Cooling Matters: Benchmarking Large Language Models and Vision-Language Models on Liquid-Cooled Versus Air-Cooled H100 GPU Systems

Imran Latif, Muhammad Ali Shafique, Hayat Ullah, Alex C. Newkirk, Xi Yu, Arslan Munir, *Senior Member*, *IEEE*

***Abstract*—The unprecedented growth in artificial intelligence (AI) workloads, recently dominated by large language models (LLMs) and vision-language models (VLMs), has intensified power and cooling demands in data centers. This study benchmarks LLMs and VLMs on two HGX nodes, each with 8× NVIDIA H100 graphics processing units (GPUs), using liquid and air cooling. Leveraging GPU Burn, Weights & Biases, and IPMItool, we collect detailed thermal, power, and computation data. Results show that the liquid-cooled systems maintain GPU temperatures between 41–50°C, while the air-cooled counterparts fluctuate between 54–72°C under load. This thermal stability of liquid-cooled systems yields 17% higher performance (54 TFLOPs/ GPU vs. 46 TFLOPs/GPU), performance-per-watt, reduced energy overhead, and greater system efficiency than the air-cooled counterparts. These findings underscore the energy and sustainability benefits of liquid cooling, offering a compelling path forward for hyperscale data centers seeking to optimize AI infrastructure. https://github.com/iscaas/Cooling-Matters.**

***Index Terms*—Large Language Models, Vision-Language Models, H100 GPU, Liquid Cooling, Air Cooling, Thermal Management, AI Benchmarking, Energy Efficiency, Performance-per-Watt, High-Performance Computing**

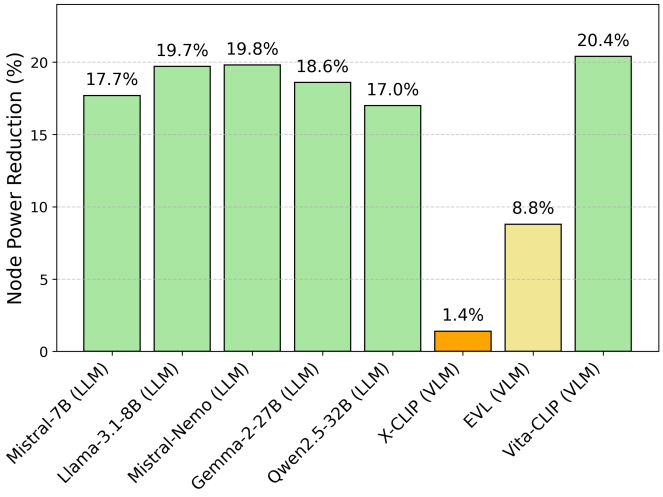


Fig. 1: Node power reduction (%) with liquid cooling over air cooling across LLM and VLM models.

I. Introduction

In recent years, the rapid development of large language models (LLMs) and vision-language models (VLMs) has dramatically transformed the landscape of artificial intelligence (AI), raising new challenges related to computational efficiency and thermal management in high-performance computing systems. As these models increase in complexity, the thermal output of the underlying hardware, particularly GPUs such as the NVIDIA H100, has become a critical factor determining energy performance. Effective cooling strategies are essential not only for maintaining optimal operating temperatures but also for ensuring consistent computational throughput during compute-intensive tasks such as model training and inference.

Today's rapid advancement of deep learning relies on algorithmic breakthroughs and high-performance computing infrastructure. As scaling laws linking model size, training data, and compute [1] continue to drive model growth, training state-of-the-art LLMs and VLMs with tens to hundreds of billions of parameters now routinely involves thousands of GPUs. Cloud providers have responded with substantial investments in AI infrastructure [2], while the resulting surge in power consumption and cumulative energy use has raised significant environmental concerns requiring active mitigation [3]. Recent empirical measurements of H100 nodes under AI training workloads reveal that actual power demand can fall well below manufacturer-rated TDP and varies substantially with model architecture and workload type [4]–[6], underscoring that the relationship between computational workload, hardware thermal state, and energy consumption is non-trivial and workload-

Imran Latif is with Johnson Controls, Milwaukee, WI 53201, USA (e-mail: imran.latif@jci.com).

Muhammad Ali Shafique is with the Intelligent Systems, Computer Architecture, Analytics, and Security Laboratory (ISCAAS Lab), Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA (e-mail: alishafique@ksu.edu).

Hayat Ullah, and Arslan Munir are with the Intelligent Systems, Computer Architecture, Analytics, and Security Laboratory (ISCAAS Lab), Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA (e-mail: hullah2024@fau.edu, arslanm@fau.edu).

Alex C. Newkirk is with the Building Technology and Urban Systems, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA (e-mail: acnewkirk@lbl.gov).

Xi Yu is with the Computing and Data Sciences Directorate, Brookhaven National Laboratory, Upton, NY 11973, USA (e-mail: xyu1@bnl.gov).

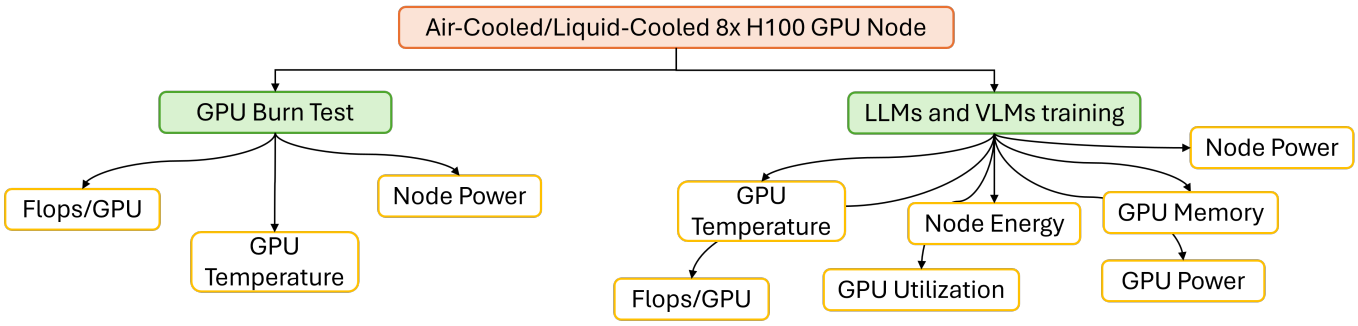


Fig. 2: Benchmarking metrics for different cooling systems across various workloads.

dependent.

Growth in IT electricity demand requires cooling infrastructure to keep pace. Cooling systems account for 30%–50% of total data center energy use, and traditional air-cooled systems have struggled to meet the escalating thermal demands of modern high-density GPU deployments [7], [8]. Liquid cooling, including direct-to-chip and immersion techniques, has been shown to handle high power densities more effectively, demonstrating improvements in thermal performance and energy efficiency [9]–[11]. While it is broadly understood that liquid cooling offers thermal advantages over air cooling [12], what remains insufficiently characterized is *how and to what degree* this thermal advantage translates into measurable differences in computational throughput, training efficiency, and energy per unit of computation under realistic, sustained LLM and VLM workloads. This throughput–cooling interaction effect is consequential: thermal throttling under air cooling can suppress GPU clock frequencies and sustained utilization, directly impacting effective training throughput and energy-per-FLOP, yet its quantitative magnitude under modern AI workloads on H100 hardware has not been concretely benchmarked in the literature.

The effective benchmarking of AI workloads therefore requires joint evaluation of thermal behavior, computational throughput, and energy efficiency. Prior work has shown that inference efficiency is highly sensitive to hardware configuration and workload geometry [13], [14], and that GPU power capping can reduce temperatures and energy consumption with minimal performance impact [15]. At the infrastructure level, co-optimized cooling and power provisioning can reduce data center TCO by up to 40% [16], making cooling strategy a consequential design decision beyond its immediate thermal effects. However, existing AI energy benchmarking studies generally do not isolate cooling modality as an explicit experimental variable under sustained LLM and VLM workloads. In this work, we treat liquid versus air cooling as a first-class system variable and quantify how cooling regime propagates into thermal stability, node-level power, performance-per-watt, and sustained throughput behavior on 8× H100 GPU systems.

