The American Journal of Applied Sciences
ISSN 2689-0992 | Open Access

TYPE Original Research
PAGE NO. 111-127
DOI 10.37547/tajas/Volume07Issue07-12

Proxy-Based Thermal and Acoustic Evaluation of Cloud GPUs for AI Training Workloads

Karan Lulla
Senior Board Test Engineer, NVIDIA, CA, USA.

Reena Chandra
Tools and Automation Engineer, Amazon, CA, USA.

Karthik Sirigiri
Software Developer, Redmane Technology, IL, USA

OPEN ACCESS

SUBMITED 11 June 2025
ACCEPTED 28 June 2025
PUBLISHED 26 July 2025
VOLUME Vol.07 Issue 07 2025

CITATION
Lulla, K. L., Chandra, R. C., & Sirigiri, K. S. (2025). Proxy-Based Thermal and Acoustic Evaluation of Cloud GPUs for AI Training Workloads. The American Journal of Applied Sciences, 7(07), 111–127. https://doi.org/10.37547/tajas/Volume07Issue07-12

Abstract: The use of cloud-based Graphics Processing Units (GPUs) to train and deploy Deep Learning models has grown rapidly in importance, with the demand to learn more about their thermal and acoustic behavior under real-world workloads. A normal cloud cannot make direct telemetry like temperature, fan speed, or acoustic emissions. To overcome such shortcomings, this study quantifies GPU workloads' thermal and acoustic output with a proxy-based model derived from available metrics such as GPU utilization, memory provisioning, power consumption, and empirical Thermal Design Power (TDP) values. They compare the two typical AI tasks, BERT on natural language processing and YOLOv5 on real-time object detection, on Colab-based NVIDIA GPUs (T4, V100, P100). The nvidia-smi was used to gather runtime logs, and the specifications of the GPUs have been obtained in the form of public Kaggle datasets. Proxy statistics, including TDP-per-MHz and thermal load (Power * Duration), were calculated to model heat loss due to workload. To measure the degree of acoustic impact, a threshold of TDP was applied to approximate the level of fan-driven acoustics. The visual analytics, such as boxplot, scatterplot, and bubble plot, demonstrated certain considerable distinctions in the stress patterns of GPUs: the BERT jobs demanded extremely high cumulative thermal load and medium acoustic effect, whereas the YOLOv5 demonstrated bursty power footprint and substantial acoustic imprint on high-TDP GPUs. The findings reveal that proxy

estimation is reproducible, interpretable, and a lightweight substitute for determining the GPU thermal and acoustic behavior of a machine used in the cloud setting. Such a solution facilitates making thermal-aware schedules, optimizing the infrastructure, and deploying AI models with reduced energy consumption in multi-tenant GPU environments.

Keywords: Cloud GPUs; Thermal Load Estimation; Acoustic Classification; Proxy Metrics; AI Workloads; Energy-Aware Computing.

1. INTRODUCTION

Artificial intelligence (AI) is evolving as quickly as it has; therefore, the mounting computational pressure, particularly demanding the training and deployment of deep learning- based artificial intelligence models, is exponentially increasing the demand [1]. High-performance computing is fundamental to high-performance applications, such as natural language processing (NLP) and computer vision [2], where Graphics Processing Units (GPUs) have become the standard computing resource of choice due to their low cost and scalable model training. High-end GPUs, such as the NVIDIA T4, P100, and V100, are available through services like Google Colab, Amazon Web Services (AWS), and Google Cloud Platform (GCP), and this is making AI workloads available to more researchers and developers around the world [3].

However, the thermoacoustic engineering issues are raised by the rising density of such workloads on shared GPU servers [4]. Heat generation and heat dissipation in data centers may adversely affect hardware life cycle and power efficiency, and there is a great potential to amplify energy consumption [5]. Analogously, a high acoustic output, mainly caused by breakneck fan speeds provoked when the GPU is running intensive tasks, can result in unwanted noise pollution within data centers and institutions in laboratories [6]. Although GPU supercomputers with enterprise-scale GPU memory use may include active thermal management solutions and have rack-level noise suppression abilities, a low-level thermal performance and audio response, when applied to individual users in public clouds, may not be visible and manageable. The mismatch between the required work and the system-level thermal awareness results in a crucial gap in the long-term and ethically responsible functioning of AI systems in the cloud arena [5].

