AI

MDPI

Article

Scalable and Energy-Efficient AI: System-Level Profiling of NVIDIA GPU Clusters for Distributed LLM Training

Muhammad Ali Shafique [1], Imran Latif [2], Hayat Ullah [3], Alex C. Newkirk [4] and Arslan Munir [3,*]

[1] Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA; alishafique@ksu.edu
[2] Johnson Controls, Milwaukee, WI 53209, USA; imran.latif@jci.com
[3] Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA; hullah2024@fau.edu
[4] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; acnewkirk@lbl.gov
* Correspondence: arslanm@fau.edu

Abstract

The rapid scaling of large language model (LLM) training has intensified demand for Graphics Processing Unit (GPU) clusters balancing throughput with energy efficiency. While NVIDIA's H100 and B200 architectures are increasingly deployed in production datacenters, their comparative behavior under distributed training remains insufficiently characterized beyond vendor specifications, leaving datacenter operators without empirical guidance on metrics such as TFLOPs/kW and tokens-per-kilojoule. This work presents a system-level evaluation of single-node 8× H100 and 8× B200 configurations using Distributed Data Parallel (DDP) training across LLMs and vision–language models (VLMs) ranging from 7B to 32B parameters, spanning various real AI workload scenarios. We benchmark end-to-end throughput, utilization, power, energy, TFLOPs/kW, and tokens-per-kilojoule, complemented by architectural analysis explaining observed behavioral differences. Across LLM workloads, B200 achieves higher utilization (1–6%), faster training (up to 15%), and greater compute efficiency (up to 32% higher TFLOPs/GPU), attributable to higher memory bandwidth and large streaming multiprocessor (SM) count. However, B200 exhibits lower TFLOPs/kW and tokens-per-kilojoule, revealing a fundamental trade-off: throughput gains come at a measurable energy cost per useful token. VLM results further expose model-dependent asymmetries, with B200 consuming disproportionately more energy for lighter compute kernels due to elevated baseline power draw. These findings provide an empirical framework distinguishing compute efficiency from energy efficiency across next-generation GPU nodes, offering practical guidance for energy-aware AI datacenter design.

Keywords: large language models; vision–language models; distributed training; benchmarking; energy efficiency; profiling

Check for updates

Academic Editor: Olufemi A. Omitaomu

Received: 11 April 2026
Revised: 30 May 2026
Accepted: 3 June 2026
Published: 23 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid progress of large language models (LLMs) and vision–language models (VLMs) has increased computational demands in high-performance computing environments. As model sizes grow and training pipelines scale to thousands of GPUs, power efficiency, thermal stability, and compute throughput have become central considerations for designing AI infrastructure. Recent scaling laws highlight clear relationships between model size, dataset volume, and training compute [1], driving the need for specialized accelerators capable of sustaining high utilization across long-running workloads. Cloud

providers and research institutions continue to expand AI capacity in response to increasing demand [2], deploying GPU architectures optimized for large-scale distributed training across scientific [3] and industrial [4] domains.

However, the growth of AI systems has also increased the power footprint of training clusters, raising concerns regarding sustainability, total cost of ownership, and overall datacenter efficiency [5,6]. Recent LLM and VLM energy studies increasingly show that accelerator evaluation should move beyond peak FLOPs and include useful system-level efficiency metrics such as TFLOPs/kW, energy-per-token, and tokens-per-kilojoule [7–11]. These metrics are important because energy behavior can vary significantly with model architecture, workload geometry, hardware platform, utilization regime, and the amount of useful work completed per unit energy. As accelerator performance continues to scale, the central infrastructure question is therefore not only whether a newer GPU is faster but whether its throughput gains translate into proportional improvements in energy efficiency.

This issue reflects a broader compute–energy misalignment in modern AI systems: higher peak throughput, larger memory bandwidth, or higher GPU utilization do not automatically imply better energy efficiency under real workloads. Workload characteristics such as arithmetic intensity, memory access behavior, communication overhead, and tensor-core saturation can strongly affect whether a system converts additional compute capability into useful work per unit energy. Therefore, benchmarking methodologies that capture real workload behavior are critical. Prior studies have evaluated thermal and power behavior of GPU systems under traditional machine learning workloads [12,13], while recent AI energy studies have shown that training and inference energy consumption can deviate substantially from vendor specifications and depend strongly on workload and hardware configuration [7,8,14,15]. Yet systematic cross-architecture measurements for modern LLM and VLM training pipelines remain limited.

Unlike earlier generations of deep learning models, LLMs and VLMs present heterogeneous compute characteristics. Autoregressive text models typically stress transformer and tensor-core pipelines, whereas multimodal models often include a mixture of convolutional, attention, pooling, and vision-specific kernels with more irregular utilization profiles. As a result, accelerator performance cannot be understood solely from peak FLOPs or isolated microbenchmarks. Instead, end-to-end system metrics must be measured under real training loads with consistent software configurations, shared workload settings, and energy-aware metrics such as TFLOPs/kW and tokens-per-kilojoule.

In this work, we present a comprehensive benchmarking study comparing single-node 8× NVIDIA H100 and 8× NVIDIA B200 configurations under Distributed Data Parallel (DDP) training across five LLMs, Mistral-7B-v0.3 [16], LLaMA-3.1-8B [17], Mistral-NeMo-Base-2407 [18], Gemma-2-27B [19], and Qwen2.5-32B [20], and three VLMs: X-CLIP [21], EVL [22], and Vita-CLIP [23]. Models were selected to reflect realistic datacenter workloads spanning sequence-intensive, vision-heavy, and compute-bound regimes, ranging from 7B to 32B parameters. All experiments use real end-to-end training workloads under identical hyperparameters and software configurations, ensuring that observed differences are attributable to architectural distinctions between the two GPU generations.

We measure system-level characteristics including training duration, GPU utilization, power draw, energy consumption, and efficiency metrics such as TFLOPs/GPU, TFLOPs/kW, and tokens-per-kilojoule. Telemetry collected through NVIDIA-SMI and Weights & Biases (W&B) logs provides fine-grained insight into utilization and power dynamics during sustained multi-GPU operation. Beyond reporting raw measurements, we position the H100-to-B200 comparison as an empirical study of compute–energy misalignment across GPU generations. Specifically, we analyze when B200's higher memory bandwidth and increased streaming multiprocessor (SM) count improve throughput and

https://doi.org/10.3390/ai7070232

when its higher power floor reduces useful energy efficiency. This framing distinguishes compute efficiency from energy efficiency across both LLM and VLM workloads.

Overall, this study provides a literature-grounded, system-level empirical comparison of H100 and B200 accelerators under real LLM and VLM training pipelines. The results offer practical guidance for researchers and datacenter architects seeking to optimize throughput, energy efficiency, workload placement, and resource allocation in modern AI infrastructure.

