

White paper

AI Workload Management: Designing efficient cooling and power architectures

Highlights the dynamic link between GPU/AI spiky workloads, cooling demand, and power distribution.





AI WORKLOAD MANAGEMENT

Designing Efficient Cooling and Power Architecture

Executive Summary

The rapid proliferation of artificial intelligence (AI) applications is transforming data centers in Asia, Pacific region into high-density computing environments, where graphics processing units (GPUs) for AI drive unprecedented computational demands. However, AI workloads are inherently dynamic, characterized by spiky and bursty patterns that create cascading effects on power distribution and thermal management systems.

In this paper, we explore the interplay between these workloads, the power train (from grid to rack-level delivery) and the thermal chain (from heat generation to dissipation). By examining a sample AI training scenario, we demonstrate how data center infrastructure solutions regionally available in APAC region —such as Vertiv's large power converters, the Vertiv™ CoolChip CDU, and advanced chillers—mitigate spikes and support efficiency, and scalability. These integrated approach can reduce peak demands, minimize thermal shocks, and maintain operational envelopes, helping data centers support next-generation AI workload management without compromising grid stability or equipment longevity.

1. Understanding AI workload profiles: The rise of spiky and bursty demands

AI workloads diverge significantly from traditional steady-state computing tasks, such as web hosting or database operations, which exhibit predictable, linear power and thermal profiles. In contrast, modern AI applications, particularly large language model (LLM) training and inference, feature highly variable, non-linear patterns driven by parallel GPU utilization. Steady profiles might involve consistent low-to-moderate loads (e.g., 20-50% utilization during routine simulations), but spiky and bursty profiles dominate in AI scenarios.

Currently, the aggregated AI IT load is likely dominated by training large models, which demands significant computational resources. However, as AI adoption accelerates and new use cases emerge, running trained models for real-time applications—will likely surpass training in resource consumption. This shift will make spiky, unpredictable inferencing workloads increasingly critical, requiring robust infrastructure to handle dynamic demand efficiently.

