Intelligent Control and Automation, 2026, 17(3), 77-107
https://www.scirp.org/journal/ica
ISSN Online: 2153-0661
ISSN Print: 2153-0653

Workload-Dependent Thermal Locality in AI GPU Nodes: Evidence from H100 and B200 Intra-Node Thermal Divergence

Daniel Weon Seok Ko, Johnathan Mun

School of Business, San Francisco Bay University, Fremont, USA
Email: johnathan.mun@sfbu.edu

How to cite this paper: Ko, D.W.S. and Mun, J. (2026) Workload-Dependent Thermal Locality in AI GPU Nodes: Evidence from H100 and B200 Intra-Node Thermal Divergence. *Intelligent Control and Automation*, **17**, 77-107.
https://doi.org/10.4236/ica.2026.173004

Received: June 27, 2026
Accepted: July 21, 2026
Published: July 24, 2026

Abstract

The increasing power density of AI infrastructure places critical importance on the thermal behavior of data-center GPUs. Measurements of these GPUs' temperatures, however, rely on aggregate metrics such as hardware power ratings, data-center efficiency metrics, and node temperatures, none of which provide information about the temperature distribution across the GPUs within that node. As a result, it is important to determine whether thermal divergence among GPU nodes varies with the hardware model alone or with the workloads assigned to those nodes. An analysis of 32 public sessions, each with eight GPUs assigned to either image-generation or language-model and text-generation workloads, shows that the mean temperature spread between the hottest and coolest GPU within each session is 13.58˚C, a value that far exceeds the 1˚C resolution of each GPU's temperature measurements. Furthermore, analysis of the interaction between workload and hardware reveals that H100 GPUs exhibit a greater temperature spread during image-generation workloads than during language-model and text-generation workloads. In comparison, B200 GPUs exhibit a greater temperature spread during language-model and text-generation workloads than during image-generation workloads (interaction effect: 16.77˚C, $F(1, 28) = 67.88$, $p < 0.001$). These results are statistically significant even after adjusting for heteroskedasticity, permutation tests, bootstrap resampling, leave-one-session-out model refitting, and total GPU power within each session. Furthermore, because each set of GPU workload data was collected on different dates, these results indicate only a relationship between workload and GPU thermal divergence, not between the workload itself and that divergence. Thus, thermal assumptions based solely on the hardware model or on the average temperature of GPUs within a node provide incomplete descriptions of the thermal behavior of those nodes. As such, decisions regarding data-center cooling, scheduling workloads across nodes, hardware procure-

ment, and managing the thermal behavior of those GPUs should account for thermal divergence based on each GPU's workload. Furthermore, while this study was unable to provide evidence of any effect of these workloads on the throttling or useful computation of those GPUs, such conclusions would require validation of these assumptions.

Keywords

AI Infrastructure, H100, B200, Thermal Locality, Node Thermal Spread, Workload Measurement, Data-Center Cooling, Thermal Management, GPU Heat Management

1. Introduction

Artificial intelligence (AI) infrastructures are among the most power-dense and power-variable workloads in data centers today. According to estimates, the world's data centers currently consume 415 TWh of electricity per year and are projected to consume more than double that amount by 2030; data centers powered by AI servers are among the major contributors to this projected growth [1]-[3]. Furthermore, as AI accelerator platforms are designed with power levels in the kilowatt range [4], the problem of how to effectively generate, distribute, and remove the heat they produce has become a first-order design and operational concern.

Many thermal measurements performed within AI accelerator platforms are based on power ratings for the AI accelerators, the density of AI accelerator nodes within data-center racks, and facility-level measurements of the efficiency with which data centers utilize the power they consume [5]-[7]. These levels of measurement are essential to the operation of AI accelerator platforms and to understanding their thermal behavior; however, they do not indicate whether any one GPU within an AI accelerator node is hotter than any other GPU in that same node at any given time. In other words, such data do not provide insight into the concept of "locality" within AI accelerator nodes. Thermal locality is the temperature difference between the hottest and coolest GPUs within an AI accelerator node at a given timestamp.

Within the AI accelerator platform, it is already established that the temperature within the dies of the GPUs is often not distributed uniformly, that the power dissipated by the GPUs is not an accurate indicator of the temperature of those GPUs, and that the way that heat moves within those GPUs is not always as expected based on the physical platform of each AI chip [8]-[11]. Additional work has been published on concepts related to AI GPUs, AI heat, AI cooling systems, AI platform power densities, AI GPU throttling, workload-dependent AI GPU power measurements, and the positioning of heterogeneous GPUs within AI systems [12]-[16]. However, in the public literature on these topics, it is less clear how observed differences in thermal spread across AI accelerator nodes vary with the AI workloads performed by the processors of each AI accelerator generation

(platforms, cards, etc.). Furthermore, in the public records of each manufacturer, it is less clear whether the rank order of GPU temperatures within a node may reverse depending on the type of workloads performed by those AI processors.

To gain an understanding of these topics, this article examines whether thermal divergence among GPUs within an AI accelerator node during public AI model sessions varies with the workloads those processors perform across AI accelerator generations. Specifically, 32 eight-GPU node sessions were collected from AI processors based on the NVIDIA H100 and B200 AI accelerators, and the sessions were performed with two categories of AI workloads: image-generation workloads and large-language-model (LLM) and text-generation workloads [17]. The sessions were the unit of analysis for this research.

The primary research question is: Does the temperature divergence among GPUs within an AI accelerator node depend on the workloads they perform, rather than on the identity of the accelerator hardware itself? Additionally, three other questions will be answered that relate to the main question: 1) Do the observed thermal divergence measurements between AI GPUs within nodes exceed a threshold of possible thermal floor measurements within those processors; 2) Does the thermal divergence within each node of AI processors as a function of the type of workloads performed by each hardware brand and generation differ between the various groups of AI workloads performed; and 3) What implications do these findings have for the management of AI data centers? An answer to each of these questions will be provided in the article. Namely, we have two main research hypotheses:

H1 tests the significance of the mean thermal divergence between GPUs within AI accelerator nodes, relative to a 1˚C sensor-resolution threshold in those chips.

H2 tests whether the cross-product term for workloads and hardware within AI processor nodes is statistically significant (as a measure of the interaction between these two variables).

In each of the 32 public AI accelerator nodes analyzed in this study, the higher-spread hardware (*i.e.*, greater thermal variation between the hottest and coolest GPU within the node) varied with the type of AI workload each node performed. For example, on nodes that performed image-generation workloads, the H100 models exhibited greater thermal divergence between GPUs within each node than the B200 models did for the same workload. In contrast, within nodes that performed language-model and text-generation workloads, the B200 models exhibited greater thermal divergence between GPUs than the H100 models did for the same workload. Furthermore, the mean intra-node thermal divergence across the 32 sessions was 13.58˚C, and the workload-by-hardware interaction was 16.77˚C, indicating that the higher-spread hardware reversed across the image-generation and LLM/text-generation cells; however, each brand of AI processor was performed on different dates, so this is only a statistically supported association between these

variables and not a statistical relationship between the workloads alone and the thermal divergence that results. Warm-up and cool-down periods were excluded from the data capture.

The contribution of this research is not in the suggestion of a new means of cooling AI data centers. Instead, it is intended as a bounded demonstration that the type of workloads that AI processors perform within data centers shows that there are statistical associations on the thermal locality of those GPUs within the nodes of the processors; thus, it is recommended that the thermal locality of the GPUs within the data center's AI processors should be managed in a way that takes into account the workloads that are being performed by those GPUs. Furthermore, such a recommendation can be applied to the management of data centers in that cooling decisions, scheduling decisions for AI model workloads, procurement of AI processing hardware, and even decisions regarding near-source thermal management of those AI processors should take into account not just the state of the workload that is to be performed by the hardware, but also each of the individual GPUs within those data-center AI processors, the thermal state of each of those GPUs, and the state of the useful output that is to be provided by each of those GPUs.

The remainder of this article is organized as follows. Section 2 describes and reviews relevant literature on artificial intelligence thermal locality. Section 3 describes the methodology used to complete the study. Section 4 defines the study's data and potential biases. Sections 5 and 6 present the study's results and the answers to H1 and H2. Section 7 provides the answers to the research questions. Section 8 describes the implications of these findings. Section 9 presents the conclusions and recommendations based on the study's findings and implications, as well as recommendations for their application. Finally, Section 10 discusses the limitations of this study and steps that can be taken to extend its impact and implications.

2. Literature Review

Five main bodies of literature were reviewed: that which relates to the thermal pressure of AI infrastructure, the cooling and efficiency metrics of data centers, the power and telemetry data of GPUs under different workloads, the concept of thermal locality within the GPUs of a single node, and the types of responses that can be made to those thermal parameters. Each of these topics reviews the current literature to make clear that the metrics for AI data centers and GPU hardware are not without their uses and importance. However, the literature review reveals a gap regarding whether thermal locality within a single node's GPUs is transferable across different workloads on those GPUs, as none of the topics explored address this issue.

Addressing this gap is important because workload-dependent thermal variation may affect the reliability of cooling strategies, performance optimization, and infrastructure planning. Accordingly, this study examines whether thermal locality remains consistent across workloads or changes in response to differing computational demands.

2.1. AI Infrastructure Density and Accelerator Thermal Pressure

AI infrastructure has become a thermal and energy-planning issue due to the increased power requirements of the computational processes necessary to perform AI tasks. Analyses by Lawrence Berkeley National Laboratory, the International Energy Agency, and the Electric Power Research Institute all report that data-center electricity requirements are becoming a material concern for the nation's infrastructure [1]-[3]. Previous analyses of data-center power requirements suggest that power estimates for such facilities require detailed, bottom-up analyses of the components to be deployed within them [18]. Thus, the literature on these topics helps establish the "why now" factor that supports the significance of the research presented in this article.

Furthermore, the recommendations regarding the hardware components of AI systems also indicate the importance of this topic. For instance, both the specifications of current data-center GPUs and the published recommendations for cooling such GPUs indicate that AI systems will generate heat at high rates within data centers [4] [19]-[21]. While these publications motivate the study and its relevance to AI infrastructure, they do not provide evidence for the results to be published from this study. For instance, the rating of hardware components indicates their design and operational limits but does not indicate the distribution of heat generated within GPUs during AI tasks. Thus, while related, the existing reports from studies on the hardware and its cooling requirements do not provide evidence for the result that this study will publish.