We present a detailed benchmarking study that evaluates the impact of cooling strategy, liquid versus air, on the performance and power efficiency of AI workloads using 8× NVIDIA H100 GPU nodes. Our analysis includes both synthetic stress testing with *GPU-Burn* and real-world model training across five LLMs: Mistral-7B-v0.3 [17], LLaMA-3.1-8B [18], Mistral-NeMo-Base-2407 [19], Gemma-2-27B [20], and Qwen2.5-32B [21], as well as three VLMs: X-CLIP [22], EVL [23], and Vita-CLIP [24]. We measure GPU power draw, temperature, utilization, training duration, energy consumption, achieved FLOPs, and GPU-level and node-level power consumption. By jointly analyzing thermal, computational, and power behavior across cooling architectures under representative AI workloads, our results provide quantitative insight into how cooling infrastructure shapes the throughput–efficiency trade-off, contributing empirical evidence that goes beyond validating known thermal principles to characterizing the performance–cooling interaction at the system level, as illustrated in Figure 1.

Contributions: The main contributions of this work are as follows:

1) We introduce a cooling-aware AI benchmarking methodology that treats cooling regime as an explicit experimental variable and jointly profiles thermal behavior, GPU power, node-level power, utilization, and computational throughput under sustained AI workloads.
2) To the best of our knowledge, we present one of the first empirical characterizations of the throughput–cooling interaction on paired 8× NVIDIA H100 nodes under QLoRA-based LLM fine-tuning, VLM training, and GPU stress testing.
3) We extend recent empirically calibrated H100 energy-measurement studies by quantifying how the measured thermal and power envelope shifts under liquid versus air cooling, including sustained node-level power differences of approximately 1–1.5 kW under high-utilization workloads.
4) We provide actionable insights showing that cooling strategy is a consequential design variable for sustained thermal behavior, node-level energy demand, performance-per-watt, data center TCO, and sustainable AI deployment.

Significant Results: The most significant results of our benchmarking study are summarized as follows:

1) Liquid-cooled systems maintained GPUs temperature between 41–50°C under peak load as compared to 54–72°C in air-cooled systems.

2) During GPU Burn, the liquid-cooled node achieved an average of 54 TFLOPs/GPU, 17% higher than the 46 TFLOPs/GPU observed on the air-cooled node.
3) During LLM fine-tuning and VLM (specifically ViTA-CLIP) training, node-level power consumption on the liquid-cooled system is consistently lower by 1–1.5 kW, while maintaining equal or improved training duration.
4) All LLM and VLM models demonstrated higher performance-per-watt on the liquid-cooled system, as compared to air-cooled system.
5) All models maintained equal or improved training durations on liquid-cooled systems and exhibited higher throughput compared to their air-cooled counterparts.

The rest of this paper is structured as follows. In Section II, we provide background on the evolution of AI infrastructure and the motivation for evaluating cooling strategies in high-performance training environments. Section III outlines the methodology used to benchmark system performance, power, and thermal behavior. Section IV describes the experimental setup, including hardware specifications, software configuration, and workload design. In Section V, we present the measured power, temperature, and performance metrics across both cooling configurations. Section VI analyzes the findings in the context of system design and energy efficiency. Finally, Section VII concludes this paper and discusses broader implications for sustainable AI infrastructure and directions for future work.

II. Background and Motivation

The performance and efficiency of large language models (LLMs) and vision-language models (VLMs) are deeply influenced by the thermal and energy behavior of the underlying hardware. As these models grow in complexity and scale, effective thermal management becomes increasingly critical to avoid bottlenecks associated with heat dissipation and power consumption. While it is broadly understood that liquid cooling provides superior thermal performance relative to air cooling [12], what remains insufficiently characterized is *how cooling infrastructure interacts with computational throughput and energy efficiency under realistic, sustained AI workloads*, a gap our study directly addresses through concrete, system-level benchmarking on H100 hardware under representative LLM and VLM training conditions.

Cooling systems account for 30%–50% of total data center energy use, with air cooling remaining the dominant technique [7], [25]. The rising thermal output of modern GPUs such as the NVIDIA H100 has exposed the limitations of air-cooled systems in high-density environments. Recent empirical measurements show that even computationally intensive AI training workloads operate at only $\sim$76% of the manufacturer-rated 10.2,kW TDP, with actual power demand varying significantly across model architectures [4]–[6], and TDP-based energy estimates carrying 27–37% error relative to empirically calibrated models [6]. Liquid cooling, particularly direct-to-chip and immersion techniques, has emerged as a thermally efficient alternative capable of reducing energy consumption by up to 50% [26] while preventing thermal throttling that degrades GPU performance under sustained load [15], [27]. Ramakrishnan et al. [12] demonstrated that direct liquid cooling (DLC) improves GPU floating-point throughput by 2.7%, reduces power consumption by 12%, and lowers chip temperatures by $20^{\circ}\mathrm{C}$ relative to air cooling, though these results were not obtained under large-scale LLM and VLM training conditions, where sustained high utilization over many hours produces qualitatively different thermal profiles. Zhu and Wang [28] further showed that immersion cooling introduces constraints, including coolant boiling state and heat removal rate mismatches, that can trigger thermal throttling even in liquid-cooled environments, underscoring the need for workload-aware thermal analysis.

The energy implications of the throughput, cooling interaction extend beyond power draw. Inference efficiency is highly sensitive to workload geometry, software stack, and hardware configuration, with FLOPs-based estimates significantly underestimating real-world consumption [13], [14], [29], [30]. At the training level, workload scaling, including token count and batch size, interacts non-trivially with energy cost [4], [31], and test-time compute strategies have been shown to offer superior accuracy–energy trade-offs over model scaling, particularly for reasoning-intensive tasks [32]. GPU power capping at HPC scale yields significant reductions in temperature and power draw with minimal performance impact [15], while adaptive cooling control via machine learning and evolutionary algorithms enables real-time optimization without compromising throughput [33]–[35]. At the infrastructure level, co-optimized power provisioning, cooling design, and operational scheduling can reduce data center TCO by up to 40% [16], and PUE remains a standard comparative metric for cooling efficiency [7]. Energy-efficient LLM deployment further depends on the interplay between quantization, accelerator architecture, and serving strategy [36]–[38], while heterogeneous cluster configurations and workload-aware scheduling offer additional pathways to reduce consumption [39]–[42]. Advanced cooling architectures, including phase-change and immersion systems, offer superior thermal regulation for high-density workloads [28], [43], [44], and next-generation data center design must jointly optimize FLOP capacity, memory bandwidth, network topology, and cooling provisioning [16], [45]. Gupta et al. [46] further emphasized that intelligent cooling strategies influence both energy and power efficiency, optimizing resource allocation at the data center level.

Prior liquid-cooling studies have demonstrated thermal and power advantages on different GPU platforms and AI/ML workloads [47], while MLPerf Power provides a standardized framework for energy-aware AI benchmarking [48]. Recent H100 measurement studies have further shown that empirically observed node power can differ substantially from TDP-based estimates [4], [49], [50].However, these studies do not explicitly quantify how the empirical power, thermal, and throughput envelope shifts when cooling modality changes under sustained LLM and VLM workloads on paired $8\times$ H100 systems. Our work fills this gap by adding cooling regime as a controlled benchmarking dimension and jointly analyzing its effect on temperature, node-level power, GPU power, utilization, and achieved throughput.

III. Methodology

This section outlines the benchmarking methodology used to evaluate air-cooled and liquid-cooled system performance, power consumption, and thermal behavior under both synthetic stress testing and realistic LLM fine-tuning and VLM training workloads. Our metric selection is directly motivated by and consistent with established methodologies in empirical AI energy and thermal benchmarking [4], [5], [12], [14], [15], extended here to jointly characterize the throughput, cooling interaction effect under current-generation workloads, a combination not addressed in prior studies that treat these dimensions in isolation [12], [14]. Our power and energy measurements are limited to the IT side of the system. Specifically, GPU-level metrics are collected using *Weights & Biases* and *CodeCarbon*, while node-level power is collected using *ipmitool*. We do not directly measure facility-side cooling energy, including chiller power, pump power, heat rejection, CRAH/CRAC units, rear-door heat exchanger fans, or secondary cooling-loop infrastructure. Therefore, the reported power and energy results should be interpreted as node-level and GPU-level measurements rather than a full PUE-inclusive facility-level energy comparison. We believe the IT side numbers reported in Tables I and II represent a conservative lower bound on the true energy benefit.

For all workloads, we captured the following node-level and GPU-level metrics, as shown in Figure 2, spanning thermal, computational, and energy efficiency dimensions consistent with multi-dimensional profiling approaches in recent literature [6], [13], [30]:

- **GPU Temperature:** Per-GPU readings collected continuously to monitor thermal stability, consistent with established GPU thermal benchmarking [12], [15].
- **Memory Usage:** Peak and average utilization per GPU to assess memory load distribution across model architectures.
- **GPU Utilization:** Real-time metrics to ensure workload saturation and compare operational efficiency across nodes.
- **Individual and Average GPU Power Draw:** Per-GPU and aggregate power consumption, consistent with empirical H100 power studies [4]–[6] demonstrating that actual power requires direct measurement rather than TDP-based estimation.
- **Node-Level Power:** Total node power sampled via *ipmitool* at 20-second intervals, providing an infrastructure-level view consistent with data center PUE analysis [7], [16].
- **Computational Throughput (FLOP/s):** TFLOPs and GFLOPs per GPU measured using DeepSpeed FLOP Profiler and THOP for training workloads, and *GPU-Burn* for stress tests, central to characterizing how thermal conditions propagate into effective computational performance [13], [14].