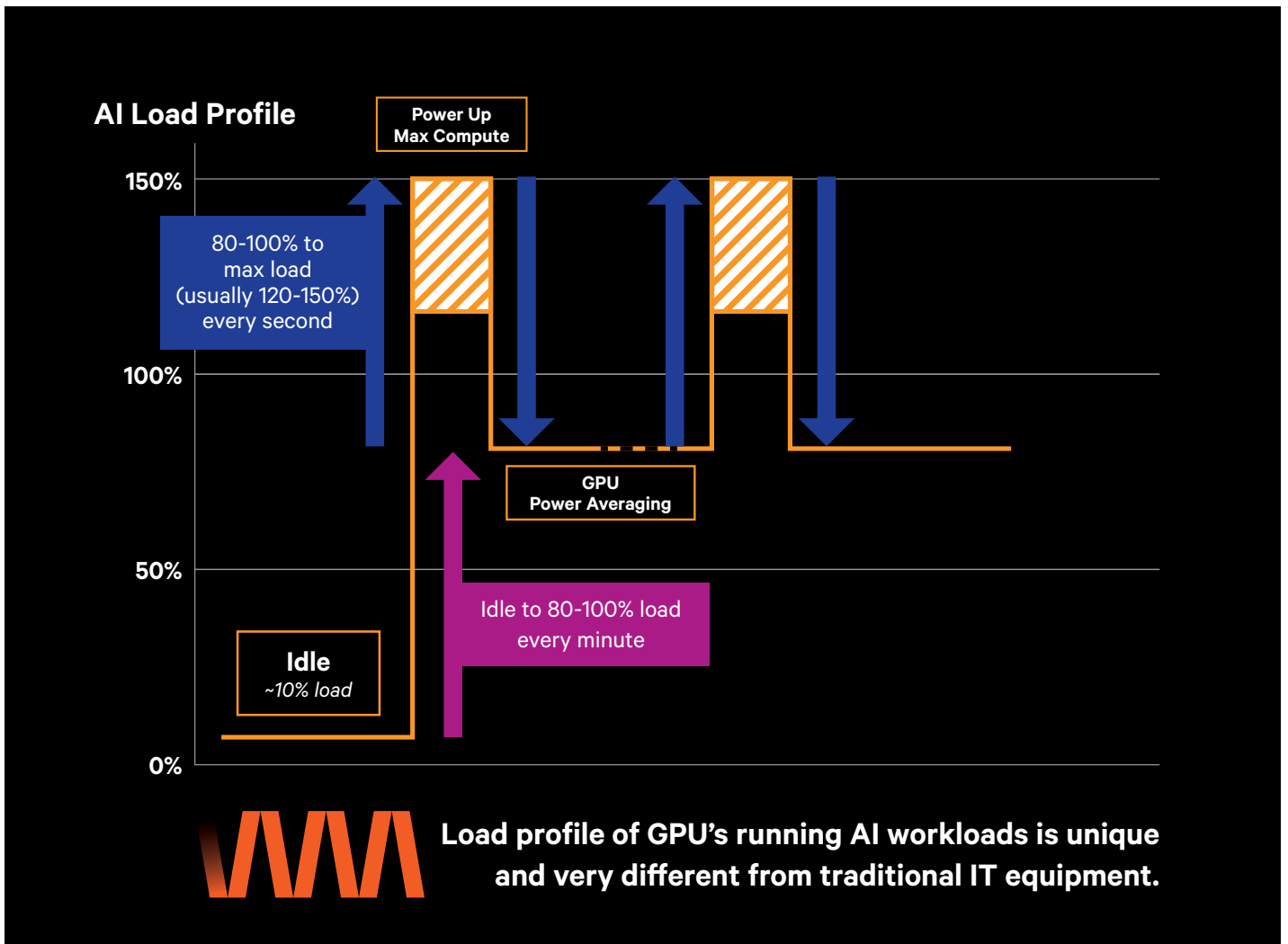


Figure 1. Workload profile in typical AI application

Bursty workloads occur during intensive phases like data ingestion or gradient computations, where GPU clusters ramp up to 150% utilization in seconds. Spiky profiles manifest as short, intense peaks, often lasting milliseconds to minutes, triggered by AI's parallel processing nature: thousands of GPUs synchronize for matrix multiplications, causing instantaneous power draws. In data centers, this translates to grid strain, with peaks overwhelming transformers, generators, batteries and feeders, while thermally, it generates localized hotspots exceeding 80°C on GPU dies. Without mitigation, such variability risks voltage sags, frequency instability, and premature hardware failure, underscoring the need for adaptive infrastructure.

2. The Thermal Chain and Power Train: A sample AI workload implementation

To illustrate the dynamic linkage and addressing the rise in AI power demand, consider the power train and thermal chain in a hyperscale data center supporting AI model training. The ****power train**** encompasses the electrical pathway from utility grid to end-use: high-voltage transmission, step-down transformers, uninterruptible power supplies (UPS), power distribution units (PDUs), busways and rack-level PSUs. It must handle transient currents without cascading failures. The ****thermal chain****, conversely, traces heat flow from GPUs through cold plates, coolant distribution units (CDUs), facility water loops and chillers to ambient rejection.