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

1. We present a controlled system-level comparison of LLM and VLM training workloads on single-node 8× NVIDIA H100 and 8× NVIDIA B200 configurations under identical software, dataset, precision, and training setups, providing empirical evidence beyond vendor peak specifications.
2. We frame the H100-to-B200 transition as a compute–energy misalignment problem, showing that higher TFLOPs/GPU and faster training do not necessarily translate into improved TFLOPs/kW or tokens-per-kilojoule under real distributed training workloads.
3. We provide a workload-sensitive analysis across sequence-intensive, vision-heavy, and compute-bound workloads, showing that B200's efficiency depends strongly on workload saturation: compute-intensive models better amortize its higher power floor, while lighter or partially saturated workloads can lose useful energy efficiency.
4. We connect node-level measurements to datacenter-scale implications through facility-level energy and cost modeling, highlighting how accelerator selection depends on workload mix, power density, cooling assumptions, and operational priorities.
5. We provide a reproducible benchmarking framework for next-generation GPU nodes by reporting telemetry methodology, software configuration, precision datatype, fixed seed setting, and public code/environment files, enabling future work on multi-node scaling and workload-aware accelerator selection.

Significant Results: The most impactful findings of our benchmarking study are summarized below:

1. Across LLM workloads, B200 systems achieve 1–6% higher GPU utilization, up to 15% faster training time, and up to 32% higher TFLOPs/GPU compared with H100 nodes, driven by B200's higher memory bandwidth and increased SM count enabling greater sustained kernel occupancy.
2. Despite raw compute gains, B200 systems consistently trail H100 in TFLOPs/kW and tokens-per-kilojoule across LLM workloads, revealing that generational throughput improvements do not translate proportionally into energy efficiency under real distributed training conditions.
3. VLM workloads exhibit model-dependent behavior: sequence-intensive models benefit from higher utilization on B200, while lighter compute kernels incur disproportionately higher energy consumption due to B200's elevated baseline power draw relative to H100.
4. Vita-CLIP presents a notable outlier, where B200 achieves nearly 2× faster training with lower total energy consumption, suggesting that sufficiently compute-dense VLM workloads can amortize B200's higher power envelope into net efficiency gains.
5. According to facility-scale analysis, B200 deployment incurs a measurable energy cost premium over H100 across both load profiles: 76.6 GWh versus 70.4 GWh (+$0.62M/year) under moderate load (2000 nodes, 50% low/50% high intensity), widening to 217.8 GWh versus 185.0 GWh (+$4.26M/year) under high load (5000 nodes, 15% low/85% high intensity), confirming that B200's throughput advantages come at a significant operational energy cost that scales with deployment size and workload intensity.

The remainder of this paper is organized as follows: Section 2 reviews prior work on large-scale AI infrastructure and the motivation for evaluating next-generation GPU architectures. Section 3 details the benchmarking methodology used to measure system performance, power consumption, and efficiency under real training workloads. In Section 4, we present the comparative performance, utilization, and energy-efficiency results for H100 and B200 GPU nodes. Section 5 provides an analysis of these findings and discusses their implications for scalable and energy-aware AI system design. Finally, Section 6 concludes the paper and outlines opportunities for future research in benchmarking and optimizing emerging accelerator architectures.

2. Background and Motivation

The rapid advancement of LLMs and VLMs has increased computational and energy demands on modern AI infrastructure. As models scale in size and training duration, achieving high hardware utilization and power efficiency has become essential to both performance and sustainability. Accelerator behavior, including throughput, utilization patterns, memory bandwidth, and energy requirements, directly influences the practicality and cost of large-scale AI deployment [14,15]. Recent scaling laws highlight strong relationships between model capacity, compute budget, and achievable performance [1], motivating the development of increasingly specialized GPU architectures capable of sustaining high compute density under distributed training.

This growth has substantially increased power demand and operational costs, elevating energy efficiency and total cost of ownership (TCO) to central considerations in AI datacenter design [5,6]. More broadly, this aligns with the green computing principle of energy-proportional computing, where efficient systems should consume power in proportion to useful work rather than maintaining a high baseline power floor [24,25]. In modern AI systems, this concern appears as compute–energy misalignment: newer accelerators may improve peak FLOPs, memory bandwidth, and training speed, but these gains do not necessarily translate into proportional improvements in useful energy efficiency. Recent benchmarking efforts such as MLPerf Power further emphasize the need for standardized energy-efficiency evaluation across hardware platforms and workload classes [26]. Metrics such as TFLOPs/kW, energy-per-token, and tokens-per-kilojoule are therefore increasingly important for determining whether additional compute capability produces more useful work per unit energy.

Recent studies on LLM training and inference show that observed power draw and throughput can deviate significantly from vendor specifications, remaining highly sensitive to workload characteristics, batch size, sequence length, and system configuration [7,10,14]. Recent token-level LLM power benchmarks further show the importance of measuring energy-normalized metrics such as joules-per-token, tokens-per-joule, or tokens-per-kilojoule rather than throughput alone [27]. Complementary work demonstrates that increasing training token counts can be energetically inefficient even when marginal performance gains are observed [11] and that inference energy consumption is highly sensitive to workload geometry, model family, and hardware accelerator choice [8,9]. Together, these findings reinforce the need for efficiency-aware evaluation beyond conventional performance metrics, motivating holistic system-level benchmarking rather than reliance on peak hardware specifications alone.

Beyond raw compute, system efficiency is shaped by memory bandwidth, communication overheads, thermal behavior, utilization regime, and power-management policies, factors often obscured in synthetic benchmarks [28,29]. From a computer architecture perspective, accelerator efficiency depends not only on available compute units but also on whether the workload can keep those units saturated without incurring excessive memory stalls, communication delays, or idle power draw. This is particularly important for LLM and VLM workloads: transformer-heavy LLMs may better exploit tensor-core pipelines, whereas multimodal VLMs often combine attention, convolutional, pooling, and vision-specific kernels that can produce more irregular utilization patterns.

At the infrastructure level, cooling, scheduling, and power-management mechanisms further influence realized efficiency. Advances in cooling technologies have been shown to improve sustained performance and energy efficiency for LLM workloads [30,31], while scheduling strategies, power capping, and distributed training design influence cluster-level efficiency through tightly coupled hardware–software–workload interactions [32–35]. These studies collectively show that energy-aware AI system design requires coordinated evaluation across hardware architecture, workload behavior, and datacenter operating conditions.