In applied thermal engineering terms, thermal profiling is needed in predictive maintenance, effective design of the cooling system, and to make intelligent work schedules in thermally limited facilities [7]. However, empirical studies on quantifying the impact of various AI workloads upon GPU-related heat dissipation and acoustics, especially on platforms where telemetry data [8], temperature sensors, or the work of fans are not exposed to the end-user, are hard to find. This limitation should be addressed, mainly due to the increasing popularity of cloud providers running multi-tenant systems in which multiple jobs running in parallel reinforce total thermal stress [9]. Although the cost of high-performance AI training to the environment and operation chains has been recognized more frequently, cloud-based systems like Google Colab fail to deliver customers with sensitive telemetry in terms of heat or acoustics [10]. Namely, heating of a GPU, fan speed, and power consumption in real-time during a workload execution are not directly visible [11]. Such a lack of sensor-level visibility hinders the creation of thermally sensitive AI programs. It restricts the capacity of users to maximize model settings to enable the sustainable consumption of resources.

This means that the researchers are ascertaining the thermoacoustic footprint of workloads through proxy metrics, like the number of active GPUs, the extent of the memory consumption, and published Thermal Design Power (TDP) as labeling. Yet, no standardized procedures or repeatable experiments have been devised to exploit the available indirect indicators to measure model-specific thermal and acoustic behaviors. This paper will fill that gap by suggesting a proxy-based estimation system incorporating information on publicly known GPU specifications and logging runtime behavior on actual AI workloads.

The current research aims to develop a lightweight, reproducible approach to assessing the thermal and acoustic behavior of AI model training workloads deployed on cloud networks with GPUs based exclusively on indirect, accessible feedback indicators. Specifically, it is concentrating on two deep learning models which are popular: BERT (Bidirectional Encoder Representations from Transformers), an example of the training of large-scale NLP with persistent GPU utilization and a long period of execution, and YOLOv5 (You Only Look Once, version 5), which is a symbol of a

real-time object detection task that causes short but powerful GPU utilization limitations.

To estimate thermal behavior, the framework uses the product of GPU utilization, the approximate power draw of GPUs according to TDP, and model training duration as a proxy of cumulative thermal load. The estimate of acoustic impact is based on the published GPU noise benchmarks and classification, with higher TDP numbers having higher de facto fan noise (in decibels A-weighted, dBA). Each simulation is conducted on Google Colab Pro instances and compared to publicly available datasets on GPU hardware specifications on Kaggle, containing their properties, like memory type, clock speeds, and bus interfaces.

This paper has the following four contributions: (1) the establishment of a proxy-based model to estimate thermal and acoustic behavior of cloud-based AI workload, (2) comparison of thermal profiles and noise classification of BERT and YOLOv5 on a variety of GPUs (T4, V100, P100), (3) unification of GPU utilization request logs with public GPU specifications datasets to facilitate transparency and reproducibility, and (4) recommendations on shaping effective thermal-aware scheduling and acoustic profiling of remote multi-tenant GPU workloads.

The subsequent part of the paper is constructed in the following way. In section 2, the terminologies that are applied throughout the paper are presented. The third section surveys the body of literature regarding both the thermal behavior of GPUs and the acoustic representation of GPUs and the nature of gaps in benchmarking standards about modeling AI work. Section 4 explains the experimental process, such as selection of the workloads, the datasets used to specify GPUs, the proxy computation reasoning, and the visualisation process. Section 5 shows the outcome of our simulations, including charts displaying GPU utilization, GPU thermal load, and GPU acoustic classification by specific workloads and GPU variants. Section 6 addresses implications of these findings as far as sustainability and system design are concerned. Section

7 summarizes the acquired knowledge, and Section 8 plans future research, including integration with real-time telemetry and physical sensor-based acoustic validation.