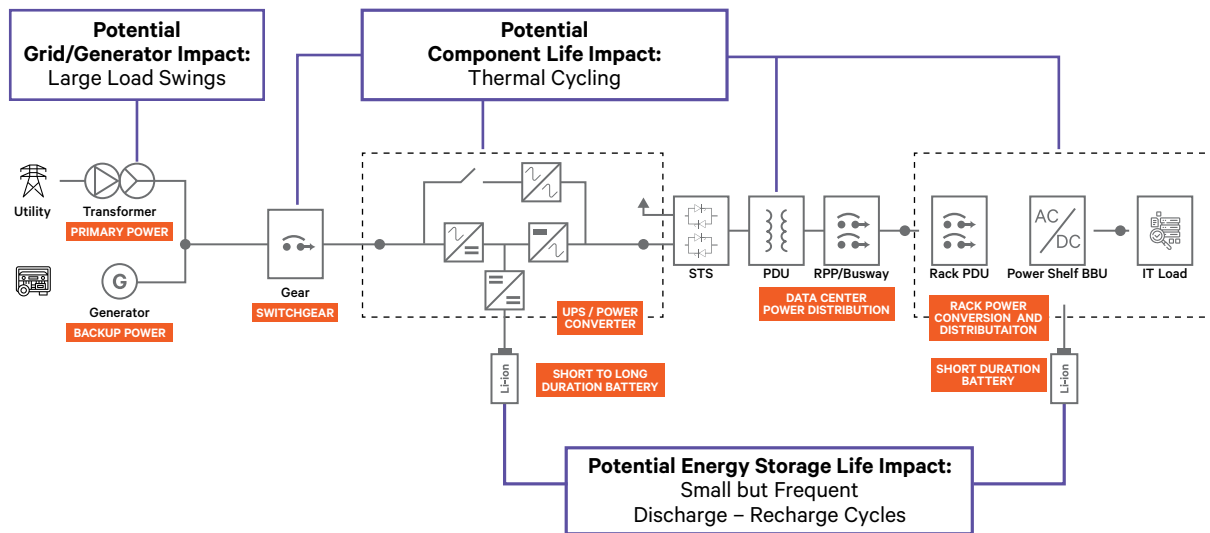


Figure 2. Impact of spike IT load in different component of power train

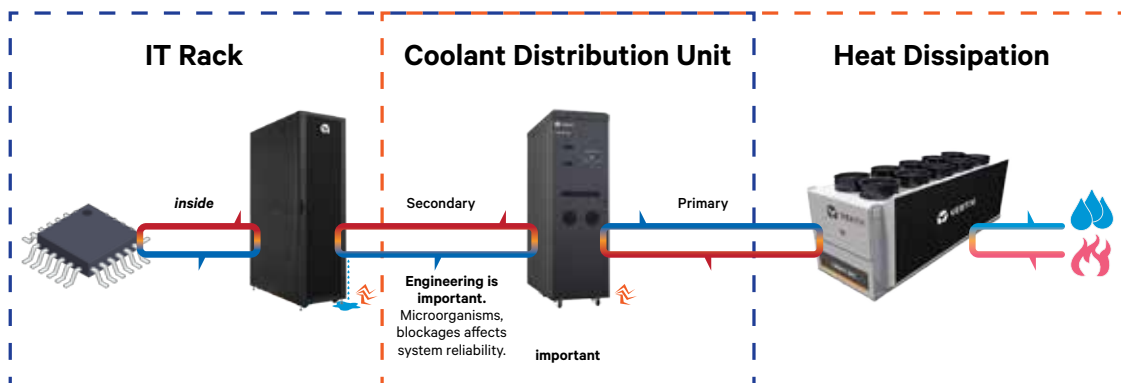


Figure 3. Thermal chain, from Chip to heat dissipation

3. Impacts on the Power Train: Rack-level mitigation with NVIDIA GB300 PSU

Addressing AI power demand first starts manifest at the rack level, where GPU clusters demand rapid power ramps that traditional PSUs cannot sustain without oversizing infrastructure by 50% or more. The NVIDIA GB300 NVL72 platform addresses this through its innovative PSU design, incorporating onboard energy storage, ramp-rate control, and GPU burn-in features to smooth transients. This advancement effectively mitigates power spikes resulting from AI workloads and contributes to a remarkable reduction in peak grid demand by as much as 30%. Excitingly, this technology is soon to be implemented in GB200 NVL72 systems, further enhancing their efficiency and performance.