Despite this growing body of work, empirical cross-architecture comparisons of next-generation GPU nodes under consistent distributed training configurations remain limited. Existing online comparisons and vendor reports focus predominantly on raw throughput without distinguishing compute efficiency from energy efficiency or providing architectural reasoning for observed differences. In this work, we address this gap by benchmarking single-node 8× NVIDIA H100 and 8× NVIDIA B200 configurations across LLM and VLM workloads spanning sequence-intensive, vision-heavy, and compute-bound regimes. By grounding the comparison in TFLOPs/GPU, TFLOPs/kW, tokens-per-kilojoule, and facility-level cost modeling, this study provides practical and literature-grounded insights for developers, researchers, and infrastructure architects optimizing large-scale AI deployment.

3. Methodology and Experimental Setup

This section describes the methodology used to benchmark system performance, power consumption, and computational efficiency across 8× NVIDIA H100 and 8× NVIDIA B200 GPU nodes. Our evaluation focuses exclusively on real training workloads for LLMs and VLMs, ensuring that all measurements reflect realistic multi-GPU behavior under Distributed Data Parallel (DDP) execution. Each reported measurement corresponds to a single sustained training run for each workload and hardware configuration with fixed seed value for reproducibility.

3.1. Training Workloads

We evaluate five LLMs, Mistral-7B-v0.3 [16], LLaMA-3.1-8B [17], Mistral-NeMo-Base-2407 [18], Gemma-2-27B [19], and Qwen2.5-32B [20], and three VLMs: X-CLIP [21], EVL [22], and ViTA-CLIP [23]. All models were trained under identical hyperparameters across both GPU systems to ensure fair comparison. Training experiments were conducted using BF16 precision. We explicitly report the precision format because achieved TFLOPs and compute-efficiency metrics are precision-dependent metrics.

3.1.1. LLM Training Setup

All LLMs were fine-tuned using the Alpaca dataset [36] in a supervised setting. Parameter-efficient fine-tuning (PEFT) with Quantized Low-Rank Adaptation (QLoRA) was applied using a rank $r = 128$. Training was conducted for one epoch with a micro-batch size of 4 per GPU, yielding an effective global batch size of 32 under 8-way DDP. This uniform configuration allows consistent cross-model and cross-architecture comparisons.

3.1.2. VLM Training Setup

For VLM evaluation, we trained all models on the UCF101 dataset [37]. Each model was trained for 50 epochs using a global batch size of 32 (4 samples per GPU) and a sequence length of 16 frames. Training was performed under 8-way DDP across both the H100 and B200 systems, with no variation in hyperparameters across architectures.

3.2. Measurement and Profiling Methodology

To characterize the performance and energy behavior of both GPU architectures, we captured a comprehensive set of GPU-level and node-level metrics throughout each training run. GPU measurements were collected using *NVIDIA-SMI* version 570.195.03, while node-level power was logged using *ipmitool* at 20-s intervals.

The following metrics were recorded continuously or at fixed sampling intervals:

- GPU Utilization: Real-time compute utilization per GPU to assess saturation and workload efficiency.
- GPU Power Draw: Instantaneous and averaged per-GPU power measurements for computing performance-per-watt.
- Achieved FLOPs: Effective TFLOPs/GPU measured during training using runtime FLOP estimators.
- Training Time: End-to-end training time as shown in Figure 1.

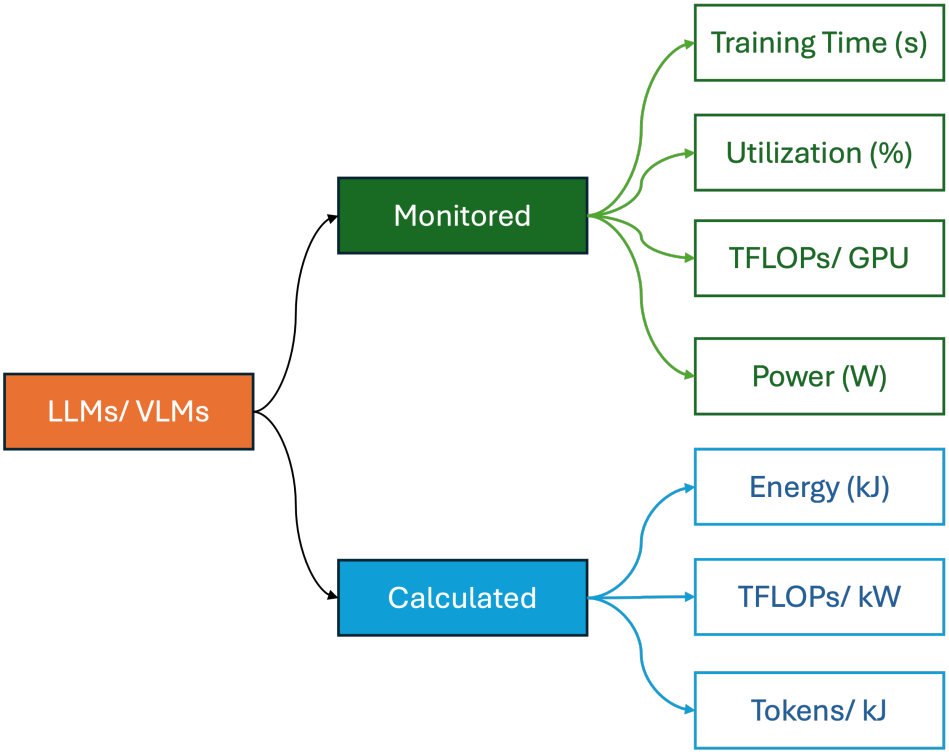


Figure 1. Metrics classification based on measurement and calculation in this study.

3.3. Benchmarking Objective

This unified methodology enables direct comparison of sustained multi-GPU performance across architectures, focusing on:

- Training time differences across workloads;
- GPU utilization and compute efficiency;
- Power consumption and energy-per-token;
- TFLOPs/GPU and TFLOPs/kW efficiency;
- Architectural behavior under realistic LLM/VLM training patterns.

This framework provides a reproducible and architecture-agnostic process for evaluating next-generation accelerators on end-to-end AI training workloads.

We use Distributed Data Parallel (DDP) across all experiments to maintain a controlled and consistent comparison between the H100 and B200 nodes. Although advanced sharding strategies such as Fully Sharded Data Parallel (FSDP) and ZeRO are commonly used in production training for larger models, we do not enable them in this study. These techniques introduce additional variables, including optimizer-state sharding, parameter partitioning, memory offloading, and modified communication patterns, which could obscure the hardware-level comparison between GPU architectures. Therefore, DDP is selected as a uniform distributed training baseline to isolate differences in throughput, utilization, power draw, and energy efficiency between the two systems; multi-node techniques FSDP/ZeRO should be considered for future work.