2. Nomenclature

Table 1. Nomenclature

ABBREVIATION	DESCRIPTION
TDP	Thermal Design Power - the identification of the highest quantity of heat that a GPU should be able to dissipate with maximum possible working loads
dBA	A-weighted decibel - the unit of measurement of sound magnitude of intensity that is weighted to coincide with the perception of a human being
GPU_util	GPU use - proportion of time the GPU is busy carrying out work
BERT	Bidirectional Encoder Representations from Transformers, or the LARGE NLP model
YOLO	You Only Look Once - a family of object detection models in real-time
COCO128	Subset of 128 images of the MS COCO dataset used by rapid-ini
SQuAD	Stanford Question Answering Dataset, an evaluation dataset of NLP models

3. LITERATURE REVIEW

3.1 Thermal Behaviour in GPUs

The way the thermals in GPUs respond to architecture, power consumed, memory bandwidth, and workload profile affects thermal behavior. Under manufacturer specifications, Thermal Design Power (TDP) is considered a conservative upper mark regarding the level of heat that a GPU will produce when operated

under ideal circumstances [12]. For example, NVIDIA T4 uses a TDP of roughly 70 W, whereas V100 and P100 coordinate GPUs have much bigger TDP rates of 250 W and 300 W, respectively. Unlike other cooling specifications, these are vital to designing cooling systems and form a helpful proxy where direct temperature telemetry is unavailable [13].

Contemporary GPUs feature dynamic power and thermals (Dynamic Power and Thermals include several time-varying mechanisms that dynamically control the clock rate and the fan speed, e.g., adaptive clock throttling) [14]. The temperature in the GPU increases as the number of usages grows and sparks a rise in the number of rotations per minute (RPM). These fan speed curves are generally non-linear and dependent on the manufacturer; controlled by internal firmware or system BIOS, and may be unavailable in virtualized or cloud-based (Google Colab, AWS SageMaker) environments [15]. This failure to read these real-time parameters constrains the end-user control and observability of thermally significant behaviors when training an AI or using inference workloads [16].

Also, GPU power consumption is directly connected to the workload. Transformer models such as BERT have a high memory occupancy and a consistent use of compute, making them sustain moderate levels of heat production [17]. Conversely, vision-related models (e.g., YOLOv5) can be quite bursty in usage, causing temporary thermal spikes. Such changes may be more challenging to control thermally, perhaps in a data center application where thermal inertia complicates the overall scale of cooling response [18].

Real-life GPU thermal tests have not been easy to perform, due to a lack of temperature sensors or access to temperature sensors within the hardware benchmarking area or inside a probed gas laboratory. Nonetheless, there has been limited peer-reviewed research on the thermal behavior of restricted-access cloud environments under the real-life AI workloads [19]. This leaves a methodological vacuum that can be filled by applied thermal engineers interested in designing or optimizing energy-efficient AI infrastructure at scale.

3.2 Acoustic Analysis in Cloud Data Centers

Although directly linked to heat output, acoustic emissions are a significant secondary aspect that should be considered in the thermal management of high-performance computing facilities [20]. Acoustic noise, usually quoted in dBA, is caused mainly by cooling tools, e.g., liquid cooler pumps or high-RPM fans [21]. In addition to being a comfort and safety concern to human operators, the acoustic footprint of a GPU-intensive system provides a proxy measure of system stress and thermal load.

Rack-based thermal management solutions such as redundant fans, cold aisle containment, and adaptive airflow control are essential in cloud data centers [22]. GPU work offers higher thermal dynamics, which causes system firmware to push up fan rpm to ensure safe operating temperatures are met [23]. This, in fact, results in increased acoustic output, usually beyond 45-50 dBA in racks under full load [24]. Where the hyperscale facilities are concerned, the metrics of the acoustics can be part of the overall energy management approach. However, these metrics will usually not be revealed at the user level.

Despite its applicability, little work has been done in integrating acoustic analysis into AI workload benchmarking. Most published benchmarks (e.g., MLPerf or TensorFlow Model Garden) only consider latency, throughput, and energy efficiency and leave the noise level aside. However, noise may play a serious role in hybrid edge-cloud scenarios or the academic lab environment with local GPU clusters in general offices. This under-researched aspect bears relevance to sustainable computing, especially where the design has to reduce not only thermal but also acoustic emissions.