This unified, multi-dimensional profiling framework provides a reproducible approach for cooling-aware AI infrastructure evaluation, with metric selection grounded in and advancing beyond prior benchmarking methodologies.

IV. Experimental Setup and Equipment

This section outlines the hardware infrastructure and software frameworks used for benchmarking. It includes the details of liquid-cooled and air-cooled nodes, followed by the software stack used to monitor, and profile metrics across all experiments.

A. Experimental Design Overview

Our benchmark consists of four complementary experiment classes, each designed to isolate a different aspect of the cooling–performance interaction. First, GPU Burn provides a compute-bound stress test that characterizes the upper thermal, power, and throughput envelope of each gpu, consistent with prior empirical H100 power studies that use stress workloads to bound maximum system behavior [4], [49]. Second, QLoRA-based fine-tuning across five LLMs from 7B to 32B parameters evaluates sustained language-model workloads under realistic multi-GPU execution, following the broader practice of energy-aware LLM benchmarking under practical model workloads [14], [51]. Third, VLM training across X-CLIP, EVL, and Vita-CLIP provides a distinct workload family with different utilization, memory-access, and runtime characteristics. Fourth, all workloads are monitored using layered instrumentation, including node-level power from *ipmitool*, per-GPU power, temperature, utilization, and memory from *Weights & Biases*, and GPU-level energy estimates from *CodeCarbon*. Together, these experiments form a cooling-aware benchmark suite rather than a single stress-test comparison.

B. Computational Hardware

Our experiments were conducted on two high-end GPU nodes with identical GPU configurations but different cooling systems, that is, liquid-cooled and air-cooled setups.

1) Liquid-Cooled Node: : This system contains 8× NVIDIA H100 80GB HBM3 GPUs, supported by an AMD EPYC 9474F [52] 48-core processor with 192 threads. This node includes a liquid-cooled system to maintain optimal thermal regulation.

2) Air-Cooled Node: : This node was also featured with 8× NVIDIA H100 80GB HBM3 GPUs, but it uses an AMD EPYC 9534 [53] 64-core processor with 256 threads. It provided an identical 1.5 TB RAM and operated on air-cooled system.

3) Cooling Configuration: : The liquid-cooled node uses direct-to-chip (D2C) cooling for both CPU and GPU, with CoolIT Systems OAT PG-25 coolant circulated through a closed-loop system. The circuit volume is 16.3 liters (4.3 US gallons), and the system operates at a maximum return pressure of 50 PSI, with a secondary loop pressure relief cap at 86 PSI. Peripheral components such as memory and storage remain air-cooled in both configurations. The air-cooled node contains 8 fans, while the liquid-cooled node uses only 4 fans, as fewer components require active airflow. This reduction contributes to lower power consumption and reduced acoustic noise.

4) Datacenter Cooling Infrastructure: : The datacenter utilizes a shared chilled water loop maintained at 20°C, produced using conventional chillers with cooling towers. There is no dedicated chiller exclusively for the DLC nodes. The air-cooled system uses a rear-door heat exchanger, while the DLC node connects directly to the chilled water loop via cold plates. All tests were conducted at ambient room temperatures of 21–22°C.

C. Software Configuration

We used a variety of tools to monitor key metrics at the GPU level and the node level. It includes GPU memory, power, temperature, utilization, and power consumption at the GPU level and node level. Each tool provides specific measurements essential for comparing air-cooled and liquid-cooled systems.

1) IPMITools: It is a command-line utility to access hardware-level metrics using the Intelligent Platform Management Interface (IPMI) [54]. In our benchmarking across all workloads, it was used to log node-level power consumption at the interval of 20 seconds during GPU Burn test and LLM fine-tuning and VLM training. Although GPU-specific metrics were measured separately, ipmitool provides a node-level power consumption for the analysis of performance-per-watt across liquid-cooled and air-cooled configurations.

2) Weight & Biases: It is the machine learning platform that provides tools for training, tracking, and visualizing experiments data, model, and metrics [55]. In our work, it was used to capture GPU hardware metrics such as memory, temperature, utilization, and power consumption during LLM fine-tuning and VLM training. These measurements helped in comparing the performance of nodes with different cooling configurations.

3) CodeCarbon: It is an open-source tool for estimating carbon emission with average power and energy consumption for CPU, GPU and RAM. In our experiments, it was used to track average GPU power and energy during LLM fine-tuning and VLM training. Unlike node-level power measurement, *CodeCarbon* [56] provides GPU-specific average power and energy metrics in Watt and Watt-hours (Wh) respectively, which were used for comparing of air-cooled and liquid-cooled nodes for various models.

4) FLOPS Profiler: To measure the FLOPS for LLMs and VLMs, we used the *DeepSpeed FLOP Profiler* and *THOP*. The DeepSpeed profiler provided runtime FLOP estimates during the training. This tool enabled us to estimate FLOPS per GPU during LLM fine-tuning and VLM training for air-cooled and liquid-cooled node. While GPU Burn test scripts provides the FLOPS per GPU automatically. The results collected from profilers discussed above, allowed us to compare the compute efficiency and the performance-per-watt across various cooling configurations.

D. Workload Configuration

1) GPU Burn Test Setup: To compare the performance difference of air-cooled and liquid-cooled nodes under maximum workload, we employed the GPU Burn [57] utility. We operated this test by enabling double-precision computation.

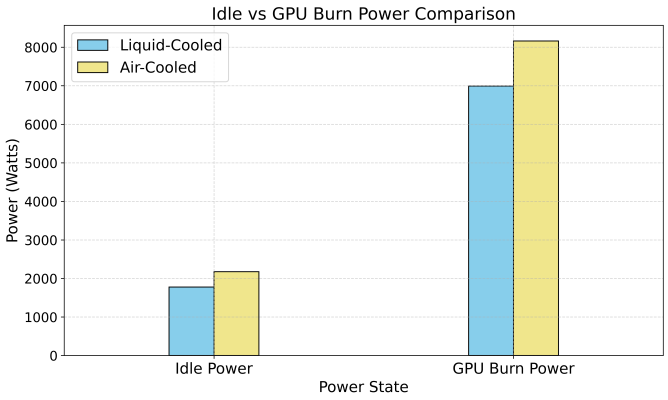


Fig. 3: Power comparison for GPU Burn Test on air-cooled and liquid-cooled nodes.

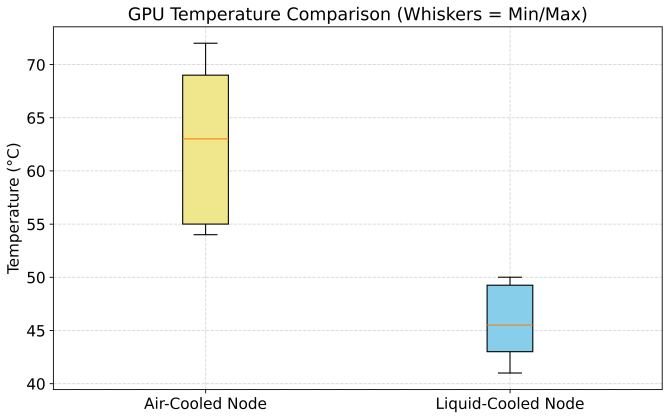


Fig. 4: Temperature comparison for GPU Burn Test on air-cooled and liquid-cooled nodes.

The test was run for two hours on all 8 GPUs on each node using the following configuration:

gpu-burn -d 7200

In above command, *-d 7200* sets the duration of GPU Burn test to 7200 seconds (2 hours). This configuration was selected to test each node under continuous high-intensity workloads. During the test, we measured each GPU temperature and FLOPS by logging GPU Burn test results. Total node power was also monitored during this test using *ipmitool*. These results served as an important reference point for evaluating the efficiency and thermal resilience of each cooling node under maximum workload.