2.2. Static and Aggregate Cooling Metrics

Data-center operators require aggregate metrics. Metrics such as PUE, WUE, rack power, temperature, and facility capacity are all important for data-center operators [5] [6] [22]. Furthermore, data on power provisioning helps explain why these aggregate metrics cannot be applied to the various populations of GPUs housed in data centers [7]. These metrics are useful to data center operators and owners for planning energy, capacity, resilience, and efficiency within those data centers.

The limitation of these metrics, though, is not the metrics themselves. It is that they do not answer the question of the temperature difference between the hottest and coolest GPUs within a given server node. While it is possible for a server to have an average temperature acceptable for that data center, some GPUs within that node may reach temperatures detrimental to their lifespan and reliability. Thus, these aggregate metrics are inherently different from, and not a substitute for, measuring each GPU within a node to determine whether such a temperature difference exists.

2.3. Workload-Dependent Power, Performance, and GPU Telemetry

Workloads within AI data centers differ in their power draw, performance characteristics, and the types of useful output they produce. Work on deep learning (LLM)

power management, power capping, inference efficiency, and energy benchmarking has shown that the characteristics of these AI workloads affect the power, performance, and energy of the GPUs executing them [12] [13] [15] [23]-[25]. Furthermore, work on benchmarking AI models and workloads indicates that the useful output of those models should be used to evaluate their performance, rather than the power they draw to produce that output [24].

Work on GPUs and GPU telemetry also indicates that each GPU's power draw and performance characteristics must be determined [26]-[30]. Furthermore, work on sensors built into GPUs suggests that the telemetry systems used to measure GPU power draw may be of low quality [31]. Thus, while the literature on GPUs supports the idea that each GPU within a node should be measured for performance and power characteristics, the work done on those GPUs to measure such characteristics warrants caution and should not be accepted unquestioningly.

2.4. Intra-Node Thermal Locality and Performance Mechanisms

The behavior of heat in electronic systems is local, layered, and time-dependent. Work on modeling heat in electronic systems indicates that heat originates at various components of a system, travels through it due to its physical properties, and is emitted by the system itself [8]-[11] [32]. Furthermore, work on the transient behavior of heat within electronic systems and the compact models that can be used to model that heat indicates that the heat models of electronic systems are additionally influenced by factors such as the time that elapses within the system, the impedance within the system, and the location within the system at which the heat is generated or emitted [33]-[37].

These different types of work on heat in electronic systems provide the foundation for investigating locality within a node's GPUs. Such work does not, however, indicate that the GPUs on a node are being measured for parameters as fine-grained as the die-level heat flux of GPUs in public H100 and B200 accelerator models. Instead, this study intends to measure the temperature difference between the hottest and coolest of the eight GPUs in each node. Furthermore, such an investigation of the locality of heat within the GPUs of each node is supported by the existing literature on GPUs that helps to explain the reasons for investigating such a factor as the difference between the hottest and coolest GPUs and why it may be important to measure such a factor within AI data-center nodes.

That literature on GPU nodes also supports the idea that the local heat within those nodes affects GPU performance. For instance, research on the thermal characteristics of distributed deep learning training indicates that factors such as thermal hotspots, power capping, and frequency throttling in GPUs are common in AI distributed training models [14]. Additionally, work on the positional thermal characteristics of GPUs within multi-GPU systems indicates that, for instance, GPUs positioned at the rear of a server may reach higher temperatures than front-facing GPUs on the same node [16]. Work on individual GPUs indicates that GPU

temperature can affect GPU efficiency [38] [39]. Furthermore, because the rate at which AI distributed training workloads can train AI models is limited by the rate at which the slowest of the GPUs within that distributed node completes its work, local differences in temperature that impact the performance of those GPUs can have an impact on the efficiency with which the entire training node can complete its work [40]-[42]. These mechanisms for the impact of local heat within nodes indicate that the performance characteristics of each GPU within a node can be affected by local thermal conditions, even if these conditions are not directly measured in this investigation.

2.5. Candidate Responses: Scheduling, Cooling, Near-Source Management, and Buffering

Several response categories for localizing thermal characteristics in data-center GPUs have been established in the existing literature. For instance, work on workload placement within data centers indicates that workloads can be placed based on the thermal characteristics of the hosts on which they will execute [43]. Furthermore, work on thermal- and power-aware scheduling in data centers indicates that scheduling algorithms exist for data-center GPU clusters and for deep learning (LLM) inference data centers [44] [45]. Thus, the scheduling of workloads to GPUs is one of the response categories to consider in light of each GPU's localized thermal characteristics.

Furthermore, work within cooling architectures of data centers indicates that architectural considerations for data-center cooling systems, cold plates, liquid cooling standards for AI systems (OAI/OAM), the resiliency of cooling systems within the data centers, and research into alternative cooling systems for those data centers help to establish that temperature management is one of the factors concerning data centers that is actively managed by data-center engineers [21] [46]-[48]. Work on near-source thermal management and buffering also helps to establish that these response categories exist within the data-center field. For instance, work on heterogeneous integration within electronic systems indicates that thermal management at the die, package, module, and system levels is a consideration for data-center and AI-chip architects [11]. Work on phase-change buffering in power-electronic modules highlights the need to consider methods for heat buffering in electronic systems [49]. While the literature on each of these response categories supports their consideration, it does not provide evidence, using any method, of a reduction in the temperature spread between the hottest and coolest GPUs within a node, or of an improvement in the useful output of the H100 and B200 GPU models.

Thus, while the existing literature on the locality of heat within nodes indicates each of the response categories to that locality, it also highlights the need to investigate each of those response categories and to justify their consideration over alternative responses to the issues of local thermal characteristics within data-center GPUs.

2.6. Literature Gap and Study Position

As established by the literature on locality and performance of GPUs in AI data centers, several facts are known about GPUs and their applications in those data centers. First, AI data centers tend to generate significant amounts of energy and heat. Furthermore, while metrics on data-center power, energy, and efficiency are helpful to data-center operators and owners, they are often aggregate measures of GPU performance within the data center [5] [6] [22]. Additionally, work on deep learning (LLM) power management, power capping, inference efficiency, energy benchmarking, and even benchmarking of AI models and workloads indicates that the power, performance, and energy of GPUs are affected by the workloads executed on them [12] [13] [15] [23]-[25]. Furthermore, work on heat in electronic systems indicates that the heat generated by these systems is local, transient, and often affects their performance [8]-[11] [32]. Previous research also indicates that within GPU nodes, some GPUs may experience higher temperatures than others, and these localized temperature differences can affect their overall performance [14]. Finally, work on GPU performance also indicates that each of these performance issues has response categories related to their management and correction [21] [43]-[49].

As such, since each aspect covered in the literature is well known in the field, the knowledge gap to be addressed by the present study concerns the differences among current GPU generations in the temperature spread between the hottest and coolest GPUs within those nodes. Furthermore, within each GPU generation, one model may exhibit a greater spread between the hottest and coolest GPUs than another model within the same generation. Thus, one of the contributions of the present study is to establish that current GPUs exhibit these spread differences between the hottest and coolest GPUs on each node. The response to this established issue is not a suggestion to address scheduling, cooling, near-source management, or GPU buffering within those data centers. Instead, this established need to measure the hottest and coolest GPUs within a GPU node indicates that decisions regarding cooling, scheduling, near-source management, and buffering should also be informed by the workload executed on those GPUs.

3. Research Methodology

3.1. Study Design

The main purpose of this research is to investigate whether intra-node temperature divergence, measured during public GPU node sessions, varies across workload-hardware cells. The study's methodology is designed to fulfill that purpose. Specifically, each node contains eight GPUs, and the temperature divergence among them can be measured. These temperature divergences are aggregated to produce a temperature divergence value for each GPU node session, and it is to be determined whether that value exceeds a certain measurement floor, which varies by workload-hardware cells.

Two main hypotheses are tested in this study. The first hypothesis (H1) is tested to determine whether the intra-node temperature divergence within each public GPU node session exceeds a specified low measurement floor, given the GPUs' physical characteristics. The second hypothesis (H2) is tested to determine whether intra-node temperature divergence within each session varies with the workload-hardware cell to which the session belongs. Each of these hypotheses is tested using session-level data.

3.2. Data and Inclusion

The data collected and used in this study comprise 32 public GPU node sessions recorded by Elsayed, Al-Obaidi, and Farag [17] on NVIDIA's H100 and B200 AI GPUs. Each of these sessions was recorded on a full GPU node (eight GPUs per node) and contained either image-generation or LLM/text-generation workloads. Thus, there are four workload-hardware cells within the dataset: H100 image generation, B200 image generation, H100 LLM/text generation, and B200 LLM/text generation.

The results table is created by calculating each session's divergence and using that value to test the hypotheses. The table used for the descriptive summaries and figures of the data is aggregated by each workload-hardware cell. A detailed overview of the collected data is presented in Section 4.

3.3. Unit of Analysis

Although the data consist of 45,000 individual timestamp records per session, the unit of analysis is the session itself. Each session contains timestamp records for the same run of workloads on the same GPU node; thus, treating these timestamp records as independent observations would overstate the effective sample size. However, the timestamps within each session are used to calculate various parameters for that session, yielding 32 observations of intra-node temperature divergence for testing the hypotheses.

3.4. Thermal-Divergence Metric

Let $T_{s,g,t}$ denote the temperature of the GPU g in session s at timestamp t, recorded in degrees Celsius for the eight GPUs in the node. The intra-node thermal spread at a timestamp is the difference between the hottest and coolest GPU at that instant:

$$D_{s,t} = \max_{g \in G_s} \left(T_{s,g,t} \right) - \min_{g \in G_s} \left(T_{s,g,t} \right) \tag{1}$$

where $D_{s,t}$ is the intra-node spread and G_s is the set of GPUs in session s. In other words, this calculation subtracts the coolest GPU from the hottest GPU inside the same node at the same timestamp.

