely from a thermodynamic perspective.

6. Intelligent regulation and control architectures

When it comes to providing air conditioning in AI data centers, the operational stability of the overall system depends primarily on how quickly and coherently the air-conditioning system as a whole is able to respond to rapid changes in load. Regulation and control technology is therefore evolving from a classic building technology discipline to become a system-critical feature that determines the level of performance, availability and scalability of an AI system. While the focus in traditional data centers was often on static optimization targets, AI workloads require more advanced strategies that are designed to deliver minimal response times and a coordinated response from all subsystems to prevent any instabilities at crucial system interfaces during the planning phase.

The control logic of pumps and valves in modern cooling circuits is based essentially on the classic PID (Proportional-Integral-Differential) controller. These systems continuously compare process variables (actual values) with corresponding setpoint variables to minimize deviations and ensure stable operation when loads change. Advances in switching operation and control for air-conditioning technology are seen not so much in a revolutionary new development of the controller architecture itself, but rather in optimum parameterization and intelligent interaction between all components. Different primary control modes apply, depending on the operational focus and the operator's specific requirements. While differential pressure control (**Δp**) provides stability in particular during intensive training phases where there are massive load spikes, differential temperature control (**ΔT**) offers the advantage of particularly energy-efficient operation under partial load. On the other hand, for dynamic workloads, especially in continuous inference mode, flow control has become established as the preferred method. As a universal guideline for the ratio of the required flow rate to the thermal load, planners use a tried-and-tested anchor value for achieving the correct level of flow rates of approx. 1.5 liters per minute per kilowatt (1.5 LPM/kW). In practice, this requires the conveying capacity to be designed to be close to the maximum load peaks so that the semiconductor layer (GPUs/TPUs) does not overheat even at extreme peaks. Although this will sometimes result in a thermodynamically less efficient mode of operation in real partial load mode, this remains technically essential to avoid any heat-induced reduction in performance.

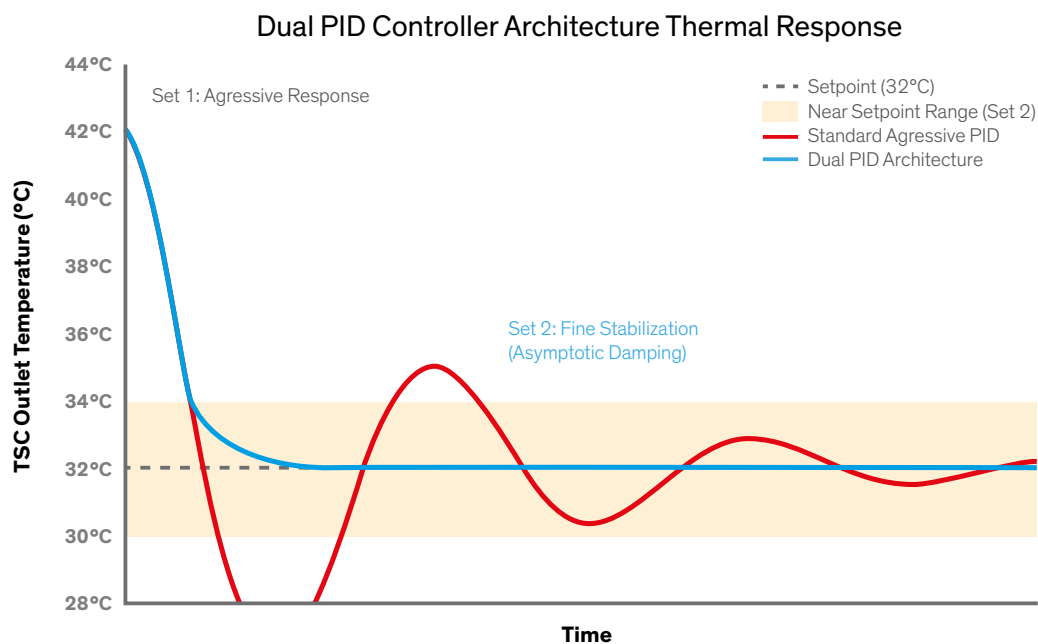
Dual PID parameter sets for flexible control tasks

To strike the right balance between rapid readjustment when large deviations occur and a sufficient level of stability in the near setpoint range, advanced control systems work with dual PID parameter sets. One parameter set is optimized for a rapid approach for the range far away from the setpoint, while the second parameter set provides granular, vibration-free stabilization in the range close to the setpoint. This is particularly crucial when multiple Coolant Distribution Units (CDUs) are connected to a shared facility water system

(FWS). Without this differentiated control, there is a risk of mutual hydraulic interference. This can cause a phenomenon that is also commonly referred to as “water theft”, which is where sudden valve movements for one unit produce uncontrolled vibrations in the system as a whole.