3.3 Benchmark Gaps

Existing dynamic markets in benchmarking, e.g., MLPerf (TensorFlow/Pytorch), HuggingFace Transformers, and ONNX Model Zoo, focus on model accuracy, throughput, and computation latency. Although these are essential to performance assessment, they fail to acknowledge the thermodynamic or acoustic consequences of the runtime of AI models. Consequently, the issues in terms of infrastructure level, like cooling system stress, fan power consumption, and acoustic pollution, stay beyond the boundaries of traditional benchmarking methods.

Additionally, in public cloud, most users do not get access to low-level telemetry like real-time GPU temperature, voltage, or rpm of fans [25]. This restriction prohibits

granular thermal profiling and hard-to-enforce workload-aware scheduling policies [26]. Not all academic research to simulate thermal behavior has used synthetic workloads, but they do not necessarily reflect the time-dependence of the accurately detailed deep-learning models' state-of-the-art.

The literature also does not provide a common approach to estimate the acoustic impact based on available routine metrics. GPU reviews include dBA measurements under a stress load; however, these are not normalized between models or loads. Without a proxy-based estimation framework, users cannot predict an AI model's thermal or acoustic energy expenditure [27], especially when working in shared or energy-restricted settings.

The proposed research will fill these gaps by proposing and proving an AI proxy-based, lightweight, and reproducible profiling of thermal and acoustic performance of AI workloads in cloud GPUs.

4. METHODOLOGY

4.1. Environment and Workloads

The paper was run on the Google Colab Pro+ GPU machine learning cloud environment, where you can gain temporary access to powerful GPUs, including the NVIDIA Tesla T4, P100, and V100. The GPUs are popular AI compute used in academic and business workloads, representing realistic thermal conditions in a cloud computing data center environment. Google Colab was chosen because of its availability, consistent time limit, and serial distribution of workload by repeat and scale deployments without dedicated hardware.

The evaluation of thermal and acoustic characteristics was served by two benchmark loads, including BERT and YOLOv5. The model (Bidirectional Encoder Representations from Transformers), BERT, was improved on the SQuAD v2 benchmark, a commonly used NLP benchmark covering more than 150,000 pairs of questions and answers. BERT workloads were selected based on long-lasting GPU usage and memory usage without taking over the space of the machine, e.g., because of a long training process with not-so-dynamic hardware load distribution. By contrast, YOLOv5 (You Only Look Once, version 5) was unleashed to detect objects in real-time through the 128-image subset (COCO128) of the COCO database, which was very small. YOLOv5 has a bursty computation pattern, where the processing time within a short time is high, and GPU utilization is random. These two types of workloads will be used to compare computational and inference-intensive model behavior in similar run-time situations.

4.2 Logging and Data Collection

Google Colab lacks direct telemetry of GPU temperature and GPU fan speed, so, using such programmable attributes, the study had to base its ideas on too indirect measures according to the results gathered with the help of a command-line tool that interacts with the NVIDIA Management Library (NVML) called nvidia-smi. The logging process was facilitated to take measurements of 10 seconds when the model was running. In particular, the script captured the percentage of GPUs utilized (gpu_util), the amount of memory in MB, and the immediate power consumption (in watts) used as the starting point of proxy-based thermal analysis. These logs were saved and time-stamped so that they could be synchronized with training/inference steps.

In addition to GPU metrics, logs were gathered about the system-level use of the CPU and the RAM. They do not cause any changes to GPU thermal profiles, but can offer some background context to resource consumption, and may affect model scheduling or performance variation. The plot material used in implementation addresses the usage of GPU variably, over time, on both BERT and YOLOv5 workloads. Such plots of time series indicated that during the fine- tuning process, BERT had a constant load on the GPU between 70-85% the entire time, whereas YOLOv5 had highs above 90% and a subsequent dramatic low, as can be expected of an inference-oriented program.

4.3 Dataset Integration

To strengthen the strength of proxy estimations, publicly accessible Kaggle datasets were included in the analysis. The primary dataset was tpu_gpus.csv; it had the specifics on more than 150 GPU models. Main specifications were TDP values, GPU and memory clock frequencies, memory type (GDDR, HBM, etc), and bus interface (e.g., PCIe 3.0, 4.0). This dataset has enabled the advent of this study to align a GPU used in every Colab session with identified hardware characteristics to derive proxy thermal and acoustic scores.