Each session is then summarized by its mean spread:

$$\bar{D}_s = \frac{1}{n_s} \sum_{t=1}^{n_s} D_{s,t} \tag{2}$$

where $\bar{D}_s$ is the session-level mean spread and n_s is the number of timestamp samples in the session. The session-level 95th percentile (P95) is also computed as an upper-tail descriptor. P95 is reported descriptively in Section 5, but it is not treated as direct evidence of sustained operational harm.

3.5. Statistical Analysis

All formal tests use a significance level of $\alpha = 0.05$. H1 is tested with a single-variable and one-sided t-test of the session-level mean spread against the practical measurement-resolution floor, $\delta_{\text{floor}} = 1\,^{\circ}\text{C}$, defined in Section 4.2:

$$t = \frac{\bar{D} - \delta_{\text{floor}}}{s_D / \sqrt{N}} \tag{3}$$

where $\bar{D}$ is the mean of the 32 session-level spreads, s_D is the standard deviation of those spreads, and N is the number of sessions. The test asks whether the observed spread exceeds a conservative measurement-resolution floor. It does not ask whether the spread harms performance, and the floor is not a throttling threshold.

H2 is tested using a two-factor ordinary least squares model of session-level mean spread that accounts for each hardware indicator, the workload-family indicator, and their interaction term, to test whether the direction of the spread varies across workload-hardware cells. The same effect may also be expressed as a double-difference contrast statistic. The model, its coefficient table, and its double-difference equation are presented in Section 6, where the results are applied and interpreted.

In addition to the main model, a series of robustness tests is performed on the H2 interaction term: using HC3 heteroskedasticity-robust standard errors, performing a permutation test with 20,000 permutations and seed 20260625, performing a cell-stratified bootstrap with 20,000 resamples and the same seed, performing a leave-one-session-out model refit, and performing a power-covariate model in which the mean total GPU power for each session is included as a Z-scored covariate. Each of these models evaluates the robustness of the interaction effect in the collected data; none of them addresses the confounding effect of the data collection date, as described in Section 4.4.

3.6. Software and Reproducibility

The study is reproducible using the public data collection source and the session- and group-level summary tables published in Section 4. The analysis and figure-producing software utilizes the Python programming language and its standard scientific packages, pandas, NumPy, and SciPy. The figure-producing software utilizes session- and grouped-level tables to create both a figure of workloads and hardware that exhibit a reversal of the expected spread across nodes, and a figure of the distribution of session-level spreads.

To fully reproduce the study, others must ensure they use the same unit of anal-

ysis as the original study. Additionally, they must report the criteria for inclusion in the study, the means of de-duplication, the labeling of the hardware models, workload families, dates of collection, timestamps of data collection, the definition of spread, and the fields for missing telemetry data.

3.7. Methodology Limits

The method of this study can only support the claim that the H100 and B200 GPUs exhibited thermal spread within each graphics processing unit node during the analyzed sessions, and that the direction of that spread did not remain consistent across workloads and hardware models. Any additional claims made outside of this narrow parameter, such as that the sessions were thermally throttled, that the GPUs lost useful computing power during those sessions, that the workloads were the cause of the reversal of spread within the nodes, or that any intervention in cooling, scheduling of workloads, near-source thermal management, or thermal buffering of the GPUs was performed during those sessions would require a proof-grade data collection methodology that co-measured those fields; currently, those fields are all represented in the study in their missing form (as presented in Section 4).

4. Data Overview, Measurement Scales, and Biases

The analytic data to be used in this study are defined in this section, as are the limits of measurement that can be applied to that data prior to interpreting the results of the analysis. The analysis will use 32 public sessions on H100 and B200 servers, each with eight GPUs, as reported by Elsayed *et al.* [17]. Within those sessions, two different workloads were used: sessions that used GPUs to generate images and sessions that used GPUs to run LLMs and perform text-generation tasks. Each of those sessions recorded metrics related to power consumption, utilization, memory usage, and GPU temperature on the servers. However, metrics such as clock speed, throttling reasons (if any), cooling loop status and temperatures, and GPU server performance metrics were not recorded during those public sessions. Thus, while the analysis can determine the temperature divergence within the individual server node, it cannot determine whether that divergence led to throttling, reduced performance, or any response to those thermal differences.

The data from those public sessions were collected as many "rows" that were timestamped during the sessions (see **Table 1**). However, the rows at those timestamps are not individual data points for analysis, as the GPU measurements were taken over time within each session. Thus, each session is used as the unit of analysis for the hypotheses tested in this paper. The timestamped rows will be used to calculate measurements within each session, but those session-level measurements will be the metrics used in the statistical analyses described in sections 5 and 6. This aggregation ensures that repeated measurements in the same session are not incorrectly treated as statistically independent observations.

Table 1. List of variables, measurement scales, and boundaries.

Variable	Role in Analysis	Measurement Scale	Interpretation Boundary
Session ID	Unit for hypothesis testing	Identifier/nominal	Used as independent observation; timestamp rows inside a session are not treated as independent cases.
Hardware	Grouping variable	Categorical/nominal	H100 or B200; not a basis for a universal hardware ranking.
Workload family	Grouping variable	Categorical/nominal	Image generation or LLM/text generation; not a causal treatment assignment.
Workload-hardware cell	Main comparison cell	Categorical/nominal	Central to H2 but aliased with the collection date in the public data.
GPU index/slot label	Within-node position label	Categorical, possibly ordered	Used to locate GPUs within a node record; physical slot, airflow path, and board-position interpretation should not be inferred unless documented.
Timestamp	Within-session time order	Ordered/interval-like	Used to align per-GPU readings within a session; not treated as an independent observation for hypothesis testing.
Per-GPU temperature	Primary measured thermal field	Interval, Celsius	Recorded at integer Celsius resolution; does not by itself show throttling or useful-output loss.
Intra-node spread	Main derived thermal metric	Quantitative difference, Celsius	Hottest minus coolest GPU at the same timestamp; node-level metric, not on-die, package-level, or heat-flux measurement.
Session-level mean spread	Main session summary	Quantitative, Celsius	Used for H1 and H2; supports within-dataset inference, not population estimation for all deployments.
Session P95 spread	Upper-tail descriptor	Quantitative, Celsius	Describes high-spread conditions within sessions; not a direct measure of sustained operating harm.
GPU power	Context and robustness variable	Ratio, watts	Useful for testing whether total power explains the spread reversal, not the thesis by itself.
GPU utilization and memory behavior	Context variables	Ratio or percentage fields, depending on the source column	Help describe workload state; do not close the useful output chain.
Clock frequency	Missing proof field	Not available in the analyzed telemetry	Prevents claims about clock reduction.
Throttle or violation reason	Missing proof field	Not available in the analyzed telemetry	Prevents claims about measured thermal throttling.
Cooling condition	Missing or incomplete proof field	Not available for the H100/B200 node analysis	Prevents claims about cooling-loop response or cooling-device causality.
Useful output	Missing proof field	Not available in the analyzed telemetry	Prevents claims about throughput, latency, tokens per second, completed work, or useful-compute loss.

4.1. Analytic Scope

Within the narrowed dataset to be used, there are four cells in the analysis: each of the H100 and B200 servers will be analyzed for image generation and LLM/text generation tasks. Each of the image-generation tasks was performed during 7 sessions of each type of GPU, and each of the LLM and text-generation tasks was performed during 9 sessions of each type of GPU. Each session contained approximately 45,000 rows of timestamped data. Thus, the image-generation tasks contained 315,000 rows on each of the H100 and B200 GPUs, and the LLM and text-generation tasks contained 405,000 rows on each GPU type. Each session represented running the eight GPUs on a server. These sessions were not randomly sampled from the available servers for each GPU type. Instead, they are public sessions used to determine whether the thermal spread within each node is within an expected measurement floor and whether any differences exist among the various workloads and GPU types.

The main variable to be analyzed for each server is the temperature spread within the node, defined as the difference between the hottest GPU's temperature at the same timestamp and the coolest GPU's temperature on that server. This metric will be calculated for each session, and the mean spread within each session will be used to test the hypotheses formulated in this paper. Additional data to be published in this paper include the mean P95 within each session. The session-level P95 spread is reported as an upper-tail descriptor of intra-node spread; it is not a direct identifier of the hottest GPU. However, that statistic should not be used to assess the performance of the GPUs under analysis under normal operating conditions.

4.2. Scales of Measurements

The practical measurement floor for the temperature-spread test is 1˚C, a conservative estimate given the integer resolution of public temperature measurements. This floor is not a threshold for system performance, nor is it a threshold for throttling or sensor calibration. Finally, this threshold is not a performance threshold in the context of this study; it is only a floor for performance measurements.

4.3. Missing Fields and Claim Boundaries

The dataset is strong for one task: measuring the thermal divergence of GPUs within a single node under the observed workload-hardware cells. However, it is not proof-grade for establishing causation across the entire chain from workload to useful-output loss. Such a proof-grade study would have required data for each of the following fields for each session in the dataset: workload, per-GPU power, per-GPU temperature, clock frequency for each GPU, throttle violation for each GPU, cooling condition for each GPU, and useful output for each GPU.

Due to a lack of data on clock frequency and throttle violations for each GPU, this paper cannot make claims about thermally throttled sessions. Because it lacks

data on the useful output of each GPU, this current research cannot claim any loss of useful output due to heat. Finally, because we lack data on cooling conditions for each GPU, we cannot quantify the impact of cooling-related factors on thermal divergence within the node. Each of these factors, however, remains a candidate for investigation in the future in baseline-versus-intervention studies.

Finally, in this analysis, the maximum temperature across all GPUs on the node was 77˚C in the grouped data. This value represents the ceiling temperature of the GPUs in this dataset, and its presence in the report does not indicate that any of the work sessions were thermally limited.

4.4. Biases, Confounds, and Design Limits

Perhaps the most important limitation of the dataset is the confusion about the dates on which the data were collected. The four different workload-hardware cells were each collected on four dates: the H100 image-generation sessions on 2025-09-16; the B200 image-generation sessions on 2025-09-17; the H100 LLM/text-generation sessions on 2025-09-20; and the B200 LLM/text-generation sessions on 2025-09-22. This is a mix of midweek, weekend, and early workweek, with unknown times of day. Thus, each workload-hardware cell is aliased to the date on which it was collected, as well as to any confounding factor that differed across these collection dates. Consequently, while the interaction effect can be analyzed to determine whether there is a statistically supported relationship between the workload-hardware cells, no conclusions can be drawn about the cause of the reversal among those workloads.