2) LLM Fine-tuning Setup: All pre-trained large language models were finetuned using the Alpaca [58] dataset in a supervised manner. Distributed data parallel (DDP) technique was used across 8× NVIDIA H100 air-cooled and liquid-cooled nodes. Due to limited computational resource, parameter-efficient finetuning (PEFT) such as QLoRA was employed, with a low-rank value $r = 128$. Each model was trained for one epoch with a micro batch size of 4. With DDP running on 8× GPU, the effective batch size was 32. This training setup was consistent, with each LLM finetuning, allowing consistent comparisons across air-cooled and liquid-

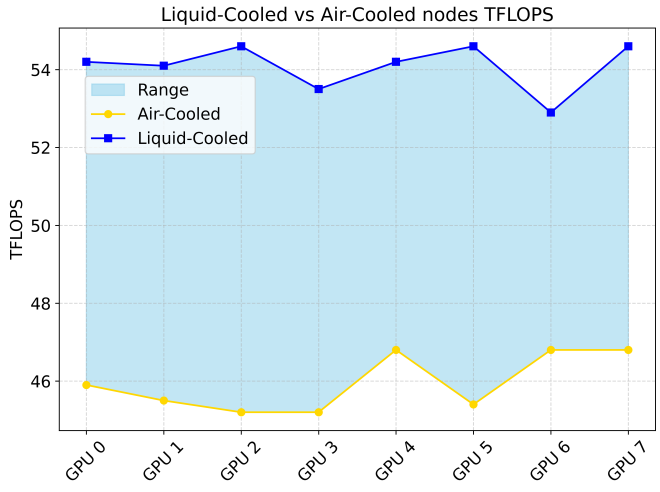


Fig. 5: TFLOPS difference for GPU Burn test on air-cooled and liquid-cooled nodes.

cooled nodes.

3) VLM Training Setup: To systematically evaluate and compare the performance capabilities of liquid-cooled and air-cooled computing nodes, we conducted comprehensive training experiments using three VLM models: X-CLIP [22], EVL [23], and ViTA-CLIP [24] on UCF101 [59] dataset. All models are trained with the same hyper-parameters on both nodes to ensure a fair comparison. Specifically, each model is trained in Distributed Data-Parallel (DDP) settings across 8× NVIDIA H100 liquid-cooled and air-cooled nodes.

Each model is trained for 50 epochs with a global batch size of 32 and a mini-batch size of 4 (i.e., 4 samples per GPU). Additionally, we set the video sequence length to 16 frames for each model. These hyperparameters are consistently maintained across both liquid-cooled and air-cooled nodes to ensure a fair and accurate performance comparison. This experimental setup enables a detailed assessment of the training efficiency, computational performance, and resource utilization differences between the liquid-cooled and air-cooled nodes.

V. Results

This section presents results across the four experiment classes described in Section IV-A: GPU Burn stress testing, QLoRA-based LLM fine-tuning, VLM training, and layered node/GPU instrumentation. GPU Burn provides the upper-envelope stress case, while the LLM and VLM workloads evaluate sustained model-training behavior under realistic workloads.

A. GPU Burn Test

To evaluate peak throughput and thermal performance of both cooling systems, we conducted GPU Burn test on both air-cooled and liquid-cooled nodes. We observed that during the idle state, the liquid-cooled node consumed 1.78 kW on average, as compared to air-cooled node which consumed 2.17 kW. For full load, power consumption for liquid-node was 6.99 kW on average while air-cooled node power usage was 8.16 kW. These results showed that as the load increased, the power difference between liquid-cooled and air-cooled node increased significantly as shown in Figure 3.

Similar trend was observed for GPU temperatures during peak workload. The GPUs in air-cooled node reached up to 72°C, whereas GPUs of liquid-cooled node remained between 41–50°C. In addition, temperature of air-cooled GPUs range from 54°C to 72°C, larger than the temperature range of liquid-cooled GPUs as depicted in Figure 4. This temperature peak and range difference across both cooling configurations translated into higher throughput on liquid-cooled node, with per-GPU FLOPS ranging from 52.9–54.6 TFLOPS as compared to 45.2–46.8 TFLOPs for air-cooled node, exhibiting a clear thermal advantage for the liquid-cooled system as illustrated in Figure 5.

These results underscore the performance advantage of the liquid cooling system, particularly for high-intensity workloads. Not only liquid-cooled node consumed less power at peak workload, but it also delivered approximately 17% more compute per GPU while maintaining significantly lower temperatures.

B. LLM Results

We benchmarked five large language models ranging from 7B to 32B parameters on both the liquid-cooled and air-cooled nodes. The liquid-cooled system demonstrated lower power draw and higher performance efficiency across all large language models as illustrated in Table I. The average GPU power consumption was consistently lower on liquid-cooled node, with values ranging from 3503 W (Mistal-7B-v0.3) to 3742 W (Gemma-2-27B), compared to 3653–3912 W on air-cooled system. This reduction in power consumption corresponded to improved performance-per-watt for liquid-cooled node.

In terms of throughput, TFLOPS per GPU were slightly higher on the liquid-cooled system, despite identical training configurations. The largest measurement was observed for the Mistral-Nemo-Base-2407 model, which achieved 35.71 TFLOPs/GPU on liquid-cooled node as compared to 35.52 on air-cooled node. Mistral-7B-v0.3, Llama-3.1-8B, and Gemma-2-27B followed similar trends.

Qwen2.5-32B, the largest model in our benchmark, also exhibited high performance for liquid-cooling system. The liquid-cooled node achieved slightly faster fine-tuning, higher throughput with lower average GPU power consumption (3671 W vs. 3844 W). GPU utilization remained high on both nodes (93.4% vs. 92.9%), confirming consistent workload execution as shown in Figure 6. These results reinforce the efficiency benefits of liquid cooling, particularly in better power and thermal performance for large-scale models.

Node-level power consumption was also notably lower on the liquid-cooled system with averaging 1000–1200 W less per model. These results confirm that liquid-cooling node delivers significant efficiency gains during real-world LLM finetuning workloads.

TABLE I: Training metrics across various LLM models for air/liquid-cooled nodes.

Model	Node	Trainable (%) / Params	Duration (sec)	Avg. 8× GPU Utilization (%)	Avg. 8× GPU Power (W)	Avg. Node Power (W)	TFLOPS/GPU
Mistral-7B-v0.3 [17]	Liquid-Cooled	4.42% / 7B	2702.03	95.8	3503	5051	35.19
	Air-Cooled	4.42% / 7B	2721.13	95.3	3658	6135	35.09
Llama-3.1-8B [18]	Liquid-Cooled	4.01% / 8B	2466.19	94.9	3505	4975	35.12
	Air-Cooled	4.01% / 8B	2479.83	95.2	3653	6193	34.76
Mistral-Nemo-Base-2407 [19]	Liquid-Cooled	3.59% / 12B	3827.89	95.8	3556	5052	35.71
	Air-Cooled	3.59% / 12B	3843.79	95.6	3717	6296	35.52
Gemma-2-27B [20]	Liquid-Cooled	3.24% / 27B	8393.51	92.7	3742	5333	39.17
	Air-Cooled	3.24% / 27B	8418.65	93.1	3912	6550	39.15
Qwen2.5-32B [21]	Liquid-Cooled	3.17% / 32B	9738.14	93.4	3671	5344	37.42
	Air-Cooled	3.17% / 32B	9783.15	92.9	3844	6440	37.40