4. Results

This section presents the performance, power, and efficiency results obtained from training LLM and VLM workloads on 8× NVIDIA H100 and 8× NVIDIA B200 GPU nodes. All measurements were collected during real training runs under identical hyperparameters and DDP configurations. The evaluation focuses on end-to-end training duration, GPU utilization, achieved FLOPs, and efficiency metrics such as TFLOPs/GPU, TFLOPs/kW, and tokens-per-kilojoule.

Before interpreting the energy-efficiency results, it is important to note that the B200 has a higher baseline power floor than the H100. This is consistent with architectural differences between Hopper and Blackwell, including B200's larger dual-die design, higher transistor count, expanded HBM3e memory subsystem, higher memory bandwidth, and additional hardware engines [38,39]. These factors help explain why B200 improves throughput and TFLOPs/GPU under compute-intensive workloads but can show lower energy efficiency for lighter or partially saturated workloads.

4.1. LLM Results

Table 1 summarizes the performance of five LLMs ranging from 7B to 32B parameters. Across all models, the B200 consistently delivers faster training throughput compared with H100. Smaller models, such as Mistral-7B, LLaMA-3.1-8B, and Mistral-NeMo, complete training 300–500 s faster on B200, while larger models, including Gemma-2-27B and Qwen2.5-32B, show reductions of 1282–1436 s, reflecting up to 15% end-to-end speedup.

GPU utilization is increased on B200 by 1–6% relative to H100, driven by greater memory bandwidth and increased SM count enabling sustained kernel occupancy under transformer-heavy workloads. Correspondingly, B200 achieves up to 32% higher TFLOPs/GPU (e.g., 50.96 TFLOPs/GPU for Gemma-2-27B vs. 39.17 on H100).

Table 1. Training and energy-efficiency metrics across LLM and VLM models on H100 and B200 nodes.

Model	Node	Duration (s)	Avg. 8× GPU Util. (%)	Avg. 8× GPU Power (W)	Avg. GPU Energy (kJ)	TFLOPS/ GPU	TFLOPs/ kW	Tokens/ kJ
Large Language Models (LLMs)								
Mistral-7B-v0.3 [16]	H100	2702	95.8	437.8	1182.9	35.19	80.4	5634
	B200	2370	96.6	579.8	1374.1	46.56	80.3	4850
Llama-3.1-8B [17]	H100	2466	94.9	438.1	1080.3	35.12	80.1	5217
	B200	2191	96.2	580.4	1271.6	40.64	70.0	4432
Mistral-Nemo [18]	H100	3827	95.8	444.5	1701.1	35.71	80.3	3385
	B200	3323	97.1	604.3	2008.0	42.85	70.9	2867
Gemma-2-27B [19]	H100	8393	92.7	467.7	3925.4	39.17	83.8	1517
	B200	7111	98.5	616.3	4382.5	50.96	82.7	1358
Qwen2.5-32B [20]	H100	9738	93.4	458.8	4467.7	37.42	81.6	1270
	B200	8302	98.70	605.1	5023.5	48.7	80.5	1129
Vision–Language Models (VLMs)								
X-CLIP [21]	H100	1503	58.6	408.5	613.9	0.377	0.923	99,414
	B200	2094	56.1	400.6	838.8	0.376	0.938	72,759
EVL [22]	H100	14,530	60.0	290.0	4213.7	0.119	0.410	14,483
	B200	16,749	66.1	279.0	4673.0	0.119	0.426	13,060
Vita-CLIP [23]	H100	91,654	97.4	536.9	49,209.0	0.106	0.198	1240
	B200	47,911	78.5	518.5	24,841.8	0.106	0.205	2456

However, B200's increased instantaneous power draw results in lower TFLOPs/kW and reduced tokens-per-kilojoule across all five LLMs, confirming that compute efficiency gains do not translate proportionally into energy efficiency, a critical consideration for power-constrained datacenter deployment.

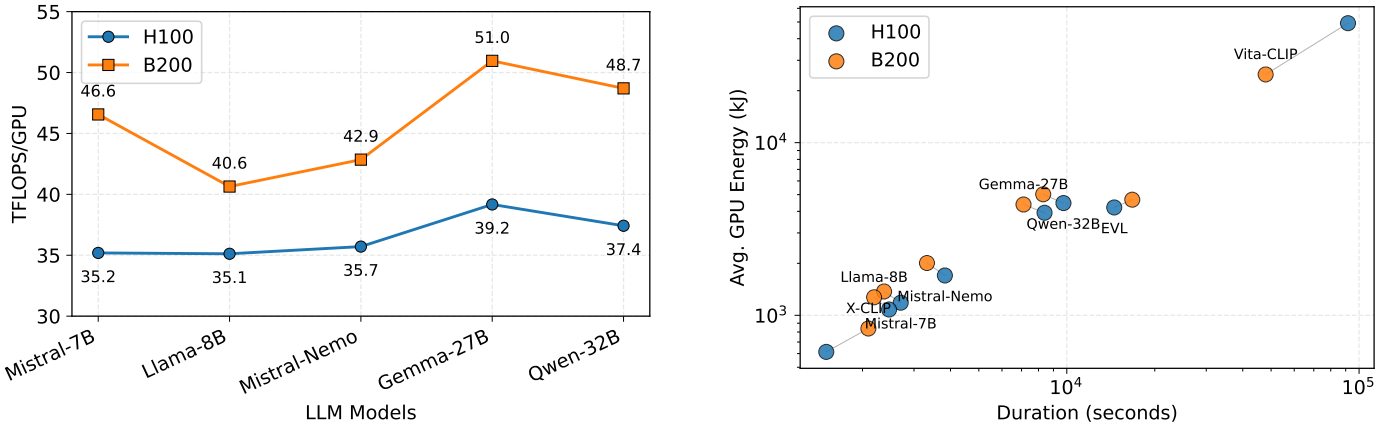
4.2. VLM Results

We benchmarked three VLMs, X-CLIP, EVL, and Vita-CLIP, representing diverse multimodal training patterns. Results are summarized in Table 1. In contrast to LLMs, VLM workloads exhibit more heterogeneous behavior due to differences in convolutional, attention, and vision-specific kernel compositions.

X-CLIP and EVL show modest utilization gains on B200 (1–6%) with marginally longer training durations. These models rely heavily on convolutions, pooling, and sparse attention kernels that do not fully saturate B200's tensor cores. Combined with B200's elevated baseline power draw, this results in disproportionately higher energy consumption despite comparable throughput.

Vita-CLIP presents a notable outlier: B200 achieves nearly 2× faster training with lower total energy consumption, attributable to its transformer-style vision–language fusion layers with higher arithmetic intensity that align well with B200's tensor-core, optimized design. This demonstrates that sufficiently compute-dense VLM workloads can amortize B200's higher power envelope into net efficiency gains.