During a burst in AI power demand, The NVIDIA GB300 NVL72 rack system, which includes the PSU, mitigates significant power spikes by primarily utilizing large-scale energy storage (capacitors and potentially integrated battery systems) and advanced power management algorithms. The energy storage acts as a high-speed buffer, rapidly discharging power to meet the sudden, brief (transient) power demands GPUs, thereby smoothing out these extreme peaks so they are not immediately drawn from the utility grid. Concurrently, the smart power management software implements features like controlled power ramping during workload startup and shutdown (sometimes referred to as a “power burner” mode) to prevent massive, instantaneous surges or

drops in power consumption, ensuring the entire data center's power draw is presented to the grid as a much flatter, more consistent, and predictable load, which is critical for maintaining grid stability.

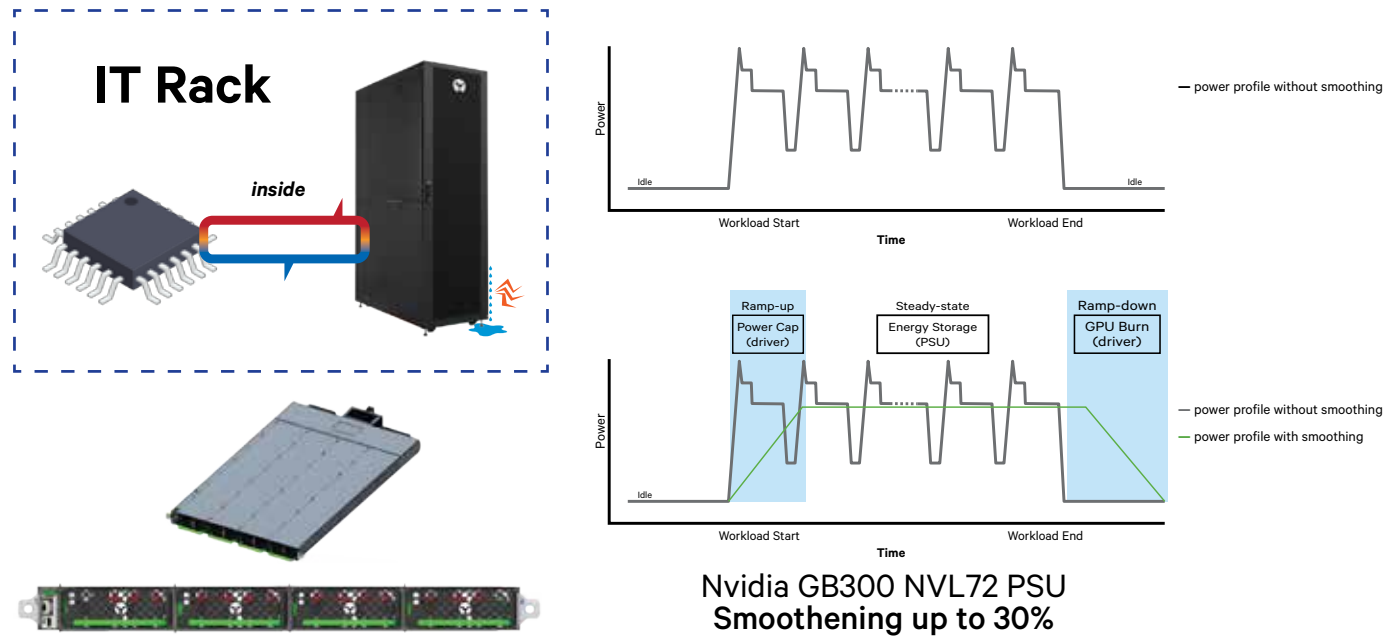


Figure 4. Nvidia GB300 NVL72 PSU, IT load spike mitigation process in rack level

4. Upstream Power Stabilization: Vertiv™ PowerUPS 9000 for spike load mitigation

While rack-level solutions at NVIDIA GB300 handle initial transients, facility-scale spikes propagate to the UPS for NVIDIA GB200 and H100, where sustained bursts can deplete batteries or trigger bypass modes, risking downtime. The Vertiv PowerUPS 9000, a modular double-conversion online UPS, helps mitigate these by providing instantaneous power conditioning and transient ride-through.