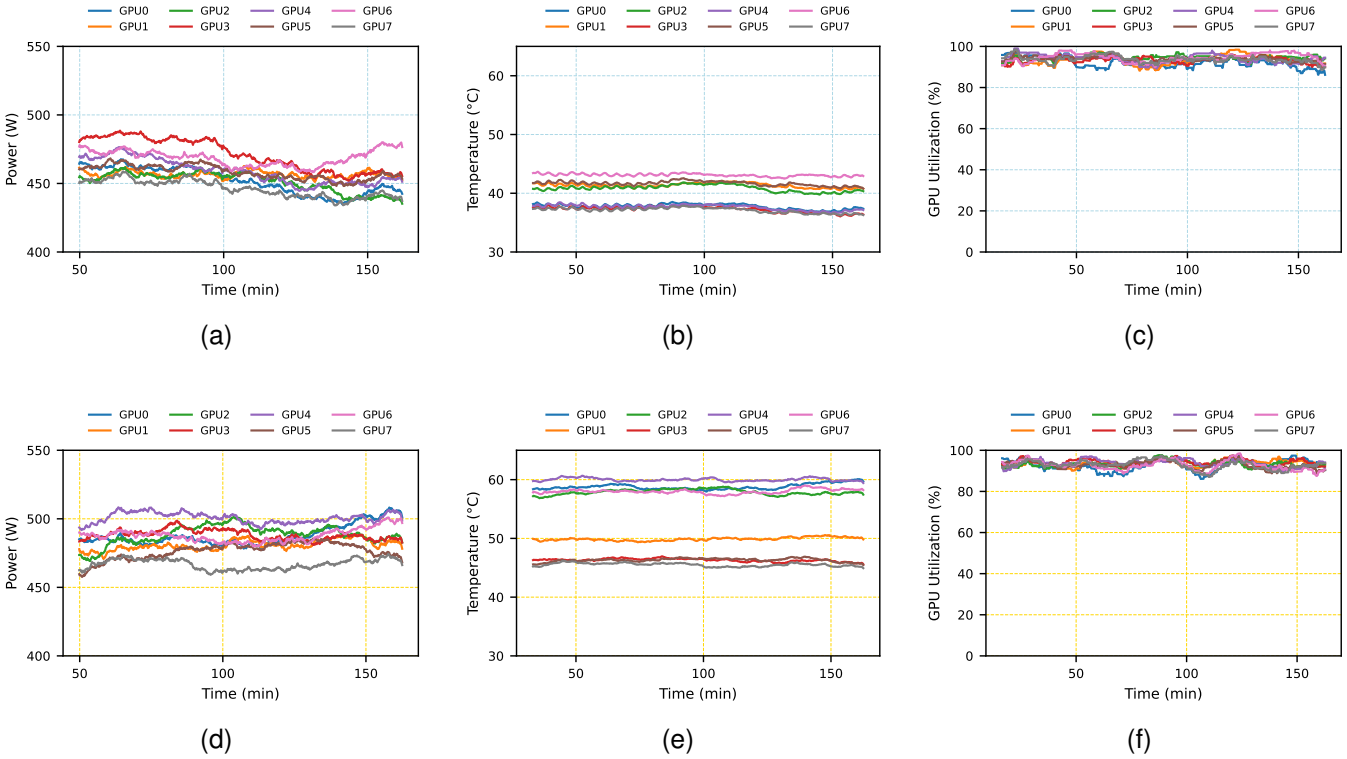


Fig. 6: GPU metrics comparison for liquid-cooled (a–c) and air-cooled (d–f) nodes for Qwen2.5-32B [21].

C. VLM Results

We benchmarked three VLM models including X-CLIP [22], EVL [23], and ViTA-CLIP [24] on two distinct computing nodes: liquid-cooled and air-cooled. Through comprehensive training experiments conducted on both nodes in the same training settings, we observed that the liquid-cooled node consistently exhibited lower power consumption and higher performance efficiency across all three VLM models, as presented in Table II.

As listed in Table II, the liquid-cooled node consistently outperformed the air-cooled node by exhibiting lower average GPU power for each VLM model used in this work. For instance, X-CLIP consumed 3268 W on the liquid-cooled node and 3383 W on the air-cooled node. Similarly, for EVL and Vita-CLIP models, the liquid-cooled has a lower average GPU power consumption of 2320 W and 4295 W in comparison with the air-cooled which has an average GPU power consumption of 2351 W and 4529 W, respectively.

Following the same trend, the liquid-cooled node demonstrates higher GFLOPS/GPU across each VLM model. More specifically, for the X-CLIP mode, the liquid-cooled node achieved 377.13, slightly higher than the air-cooled node's 376.92 GFLOPS/GPU. Similarly, for the EVL model, the liquid-cooled node obtained 119.87 GFLOPS/GPU in comparison to an air-cooled node which has slightly lower GFLOPS/GPU (119.41). Lastly, Vita-CLIP exhibited 1064.06 on the liquid-cooled node and 1063.81 GFLOPS/GPU on the air-cooled node, thereby, consistently showing the computational efficiency of the liquid-cooled node.

From the node-level power consumption results presented

TABLE II: Training metrics across various VLM models for liquid-cooled and air-cooled nodes.

Model	Node	Trainable% / Params	Duration (sec)	Avg. 8× GPU Utilization (%)	Avg. 8× GPU Power (W)	Avg. Node Power (W)	GFLOPS/GPU
X-CLIP [22]	Liquid-Cooled	100% / 97.95M	1503	58.6	3268	5890	377.13
	Air-Cooled	100% / 97.95M	1549	58.1	3383	5975	376.92
EVL [23]	Liquid-Cooled	100% / 28.43M	14530	60.0	2320	4053	119.87
	Air-Cooled	100% / 28.43M	15296	59.3	2351	4444	119.41
Vita-CLIP [24]	Liquid-Cooled	100% / 146.24M	91654	97.4	4295	5867	1064.06
	Air-Cooled	100% / 146.24M	91941	97.8	4529	7375	1063.81

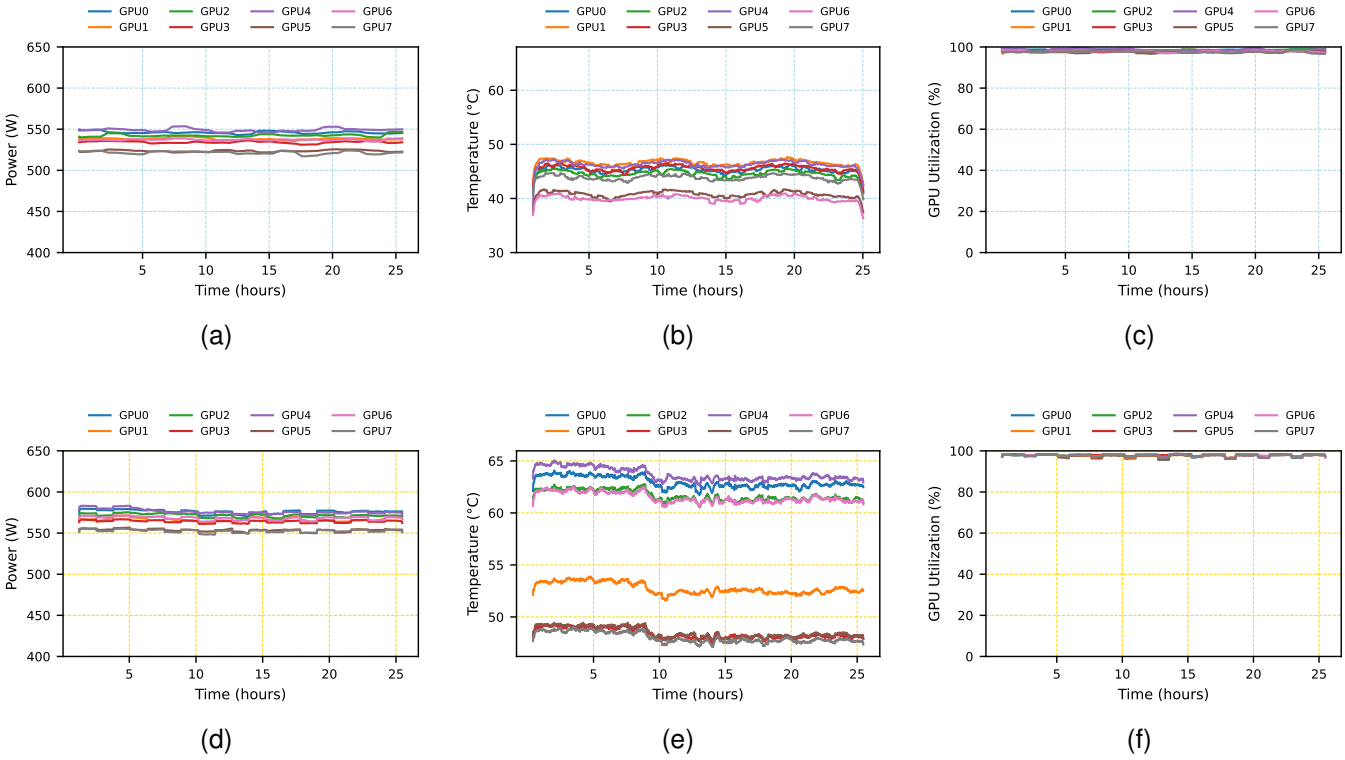


Fig. 7: GPU metrics comparison for liquid-cooled (a,b,c) and air-cooled (d,e,f) nodes for Vita-CLIP [24].

in Table II, it can be observed that the liquid-cooled node consistently demonstrates lower power usage compared to the air-cooled node for each VLM model. For example, in the case of the X-CLIP model, the liquid-cooled node consumed 5890 W average power, which is 85 W less than the 5975 W consumed by the air-cooled node. This difference in node-level power consumption between the two nodes increased for models with longer training durations. Specifically, for the EVL model, the liquid-cooled consumed 4053 W average power, which is 391 W less than the 4444 W consumed by the air-cooled node. Similarly, for the Vita-CLIP model, the liquid-cooled node consumed 5867 W, which is 1508 W lower than the 7375 W consumed by the air-cooled Node.