Together, these results characterize the practical performance impact of transitioning from H100 to B200 across both language and vision–language model families, as shown in Figure 2, and highlight that next-generation accelerator benefits remain strongly workload-dependent.



(**a**) (**b**)

Figure 2. Compute throughput comparison between H100 and B200 GPU nodes for LLM workloads. (**a**) TFLOPS-per-GPU across different LLM models, highlighting the per-GPU throughput advantage of B200. (**b**) Average GPU energy consumption as a function of training duration, showing the energy-runtime behavior of H100 and B200 across model scales.

5. Discussion

The empirical results across LLM and VLM workloads reveal several important insights regarding architectural behavior, performance scaling, and energy characteristics when transitioning from NVIDIA H100 to B200 GPU systems. This section critically examines these findings, highlighting both the strengths and limitations of the B200 under real distributed training workloads.

5.1. Implications of Higher Utilization and Compute Throughput

Across all LLM workloads, the B200 consistently achieves higher GPU utilization and effective TFLOPs/GPU, driven by greater memory bandwidth and increased SM count enabling sustained kernel occupancy under transformer-heavy workloads. The observed reductions in training time, up to 15%, demonstrate that these architectural enhancements yield substantial practical benefits. However, these gains come with increased instantaneous power per GPU. B200's higher utilization amplifies its power footprint, leading to lower TFLOPs/kW and reduced tokens-per-kilojoule across all LLMs, indicating that compute efficiency improvements do not translate proportionally into energy efficiency. This distinction is critical for datacenters operating under strict power and cooling constraints.

5.2. Workload Sensitivity in Vision–Language Models

VLM workloads exhibit more heterogeneous scaling behavior compared with LLMs. X-CLIP and EVL show modest utilization gains and marginally longer training durations on B200, as their reliance on convolutions, pooling, and sparse attention cannot fully saturate B200's tensor cores. Combined with B200's elevated baseline power draw, this results in disproportionately higher energy consumption for lighter compute kernels, as shown in Figure 3.

In contrast, Vita-CLIP benefits considerably, achieving nearly 2× faster training with lower total energy consumption, owing to its transformer-style fusion layers with higher arithmetic intensity that align well with B200's tensor-core, optimized design. This highlights a key insight: next-generation accelerator benefits are strongly workload-dependent, and models with irregular compute patterns may not benefit uniformly.

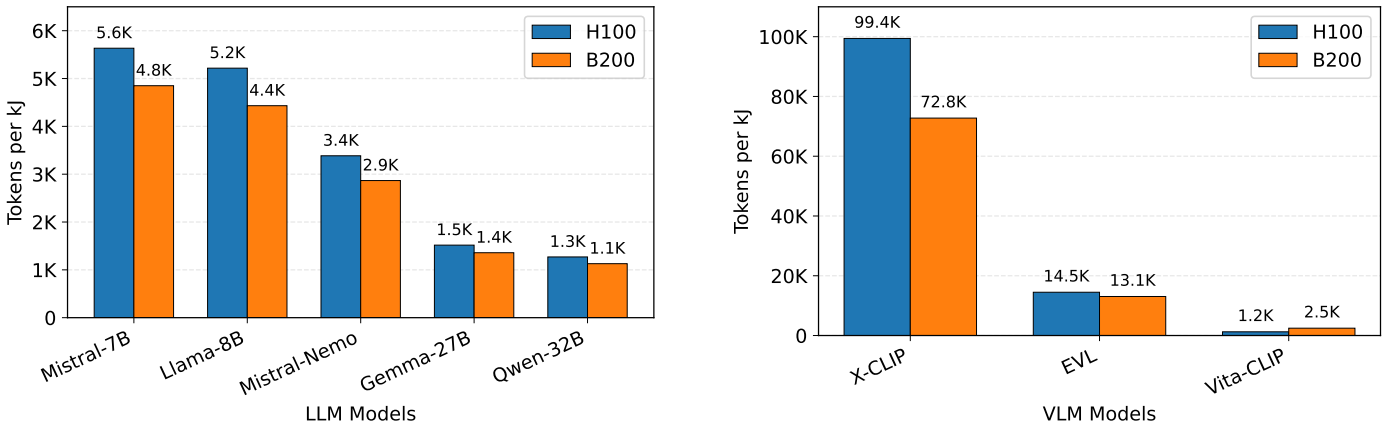


Figure 3. Energy-efficiency comparison between H100 and B200 GPU nodes for (**a**) LLM workloads and (**b**) VLM workloads. Tokens-per-kilojoule highlights differences in energy efficiency across model families.

5.3. Energy Efficiency and System-Level Characteristics

The energy results reveal a clear divergence between throughput improvements and energy metrics. Despite accelerating training substantially, B200's higher instantaneous power draw leads to lower TFLOPs/kW and reduced tokens-per-kilojoule, confirming that generational throughput gains do not translate proportionally into energy efficiency. B200's lower energy efficiency in some workloads is not simply due to higher measured power but to the interaction between architectural power floor, workload saturation, and the amount of useful compute extracted from the available hardware resources.

It is important to note that we did not evaluate power-capped configurations in this study. Power capping may reduce B200's instantaneous power draw and improve tokens-per-kilojoule efficiency, potentially narrowing the energy-efficiency gap with H100 at the cost of some training throughput. Evaluating B200 under different power caps is therefore an important direction for future work, particularly for datacenters operating under strict power or cooling constraints.

5.4. Architectural Considerations and Practical Trade-Offs

B200's advantages appear most prominently in workloads that fully saturate its tensor-core pipelines. Developers targeting multimodal or vision-centric workloads may need to optimize kernel composition, through operator fusion, dynamic shape tuning, or data pipeline optimizations, to realize substantial gains.

A limitation of this study is that our experiments are restricted to DDP-based single-node training. While this setup enables a controlled comparison between H100 and B200 systems, the production-scale training of larger models often relies on memory-efficient sharding approaches such as FSDP and ZeRO. These methods can alter memory pressure, communication overhead, GPU utilization, and power behavior. Evaluating H100 and B200 under FSDP, ZeRO, and hybrid parallelism across multi-node deployment is therefore an important direction for future work.

5.5. Cross-Generational Implications for Energy-Aware Accelerator Selection

These findings contribute to broader discussions on compute–energy misalignment across GPU generations. The H100-to-B200 transition shows that newer accelerators can substantially improve raw throughput while still reducing useful energy efficiency for some workloads. In particular, higher TFLOPs/GPU and faster training do not automatically

imply higher TFLOPs/kW or tokens-per-kilojoule. Instead, realized efficiency depends on workload saturation, memory behavior, communication overhead, power-management policy, and the baseline power floor of the accelerator.