Beyond this important confound, other biases are introduced by the use of this publicly available secondary dataset. For instance, the sessions are a publicly available sample of those conducted in the establishment, rather than a sample collected in a way that minimizes selection and availability biases. Additionally, while the dataset contains two accelerator generations and two workload families, any conclusions drawn from the data should not be generalized to other GPU types, workloads, node designs, rack designs, or cooling architectures. Furthermore, since the 7 or 9 sessions within each workload-hardware cell were collected on the same day, across four different days overall, and may share common environmental conditions, it is appropriate to generalize within each workload-hardware cell. However, any generalization beyond those categories would be invalid. Additionally, while a slot within the node labels the GPUs, those labels do not indicate either the cooling condition of each GPU or the airflow from the node to each GPU. Thus, while it is possible to determine the temperature of each GPU within the node, it is not possible to determine its cooling conditions or airflow.

Finally, an issue that arises from using these different aggregated measurements is that the data for nodes, facilities, and GPUs can inherently differ. For instance, a data center has GPUs with acceptable average temperatures, yet still has a temperature difference between the hottest and coolest GPU within a single node. Additionally, while higher total power consumption by a data center may

indicate it is hotter than other data centers with lower total power, this does not imply a difference between the hottest and coolest GPU within each data center. For these reasons, while the power of each data center can be measured, the study directly measures the temperature differences between GPUs within each node, treats total power as a context variable, and tests the robustness of the findings.

5. Descriptive Statistics

Table 2 reports the magnitude and direction of divergence within each node prior to hypothesis testing. Measuring the temperature difference between GPUs within a node allows evaluation of whether the hardware with the greatest thermal divergence remains consistent across different workload families. The temperature spread across the eight GPUs in a node is represented by the difference between the hottest and coolest GPUs (Equations (1) and (2)). Tests of these statistics are presented in Section 6. All measurements presented in this table are from the 32 analyzed public H100/B200 sessions [17].

5.1. Group-Level Summary

Table 2 summarizes the four workload-hardware cells. Each cell includes either 7 image-generation sessions or 9 LLM/text-generation sessions, for a total of 32 sessions. Be advised that the effective number of independent collection environments may be closer to 4 collection blocks than to 32 fully independent sessions. Thus, the reported p-values may reflect small-sample effects and should be interpreted with caution.

Table 2. Session-level descriptive statistics by workload family and hardware (Values are means across sessions within each workload-hardware cell unless otherwise noted.).

Workload Family	Hardware	Sessions (n)	Mean GPU Temp (˚C)	Max GPU Temp (˚C)	Mean Intra-Node Spread (˚C)	Mean Session P95 (˚C)
Image generation	H100	7	42.19	73	18.45	19.71
Image generation	B200	7	38.50	61	8.72	10.44
LLM/text generation	H100	9	51.90	70	10.06	13.78
LLM/text generation	B200	9	51.08	77	17.09	20.89

The mean session P95 spread is the average of each session's 95th percentile. This metric does not represent the system's typical conditions.

5.2. Magnitude of Divergence

Across all 32 sessions, the mean session-level intra-node spread is 13.58˚C (SD = 5.03˚C). Individual sessions range from 5.90˚C to 23.89˚C, and all 32 sessions have mean spreads that exceed the 1˚C floor for practical resolution of temperature

measurements (Section 4). The smallest cell mean (8.72˚C) is approximately nine times this floor. These measurements are within-node temperature differences; Figure 1 presents the distribution of these session-level measurements. A statistical test to determine whether these differences are significantly above the measurement floor is presented in H1 in Section 6.

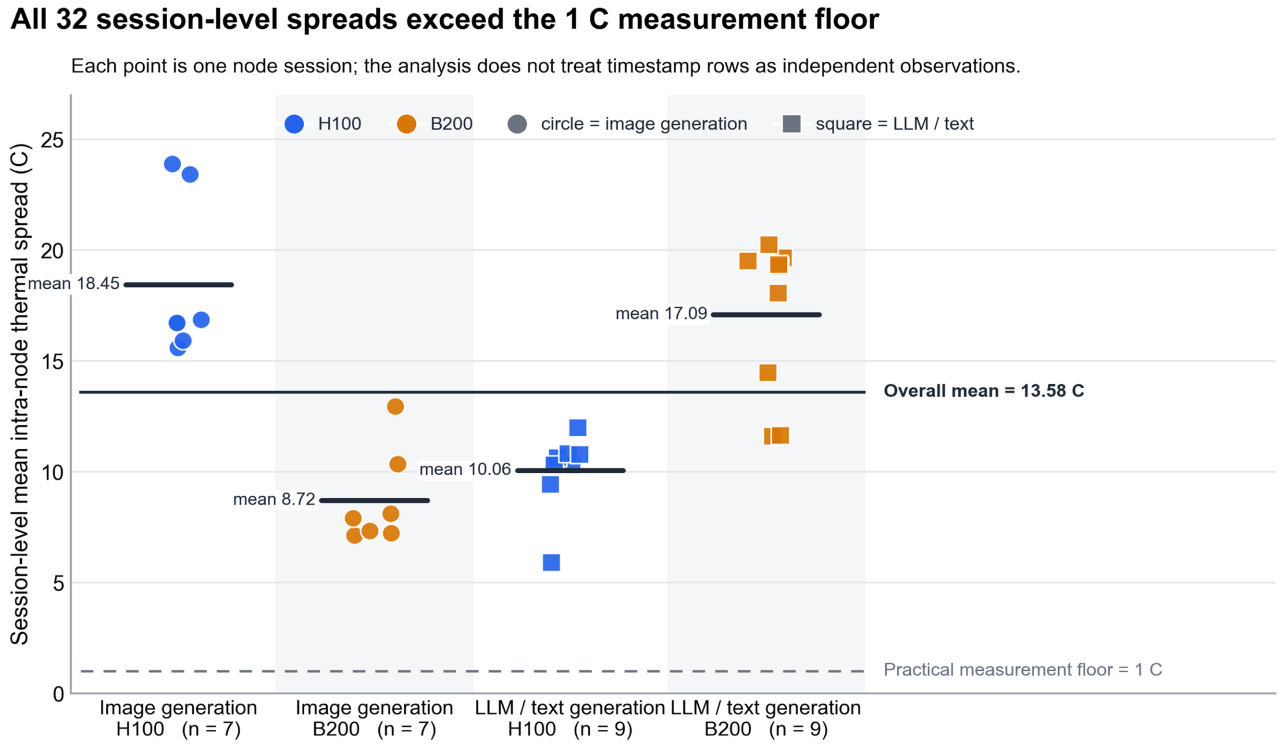

Figure 1. Distribution of session-level intra-node thermal spread across 32 H100/B200 node sessions. Each point represents the mean of the hottest-minus-coolest GPU temperature spread for that session. Each session exceeds the 1˚C floor for practical measurement of temperature differences, with an overall mean of 13.58˚C and a 95 percent confidence interval of 11.77˚C to 15.39˚C. The figure uses sessions rather than individual rows in the data table to reflect that measurements were taken over time within each session.

5.3. Direction of the Divergence

The mean spread across nodes of the H100 and B200 hardware components is higher for image-generation workloads than for LLM/text-generation workloads (Figure 2). For the image-generation workloads, the H100 has a mean intra-node spread of 18.45˚C compared to the B200's 8.72˚C (a difference of 9.74˚C using the unrounded session-level spread measurements; the rounded means differ by 9.73˚C). For the LLM/text-generation workloads, the spread within the B200 is higher than within the H100; the B200's mean spread is 17.09˚C compared to the H100's 10.06˚C (a difference of 7.03˚C). Thus, the difference between the H100 and B200 within the image-generation workloads is reversed relative to the LLM/text-generation workloads; the double difference between the hardware components is 16.77˚C. Because the sessions within each workload family were collected on separate dates (as noted in Section 4), these differences cannot be attributed solely to within-family variation.