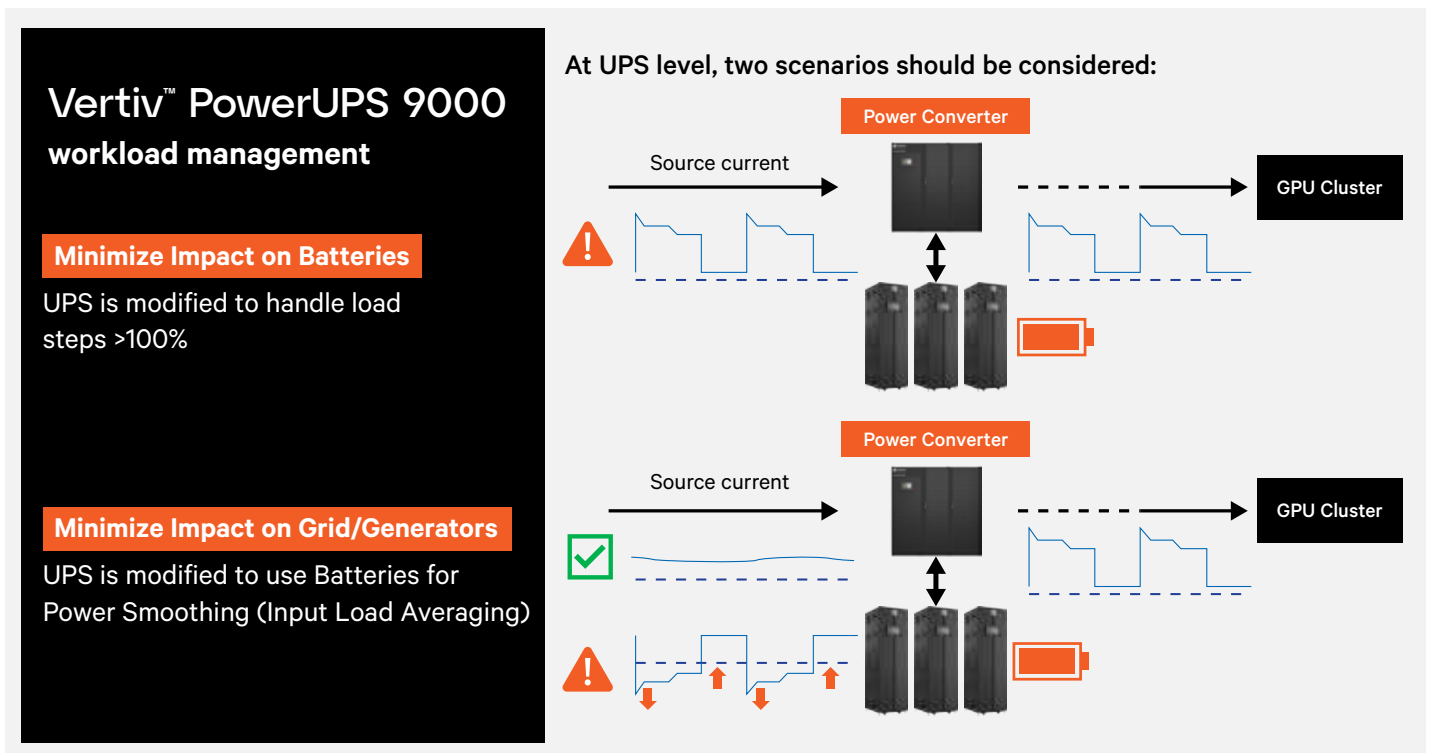


Figure 5. Vertiv™ PowerUPS 9000, IT load spike mitigation solution

Rated for 100-1200 kVA, the Vertiv™ PowerUPS 9000 achieves 97.5% efficiency in double-conversion mode (up to 99% in dynamic online), helping maintain minimal heat generation. Its rectifier draws from a wide input voltage window (e.g., $\pm 20\%$ nominal), absorbing IT load fluctuations before reflecting on grid, which supplies clean sine-wave output to the load with low THD (total harmonic distortion). In the sample AI burst scenario, a rack spike is absorbed via the UPS's intelligent power conversion, helping reduce battery discharge cycles that could shorten lifespan. Enhanced history logs and validity tests further allow predictive maintenance, reducing MTTR (mean-time-to-repair). Thus, the Vertiv PowerUPS 9000 acts as a resilient buffer, harmonizing rack variability with grid supply.

5. Addressing the thermal chain: CDU remedies for thermal spikes

Shifting to the thermal chain, AI bursts generate acute heat fluxes (up to 1 kW/cm^2 on GPUs), demanding responsive coolant modulation to avoid hotspots. Coolant distribution units (CDUs) serve as the nexus, interfacing rack-level direct-to-chip cooling with facility loops. Key remedies begin with inertia buffering and valve precision.

A primary safeguard is the integrated buffer tank (or reservoir), which stores 50-200 liters of coolant to dampen flow temperature surges. During a spike, it supplies excess volume (e.g., +30 LPM) without pump cavitation, stabilizing pressure at 2-3 bar and preventing ΔP (pressure drops) that could starve GPUs. Complementing this, fast-response two-way modulating valves—actuating in <30 seconds—throttle flow bidirectionally, throttling hot return to maintain outlet temperatures. In three-way configurations, these valves bypass excess coolant during bursts, reducing thermal inertia lag by 50% and eliminating shocks that degrade GPU yields by 10%.

