

Comparison of Air-Cooled versus Liquid-Cooled NVIDIA GPU Systems

Liquid-Cooled GPU Servers Reduce Power Consumption and Increase Performance

OCTOBER 2025



TABLE OF CONTENTS

Executive Summary.....	2
Introduction.....	2
Methods.....	3
Results.....	3
Facility Analysis.....	7
Key Takeaways.....	8
References.....	8

EXECUTIVE SUMMARY

This study benchmarks identical artificial intelligence (AI) workloads on 8× NVIDIA HGX™ GPU systems with direct-to-chip liquid cooling versus traditional air cooling. Key findings:

- **Thermal Performance:** Liquid cooling keeps GPUs at 46–54°C, compared to 55–71°C for air cooling.
- **Performance:** Liquid-cooled systems deliver up to 17% higher computational throughput during stress tests and 1.4% faster training times for real-world AI models.
- **Energy Efficiency:** Liquid cooling reduces node-level power consumption by an average of 1 kW (16%), with savings scaling from 100–400 W for low-utilization tasks up to 1.5 kW for compute-intensive workloads.
- **Operational Savings:** At 2,000 nodes, this yields an estimated \$2.25 million in annual energy cost reduction; at 5,000 nodes, savings reach \$11.8 million per year.
- **Facility Metrics:** Liquid cooling shifts cooling load from embedded fans (counted as IT load) to centralized systems, providing more accurate infrastructure efficiency metrics.

As AI models continue to grow in size and computational demand, direct-to-chip liquid cooling presents a strategic path to sustainable, high-performance AI infrastructure.

INTRODUCTION

The rapid advancement of AI models has intensified power and cooling demands in data centers. As models grow in size and complexity, effective thermal management becomes critical for both performance and sustainability. This whitepaper overviews how cooling architecture impacts computational efficiency and energy consumption during real-world AI workloads.

Modern AI infrastructure presents unprecedented thermal challenges. The concentrated heat output of densely packed GPU clusters must be efficiently managed not just to prevent hardware damage, but to maintain optimal computational throughput during extended training runs. Cooling is no longer merely an operational necessity—it has become a key design variable with a direct impact on system performance, energy efficiency, and training costs.

High-performance computing has long relied on traditional air cooling, where ambient air is forced across components to dissipate heat. This approach, while straightforward to implement, introduces significant power overhead as computational loads increase. The embedded cooling fans contribute to both direct energy consumption and less obvious inefficiencies in thermal management that compound at scale.



By contrast, direct-to-chip liquid cooling systems leverage the superior thermal conductivity of liquid coolants to remove heat more efficiently from processing components. This approach promises both improved thermal stability and reduced infrastructure power demands—benefits that could translate into meaningful advantages in performance per watt for large-scale AI deployments.

To quantify these potential advantages, we benchmarked two identical 8× NVIDIA GPU systems with different cooling technologies (air-cooled vs. liquid-cooled) under both synthetic stress tests and production AI workloads. Our analysis captures the thermal and power differences between the systems, along with their implications for facility-level energy consumption and operational expenditure—metrics that directly influence both cost efficiency and environmental impact as AI workloads continue to grow in scale and complexity.

METHODS

We benchmarked identical AI training workloads on two comparable 8x NVIDIA GPU systems with different cooling technologies, as described below.

Hardware Comparison Table

		
	Liquid-Cooled	Air-Cooled
GPUs	8x NVIDIA H100 80GB	8x NVIDIA H100 80GB
CPUs	Dual AMD EPYC 9474F <i>(48-cores per CPU, 192 threads total)</i>	Dual AMD EPYC 9534 <i>(64-cores per CPU, 256 threads total)</i>
Memory	1.5TB HBM	1.5TB HBM

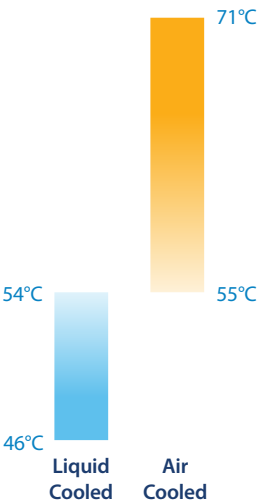
Our testing included:

- Synthetic stress testing using GPU Burn
- Real-world model fine-tuning of five open-weight LLMs (Mistral-7B-v0.3, LLaMA-3.1-8B, Mistral-NeMo-Base-2407, Gemma-2-27B, and Qwen2.5-32B)
- Pretraining of three multi-modal vision-language models (VLMs) relevant for scientific computing (X-CLIP, EVL, and ViTA-CLIP)

Metrics captured: GPU temperatures, power consumption, utilization, computational throughput, and training duration.

RESULTS

- Temperature Control: Liquid cooling maintains GPUs at 46-54°C vs 55-71°C for air cooling, enabling higher sustained performance
- Peak Performance: 17% higher computational throughput during stress tests due to improved thermal headroom and DVFS optimization
- Training Efficiency: 1.4% faster training times across real-world AI workloads (LLMs and vision-language models)
- Power Reduction: Consistent 1 kW (16%) node-level power savings, scaling from 100W for low-intensity tasks to 1.5 kW for compute-intensive workloads
- Infrastructure Accuracy: Liquid cooling provides more accurate PUE calculations by moving cooling load from embedded fans (counted as IT load) to centralized systems



PERFORMANCE SUMMARY TABLE

	Air-Cooled	Liquid-Cooled	Improvement
GPU Temperature Range	55-71°C	46-54°C	9-17°C Reduction
Stress Test Throughput	Baseline	+17%	17% Increase
Average Training Time	Baseline	5.2kW	1.4% Faster
Node Power (Production Workload)	6.2kW	5.2kW	1.0kW (16%) Reduction
Power Savings Range	-	100W-1.5kW	Scales with Workload
GPU Power Reduction	-	~150W Average	Per 8-GPU node

While production workloads showed modest throughput improvements averaging 0.3% (resulting in 1.4% faster training times), the dramatic 17% increase during stress tests highlights the crucial role of NVIDIA GPU systems' dynamic voltage-frequency-scaling (DVFS) mechanism. This technology automatically adjusts GPU frequency based on temperature, allowing liquid-cooled systems to maintain higher clock speeds without compromising hardware health.

Lower and more stable GPU temperatures directly translate to improved computational efficiency. This advantage cannot be attributed to GPU variance alone but rather reflects the systemic benefit of replacing inefficient air cooling with direct-to-chip liquid cooling. This reinforces that cooling infrastructure is not merely a passive operational concern, but an active design variable that significantly influences performance per watt in high-density compute environments.

Comparing GPU utilization for the same workload, systems differed on average by only $\pm 0.04\%$, and a paired t-test did not find this difference statistically significant. Examining only the LLM training runs where GPU utilization remained tightly grouped (93-95%), we observe comparable workload-to-workload variation in power draw across cooling modalities, with both liquid and air-cooled nodes having a standard deviation of 170 watts. These results suggest that, beyond the observed differences in throughput, the core computational components of the system were meaningfully comparable.

The most dramatic difference between cooling systems was in power demand, with liquid-cooled nodes drawing an average of 5.2 kW during production workloads compared to 6.2 kW for air-cooled nodes. This consistent 1 kW reduction represents a significant 16% power savings at the node level, which would compound dramatically in large-scale deployments.

Our measurements revealed that these savings scale proportionally with computational intensity—the higher the workload utilization, the greater the efficiency advantage of liquid cooling. This scaling occurs because, as computational load increases, air-cooled systems must ramp up fan speeds, creating compounding inefficiency. Specifically:

- Low-utilization workloads (EVL and X-CLIP): Showed modest differences of 100-400W
- LLM fine-tuning tasks: Demonstrated substantial reductions, averaging 1.2 kW per node
- Highest-intensity workloads (ViTA-CLIP): Achieved peak savings of 1.5 kW

The underlying cause of this efficiency gap stems from fundamental differences in cooling architecture and thermal physics. Air-cooled nodes rely on embedded fans that must work increasingly harder as heat generation intensifies. These fans not only consume significant power themselves but also provide less effective thermal management than direct-to-chip liquid solutions.