This has practical implications for AI datacenter design. For compute-intensive LLM workloads, B200's higher throughput can reduce wall-clock training time and may justify its higher instantaneous power draw. However, for lighter or partially saturated VLM workloads, the higher power floor can reduce useful energy efficiency. Therefore, accelerator selection should be workload-aware rather than based only on peak hardware specifications. These results support the need for energy-aware scheduling and workload placement policies that match model characteristics to accelerator behavior.

In summary, while the B200 delivers substantial throughput improvements and higher compute efficiency, its less favorable energy metrics and workload-dependent scaling highlight the importance of holistic, energy-aware system evaluation for future AI infrastructure planning.

5.6. Facility-Level Analysis

To contextualize node-level findings within operational realities, we extend our empirical power measurements to facility scale, modeling the cumulative energy and cost implications of deploying H100 versus B200 nodes across two representative load profiles.

5.6.1. Moderate-Load Profile

To quantify the energy and cost implications of deploying B200 versus H100 nodes at datacenter scale, we modeled a representative facility operating under a moderate workload mix. This configuration simulates an enterprise-scale AI training environment with 2000 compute nodes (16,000 total GPUs), identical in deployment architecture but differing in GPU generation. We assume 99% uptime and 10% power delivery losses. PUE values are assumed as 1.25 and 1.20 for H100 and B200 nodes respectively, reflecting the inherent thermal efficiency advantages of B200's liquid-cooled infrastructure at moderate load, consistent with documented efficiency values of centralized thermal management systems. The energy costs are selected as $0.10/kWh, based on regional industrial rates reported by the U.S. Energy Information Administration [40].

The PUE values used in this facility-level analysis are scenario assumptions, not direct facility measurements from our testbed. The lower PUE assumed for the B200 liquid-cooled deployment is motivated by prior studies showing reduced cooling overhead in dense direct-to-chip liquid-cooling infrastructure [41]. However, the exact PUE improvement depends on facility design, cooling loop, climate, workload intensity, and power-delivery infrastructure.

The workload distribution comprises 50% low-intensity jobs (VLM workloads such as X-CLIP and EVL) and 50% high-intensity training runs (LLMs including Mistral-7B, Llama-3.1-8B, Mistral-Nemo, Gemma-2-27B, Qwen2.5-32B, and the compute-dense Vita-CLIP). Node-level power draw for each tier is derived directly from our empirical benchmarks, scaled to full 8-GPU node power, yielding weighted average node powers of 3253 W and 3696 W for H100 and B200 respectively.

Under this profile, the H100 facility is projected to consume 70.4 GWh annually, resulting in an energy cost of $7.04 million. The B200 configuration, despite offering higher throughput on LLM workloads, consumes 76.6 GWh, translating to $7.66 million annually. This represents an energy cost overhead of $0.62 million per year, a direct consequence of B200's substantially elevated per-node power draw under high-intensity compute kernels, as shown in Table 2.

Table 2. Facility summary of annual energy consumption and cost metrics under moderate and high GPU cluster loads for H100 and B200 nodes. (PUE: Power Usage Effectiveness).

Metric	Moderate Load (2000 Nodes)	High Load (5000 Nodes)
Annual Energy—H100 (GWh)	70.4	185.0
Annual Energy—B200 (GWh)	76.6	217.8
PUE (H100)	1.25	1.20
PUE (B200)	1.20	1.15
Electricity Rate (USD/kWh)	$0.10	$0.13
Annual Energy Cost—H100	$7.04M	$24.05M
Annual Energy Cost—B200	$7.66M	$28.31M
Annual Cost Overhead (B200 vs. H100)	+$0.62M	+$4.26M

This finding is consistent with the fundamental trade-off identified in our node-level benchmarks: B200 achieves superior throughput and compute efficiency (up to 32% higher TFLOPs/GPU) but does so at a measurably higher energy cost per unit time. Notably, at low-intensity workloads such as EVL, B200 actually consumes slightly less power than H100 (279 W vs. 290 W per GPU), reflecting its architectural efficiency at underutilized states. However, this advantage is overwhelmed by B200's higher power draw once utilization scales toward LLM-class workloads, where the energy cost overhead becomes operationally significant.

5.6.2. High-Load Profile

To capture the energy dynamics of production-scale AI training, we extend the model to 5000 nodes (40,000 GPUs) under a workload mix dominated by high-intensity training, where 85% of compute cycles are allocated to large-scale LLM and compute-heavy VLM workloads, with the remaining 15% comprising low-intensity VLM jobs such as X-CLIP and EVL. This configuration reflects conditions in dedicated AI training clusters where throughput maximization drives scheduling decisions, yielding weighted average node powers of 3574 W and 4379 W for H100 and B200 respectively.

We adopt differentiated PUE values of 1.20 for H100 and 1.15 for B200, consistent with the higher thermal density management efficiency of next-generation liquid-cooled infrastructure at scale. Electricity costs are selected as $0.13/kWh, based on regional industrial rates reported by the U.S. Energy Information Administration [40].

Under this high-load profile, the H100 facility consumes 185.0 GWh annually, incurring an energy cost of $24.05 million. The B200 configuration consumes 217.8 GWh, resulting in an annual energy bill of $28.31 million, an overhead of $4.26 million per year, as shown in Table 2 and Figure 4. This gap reflects the compounding effect of B200's elevated per-node power draw as high-intensity workloads dominate an increasing share of cluster cycles and represents a substantial operational cost premium that scales directly with facility size and utilization intensity.

At this scale, however, B200's throughput advantages also compound meaningfully. Faster training completion times, up to 15% on LLM workloads and up to 48% on VitaCLIP, translate to higher effective cluster utilization and reduced wall-clock time per training run. For facilities operating under strict model delivery timelines, this throughput premium may justify the energy cost overhead. Conversely, for operators where energy budget and sustainability targets take precedence, H100 nodes offer a more energy-efficient path at comparable model quality, particularly for workload mixes with significant low-intensity fractions.