Besides the quantitative analysis, we also performed a visual investigation of the two nodes' performance in terms of GPU power usage (%), GPU temperature (°C), and GPU utilization (%), as depicted in Figure 7. The first row presents the results for the liquid-cooled node, while the second row illustrates the corresponding results for the air-cooled node. As it can be visually perceived, the liquid-cooled node demonstrates significantly better performance in terms of lower GPU power consumption and reduced GPU temperature while maintaining comparable GPU utilization to the air-cooled node.

Overall, these results clearly highlight the advantages of liquid cooling in reducing energy consumption and enhancing computational efficiency for training advanced deep learning models and large vision language models.

VI. Discussion

These results illustrate the complex interaction between thermal design and power efficiency in large-scale AI training infrastructure. Despite identical GPU hardware and workloads, each cooling strategy produced clear differences in thermal stability, energy consumption, and sustained compute performance, highlighting that IT hardware performance is not independent of cooling infrastructure.

The total required FLOPs of our training workloads fall within the natural run-to-run variability of neural network

TABLE III: Liquid-Cooling Performance Advantage across LLMs and VLMs.

System/Model	Duration Reduction (%)	GPU Power Reduction (W)	Node Power Reduction (W)	Performance (FLOPs) Improvement (%)
Large Language Models (LLMs)				
MISTRAL-7B-V0.3 [17]	0.7	155	1084	0.3
LLAMA-3.1-8B [18]	0.6	148	1218	1.0
MISTRAL-NEMO-BASE-2407 [19]	0.4	161	1244	0.5
GEMMA-2-27B [20]	0.3	170	1217	0.1
QWEN2.5-32B [21]	0.5	173	1096	0.1
Vision-Language Models (VLMs)				
X-CLIP [22]	3.0	115	85	0.1
EVL [23]	5.0	31	391	0.4
VITA-CLIP [24]	0.3	234	1508	0.1
AVERAGE (LLM & VLM)	1.4	148	980	0.3

training ($\pm$0.05%), making direct hardware performance comparisons appropriate. On average, liquid cooling enabled a 0.03% increase in per-GPU operations across production workloads, with throughput gains most pronounced in fine-tuning workloads, resulting in an average 1.34% reduction in training time. The 17% throughput increase observed during stress testing is most likely attributable to NVIDIA H100's dynamic-voltage-frequency scaling (DVFS), where lower sustained GPU temperatures under liquid cooling enable higher clock frequencies without jeopardizing hardware health [12], [15]. We note that while this DVFS-based interpretation is mechanistically well-motivated and consistent with documented H100 behavior, direct clock frequency logging and throttling event measurements were not captured in this study; we explicitly identify these as priority measurements for future controlled experiments to confirm the causal mechanism.

The workload-dependent variation in performance gains reflects the utilization-driven nature of the throughput–cooling interaction. Models with higher sustained GPU utilization (e.g., Vita-CLIP, LLM fine-tuning) generate greater thermal load, producing larger temperature differentials between cooling configurations and consequently larger throughput and power differences [6], [13]. Conversely, low-utilization workloads (e.g., EVL, X-CLIP) generate insufficient thermal load to trigger meaningful throttling under either cooling modality, resulting in modest power differentials of only 100–400 W.

The most dramatic performance difference was in node-level power demand. The liquid-cooled node drew $\sim$5.2 kW during production workloads versus $\sim$6.2 kW for the air-cooled node. Power differentials scaled with computational intensity: low-utilization workloads (EVL, X-CLIP) differed by only 100–400 W, while LLM fine-tuning averaged 1.2 kW lower and Vita-CLIP reached 1.5 kW lower on the liquid-cooled node (Table III). This scaling behavior stems from air-cooled nodes relying on embedded fans that consume increasing node-level power under thermal load, whereas liquid cooling transfers this thermal burden to centralized chilled water infrastructure with superior coefficient of performance. Critically, because embedded fans are powered at the node level, their consumption is counted as IT load in standard PUE metering, meaning liquid cooling offers a dual benefit: reducing absolute power consumption while transferring cooling workload from IT load to more efficient facility overhead, improving both total energy efficiency and PUE transparency.

Workload configuration also significantly influenced energy consumption. Architectural differences among LLMs and VLMs affected both power demand and compute efficiency, and fine-tuning may present differential power characteristics compared to pre-training. This underscores the need for workload-aware optimization strategies that account for energy footprint alongside model accuracy, increasingly critical as AI scales to gigawatt-class clusters [16], [45].

Practical Deployment Considerations. The energy savings demonstrated in this study, averaging $\sim$1 kW lower node-level power draw during production workloads and up to 1.5 kW during high-utilization tasks, translate to meaningful operational cost reductions at cluster scale. However, deployment decisions must weigh these benefits against higher upfront infrastructure costs and increased maintenance complexity of liquid cooling systems. Recent data center lifecycle analyses suggest that co-optimized cooling and power provisioning can reduce TCO by up to 40% [16], providing a framework within which our empirical power measurements can inform practical deployment decisions. A full ROI analysis incorporating site-specific utility rates and cooling infrastructure amortization represents an important direction for future work.

Limitations and Confounds. These findings must be interpreted within the context of the specific configurations tested. Most importantly, the two nodes differ in CPU model: both are AMD EPYC processors on the same Zen4 microarchitecture at the same 5 nm TSMC process node, but with a manufacturer-rated TDP difference of 80 W. This bounds the direct CPU power contribution to observed node-level energy differences, which ranged from 100 W to 1.5 kW, suggesting CPU differences alone cannot account for the majority of observed differentials, particularly at high GPU utilization where GPU power dominates node-level consumption [4]–[6]. We explicitly prioritize matched-CPU, component-isolated benchmarking as a key direction for future work to cleanly separate cooling modality from CPU architectural effects.

The study is further limited to single-node analysis, whereas real-world training spans multiple nodes, introducing communication overheads and synchronization latency that may dilute single-node gains [40], [42]. Finally, the absence of

fine-grained sub-metering limits component-level power attribution; future work leveraging integrated hardware power telemetry would enable more precise identification of per-component efficiency bottlenecks [30].

Another limitation concerns workload representativeness. Our LLM experiments use QLoRA-based parameter-efficient fine-tuning rather than full-parameter fine-tuning or pre-training. Thus, they may not capture the most energy-intensive stages of the LLM lifecycle. Nevertheless, these workloads provide sustained multi-GPU fine-tuning with high GPU utilization, while GPU Burn and VLM training provide complementary high-utilization stress cases. We therefore interpret the LLM results as representative of PEFT-based fine-tuning and identify full-parameter fine-tuning and pre-training as important directions for future work.

Another limitation is that repeated end-to-end runs were not performed for each model–cooling configuration due to compute constraints. Therefore, sub-1% throughput and duration differences should be interpreted cautiously. Our primary evidence for cooling benefits instead comes from sustained temperature, power, and performance-per-watt trends observed over long-running workloads.

A further limitation is that our experiments are conducted on paired single-node 8× H100 systems rather than at multi-node or multi-rack scale. We therefore do not claim to fully characterize data-center-scale deployment behavior. However, controlled node-level measurements are necessary inputs to cluster-scale TCO, PUE, scheduling, and power-provisioning models, which rely on per-node thermal, power, and throughput characteristics [15], [16], [45]. By isolating cooling modality at the node level, our study avoids additional confounders such as network topology, synchronization overhead, workload placement, and rack-level airflow interactions. The results should therefore be interpreted as controlled node-level evidence that informs cluster-scale modeling, rather than as a complete data-center-scale deployment study.

VII. Conclusion

This study presents a comparative analysis of air-cooled and liquid-cooled GPU systems for QLoRA-based LLM fine-tuning, VLM training, and synthetic GPU stress testing. By benchmarking across identical GPU hardware configurations each equipped with 8× NVIDIA H100 GPUs, we found that liquid-cooled systems consistently improved thermal stability and power efficiency, while small throughput differences during production workloads were generally modest and should be interpreted cautiously.

Liquid-cooled system maintained GPU temperatures within a narrower and lower range (41–50°C), while air-cooled system exhibited greater variability and reached peak temperatures above 70°C. This thermal advantage translated into approximately 17% higher TFLOPS per GPU under the GPU Burn stress test, while model-training workloads showed smaller directional throughput improvements alongside clearer and more sustained power-efficiency gains. Additionally, node-level power consumption was significantly lower in liquid-cooled systems, indicating improved performance-per-watt across all workloads.