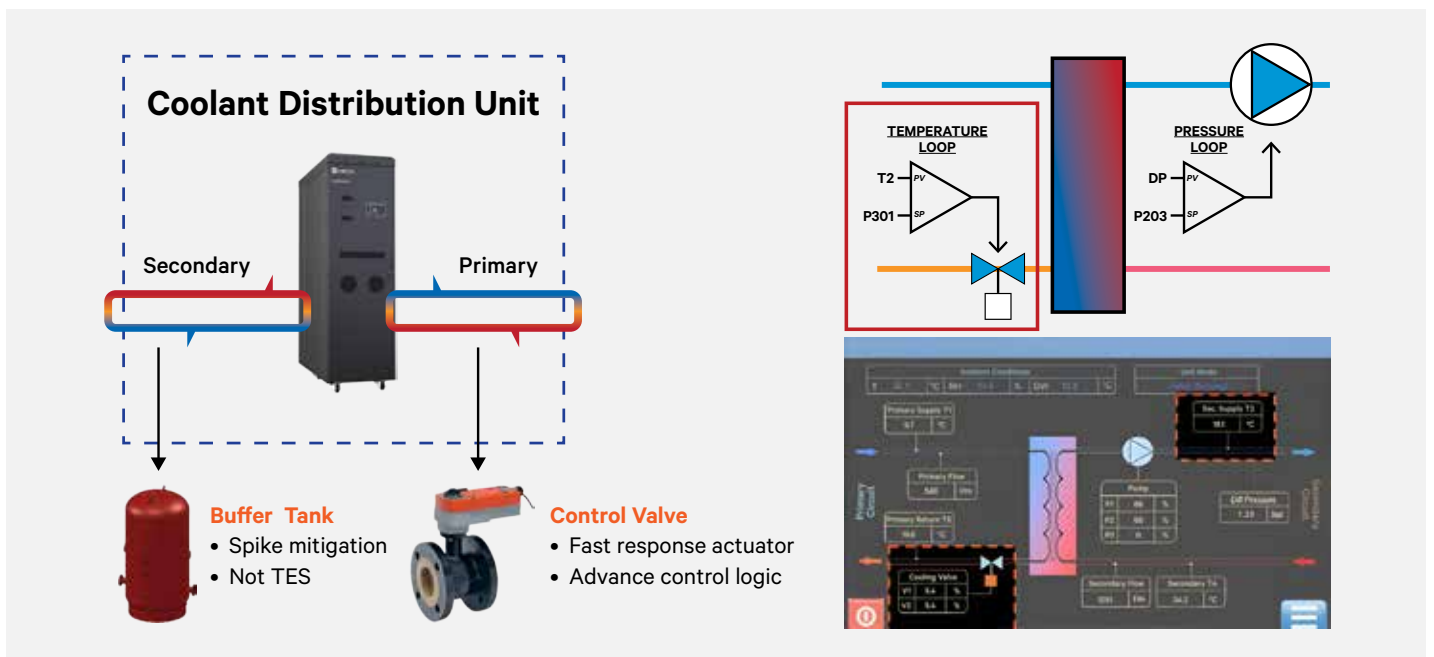


Figure 6. IT load spike mitigation solutions around CDU

6. Refining CDU Control logic

Elevating the basics with intelligent controls tailored for AI variability by employing predictive algorithms, monitoring upstream workload signals (e.g., via Modbus/IP from cluster managers) to preemptively adjust valve positions by 10-20% before spikes arrive. This proactive tuning — versus reactive feedback — aims to improve settling time, helping precise temperature control that eliminates thermal shock for CPUs and GPUs.

7. Chiller-level safeguards: Auto-protect for operational integrity

When a chiller experiences a spike in return water temperature, indicating a sudden surge in cooling demand (e.g., a +50 kW thermal load), the controller in chillers activates the auto-protect feature to manage compressor operation and maintain system efficiency. Integrated via LAN-based sensors, the controller continuously monitors return water temperature, evaporator temperature, and system load. Upon detecting a thermal spike, it initially increases loading by ramping up refrigerant flow to meet the demand while aiming to stay within safe operational parameters (e.g., 20-80% compressor part-load).



Figure 6. Chiller auto-protect feature, maintain safe compressor operation

If the evaporator temperature exceeds the safe envelope limit despite increased loading, the controller shifts to unloading for protection. It reduces compressor load by throttling refrigerant flow and lowering VSD speed to prevent overheating or surge/stall cycles, which can reduce compressor lifespan by up to 25%. This unloading action stabilizes the system by