As workloads approach maximum GPU utilization, air cooling becomes disproportionately inefficient—the fans must operate at higher speeds, consuming more power while heat dissipation capabilities diminish. In contrast, liquid cooling maintains consistent thermal efficiency across utilization levels, with the cooling medium’s superior heat transfer properties enabling more effective temperature regulation without the escalating power penalties seen in air-cooled systems. The power and throughput improvements of the liquid-cooled node are summarized below.

POWER SAVING SCALE WITH WORKLOAD INTENSITY

As AI workloads become more compute-intensive, the efficiency advantage of liquid-cooling increases

	Workload	Power Savings	Key Benefit
Low Intensity	~60% GPU Utilization	~400W <i>(per node)</i>	Operational Stability
High Intensity	~95% GPU Utilization	~1200W <i>(per node)</i>	Higher Sustained Performance

WHY LIQUID COOLING ADVANTAGE SCALES



Fan Power Scaling

Air-cooled systems require faster fan speeds as thermal load increases, consuming 400-1000W for cooling alone



GPU Efficiency

Dynamic voltage-frequency-scaling enables higher sustained clock speeds at lower temperatures



Training Time Reduction

Up to 5% faster training time means lower total energy consumption per model

The Hyperscale Advantage

For every 1,000 nodes in your AI infrastructure, liquid cooling can save approximately \$2.3 million annual in energy costs while improving computational throughput.

	Duration Reduction (%)	GPU Power Reduction (W)	Node Power Reduction (W)	Performance (FLOPS) Improvement (%)
Idle	N/A	N/A	390	N/A
GPU Burn Test	N/A	N/A	1170	17.0%
Large Language Models				
Mistral-7B-V0.3	0.7%	155	1084	0.3%
Llama-3.1-8B	0.6%	148	1218	1.0%
Mistral-Nemo-Base-2407	0.4%	161	1244	0.5%
Gemma-2-27B	0.3%	170	1217	0.1%
Qwen2.5-32B	0.5%	173	1096	0.1%
Vision-Language Models				
X-Clip	3.0%	115	85	0.1%
EVL	5.0%	31	391	0.4%
Vita-Clip	0.3%	234	1508	0.1%
Average (LLM & VLM)	1.4%	148	980	0.3%

*Note: All values show the advantage of liquid cooling over air cooling. N/A indicates metrics not applicable for the test condition.

These findings suggest that liquid cooling reduces GPU power by approximately 100-200 watts. While there is some indication that this reduction scales with utilization, as seen in the VLM models, our dataset is not granular enough to verify this. Therefore, we'll adopt the simplification of a fixed 150-watt difference. Given that each node contains 8 GPUs, this represents a modest ~20-watt difference in power demand per chip. This is likely a result of different DVFS optima under the various cooling regimes, with the higher-performing liquid cooling system enabling a more energy-efficient voltage/frequency combination.

Examining the power difference in the low-utilization workloads (EVL and X-CLIP) is illustrative: here, the differences are relatively modest, on the order of 400 watts, or in the case of X-CLIP, 85 watts. This difference likely represents a lower-intensity setting of the embedded fan modules. If the thermal load of the modest workloads didn't require fan activation, we see the minimal differences observed with X-CLIP.

This thermal load is itself sensitive to environmental conditions; variation in ambient air temperature or humidity might influence whether thermally modest workloads require fan activation. Similarly, higher ambient air temperature due to weather variations or waste heat from nearby nodes might explain the observed differences in idle power.

At the higher end of the utilization spectrum, the liquid-cooled node demands 1200 fewer watts on average. The VITA-CLIP workload demonstrated the most significant comparative advantage, exceeding even the stress test results. Given that the stress-test value was consistent with what was observed in the LLM fine-tuning workloads, and considering run-to-run variability in computational intensity for deep learning workloads, we adopt the more conservative estimate of 1200 watts reduction in node-level power draw at saturation.