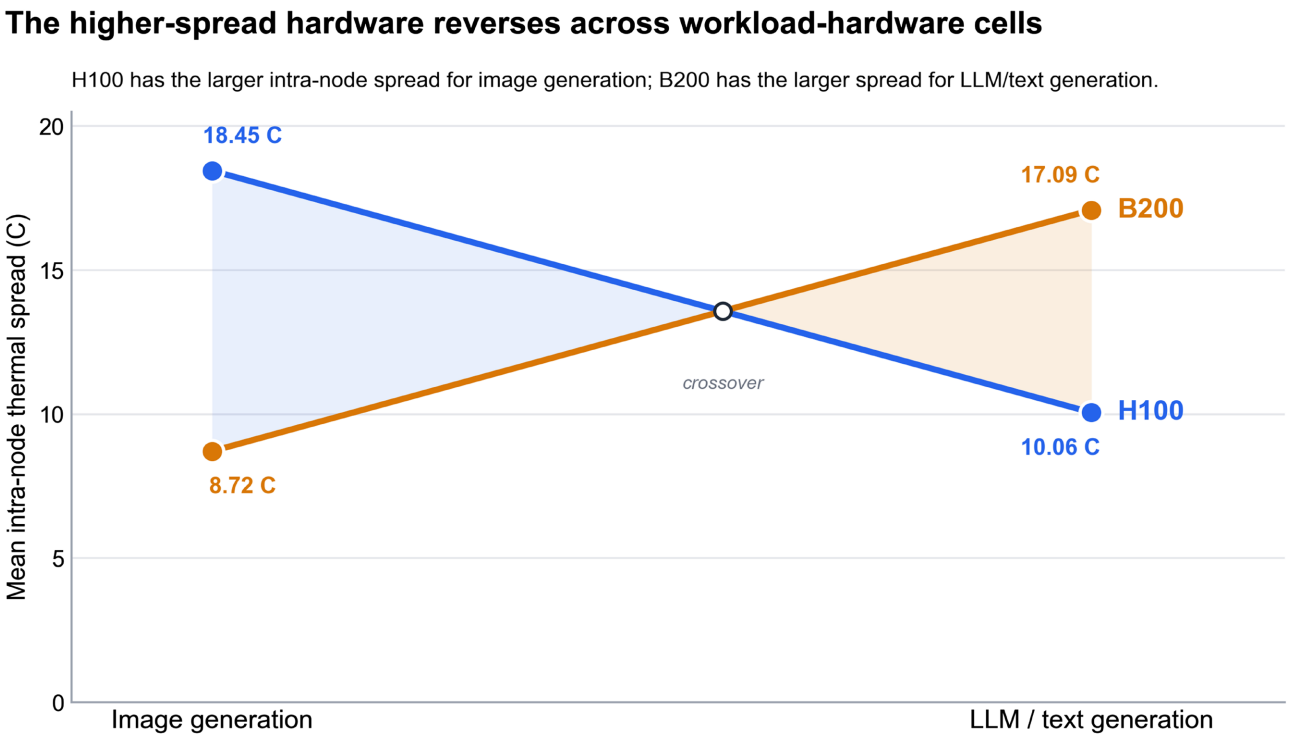

Figure 2. Workload x hardware reversal in mean intra-node thermal spread. The figure plots the mean hottest-minus-coolest GPU temperature spread within each group of sessions held on each eight-GPU node. For image-generation sessions, the H100 GPUs exhibit a greater spread than the B200 GPUs. For LLM/text-generation sessions, however, the B200 GPUs exhibit a greater spread than the H100 GPUs. The difference between these amounts is 16.77˚C. This contrast does not indicate that session type caused the reversal of the trend in GPU temperature spreads, as each workload-hardware combination is aliased with the date on which the GPUs were measured.

5.4. Neither Node-Average Temperature nor Total Power Orders the Spread

Two additional aggregate perspectives of the same sessions do not reproduce the spread ordering. First, the node-average temperature does not reproduce the ordering of spread within the nodes. For instance, the cell with the highest mean GPU temperature was LLM/H100 at 51.90˚C, with a mean spread of 10.06˚C across its GPUs, one of the lowest spreads within the cluster. By contrast, the cell with the highest spread (image/H100) had a mean temperature of 42.19˚C. This relationship between spread and temperature is consistent with the literature on the thermal spread of GPUs [8] [10]. The temperature measurements were performed at the node level, not the die level, meaning that the spread was calculated for the eight GPUs within each node.

Second, total GPU power does not have a monotonic relationship with spread within the cluster. The cell with the highest spread (image/H100) had the lowest mean total GPU power among the four compute cells in the cluster (1995 W). Both B200 models had higher total GPU power than the H100 models for both workload families: image and LLM/text generation. B200 used 3165W and 4712W compared to 1995W and 3717W, respectively. B200, however, had a wider spread

within the LLM/text-generation workload. Section 6 presents a model that predicts spread within compute cells, reversing the relationship between total GPU power and spread. Thus, the spread among GPUs on a compute node cannot be predicted from the average temperature or total power of the GPUs in the node.

5.5. Observed Temperature Regime

The maximum temperature of each GPU in each of the four compute cells within the data center was measured. Each GPU reached a maximum of 73, 61, 70, or 77˚C (Table 2). The highest maximum temperature among GPUs was 77˚C, achieved by GPUs in the LLM/B200 compute cell. Thus, 77˚C is the maximum temperature of any GPU in the cluster. It is important to reiterate, however, that this temperature does not indicate that the GPUs were at the thermal throttling threshold. GPU clock frequency, the reason for any potential throttling, and any metrics related to the usefulness of the computations performed by the GPUs were not available for these measurements. Determining whether thermal throttling occurred on any GPU is therefore outside the scope of this paper (see Section 4).

6. Statistical Inference and Hypothesis Tests

This section presents the formal tests of the two hypotheses introduced in Section 1. As previously stated (Section 3), the unit of analysis is the session; each test employs N = 32 session-level values as units of analysis. The analyses ignore the autocorrelation of the sub-second telemetry measurements within each run. Each test employs a significance level of 0.05. Hypothesis 1 (H1) relates to whether the intra-node spread is valid as a measure of heat distribution within the node by determining whether the spread exceeds the resolution of the temperature sensors; Hypothesis 2 (H2) pertains to the main research question of whether workload distribution within hardware cells explains the intra-node spread rather than the identity of the hardware from which the measurements were collected. Each of these tests focuses exclusively on the spread of heat within the nodes and does not attempt to investigate thermal throttling, the amount of useful compute work performed by the nodes, or any interventions to improve cooling or scheduling of work to the nodes, all of which are not recorded in the dataset (Section 4).

6.1. H1: Intra-Node Divergence Exceeds the Measurement Floor

H1 asks whether the mean session-level spread exceeds the 1˚C practical measurement-resolution floor, δ_{floor}, defined in Section 4. Because the per-GPU temperatures are recorded at integer resolution, $\delta_{\text{floor}} = 1\ ^{\circ}\text{C}$ corresponds to one quantization step; the test therefore asks whether the observed spread exceeds a conservative measurement-resolution floor, not whether it reaches any particular operationally consequential magnitude.

- **$H1_0$:** $\mu_D \le 1\ ^{\circ}\text{C}$
- **$H1_A$:** $\mu_D > 1\ ^{\circ}\text{C}$

A one-sample, one-sided t test against the floor (Equation (3)) yields a mean

spread of 13.58˚C (SD = 5.03˚C, SE = 0.89˚C), t(31) = 14.16, p < 0.001; the two-sided 95% confidence interval for the mean spread within nodes is [11.77, 15.39]˚C. The null hypothesis $H1_0$ is rejected. These results do not rely on the pooled test result alone: individual session means range from 5.90˚C to 23.89˚C, and each of the 32 session means exceeds 1˚C. In addition, each of the four cell means for workload and hardware configurations exceeds 1˚C (the smallest mean is 8.72˚C). Thus, the within-node spread exceeds the resolution of our measuring instruments in all 32 analyzed sessions. Since H1 does not claim that any measured spread of node temperatures within a system is harmful, the confidence interval refers to the 32 analyzed sessions rather than some general population of system sessions.

6.2. H2: Workload × Hardware Interaction

H2 tests whether the direction of intra-node spread depends on the workload-hardware cell. Using the two-factor model,

$$\bar{D}_s = \beta_0 + \beta_1 B200_s + \beta_2 LLM_s + \beta_3 \left(B200_s \times LLM_s \right) + \varepsilon_s \tag{4}$$

with B200 = 1 for B200 sessions and LLM = 1 for LLM/text-generation sessions (baseline = H100 image generation), the hypotheses are:

- **$H2_0$:** $\beta_3 = 0$
- **$H2_A$:** $\beta_3 \neq 0$

The fitted coefficients are reported in Table 3.

Table 3. Two-factor OLS coefficients for session-level intra-node thermal spread (N = 32; baseline = H100 image generation).

Term	Estimate (C)	SE	$t(28)$	p
Intercept (H100, image)	18.45	1.08	17.10	<0.001
B200 (vs H100, within image)	−9.74	1.53	−6.38	<0.001
LLM (vs image, within H100)	−8.40	1.44	−5.83	<0.001
B200 x LLM (interaction)	16.77	2.04	8.24	<0.001

The interaction term is $\beta_3 = 16.77$ ˚C, $F(1, 28) = 67.88$, $p = 5.75 \times 10^{-9}$, $R^2 = 0.71$, adjusted $R^2 = 0.68$, and partial $\eta^2 = 0.71$. $H2_0$ is rejected. The interaction estimate matches the descriptive double difference reported in Section 5:

$$\left(17.09 - 10.06\right) - \left(8.72 - 18.45\right) = 16.76^{\circ}\text{C} \tag{5}$$

These cell means are rounded to two decimal places. Thus, the subtraction that defines the contrast within the cell yields 16.76˚C; the exact value based on the session means is 16.77˚C, which matches the interaction coefficient. Furthermore, the gap between the H100 and B200 for image generation is reported as the rounded value of 9.73˚C rather than the unrounded contrast coefficient of 9.74˚C (Section 5).

The effect is carried almost entirely by the interaction effect. A Type-II analysis of variance decomposition indicates that the sum of squares for the main effect of the type of hardware was 0.74, that for the main effect of workload was 0.001, and

that for the interaction effect was 553.65 (with a residual sum of squares of 228.37). Thus, each of the main effects is very small compared with the interaction effect. The reason for these small main effects is that the effect of the hardware type varies with the workload: B200 processors are 9.74˚C less spread out in their performance in image generation than on H100 servers, yet are 7.03˚C more spread out in their performance in LLM/text generation. Furthermore, in H100 systems, image generation exhibits an 8.40˚C spread difference compared to text generation. In contrast, within the B200 systems, text generation exhibits an 8.37˚C difference in spread relative to image generation. These two sets of differences cancel each other out, an effect that is accounted for by the interaction effect. Thus, the coefficients that are statistically significant for B200 and LLM in Table 3 are each conditional on the H100 baseline for image generation, rather than on main effects generalizable to all workloads.

Lastly, it should be noted that this study investigates only the association between workload type, hardware type, and intra-node thermal spread within each system; no claims can be made regarding causation. Moreover, because each workload was tested on separate dates, any performance differences may be attributable to the testing date. Thus, no conclusions can be drawn from this analysis regarding whether the type of workload caused the differences in hardware performance.

6.3. Robustness of the Interaction

The following four checks are used to examine whether the interaction estimate is sensitive to distributional assumptions or to individual sessions:

- **Heteroskedasticity-robust standard error (HC3):** SE = 2.20, $t = 7.62$ ($p = 2.70 \times 10^{-8}$), accommodating unequal cell variances.
- **Permutation test** (20,000 permutations, seed 20260625; workload labels shuffled within each hardware group, preserving group sizes): the observed β_3 exceeds every permuted value in both tails (permuted range −13.84˚C to 12.08˚C), so the plus-one-corrected two-sided permutation p is $1/20,001 \approx 5.0 \times 10^{-5}$.
- **Stratified bootstrap** (20,000 resamples stratified by cell): 95% CI for β_3 = [13.08, 20.54]˚C, excluding zero.
- **Leave-one-session-out:** refitting after dropping each session in turn yields $\beta_3 \in [15.86, 17.47]$ ˚C, positive in all 32 refits.