capping compressor lift at 80%, ensuring operational safety. The controller typically allows this condition to persist for a short duration, as it assesses whether the system can return to the safe envelope through these adjustments. If the evaporator temperature remains above the limit beyond this period, the controller may initiate a controlled shutdown (trip) to protect the compressor, depending on the severity and duration of the exceedance.

Conclusion

The symbiotic challenges of spiky AI workloads demand holistic power and thermal strategies, from NVIDIA's GB300 PSU clipping peaks at the edge to Vertiv's ecosystem—the Vertiv™ PowerUPS 9000, Vertiv™ CoolChip CDU, and auto-protected chillers—providing layered resilience. Implementing these in our sample scenario yields peak reductions, making data centers become future-ready for exascale AI. These innovations can support data centers in meeting current and evolving demands, balancing innovation with sustainability, inviting operators to pilot integrated designs for bespoke optimization.



Vertiv.com

© 2026 Vertiv Group Corp. All rights reserved. Vertiv™ and the Vertiv logo are trademarks or registered trademarks of Vertiv Group Corp. All other names and logos referred to are trade names, trademarks or registered trademarks of their respective owners. While every precaution has been taken to ensure accuracy and completeness here, Vertiv Group Corp. assumes no responsibility, and disclaims all liability, for damages resulting from use of this information or for any errors or omissions. Specifications, rebates and other promotional offers are subject to change at Vertiv's sole discretion upon notice.