For low-utilization workloads that do not require sufficient cooling to trigger embedded fans, the differences between nodes are modest, with only a fixed 150-watt difference. Beginning at moderate computational intensity, in addition to this difference in GPU power, we observe marginal power demand from the embedded fan units. Our results indicate that the embedded fan unit power draw ranges from approximately 400 to 1000 watts, increasing with utilization. This means that at the node level, differences in IT-power demand between liquid and air cooling are modest at low utilization but increase to 1.2 kW during computationally intensive workloads.

An additional consideration for data center planning relates to power usage effectiveness (PUE) calculations. PUE is defined as the total electricity use of the facility divided by the information technology (IT) electricity within the facility. It is a ratio that is always

larger than 1.0, and the fraction above 1.0 represents the losses and electricity consumed by external fans, pumps, power distribution, and cooling (generally characterized as “infrastructure electricity use”).

The embedded cooling fans in air-cooled nodes are supplied power at the node level, meaning their consumption is counted as IT load rather than cooling overhead in standard PUE metrics. Liquid cooling effectively transfers this thermal management workload from inefficient node-level fans to more efficient centralized cooling systems, providing both absolute energy reduction and a more accurate representation of true infrastructure efficiency. This shift creates a more transparent assessment of facility energy efficiency that better reflects the actual relationship between productive computation and supporting infrastructure.

FACILITY ANALYSIS

Conservative Scenario

Using these results, we can estimate the expected energy savings an AI training data center could achieve from transitioning to direct-to-chip liquid cooling. Our initial scenario is built on conservative assumptions to develop a lower-bound case, and we’ll then explore what conditions could unlock even higher savings.

We assume the facility operates a mixture of workloads with varying utilization. Low-intensity workloads, which don’t trigger the fans, comprise 15% of annual jobs, moderate-intensity workloads with moderate fan usage comprise 50% of annual demand, and high-intensity workloads, which saturate demand (such as an extended foundation model pre-training), constitute the remaining 35% of annual utilization. For each of these workloads, we estimate node-level power draw based on our empirical results:

UTILIZATION LEVEL

	Air-Cooled (W)	Liquid-Cooled (W)
Low-UTIL	4300	4150
Mid-UTIL	6000	5000
High-UTIL	7400	6200

FACILITY-SCALE ANNUAL SAVINGS

Conservative Scenario

Deployment Size: 2,000 nodes

Energy Savings: 22.7 GWh/year

Cost Reduction: \$2.25M/year

Cost Reduction: 15%

Aggressive Scenario

Deployment Size: 5,000 nodes

Energy Savings: 91.0 GWh/year

Cost Reduction: \$11.8M/year

Energy Reduction: ~20%

Savings compound at scale, with higher returns for compute-intensive AI applications and large deployments

We’ll also make some assumptions about the operational characteristics of the facility. We’ll assume that the facility has an uptime of 99% and only modest power switching losses of 10%. We’ll assume all servers are identical 8× DGX nodes, with the only difference between the two facilities being cooling modality. We’ll model a facility with 2,000 total servers (16,000 GPUs), their associated interconnect fabric, and required data storage infrastructure. We estimate such a facility would have a nameplate capacity of 27 MW. We’ll also assume that both facilities operate with an annualized PUE of 1.25, a value within the estimated operating range of both high-performing air and low-performing liquid cooling systems (Shehabi et al. 2024). We can then model each facility under these assumptions and estimate the comparative energy performance of the liquid and air-cooled systems.

We estimate the annual energy consumption of the air-cooled facility under this scenario to be 150.2 GWh. Assuming a commercial electricity price of \$0.10/kWh, the operating energy cost of the facility would be approximately \$15 million USD per annum. We estimate that under the same scenario, the liquid-cooled facility would consume 127.5 GWh annually, with an associated cost of \$12.75 million USD. Even under this conservative scenario, the liquid-cooled nodes result in a 15% reduction in facility energy consumption and a direct annual savings of \$2.25 million.