These results highlight the critical role of thermal design in optimizing AI infrastructure at scale. As large model training continues to grow in complexity and intensity, the adoption of liquid cooling offers a promising path forward for sustainable and energy-efficient data center operations.

Future work can extend this analysis to full-parameter LLM fine-tuning and pre-training using multi-GPU/ multi-node distributed training strategies such as DDP, FSDP, and ZeRO-style optimization to evaluate whether higher memory pressure and optimizer-state communication amplify the observed cooling benefits.

Future work can also include facility-level submetering of chiller power, pump power, CRAH/CRAC units, rear-door heat exchanger fans, and cooling-loop infrastructure to enable a complete PUE-inclusive comparison of liquid- and air-cooled AI systems.

Acknowledgment

During the preparation of this work, the author(s) used ChatGPT-4o to improve language clarity and typesetting. After using this tool, the author(s) reviewed and edited the content as necessary and took full responsibility for the final version of the publication.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[2] N. Rattner and T. Dotan, "The ai spending boom in charts," The Wall Street Journal, September 2024, tech Section. [Online]. Available: https://www.wsj.com/tech/ai/artificial-intelligence-investing-charts-7b8e1a97

[3] L. H. Kaack, P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig, and D. Rolnick, "Aligning artificial intelligence with climate change mitigation," Nature Climate Change, vol. 12, no. 6, pp. 518–527, 2022. [Online]. Available: https://doi.org/10.1038/s41558-022-01377-7

[4] I. Latif, A. C. Newkirk, M. R. Carbone, A. Munir, Y. Lin, J. Koomey, X. Yu, and Z. Dong, "Empirical measurements of AI training power demand on a GPU-accelerated node," arXiv preprint arXiv:2412.08602, 2024.

[5] ——, "Single-node power demand during ai training: Measurements on an 8-gpu nvidia h100 system," IEEE Access, 2025.

[6] A. C. Newkirk, J. Fernandez, J. Koomey, I. Latif, E. Strubell, A. Shehabi, and C. Samaras, "Empirically-calibrated h100 node power models for accurate ai training energy estimation," Environmental Research: Energy, vol. 2, no. 4, p. 045016, 2025.

[7] J. H. Kim, D. U. Shin, and H. Kim, "Data center energy evaluation tool development and analysis of power usage effectiveness with different economizer types in various climate zones," Buildings, vol. 14, no. 1, p. 299, 2024.

[8] A. Narayanan, Q. Wang, S. Ozguc, and R. W. Bonner III, "Investigation of server level direct-to-chip two phase cooling solution for high power gpus," in International Electronic Packaging Technical Conference and Exhibition, vol. 88469. American Society of Mechanical Engineers, 2024, p. V001T02A009.

[9] P. E. Ohenhen, O. Chidolue, A. A. Umoh, B. Ngozichukwu, A. V. Fafure, V. I. Ilojianya, and K. I. Ibekwe, "Sustainable cooling solutions for electronics: A comprehensive review: Investigating the latest techniques and materials, their effectiveness in mechanical applications, and associated environmental benefits," World Journal of Advanced Research and Reviews, vol. 21, no. 1, pp. 957–972, 2024.

[10] K. Zhou, X. Yu, B. Xie, H. Xie, and W. Fu, "Immersion cooling technology development status of data center," Science and Technology for Energy Transition, vol. 79, p. 41, 2024.

[11] F. Naduvilakath-Mohammed, M. Lebon, and A. Robinson, "Numerical modelling of a hybrid vapor compression refrigeration assisted closed loop liquid cooling system for high-performance computing systems," in Journal of Physics: Conference Series, vol. 2766, no. 1. IOP Publishing, 2024, p. 012078.

[12] B. Ramakrishnan, C. Turner, H. Alissa, D. Trieu, F. Rivera, L. Melton, M. Rao, S. Chigullapalli, T. Getachew, V. Prodanovic et al., "Understanding the impact of data center liquid cooling on energy and performance of machine learning and artificial intelligence workloads," Journal of Electronic Packaging, vol. 147, no. 2, p. 021003, 2025.

[13] J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, and E. Strubell, "Energy considerations of large language model inference and efficiency optimizations," in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 32 556–32 569. [Online]. Available: https://aclanthology.org/2025.acl-long.1563/

[14] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," in 2023 IEEE High Performance Extreme Computing Conference (HPEC), 2023, pp. 1–9.

[15] D. Zhao, S. Samsi, J. McDonald, B. Li, D. Bestor, M. Jones, D. Tiwari, and V. Gadepally, "Sustainable supercomputing for AI: GPU power capping at HPC scale," in Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC), 2023, pp. 588–596.

[16] J. Stojkovic, C. Zhang, Í. Goiri, and R. Bianchini, "Rearchitecting datacenter lifecycle for AI: A TCO-driven framework," arXiv preprint arXiv:2509.26534, 2025.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[19] Mistral AI Team, "Mistral nemo: Our new best small model," https://mistral.ai/news/mistral-nemo, July 2024, accessed March 2025. [Online]. Available: https://mistral.ai/news/mistral-nemo

[20] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju et al., "Gemma 2: Improving open language models at a practical size," 2024. [Online]. Available: https://arxiv.org/abs/2408.00118

[21] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei et al., "Qwen2. 5 technical report," arXiv preprint arXiv:2412.15115, 2024.

[22] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in European conference on computer vision. Springer, 2022, pp. 1–18.

[23] Z. Lin, S. Geng, R. Zhang, P. Gao, G. De Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in European Conference on Computer Vision. Springer, 2022, pp. 388–404.

[24] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23 034–23 044.

[25] Y. Tang, X. Zhang, and Z. Liu, "Experimental study on the thermal performance of flat loop heat pipe applied in data center cooling," Energies, vol. 16, no. 12, p. 4677, 2023.

[26] K. Haghshenas, B. Setz, Y. Blosch, and M. Aiello, "Enough hot air: the role of immersion cooling," Energy Informatics, vol. 6, no. 1, p. 14, 2023.

[27] M. Borkowski and A. Piłat, "Saving energy and efficiency increase by enabling free-cooling mode for cooling system in data center," in Proceedings of the 13th International Conference on Applied Energy (ICAE2021), ser. Energy Proceedings, vol. 23. ICAE, November 2021, iCAE International Conference on Applied Energy, Nov. 29–Dec. 5, 2021, Thailand/Virtual. Paper ID: 240.

[28] S. Zhu and D. Wang, "Energy-efficient llm training in gpu datacenters with immersion cooling systems," in Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems, 2025, pp. 407–414.

[29] M. F. Argerich and M. Patiño-Martínez, "Measuring and improving the energy efficiency of large language models inference," IEEE Access, vol. 12, pp. 80 194–80 207, 2024.

[30] C. Niu, W. Zhang, Y. Zhao, and Y. Chen, "Energy efficient or exhaustive? benchmarking power consumption of llm inference engines," ACM SIGENERGY Energy Informatics Review, vol. 5, no. 2, pp. 56–62, 2025.

[31] J. Dwyer, "Revisiting training scale: An empirical study of token count, power consumption, and parameter efficiency," 2026. [Online]. Available: https://arxiv.org/abs/2601.06649

[32] Y. Jin, G.-Y. Wei, and D. Brooks, "The energy cost of reasoning: Analyzing energy usage in llms with test-time compute," 2025. [Online]. Available: https://arxiv.org/abs/2505.14733

[33] R. Wang, X. Zhang, X. Zhou, Y. Wen, and R. Tan, "Toward physics-guided safe deep reinforcement learning for green data center cooling control," in 2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS). IEEE, 2022, pp. 159–169.

[34] J. Athavale, M. Yoda, and Y. Joshi, "Genetic algorithm based cooling energy optimization of data centers," International Journal of Numerical Methods for Heat & Fluid Flow, vol. 31, pp. 3148–3168, 2021.

[35] W. Ebirim, F. Usman, K. Olu-Lawal, N. Ninduwesuor-Ehiobu, E. Ani, and D. Montero, "Optimizing energy efficiency in data center cooling towers through predictive maintenance and project management," World Journal of Advanced Research and Reviews, vol. 21, no. 2, pp. 1782–1790, 2024.

[36] J. Kim, J. Lee, G. Park, B. Kim, S. J. Kwon, D. Lee, and Y. Lee, "An inquiry into datacenter tco for llm inference with fp8," arXiv preprint arXiv:2502.01070, 2025.

[37] R. Geens, M. Shi, A. Symons, C. Fang, and M. Verhelst, "Energy cost modelling for optimizing large language model inference on hardware accelerators," in 2024 IEEE 37th International System-on-Chip Conference (SOCC). IEEE, 2024, pp. 1–6.