These checks help ensure that the interaction effect is preserved across the different types of data. Each workload-hardware cell has only 7 - 9 data sessions. Therefore, these distribution-free tests serve as an alternative to the normality and homogeneity-of-variance assumptions underlying parametric tests. Neither of these tests addresses the confounding effect of the date on which the sessions were collected (see sections 4 and 10).

6.4. The Interaction Is Not Reducible to Total Power

In Section 5, using descriptive statistics, we established that the cell with the highest spread in GPU temperatures also had the lowest total power delivered to those

GPUs. A statistical model was created, with the mean total GPU power per session as a covariate, to test whether the interaction between the workloads and the hardware specifications was an artifact of total GPU power. The interaction term remained statistically significant in the model (β_3 = 17.21˚C, t = 13.04, p < 0.001). However, the total GPU power was also significantly related to session-level spread (3.24˚C per standard deviation in total power, t = 6.31). Thus, the total power of the GPUs does not account for the interaction between workloads and hardware types. Furthermore, as with the other tests in this section, the run-day, node, and environment confound are not accounted for by this test, nor is there any available covariate to address them. Thus, this test served only as a descriptive check for the alternative explanation of the interaction with total power, rather than as a means of causal control over that variable.

6.5. Interpretation Boundary

These tests provided evidence for both hypotheses: H1 and H2 were found to be true across the 32 GPU sessions. However, each of these statistical analyses only helps to prove that the thermal spread within the nodes of the GPUs was experienced above the thermal floor of those GPUs (H1) and that there was a statistically significant interaction between the workloads that were assigned to each GPU and the type of GPU hardware that was utilized (H2). Neither hypothesis, however, establishes a link between the workloads and the temperature spread within the GPUs, that the GPUs were experiencing thermal throttling as a result, or that their processing power was reduced. Furthermore, while these analyses used N = 32 session-level observations, the interaction between those workloads and GPUs was only established across the four blocks of dates on which the sessions were collected (Section 10). Thus, these hypotheses are supported but limited to the interpretations that can be made about these two variables at the current level of analysis. The implications of these results are discussed in sections 7 and 8.

7. Results

The answers to the research questions posed in Section 1 are presented in this section, using the descriptive statistics from Section 5 and the tests presented in Section 6. Results 1 and 2 present the findings of the research questions, and Result 3 presents an interpretation of these findings. Section 8 discusses the implications of the research for cooling, scheduling, and measurement, including in response to support for Question 3. No new analyses are performed in this section; all statistics supporting the findings are presented in Section 6.

7.1. Result 1: Intra-Node Thermal Divergence Is Observed above the Measurement Floor

The within-node temperature spread exceeds the 1˚C floor in all 32 observed sessions. The mean spread across all sessions is 13.58˚C within each node, and each of the sessions individually exhibits within-node spreads above the 1˚C floor (with

the lowest mean within any workload/hardware cell measuring 8.72˚C; Figure 1; Section 6.1: t(31) = 14.16, p < 0.001, 95% CI [11.77, 15.39]˚C). Thus, in support of Question 1, these results show that each observed session includes within-node thermal divergence exceeding the 1˚C measurement resolution floor. The smallest mean within each workload/node cell is 8.72˚C; however, these results apply only to the analyzed sessions and do not generalize to GPU systems beyond the H100 and B200 evaluated.

7.2. Result 2: The Higher-Spread Hardware Reverses across Workload-Hardware Cells

The hardware that exhibits the largest temperature spread varies with the type of workload the servers perform. For instance, during image-generation workloads, the H100 GPUs exhibit a wider temperature range than the B200 GPUs (18.45˚C vs. 8.72˚C). However, for LLM/text-generation workloads, the B200 GPUs exhibit a larger spread in temperature than the H100 GPUs (17.09˚C vs. 10.06˚C) (Figure 2; Table 2). This workload × hardware interaction is statistically significant and is driven mainly by the interaction term: β_3 = 16.77˚C, $F(1, 28) = 67.88$, $p = 5.75 \times 10^{-9}$, partial $\eta^2 = 0.71$ (Section 6.2; Table 3). The finding is robust to adjustments for heteroskedasticity-robust standard errors, permutation tests, stratified bootstrap sampling (95% CI [13.08, 20.54]˚C), leave-one-session-out model refitting, and adjustment for total GPU power (sections 6.3 and 6.4). These results support question 2: there is an association between divergence and workload-hardware cells. However, because each cell contains samples collected on a single date (Section 4), it is not possible to conclude that the workload affected the observed reversal in the direction and magnitude of the divergence.

7.3. Result 3: A Hardware-Fixed Interpretation Is Incomplete for These Sessions

Since the hardware with the higher intra-node spread within a given workload family is the hardware that reverses within workload families, and since neither the main effect of the hardware nor the main effect of the workload explains the spread within a given hardware system, the effect is best characterized as a crossover interaction (Section 6.2). Thus, within these sessions, the GPU identity alone does not indicate which configuration is likely to exhibit greater intra-node spread. Across the four workload-hardware cells, node-average GPU temperature and mean total GPU power do not reproduce the ordering of intra-node spread; notably, the cell with the highest mean spread had the lowest mean total GPU power (Sections 5.4 and 6.4).

In response to the main research question, the observed data indicate that intra-node spread is not explained by GPU hardware identity alone; instead, it differs across workload-hardware cells. The spread among GPUs is explained not solely by their manufacturers but also by the characteristics of the workloads they are assigned to execute. This finding does not invalidate the use of the aggregate met-

rics for their intended applications. The implications of this finding are discussed in Section 8.

8. Discussion

The main result of this study is that the spread of heat within a node is related to the workload-hardware cell rather than to the hardware identity alone (Section 7). This discussion considers the implications of that result for practice and measurement, within the bounds of the evidence (Section 4): specifically, that the result is based on an observational study that was confounded with the date on which the data were collected, and does not include any information about throttling, useful output, or any form of intervention.

8.1. From Hardware Identity to Workload-Dependent Locality

The reversal of the association between hardware and intra-node thermal locality implies that the intra-node locality of observed sessions cannot be understood from the identity of the accelerator hardware alone. The same hardware displays larger intra-node spreads under some workloads than under others and exhibits varying degrees of intra-node locality across different hardware brands. It is this crossover phenomenon that is observed in this study.

8.2. Why Node and Facility Averages Are Incomplete for Intra-Node Locality

It is important to provide a view of the data aggregated across the entire facility and at individual nodes within it. Metrics at the level of the entire facility help in understanding its efficiency and data center management [5] [6]. Metrics at the level of individual nodes help facility operators to understand their individual nodes' power and thermal capabilities [7]. The infrastructure's ability to integrate design and operations components ensures that thermal and power planning are coordinated in the design phase [4].

These metrics are, by construction, not capable of providing insight into how temperature is distributed across GPUs within a single node. While the analyzed sessions have provided insight into differences in average GPU temperatures within a node and the spread of those temperatures (Section 5.4), these metrics remain insufficient to understand the thermal locality of GPUs within a node. Thus, while these aggregate metrics are useful for their intended use cases, they remain incomplete as metrics for assessing the thermal locality of GPUs within a node.

8.3. Co-Measurement as the Main Practical Implication

If the thermal locality of a GPU can be understood in part through the hardware on which it runs and the workload it executes, then it can be best understood when co-measurements are made of those variables. Such a dataset would include measurements of the workload executed by each GPU, each GPU's power, and each GPU's temperature at the same time intervals.

While the available dataset contains the power and temperature of each GPU, it does not include variables for measuring each GPU's clock frequency, whether it is being throttled, the cooling condition of each GPU, or the amount of useful output. To measure these variables would require the inclusion of instrumentation to measure the clock frequency of each GPU [29], whether it is being throttled or not [30], the accuracy of the power sensors for each GPU [31], the cooling condition of each GPU [15], and the ability to measure the amount of useful output of each GPU [25].

Thus, co-measurement is proposed as a means of making decisions about thermal management and cooling. Regarding research question 3, cooling, scheduling, and thermal management decisions should be based on co-measurement of these variables rather than on the identity of the hardware or the average temperature of the GPU nodes.

8.4. Candidate Responses

There are several potential approaches to addressing intra-node thermal locality, based on the recognition that workloads can characterize this locality. One set of approaches involves the scheduling of workloads. Workload-aware scheduling has a long-standing tradition in the literature [43]. Thermal-aware and power-aware scheduling approaches have recently been proposed for data centers that deploy GPUs and LLM inference clusters for machine learning tasks [44] [45].

Additional approaches to managing thermal locality involve the data-center facility's cooling system. Approaches to cooling that introduce liquid-cooling architectures and enhance the resiliency of data-center cooling systems have been suggested [21] [50]. Additionally, public cooling research programs are considering methods of effectively removing heat from chips and GPUs in data centers [48].

A third set of approaches relates to the thermal management of the chips themselves. Methods for managing the heat generated by chips have been proposed at both the chip package level (encompassing all dies within a given package) and within the chips themselves (e.g., within processor nodes) [11] [37]. Techniques involving phase-change materials to buffer heat in power electronic circuits have been discussed in the literature, though with goals different from those for managing heat in data-center nodes housing GPUs [49]. Thus, approaches to thermal management of data-center nodes and their GPUs (especially if GPU chips include techniques for thermal buffering of their nodes) represent one potential way to address the problem of providing thermal locality for GPUs within each data-center node. Unfortunately, no studies were conducted to determine whether these thermal-buffering methods would reduce thermal spread to GPUs within each data-center node or increase their useful output.

8.5. Why the Candidate Responses Are Not Findings of This Study

None of these candidate responses to intra-node thermal locality of GPUs within data centers are supported by the findings of this study. The dataset did not con-

tain a field for measuring interventions in the cooling or operation of the GPUs. Thus, it was not possible to determine the effect of any of these candidate responses.

Any response to the problem of intra-node locality needs to be evaluated through a baseline-versus-intervention study that measures the same variables under the same workloads and on the same GPUs. Ideally, such a study would utilize the same data-center nodes to test these responses (Section 10). This study's contribution is the suggestion that the intra-node thermal locality of GPUs should be treated as a variable understood through its workload-aware, co-measured nature prior to making any cooling, scheduling, or thermal-management decisions. Thus, while each of these response categories is a potential means of addressing the problem, this study does not provide support for any particular candidate response.