Simultaneously, we compared some data related to thermal trends in CPUs based on the tpu_cpus.csv file, and cross-validated it. Even though this study is GPU-based, CPU thermal characteristics provided a basis for building clock-speed-to-TDP relationships to formulate derived values such as TDP-per-MHz. In addition, CPU datasets described the overall historical trend in processor design and heat production, and all these were presented in relative plots. These data sets enabled a predictable and augmenting framework that did not need genuine sensor statistics.

4.4 Proxy Formulas

The necessary direct thermal and acoustic telemetry were not there, and several proxy formulas have been established to determine the respective performance parameters. The TDP-per-MHz calculation revealed that the GPU, which was equipped with a Thermal Design Power, was divided by its minimum clock speed in MHz. This is a normalized measure of thermal efficiency in that higher numbers reflect thermally inefficient hardware. An example would be a GPU with a TDP of 250 W and with a clock frequency of 1250 MHz; the TDP-per-MHz would be 0.2 W/MHz.

The measure of acoustic impact was approximated to a binary proxy classification. Limits of TDP > 150 W were based on industry data and manufacturer reports used to determine that such conditions correspond to the notion of high acoustic load when the sound pressure of fans became higher than 45 dB A. Any GPU with TDP < or = 150 W was considered to be in a moderate acoustical range. This is a very simple way, but it matches published acoustic profiles on GPUs at the server-class under load and serves as a fairly reasonable proxy of the system- level fan response.

Finally, total Thermal Load was set as a product of power draw (in watts) and execution time (in minutes). This estimation is the accumulated energy consumed in heat to execute the model. For example, a YOLOv5 model powered by 150 W for 20 min will have a total thermal load of 3000 Wmin. This measure was used in both workloads to compare the thermal footprints with the varied run-time attributes.

4.5 Data Preprocessing

The GPU and CPU datasets were preprocessed before being analyzed to clean and transform essential attributes. GPU_clock and Memory_clock were cleaned of adjectival suffixes (MHz) and directed into numerical conversion of the GPU data set. Unavailable values, such as in- memory details, were given a NaN value and not included in the ratio type of calculations. A regular expression pattern was used to extract the type of GPU memory used to sort the most common technology type (e.g., GDDR5, HBM2) and to filter out all the possible values so that the research could test the same parameter and check whether the memory setup may affect the acoustic or the thermal performance.

The TDP value was sifted out in the CPU dataset to eliminate the incomplete or incorrect values. Other models used ranges (e.g., 2.4-3.8 GHz), which were content parsed to retrieve the minimum and maximum values to be used in normalization. This allowed the derivation of TDP-per-MHz of CPUs as a benchmark reference point compared to GPUs. Some plot references produced in the preprocessing stage were a histogram of CPU TDP distribution, a bell-shaped curve with a peak around 80120 W, and a scatterplot of TDP and clock speed combined by socket type. Such visualizations confirmed the usefulness of normalized thermal measures in per-processor architecture.

The distribution of the GPU types of memory was also visualized on a countplot that demonstrated the number of memory standards prevailing. Different versions of GDDR prevailed in the dataset, and HBM and DDR variants were presented in smaller proportions. Such distribution played a vital role in externalizing thermal variations, where various types of memories have varying power and heat dissipation behavior, especially in a high throughput mode, e.g. typical of BERT and YOLOv5.