[38] A. K. Kakolyris, D. Masouros, P. Vavaroutsos, S. Xydis, and D. Soudris, "throttll'em: Predictive gpu throttling for energy efficient llm inference serving," in 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2025, pp. 1363–1378.

[39] G. Wilkins, S. Keshav, and R. Mortier, "Hybrid heterogeneous clusters can lower the energy consumption of llm inference workloads," in Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems, 2024, pp. 506–513.

[40] R. B. Guo, U. Anand, K. Daudjee, and R. Sen, "Zorse: Optimizing llm training efficiency on heterogeneous gpu clusters," arXiv preprint arXiv:2507.10392, 2025.

[41] P. Patel, E. Choukse, C. Zhang, Í. Goiri, B. Warrier, N. Mahalingam, and R. Bianchini, "Polca: Power oversubscription in llm cloud providers," arXiv preprint arXiv:2308.12908, 2023.

[42] T. R. Reddy, B. Kataria, R. Gandhi, K. Tandon, D. Bhattacherjee, V. N. Padmanabhan et al., "Improving training time and gpu utilization in geo-distributed language model training," arXiv preprint arXiv:2411.14458, 2024.

[43] T. Zhao, R. Sun, X. Hou, J. Huang, W. Geng, and J. Jiang, "Simulation study of influencing factors of immersion phase-change cooling technology for data center servers," Energies, vol. 16, no. 12, p. 4640, 2023.

[44] D. Jia, W. He, C. Xu, and T. Guo, "Simulation study of modular independent cooling systems for servers," in Journal of Physics: Conference Series, vol. 2835, no. 1. IOP Publishing, 2024, p. 012066.

[45] J. J. Tithi, H. Wu, A. Abuhatzera, and F. Petrini, "Scaling intelligence: Designing data centers for next-gen language models," arXiv preprint arXiv:2506.15006, 2025.

[46] R. Gupta, S. Asgari, H. Moazamigoodarzi, D. G. Down, and I. K. Puri, "Energy, exergy and computing efficiency based data center workload and cooling management," Applied Energy, vol. 299, p. 117050, 2021.

[47] B. Ramakrishnan, C. Turner, H. Alissa, D. Trieu, F. Rivera, L. Melton, M. Rao, S. Chigullapalli, T. Getachew, V. Prodanovic et al., "Understanding the impact of data center liquid cooling on energy and performance of machine learning and artificial intelligence workloads," ASME Journal of Electronic Packaging, vol. 147, no. 2, p. 021003, 2025.

[48] A. Tschand, A. T. R. Rajan, S. Idgunji, A. Ghosh, J. Holleman et al., "MLPerf power: Benchmarking the energy efficiency of machine learning systems from μWatts to MWatts for sustainable AI," arXiv preprint arXiv:2410.12032, 2024, to appear in HPCA 2025.

[49] A. C. Newkirk, J. Fernandez, J. Koomey, I. Latif, E. Strubell, A. Shehabi, and C. Samaras, "Empirically-calibrated H100 node power models for accurate AI training energy estimation," Environmental Research: Energy, vol. 2, no. 4, p. 045016, 2025.

[50] I. Latif, A. C. Newkirk, M. R. Carbone, A. Munir, Y. Lin, J. Koomey, X. Yu, and Z. Dong, "Single-node power demand during AI training: Measurements on an 8-GPU NVIDIA H100 system," IEEE Access, vol. 13, pp. 61 740–61 747, 2025.

[51] J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, and E. Strubell, "Energy considerations of large language model inference and efficiency optimizations," in Proceedings of the 63rd Annual Meeting of the

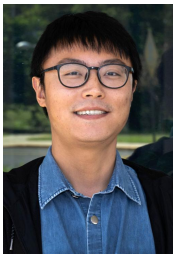
Association for Computational Linguistics (ACL), Volume 1: Long Papers, 2025, pp. 32 556–32 569.
[52] Advanced Micro Devices (AMD), "Amd epyc 9474f processor," 2022, accessed: 2026-03-16. [Online]. Available: https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9474f.html
[53] Advanced Micro Devices, "Amd epyc 9534 processor specifications," 2022, 64 cores, 128 threads, base clock 2.45 GHz, boost up to 3.7 GHz, 256 MB L3 cache, 280 W TDP. Accessed: 2026-03-16. [Online]. Available: https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9534.html
[54] ipmitool contributors, "ipmitool - utility for ipmi-enabled systems," https://github.com/ipmitool/ipmitool, 2025, accessed: April 11, 2025.
[55] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/
[56] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, L. O. N. de Araújo, JPW, and MinervaBooks, "mlco2/codecarbon: v2.4.1," May 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11171501
[57] V. Timonen, "wilicc/gpu-burn," https://github.com/wilicc/gpu-burn, October 2024, original-date: 2017-11-25. [Online]. Available: https://github.com/wilicc/gpu-burn
[58] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.
[59] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

Imran Latif (Member, IEEE) received the master's degree in mechanical engineering from The City College of New York. He is currently Vice President of Global Technology and Innovation, Data Centers, at Johnson Controls. Before joining Johnson Controls, he was an Engineering Executive and the Chief Operations Officer with the Brookhaven National Laboratory, where he led the Infrastructure Operations of High Performance Computing Centers, supported the global cutting-edge research collaborations on physics, life sciences, quantum, AI, and HPC. He has over 20 years of leadership experience delivering large scale engineering, construction, and sustainability projects. He leads worldwide sustainability research with leading academia and tech companies. His research work is focused on solving the real-life energy issues in the mission critical facilities and developing the AI and ML-based technologies to automate the sustainability processes. He is a subject matter Expert in direct to chip and immersion cooling in data centers.

Muhammad Ali Shafique received the Bachelor of Science and master's degrees from the University of Engineering and Technology, Lahore, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at Kansas State University. He also holds the position of a Research Assistant with the ISCAAS Laboratory. His research interest includes design of efficient LLMs in resource-constrained systems. He is committed to conducting research in this area with the aim of advancing the current knowledge and understanding of these fields.

Hayat Ullah received his Bachelor's degree in Computer Science from Islamia College University, Peshawar, Pakistan, in 2018, and his Master's degree in Computer Science from Sejong University, Seoul, Republic of Korea, in 2021. He is currently pursuing his Ph.D. in Computer Science at Florida Atlantic University, Boca Raton, FL, USA. He is also a Research Assistant with the Intelligent Systems, Computer Architecture, Analytics, and Security (ISCAAS) Laboratory, Florida Atlantic University. He is exclusively working on multi-model human actions modeling and activity recognition. He has published several articles in well-reputed journals, including IEEE Internet of Things Journal and IEEE Transactions on Image Processing. His research interests include human action recognition, temporal action localization, knowledge distillation, adversarial robustness, vision-language models for video analytics, distributed training, and HPC nodes performance benchmarking using LLMs and VLMs.

ALEX C. NEWKIRK received the bachelor's degree in physics from Carleton College and the Ph.D. degree in engineering and public policy researching the criticality of semiconductors from Carnegie Mellon University. He is an Energy Technology Researcher with the Lawrence Berkeley National Laboratory and a member of the Center of Expertise for Energy Efficiency in Data Centers. He has led or contributed to research on organizational barriers to energy efficiency and AI related computational demand.



Xi YU (Member, IEEE) received the master's and Ph.D. degrees in electrical and computer engineering from the University of Florida, in 2019 and 2022, respectively. He is currently a Postdoctoral Research Associate with the Computing and Data Sciences Directorate, Brookhaven National Laboratory. His research interests include domain generalization, robust machine learning, information theory, and scientific imaging applications.

Arslan Munir (M'09, SM'17) received his M.A.Sc. degree in electrical and computer engineering (ECE) from the University of British Columbia, Vancouver, Canada, in 2007, and his Ph.D. degree in ECE from the University of Florida, Gainesville, FL, USA, in 2012. From 2007 to 2008, he worked as a Software Development Engineer at the Embedded Systems Division, Mentor Graphics Corporation. He was a Postdoctoral Research Associate with the ECE Department at Rice University, Houston, TX, USA, from May 2012 to June 2014. He is currently a Professor in the Department of Electrical Engineering and Computer Science at Florida Atlantic University.

Munir's current research interests include embedded and cyber-physical systems, secure and trustworthy systems, parallel computing, quantum computing, artificial intelligence, and computer vision. He has received many academic awards, including the Doctoral Fellowship from the Natural Sciences and Engineering Research Council (NSERC) of Canada. He earned gold medals for best performance in electrical engineering and gold medals and academic roll of honor for securing rank one in pre-engineering provincial examinations (out of approximately 300,000 candidates).