9. Conclusions and Recommendations

9.1. Conclusions

The study was conducted to determine whether the thermal locality within GPU nodes of AI systems depended on the workload performed by those GPUs rather than being an inherent feature of the GPUs themselves. In each of the 32 observed public sessions on eight-GPU nodes, the higher-spread GPU hardware within each node reversed depending on the workload performed by that node.

Two results support the argument that thermal locality within GPU nodes is not an inherent feature of the GPUs. First, the within-node spread of GPU temperatures often exceeded the 1˚C floor used to measure temperatures within each node; the mean spread within each node was 13.58˚C across sessions, and the smallest mean spread within any workload-hardware cell was 8.72˚C (H1). Second, the direction of temperature spread within each node differed across hardware nodes; the H100 GPU models exhibited greater spread in workloads that performed image generation, while the B200 GPU models exhibited greater spread in workloads that performed LLM and text-generation tasks. Furthermore, the workload-by-hardware was statistically significant, with a crossover magnitude of 16.77˚C (H2). Each of these workloads-hardware cells occurred on the same dates as the data were collected, supporting only the notion that there is some relationship between the two variables, rather than providing evidence of a relationship that developed over time as a result of the performance of those workloads.

The findings of this research indicate that the thermal locality within AI GPU nodes depends on the workloads they perform. Thus, thermal locality within these nodes should be evaluated on a per-GPU basis rather than across the entire node. However, the research did not determine whether there was any loss of the GPUs' computational usefulness due to differences in thermal locality across nodes, or whether any cooling or thermal-buffering solutions were effective in mitigating these problems. Future research ought to examine these performance effects and potential thermal-management solutions.

9.2. Recommendations and Practical Applications

1) Co-measure before deciding. Before committing to any scheduling, cooling, management, or buffering effort, characterize the workloads and GPUs that are to be used to collect the co-measured data described in Section 8.3.

2) Use intra-node spread as a standing signal. Per-GPU temperature spread should be tracked in addition to within-data-center and within-node metrics (which can be derived from the within-node spread).

3) Test the target workload on the actual hardware. Because workloads differ in their thermal characteristics, only by testing the target workload on the actual hardware can thermal characterization and procurement decisions be made.

4) Consider the standard of baseline versus intervention. Any response to the problems described in this paper should be applied only in the instances in which measured baseline data indicate the response will reduce the spread between GPUs within a node and increase the amount of useful output from those GPUs, best determined through testing the workloads on the same node in interleaved jobs (Section 10).

10. Research Limitations and Recommended Next Steps

10.1. Limitations

Observational design and the collection-date confound. This is the most important limitation of the study. Each of the four cells containing combinations of workloads and hardware GPUs was collected on a single date: H100 image generation on 2025-09-16; B200 image generation on 2025-09-17; H100 LLM/text generation on 2025-09-20; and B200 LLM/text generation on 2025-09-22. Thus, each workload-hardware cell is confounded with the collection date for that workload. Consequently, it is not possible to separate any effect of the workload from the effect of the date on which it was collected. Thus, the reversal is supported by the workload-by-hardware interaction, but the interaction term does not remove the collection-date confound; the robustness checks all conditions on the observed cells and therefore cannot resolve it.

Sample size, clustering, and unbalanced cells. The study is based on 32 session-level observations. Such a sample size is appropriate for providing evidence regarding the workloads and hardware types in general. However, the public layout of the sessions provides only four distinct collection dates for the workloads and hardware types; each workload-hardware combination was collected on a single date. Furthermore, the sessions for each workload-hardware combination were collected sequentially on the same date. Thus, the sessions within each of the four cells are not independently collected. Additionally, there are seven sessions of image-generation workloads but nine LLM/text-generation workloads, indicating an unbalanced study. Finally, the sessions have varied configurations: batch size, image size, model size, sequence length, and DeepSpeed settings vary across sessions and are not independently separated from the categories of workloads and hardware.

Descriptive statistics limits. The descriptive statistics in Section 5 provide only a summary of the 32 observed sessions. They are not to be interpreted as population estimates, hardware benchmarks, or indicators of any general difference between GPU generations. Thus, the reversal stated in Section 5 applies only to the observations of workloads and hardware types; the descriptive statistics motivate the statistical test, but they are not the test itself.

Single public secondary dataset. The study was performed on a public dataset collected by authors different from those who conducted this study [17]. Thus, the findings of this study are based on the public and separate datasets.

Missing measurement fields. The temperature measurements did not include clock frequency, throttling reasons, cooling loop state, coolant temperature, or the GPUs' useful output.

Observed thermal regime. The maximum temperature of any GPU across all workload-hardware combinations was 77˚C. This is not the threshold at which the GPUs are designed to be throttled. Because the reasons for and clock frequency measurements of the GPUs are unavailable, no conclusions can be drawn about whether the GPUs were throttled during the observed sessions.

Scope of generalization and measurement scale. These results apply only to the two accelerator generations of GPUs tested and to the two types of workloads performed on them. These results cannot be generalized to other hardware, other types of workloads, or to any other data-center cooling architectures. Furthermore, temperature-difference measurements were taken only within each node of the eight GPUs; the results do not apply to within-die or within-package temperature differences or to heat-flux measurements.

No intervention validated. This study found no evidence of interventions related to the GPUs or the data center. For instance, no interventions related to thermal buffering, cooling, scheduling workloads on GPUs or any other vendor- or product-specific interventions were validated by this study.

10.2. Next Steps

The recommended next step is a study that addresses the confounding factors in this public dataset. Such a study would run both workload families on the same node or on nodes matched for cooling and time window. This design would then be repeated on both H100 and B200 nodes. Within each node, workloads would be executed in an interleaved, randomized order. In such a study, it would be necessary to measure the workload state, GPU power draw, GPU temperature, GPU utilization, memory behavior, clock frequency, throttle state, cooling condition, and the useful output of each GPU on each node. Each of these metrics is necessary to determine whether the node is experiencing thermal spread issues.

Future intervention studies would employ a baseline-versus-intervention study design. Each intervention candidate (workload-aware scheduling, cooling, thermal management, buffering) would be tested under matched workload, hardware, cooling, and time-window conditions. Each intervention would be credited only

for reducing the thermal metric and for increasing (or at least not decreasing) the nodes' useful output. Thus, future work should perform a study that measures the workload, the hardware components of the nodes, the local thermal spread of each node, the cooling condition of each node, and the useful output of each node. This study identifies the problem of insufficient measurement of these variables. Furthermore, it demonstrates the impact of workload and hardware on thermal locality within nodes. However, it does not provide solutions for establishing a causal relationship among workload, heat generation, throttling, and loss of useful output. That relationship is the problem for the next study on the topic.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] International Energy Agency (2025) Energy and AI (World Energy Outlook Special Report). IEA. https://www.iea.org/reports/energy-and-ai

[2] Lawrence Berkeley National Laboratory (2024) 2024 United States Data Center Energy Usage Report. https://eta-publications.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report

[3] Electric Power Research Institute (2026) Powering Intelligence 2026: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies (Report No. 3002034696). EPRI. https://www.epri.com/research/products/000000003002034696

[4] ASHRAE, NEMA and Pacific Northwest National Laboratory (2026) AI Data Center Energy Performance Framework. ASHRAE, NEMA and PNNL. https://www.ashrae.org/technical-resources/ai-data-center-framework

[5] The Green Grid (2012) PUE: A Comprehensive Examination of the Metric (White Paper No. 49; V. Avelar, D. Azevedo, & A. French, Eds.). The Green Grid. https://archive.thegreengrid.org/en/resources/library-and-tools/237-WP

[6] Horner, N. and Azevedo, I. (2016) Power Usage Effectiveness in Data Centers: Overloaded and Underachieving. *The Electricity Journal*, **29**, 61-69. https://doi.org/10.1016/j.tej.2016.04.011

[7] Fan, X., Weber, W. and Barroso, L.A. (2007) Power Provisioning for a Warehouse-Sized Computer. *Proceedings of the* 34*th Annual International Symposium on Computer Architecture*, San Diego, 9-13 June 2007, 13-23. https://doi.org/10.1145/1250662.1250665

[8] Skadron, K., Stan, M.R., Huang, W., Velusamy, S., Sankaranarayanan, K. and Tarjan, D. (2003) Temperature-Aware Microarchitecture. *Proceedings of the* 30*th Annual International Symposium on Computer Architecture—ISCA*'03, San Diego, 9-11 June 2003, 2-13. https://doi.org/10.1145/859618.859620

[9] Huang, W., Ghosh, S., Velusamy, S., Sankaranarayanan, K., Skadron, K. and Stan, M.R. (2006) Hotspot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design. *IEEE Transactions on Very Large Scale Integration* (*VLSI*) *Systems*, **14**, 501-513. https://doi.org/10.1109/tvlsi.2006.876103

[10] Hamann, H.F., Weger, A., Lacey, J.A., Hu, Z., Bose, P., Cohen, E., *et al.* (2007) Hotspot-Limited Microprocessors: Direct Temperature and Power Distribution Measurements. *IEEE Journal of Solid-State Circuits*, **42**, 56-65.

https://doi.org/10.1109/jssc.2006.885064

[11] Heterogeneous Integration Roadmap (2023) Chapter 20: Thermal. IEEE Electronics Packaging Society. https://eps.ieee.org/wp-content/uploads/2025/11/ch20_thermalfinal.pdf

[12] Patel, P., Choukse, E., Zhang, C., Goiri, Í., Warrier, B., Mahalingam, N., *et al.* (2024) Characterizing Power Management Opportunities for LLMs in the Cloud. *Proceedings of the* 29*th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, *Volume* 3, La Jolla, 27 April-1 May 2024, 207-222. https://doi.org/10.1145/3620666.3651329

[13] Mayr, M., Wind, S., Schroder, L., Moradi, M., Hager, G., Kostler, H. and Wellein, G. (2026) AI Application Benchmarking: Power-Aware Performance Analysis for Vision and Language Models. arXiv: 2603.16164. https://arxiv.org/abs/2603.16164