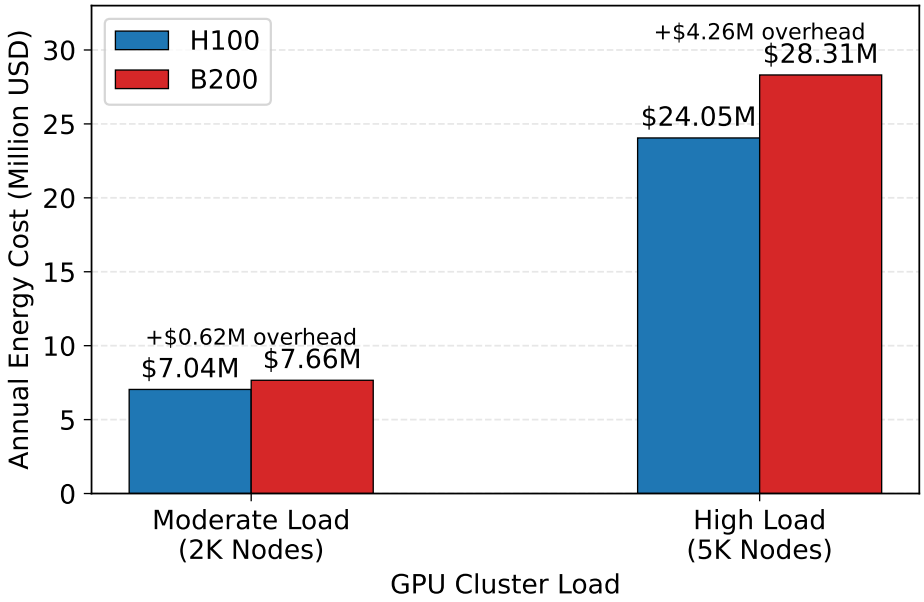


Figure 4. Annual facility-level energy cost comparison for H100 and B200 GPU clusters in moderate- and high-load scenarios. B200 incurs higher annual cost in both cases, with the overhead increasing from $0.62M at moderate load to $4.26M at high load.

These results underscore that the choice between H100 and B200 at datacenter scale is not merely a hardware procurement decision but an architectural one, requiring careful alignment of workload profiles, energy costs, throughput requirements, and sustainability commitments.

6. Conclusions

This work presented a comprehensive system-level comparison of NVIDIA H100 and B200 GPU architectures across LLMs and VLMs ranging from 7B to 32B parameters, evaluating end-to-end throughput, utilization, power, energy consumption, and compute efficiency under Distributed Data Parallel (DDP) execution on single-node 8× GPU configurations. Across LLM workloads, B200 achieved up to 15% faster training and 32% higher TFLOPs/GPU, driven by superior memory bandwidth and increased SM count. However, this throughput advantage comes at a measurable energy cost, with B200 exhibiting lower TFLOPs/kW and tokens-per-kilojoule across all LLM benchmarks. VLM results further exposed model-dependent asymmetries, where lighter compute kernels disproportionately amplified B200's elevated baseline power draw, except in the case of Vita-CLIP, where B200's compute density was sufficient to amortize its power envelope into net energy savings.

Extending these findings to facility scale reveals their operational significance. Under a moderate-load profile of 2000 nodes (50% low intensity and 50% high intensity), B200 deployment incurs an estimated $0.62 million in additional annual energy costs relative to H100 infrastructure, consuming 76.6 GWh versus 70.4 GWh. Under a high-load 5000-node configuration (85% high intensity), this gap widens substantially to $4.26 million annually, with B200 consuming 217.8 GWh versus 185.0 GWh for H100. These projections, grounded in empirically derived power profiles and realistic PUE and electricity rate assumptions, underscore that the economic case for next-generation GPU adoption is highly sensitive to workload composition, utilization intensity, and energy pricing. These results provide an empirical framework distinguishing compute efficiency from energy efficiency across GPU generations, offering actionable guidance for energy-aware AI datacenter design.

Future work may extend this analysis to multi-node training, mixed-precision scheduling, network topology effects, and emerging accelerator families and utilize advanced tools such as Nsight Compute for low-level hardware telemetry. Future work can also extend this analysis to additional GPU generations and variants, including A100, H100, and B200, under consistent LLM and VLM training workloads to better characterize long-term cross-generational efficiency trends.

Author Contributions: Conceptualization, M.A.S.; methodology, M.A.S. and H.U.; validation, M.A.S. and H.U.; results and discussion, M.A.S. and H.U.; investigation, A.M., A.C.N. and I.L.; writing—original draft preparation, M.A.S.; writing—review and editing, A.M. and I.L.; visualization, M.A.S. and H.U.; supervision, A.M. and I.L.; project administration, A.M. and I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used Alpaca [36] and UCF101 [37] datasets, which are publicly available. The code and analysis scripts are publicly available at: https://github.com/iscaas/System-Level-H100-Vs-B200-Profiling.git (accessed on 9 May 2026).

Acknowledgments: The authors used Claude-Sonnet-4.6 to assist with grammar refinement and formatting in the paper. All AI-grammar-refined content was carefully reviewed, revised, and validated by the authors, who take full responsibility for the final version of the publication.

Conflicts of Interest: Author Imran Latif was employed by Johnson Controls, Milwaukee, United States. The remaining authors declare that this research study was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *arXiv* **2020**, arXiv:2001.08361.
2. Rattner, N.; Dotan, T. The AI Spending Boom in Charts. *Wall Street J.* **2024**.
3. Schmidt, B.; Hildebrandt, A. From GPUs to AI and quantum: Three waves of acceleration in bioinformatics. *Drug Discov. Today* **2024**, *29*, 103990. [CrossRef] [PubMed]
4. Nave, O.; Elbaz, M. Artificial immune system features added to breast cancer clinical data for machine learning (ML) applications. *Biosystems* **2021**, *202*, 104341. [CrossRef] [PubMed]
5. Kaack, L.H.; Donti, P.L.; Strubell, E.; Kamiya, G.; Creutzig, F.; Rolnick, D. Aligning Artificial Intelligence with Climate Change Mitigation. *Nat. Clim. Change* **2022**, *12*, 518–527. [CrossRef]
6. Stojkovic, J.; Zhang, C.; Goiri, Í.; Bianchini, R. Rearchitecting Datacenter Lifecycle for AI: A TCO-Driven Framework. *arXiv* **2025**, arXiv:2509.26534.
7. Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergeron, W.; Kepner, J.; Tiwari, D.; Gadepally, V. From words to watts: Benchmarking the energy costs of large language model inference. In *Proceedings of the 2023 IEEE High Performance Extreme Computing Conference (HPEC)*; IEEE: New York, NY, USA, 2023; pp. 1–9.
8. Fernandez, J.; Na, C.; Tiwari, V.; Bisk, Y.; Luccioni, S.; Strubell, E. 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)*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 32556–32569. [CrossRef]
9. Argerich, M.F.; Patiño-Martínez, M. Measuring and Improving the Energy Efficiency of Large Language Models Inference. *IEEE Access* **2024**, *12*, 80194–80207. [CrossRef]
10. Niu, C.; Zhang, W.; Zhao, Y.; Chen, Y. Energy Efficient or Exhaustive? Benchmarking Power Consumption of LLM Inference Engines. *ACM Sigenergy Energy Inform. Rev.* **2025**, *5*, 56–62. [CrossRef]
11. Dwyer, J. Revisiting Training Scale: An Empirical Study of Token Count, Power Consumption, and Parameter Efficiency. *arXiv* **2026**, arXiv:2601.06649.