Air-cooled nodes introduce significant error into facility PUE calculations due to their behind-the-meter fan modules. In our modeled scenario, fans in low, mid, and high utilization workloads account for 0, 850, and 1000 watts, respectively. Correctly accounting for these embedded fan modules as infrastructure energy use reveals that the actual PUE of the air-cooled facility is 1.43, not the nameplate 1.25.

While this misattribution doesn't change total facility energy consumption, it sends misleading signals to facility operators. Improved cooling typically represents the lowest-hanging fruit for facility energy optimization, alongside high uptime and low PUE, which is an industry heuristic for proximity to the technical frontier. This measurement error creates a blind spot where operators may believe they've achieved near-optimal efficiency when significant improvement opportunities remain unexploited. Organizations pursuing targets based on these flawed metrics may prioritize investments with minimal real-world impact while overlooking more cost-effective efficiency interventions.

Aggressive Scenario

The conservative scenario presented above establishes a solid baseline for evaluating liquid cooling benefits. Still, real-world AI infrastructure environments often operate under more demanding conditions that could amplify these advantages. To explore this more aggressive scenario, we can consider how several operational factors might compound to create even greater efficiency differentials between cooling approaches. We'll adjust the workload distribution to reflect the sustained high utilization typical of foundation model training operations, with 70% high-utilization workloads and just 15% each of medium and low-intensity tasks. We'll also scale the infrastructure to 5,000 servers with 40,000 total GPU chips—a configuration more representative of hyperscale AI training clusters. We'll incorporate more realistic cooling efficiency differences, with the liquid-cooled facility achieving a nameplate PUE of 1.2 compared to 1.3 for the air-cooled facility, reflecting the inherently higher efficiency potential of liquid cooling infrastructure. Finally, we'll assume a higher energy cost of \$0.13/kWh—the U.S. Energy Information Administration (EIA) estimate for industrial electricity in California or Washington state (U.S. Energy Information Administration 2025). These parameters represent conditions commonly found in dedicated AI training facilities where maximizing computational throughput takes priority, creating an environment where cooling efficiency advantages become even more consequential.

The results of this more aggressive facility model are predictably dramatic: the liquid-cooled data center consumes 91 GWh less energy annually than the air-cooled facility, with associated operating cost savings of \$11.8 million. These scenarios illustrate the facility-level operational implications of direct-to-chip liquid cooling. For an individual node, especially at low utilization, the savings are modest. Across an entire cluster of high-intensity workloads, liquid cooling can represent tens of millions of annual savings over air-cooled equivalents.

KEY TAKEAWAYS

- Liquid cooling delivers up to 17% higher throughput in stress tests and consistent node-level power reductions of 1 kW or more.
- Facility-scale savings are substantial, reaching \$2.25M–\$11.8M annually, depending on deployment size and workload intensity.
- Liquid cooling enables more accurate infrastructure efficiency metrics by properly attributing cooling overhead.
- As AI infrastructure scales, direct-to-chip liquid cooling is a strategic investment for performance, cost efficiency, and sustainability.

REFERENCES

Shehabi, Arman, Sarah Smith, Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakkar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. "2024 United States Data Center Energy Usage Report." National Lab Report.

<https://escholarship.org/uc/item/32d6m0d1>

U.S. Energy Information Administration (EIA). 2025. "Electric Power Monthly." January 2025.

https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a

FOR MORE INFORMATION

Website: www.supermicro.com

<https://www.linkedin.com/pulse/why-cios-reverting-on-prem-what-means-enterprise-ai-fergal-mcgovern-ek1ae/>

Note: these tests were run on an NVIDIA HGX H100 system.

SUPER MICRO COMPUTER, INC.

As a global leader in high-performance, high-efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based on your requirements.

www.supermicro.com

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions.

www.nvidia.com



©Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.