[14] Go, S., Park, J., More, S., Wu, H., Wang, I., Jezghani, A., *et al.* (2025) Characterizing the Efficiency of Distributed Training: A Power, Performance, and Thermal Perspective. *Proceedings of the* 58*th IEEE/ACM International Symposium on Microarchitecture*, Seoul, 18-22 October 2025, 626-642. https://doi.org/10.1145/3725843.3756111

[15] Chung, J.W., Ma, J.J., Wu, R., Liu, J., Kweon, O.J., Xia, Y., Wu, Z. and Chowdhury, M. (2025) The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization. arXiv: 2505.06371. https://arxiv.org/abs/2505.06371

[16] Špeťko, M., Vysocký, O., Jansík, B. and Říha, L. (2021) DGX-A100 Face to Face Dgx-2—Performance, Power and Thermal Behavior Evaluation. *Energies*, **14**, Article 376. https://doi.org/10.3390/en14020376

[17] Elsayed, A.A.E., Al-Obaidi, A.A. and Farag, H.E.Z. (2026) Characterization of High-Resolution AI Data Center Training Workloads on Single and Multiple GPU Nodes. *Scientific Data*. https://doi.org/10.1038/s41597-026-07496-6

[18] Masanet, E., Shehabi, A., Lei, N., Smith, S. and Koomey, J. (2020) Recalibrating Global Data Center Energy-Use Estimates. *Science*, **367**, 984-986. https://doi.org/10.1126/science.aba3758

[19] NVIDIA (2026) NVIDIA H100 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/h100/

[20] NVIDIA (2026) Product Carbon Footprint Summary for NVIDIA HGX B200. https://images.nvidia.com/aem-dam/Solutions/documents/HGX-B200-PCF-Summary.pdf

[21] ASHRAE Technical Committee 9.9. (2024) Liquid Cooling: Resiliency Guidance for Cold Plate Deployments [Technical Bulletin]. ASHRAE. https://tpc.ashrae.org/Documents?cmtKey=fd4a4ee6-96a3-4f61-8b85-43418dfa988d

[22] Kuzay, M., Demirel, E., Bayraktar, B., Vilestad, J., Kärnebro, A., Yilmaz, C., *et al.* (2026) A Self-Assessment Framework for Evaluating Efficiency of Data Centers. *Energy Informatics*, **9**, Article No. 39. https://doi.org/10.1186/s42162-026-00652-7

[23] You, J., Chung, J.W. and Chowdhury, M. (2023) Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training. *Proceedings of the* 20*th USENIX Symposium on Networked Systems Design and Implementation* (*NSDI*'23), Boston, 17-19 April 2023, 119-139. https://www.usenix.org/conference/nsdi23/presentation/you

[24] Tschand, A., Rajan, A.T.R., Idgunji, S., Ghosh, A., Holleman, J., Kiraly, C., *et al.* (2025) MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI. 2025 *IEEE International Symposium on High Performance Computer Architecture* (*HPCA*), Las Vegas, 1-5 March

2025, 1201-1216. https://doi.org/10.1109/hpca61900.2025.00092

[25] Fadel Argerich, M., Furst, J. and Patino-Martinez, M. (2026) Watt Counts: Energy-Aware Benchmark for Sustainable LLM Inference on Heterogeneous GPU Architectures. arXiv: 2604.09048. https://arxiv.org/abs/2604.09048

[26] Hong, S. and Kim, H. (2010) An Integrated GPU Power and Performance Model. *Proceedings of the* 37*th Annual International Symposium on Computer Architecture*, Saint-Malo, 19-23 June 2010, 280-289. https://doi.org/10.1145/1815961.1815998

[27] Leng, J., Hetherington, T., ElTantawy, A., Gilani, S., Kim, N.S., Aamodt, T.M., *et al.* (2013) GPUWattch: Enabling Energy Optimizations in GPGPUs. *Proceedings of the* 40*th Annual International Symposium on Computer Architecture*, Tel-Aviv, 23-27 June 2013, 487-498. https://doi.org/10.1145/2485922.2485964

[28] Kandiah, V., Peverelle, S., Khairy, M., Pan, J., Manjunath, A., Rogers, T.G., *et al.* (2021) AccelWattch: A Power Modeling Framework for Modern GPUs. *MICRO*-54: 54*th Annual IEEE/ACM International Symposium on Microarchitecture*, 18-22 October 2021, 738-753. https://doi.org/10.1145/3466752.3480063

[29] NVIDIA (n.d.) Data Center GPU Manager Documentation. https://docs.nvidia.com/datacenter/dcgm/

[30] NVIDIA (n.d.) NVIDIA Management Library Documentation. https://docs.nvidia.com/deploy/nvml-api/

[31] Yang, Z., Adamek, K. and Armour, W. (2024) Accurate and Convenient Energy Measurements for GPUs: A Detailed Study of NVIDIA GPU's Built-In Power Sensor. *SC*24: *International Conference for High Performance Computing, Networking, Storage and Analysis*, Atlanta, 17-22 November 2024, 1-17. https://doi.org/10.1109/sc41406.2024.00028

[32] Garimella, S.V., Fleischer, A.S., Murthy, J.Y., Keshavarzi, A., Prasher, R., Patel, C., *et al.* (2008) Thermal Challenges in Next-Generation Electronic Systems. *IEEE Transactions on Components and Packaging Technologies*, **31**, 801-815. https://doi.org/10.1109/tcapt.2008.2001197

[33] Szekely, V. (1998) Identification of RC Networks by Deconvolution: Chances and Limits. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, **45**, 244-258. https://doi.org/10.1109/81.662698

[34] Lasance, C.J.M. (2008) Ten Years of Boundary-Condition-Independent Compact Thermal Modeling of Electronic Parts: A Review. *Heat Transfer Engineering*, **29**, 149-168. https://doi.org/10.1080/01457630701673188

[35] JEDEC Solid State Technology Association (2010) Transient Dual Interface Test Method for the Measurement of the Thermal Resistance Junction-to-Case of Semiconductor Devices with Heat Flow through a Single Path (JESD51-14). JEDEC.

[36] Farkas, G., Poppe, A. and Rencz, M. (2022) Theoretical Background of Thermal Transient Measurements. In: Rencz, M., Farkas, G. and Poppe, A., Eds., *Theory and Practice of Thermal Transient Testing of Electronic Components*, Springer, 7-96. https://doi.org/10.1007/978-3-030-86174-2_2

[37] Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschwiler, T. and Atienza, D. (2010) 3D-ICE: Fast Compact Transient Thermal Modeling for 3D ICs with Inter-Tier Liquid Cooling. 2010 *IEEE/ACM International Conference on Computer-Aided Design* (*ICCAD*), San Jose, 7-11 November 2010, 463-470. https://doi.org/10.1109/iccad.2010.5653749

[38] Price, D.C., Clark, M.A., Barsdell, B.R., Babich, R. and Greenhill, L.J. (2015) Optimizing Performance-per-Watt on GPUs in High Performance Computing. *Computer Science—Research and Development*, **31**, 185-193.

https://doi.org/10.1007/s00450-015-0300-5

[39] Jia, Z., Maggioni, M., Smith, J. and Scarpazza, D.P. (2019) Dissecting the NVIDIA Turing T4 GPU via Microbenchmarking. arXiv: 1903.07486. https://arxiv.org/abs/1903.07486

[40] Dean, J. and Barroso, L.A. (2013) The Tail at Scale. *Communications of the ACM*, **56**, 74-80. https://doi.org/10.1145/2408776.2408794

[41] Chen, J., Pan, X., Monga, R., Bengio, S. and Jozefowicz, R. (2016) Revisiting Distributed Synchronous SGD. arXiv: 1604.00981. https://arxiv.org/abs/1604.00981

[42] Lin, J., Jiang, Z., Song, Z., Zhao, S., Yu, M., Wang, Z., Wang, C., Shi, Z., Shi, X., Jia, W., Liu, Z., Wang, S., Lin, H., Liu, X., Panda, A. and Li, J. (2025) Understanding Stragglers in Large Model Training Using What-If Analysis. *Proceedings of the* 19*th USENIX Symposium on Operating Systems Design and Implementation* (*OSDI*'25), Boston, 7-9 July 2025, 483-498. https://www.usenix.org/conference/osdi25/presentation/lin-jinkun

[43] Moore, J., Chase, J., Ranganathan, P. and Sharma, R. (2005) Making Scheduling Cool: Temperature-Aware Workload Placement in Data Centers. *Proceedings of the* 2005 *USENIX Annual Technical Conference*, Anaheim, 10-15 April 2005, 61-75. https://www.usenix.org/conference/2005-usenix-annual-technical-conference/making-scheduling-cool-temperature-aware-workload

[44] Stojkovic, J., Zhang, C., Goiri, Í., Choukse, E., Qiu, H., Fonseca, R., *et al.* (2025) TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms. *Proceedings of the* 30*th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, *Volume* 2, Rotterdam, 30 March-3 April 2025 1266-1281. https://doi.org/10.1145/3676641.3716025

[45] Lu, R. and Wang, D. (2025) A Thermal-Aware Workload Scheduler for High-Performance LLM Inference in Cooling-Regulated Datacenters. *ACM SIGEnergy Energy Informatics Review*, **5**, 98-104. https://doi.org/10.1145/3757892.3757906

[46] Open Compute Project (n.d.) ACS Liquid Cooling Cold Plate Requirements. Open Compute Project. https://www.opencompute.org/documents/ocp-acs-liquid-cooling-cold-plate-requirements-pdf

[47] Open Compute Project (n.d.) OAI System Liquid Cooling Guidelines. Open Compute Project. https://www.opencompute.org/documents/oai-system-liquid-cooling-guidelines-in-ocp-template-mar-3-2023-update-pdf

[48] ARPA-E (n.d.) COOLERCHIPS Program. U.S. Department of Energy. https://arpa-e.energy.gov/technologies/programs/coolerchips

[49] Bhatasana, M. and Marconnet, A.M. (2025) Phase Change Materials as Thermal Buffers for Power Electronics Modules with Transient Heat Loads. *Energy Conversion and Management*, **343**, Article ID: 119931. https://doi.org/10.1016/j.enconman.2025.119931

[50] Open Compute Project (n.d.) Cooling Environments. Open Compute Project. https://www.opencompute.org/wiki/Cooling_Environments