12. Naduvilakath-Mohammed, F.; Lebon, M.; Robinson, A. Numerical modelling of a hybrid vapor compression refrigeration assisted closed loop liquid cooling system for high-performance computing systems. *J. Phys. Conf. Ser.* **2024**, *2766*, 012078. [CrossRef]
13. Shao, S.; Gao, T.; Yang, H.; Zhao, J.; Zhang, J. Evaluation of single phase immersion cooling system for high performance server chassis using dielectric coolants. In *Proceedings of the International Electronic Packaging Technical Conference and Exhibition*; American Society of Mechanical Engineers: New York, NY, USA, 2020; Volume 84041, p. V001T08A010.
14. Latif, I.; Newkirk, A.C.; Carbone, M.R.; Munir, A.; Lin, Y.; Koomey, J.; Yu, X.; Dong, Z. Empirical Measurements of AI Training Power Demand on a GPU-Accelerated Node. *arXiv* **2024**, arXiv:2412.08602.
15. Latif, I.; Newkirk, A.C.; Carbone, M.R.; Munir, A.; Lin, Y.; Koomey, J.; Yu, X.; Dong, Z. Single-Node Power Demand During AI Training: Measurements on an 8-GPU NVIDIA H100 System. *IEEE Access* **2025**, *13*, 61740–61747. [CrossRef]
16. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
17. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
18. Mistral AI Team. Mistral NeMo: Our New Best Small Model. 2024. Available online: https://mistral.ai/news/mistral-nemo (accessed on 7 December 2025).
19. Team, G.; Riviere, M.; Pathak, S.; Sessa, P.G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv* **2024**, arXiv:2408.00118.
20. Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. Qwen2.5 technical report. *arXiv* **2024**, arXiv:2412.15115.
21. Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; Ling, H. Expanding language-image pretrained models for general video recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 1–18.
22. Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; De Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; Li, H. Frozen clip models are efficient video learners. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 388–404.
23. Wasim, S.T.; Naseer, M.; Khan, S.; Khan, F.S.; Shah, M. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 23034–23044.
24. Barroso, L.A.; Hölzle, U. The Case for Energy-Proportional Computing. *Computer* **2007**, *40*, 33–37. [CrossRef]
25. Barroso, L.A.; Clidaras, J.; Hölzle, U. *The Datacenter as a Computer: Designing Warehouse-Scale Machines*, 3rd ed.; Morgan & Claypool Publishers: San Rafael, CA, USA, 2018.
26. Tschand, A.; Rajan, A.T.R.; Idgunji, S.; Ghosh, A.; Holleman, J.; Kiraly, C.; Ambalkar, P.; Borkar, R.; Chukka, R.; Cockrell, T.; et al. MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI. *arXiv* **2024**, arXiv:2410.12032.
27. Niu, C.; Zhang, W.; Li, J.; Zhao, Y.; Wang, T.; Wang, X.; Chen, Y. TokenPowerBench: Benchmarking the Power Consumption of LLM Inference. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 32582–32590. [CrossRef]
28. Tithi, J.J.; Wu, H.; Abuhatzera, A.; Petrini, F. Scaling Intelligence: Designing Data Centers for Next-Gen Language Models. *arXiv* **2025**, arXiv:2506.15006.
29. Kim, J.; Lee, J.; Park, G.; Kim, B.; Kwon, S.J.; Lee, D.; Lee, Y. An Inquiry into Datacenter TCO for LLM Inference with FP8. *arXiv* **2025**, arXiv:2502.01070.
30. Latif, I.; Shafique, M.A.; Ullah, H.; Newkirk, A.C.; Yu, X.; Munir, A. Cooling Matters: Benchmarking Large Language Models and Vision-Language Models on Liquid-Cooled Versus Air-Cooled H100 GPU Systems. *arXiv* **2025**, arXiv:2507.16781.
31. Ramakrishnan, B.; Turner, C.; Alissa, H.; Trieu, D.; Rivera, F.; Melton, L.; Rao, M.; Chigullapalli, S.; Getachew, T.; Prodanovic, V.; et al. Understanding the impact of data center liquid cooling on energy and performance of machine learning and artificial intelligence workloads. *J. Electron. Packag.* **2025**, *147*, 021003.
32. Zhao, D.; Samsi, S.; McDonald, J.; Li, B.; Bestor, D.; Jones, M.; Tiwari, D.; Gadepally, V. Sustainable supercomputing for AI: Gpu power capping at hpc scale. In *Proceedings of the 2023 ACM Symposium on Cloud Computing*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 588–596.
33. Patel, P.; Choukse, E.; Zhang, C.; Goiri, Í.; Warrier, B.; Mahalingam, N.; Bianchini, R. Polca: Power oversubscription in llm cloud providers. *arXiv* **2023**, arXiv:2308.12908.
34. Reddy, T.R.; Kataria, B.; Gandhi, R.; Tandon, K.; Bhattacherjee, D.; Padmanabhan, V.N. Improving training time and GPU utilization in geo-distributed language model training. *arXiv* **2024**, arXiv:2411.14458.
35. Wang, W.; Ghobadi, M.; Shakeri, K.; Zhang, Y.; Hasani, N. Rail-only: A low-cost high-performance network for training llms with trillion parameters. In *Proceedings of the 2024 IEEE Symposium on High-Performance Interconnects (HOTI)*; IEEE: New York, NY, USA, 2024; pp. 1–10.

36. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-Following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 24 November 2025).
37. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
38. Jarmusch, A.; Chandrasekaran, S. Microbenchmarking NVIDIA's Blackwell Architecture: An in-depth Architectural Analysis. *arXiv* **2026**, arXiv:2512.02189.
39. Ujeniya, A.; Eitzinger, J.; Hager, G.; Wellein, G. Architectural Trade-offs in the Energy-Efficient Era: A Comparative Study of power-capping NVIDIA H100 and H200. *arXiv* **2026**, arXiv:2604.11391.
40. U.S. Energy Information Administration (EIA). Electric Power Monthly. 2025. Available online: https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a (accessed on 27 January 2026).
41. Supermicro. *Comparison of Air-Cooled Versus Liquid-Cooled NVIDIA GPU Systems: Liquid-Cooled GPU Servers Reduce Power Consumption and Increase Performance*; White Paper; Super Micro Computer, Inc.: San Jose, CA, USA, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.