The higher-level coordination of N+1 CDU groups is performed via an intelligent group control system in which a master controller (lead unit) compares the sensor values and carries out a time-based rotation of the load of the units. The grading of the units is regulated using defined threshold values, typically if the active units have a valve opening or pump speed of more than 80 percent over a prolonged period of at least ten minutes or, conversely, initiate a deactivation if this percentage drops below 40 percent. To ensure a high level of availability, planners can choose between a traditional cold stand-by architecture with uniform wear caused by rotational operation and a permanently running hot stand-by in partial load mode, with the latter enabling even faster reaction times if a component fails.

These kinds of architectural protection mechanisms can be supplemented by specific control functions, such as those required for certification for demanding hyperscalers. To guarantee a TCS outlet temperature that is permanently below 35°C even when extreme load spikes occur, a configurable FWS valve preopening feature is used. This proactively absorbs thermal load peaks before they reach the sensitive IT equipment.



Summary

In AI environments, stable operation depends less on isolated individual components and much more on how quickly and coherently the overall cooling system is able to respond to rapid changes in load. Modern control strategies are consistently focused on short reaction times, a coordinated response across all subsystems and avoiding any instability at the interfaces. This transforms the control of a cooling system from a pure HVAC discipline to a central, cross-system factor in the architecture of data centers.

Engineering checklist: Intelligent regulation and control architectures

- **Control modes:** Deliberate design for Δp (training), ΔT (energy efficiency) or flow rate control (inference).
- **Dual PID parameters:** Implementation of separate parameter sets to minimize vibrations at the FWS interfaces.
- **N+1 group grading:** Configuration of defined load thresholds ($\geq 80\%$ / $\leq 40\%$) for valves and pumps.
- **Redundancy management:** Choice between cold and hot stand-by including automatic master failover.
- **Temperature cutoff:** Use of proactive control functions such as FWS preopening to maintain strict upper limits (e.g. $< 35^\circ\text{C}$).

The STULZ GmbH

STULZ is a global provider of innovative, mission-critical cooling solutions. Founded in 1947 and with its headquarters in Hamburg, Germany, it is represented by 35 subsidiaries, a network of more than 150 partners and ten production sites worldwide. Its solutions provide precise temperature and humidity control, as well as reliable cooling capacity, to safeguard the operation and longevity for data centers, telecom infrastructure and industry. STULZ prioritises energy efficiency and sustainability in its product development and manufacturing processes to deliver future proof cooling solutions.

For further information, please visit our website at: www.stulz.com

About the Author

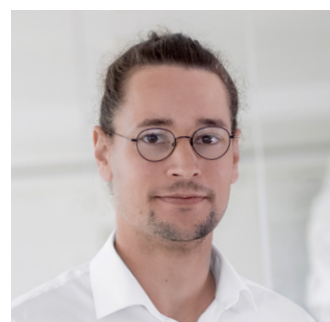
Jonathan Aaron Jürgens is Team Lead Liquid Cooling Global Product Management at STULZ GmbH and the company's in-house expert on liquid cooling infrastructures. Drawing on experience from more than 300 LC projects worldwide, he drives the global development and market implementation of scalable cooling solutions for data centers and AI factories. Jonathan has authored three white papers and co-authored the Amazon bestseller Greener Data: Volume Three, where he addresses sustainability across liquid cooling topologies.



As Co-Captain of the Talent in Digital Infrastructure initiative at Datacloud Global Congress in Cannes, he also supports the development of the industry's next generation. Through international seminars and close collaboration with customers, partners, and stakeholders, he advances high-performance, resilient, and sustainable digital infrastructure.

About the Author

Daniel Ducci is Product Owner for Liquid Cooling Software at STULZ Digital Solutions GmbH, where he drives the development of software solutions for liquid cooling infrastructures in data centers. Holding a master's degree in Energy Systems Engineering, Daniel is an intrinsically motivated engineer who has continuously broadened his technical expertise — starting in mechanical engineering, moving into electrical engineering, and now specializing in software development. Before taking on his current role, Daniel gained hands-on engineering experience on Coolant Distribution Unit (CDU) projects, giving him a deep understanding of liquid cooling systems from a mechanical, electrical, and software perspective. Today, he brings this cross-disciplinary background together to shape intuitive, reliable software that supports the next generation of liquid cooling technology.



CLOSE TO YOU AROUND THE WORLD

STULZ Company Headquarters

STULZ GmbH

Holsteiner Chaussee 283
22457 Hamburg
Tel. +49 40 5585-0
products@stulz.de

For further information, please visit our
website at www.stulz.com or download our
STULZ Products and Services app.

AUSTRALIA
AUSTRIA
BELGIUM
BRAZIL
CHINA
FRANCE
GERMANY
INDIA
INDONESIA
IRELAND
ITALY
MALAYSIA
MEXICO
NETHERLANDS
NEW ZEALAND
POLAND
SINGAPORE
SOUTH AFRICA
SPAIN
SWEDEN
THAILAND
UNITED ARAB EMIRATES
UNITED KINGDOM
USA

**+ 150 PARTNERS
WORLDWIDE**