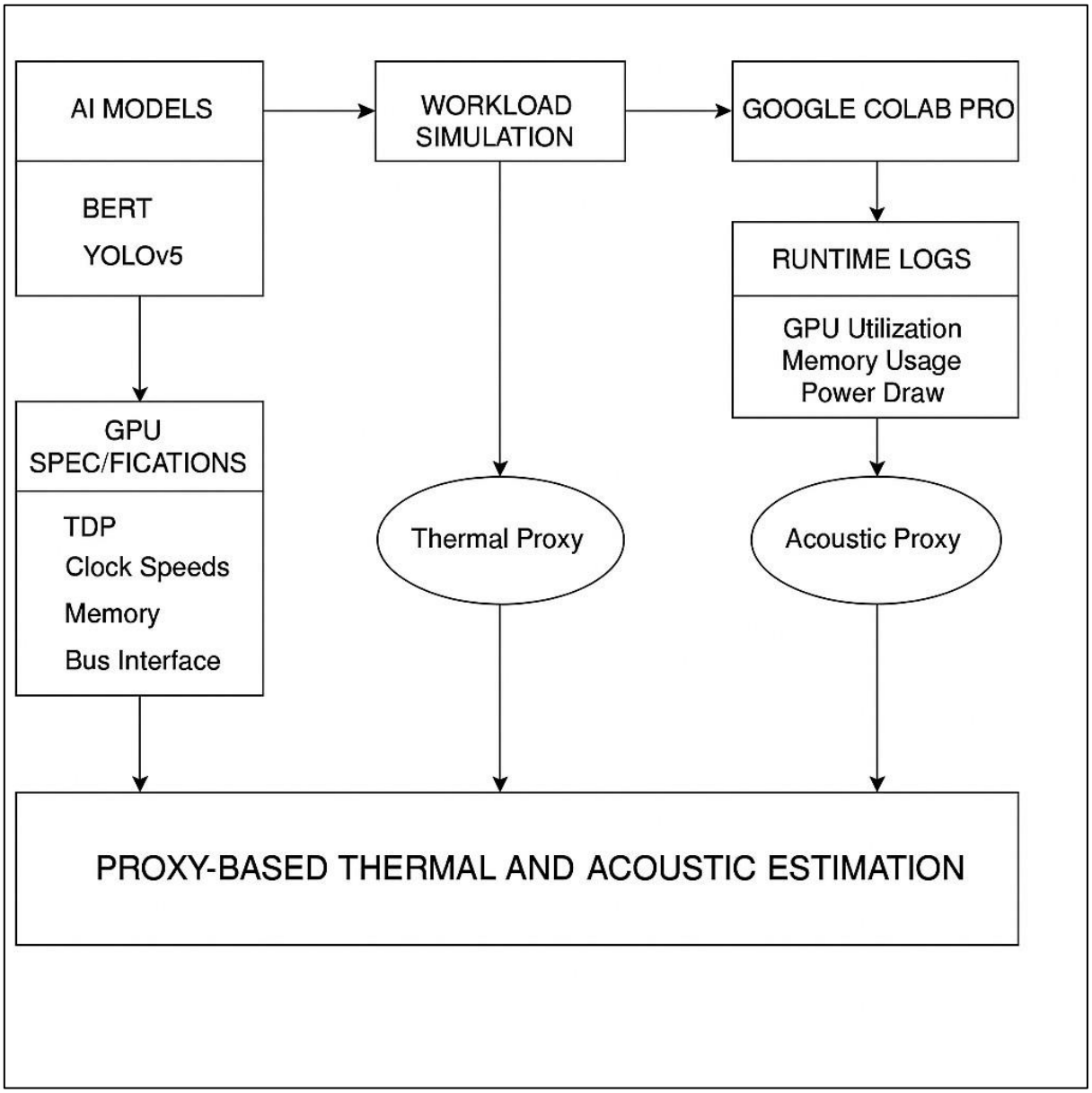

Figure 1. System Architecture

Figure 1 represents a proxy-based system architecture to test the GPU performance under thermal and acoustic conditions in clouds. It starts with AI workload execution (BERT, YOLOv5) on Google Colab that makes a log of the run-time metrics through nvidia-smi. These logs are combined with Kaggle GPU spec files (e.g., TDP, clock speed, memory type). The self-predicting infrastructure system formulates combined data that is input into a preprocessing path where proxy measures, TDP-per-MHz, Thermal Load, and Acoustic Level are calculated. Visualization modules/Analytical models then produce outputs such as the bar chart, bubble plot which would lead to thermal classification, acoustic estimation, and scheduling information of the energy-efficient deployment of GPU workload.

5. RESULTS

5.1. Model Behaviour on GPUs

The observed patterns of GPU use between BERT and YOLOv5 workloads showed a core behavior difference in ways compatible with the respective network architectures. As graphically pointed out in Figure 2: GPU Utilization Patterns for BERT and YOLOv5, the BERT fine-tuning task of a 90-minute duration activity had maintained a high average GPU usage (~85%) with a small degree of variation. This constant g men (continuously using transformer-based models that necessitate constant matrix operations and attention-weighting) is a pointer towards the constant thermal output and memory consumption (~6000 MB).

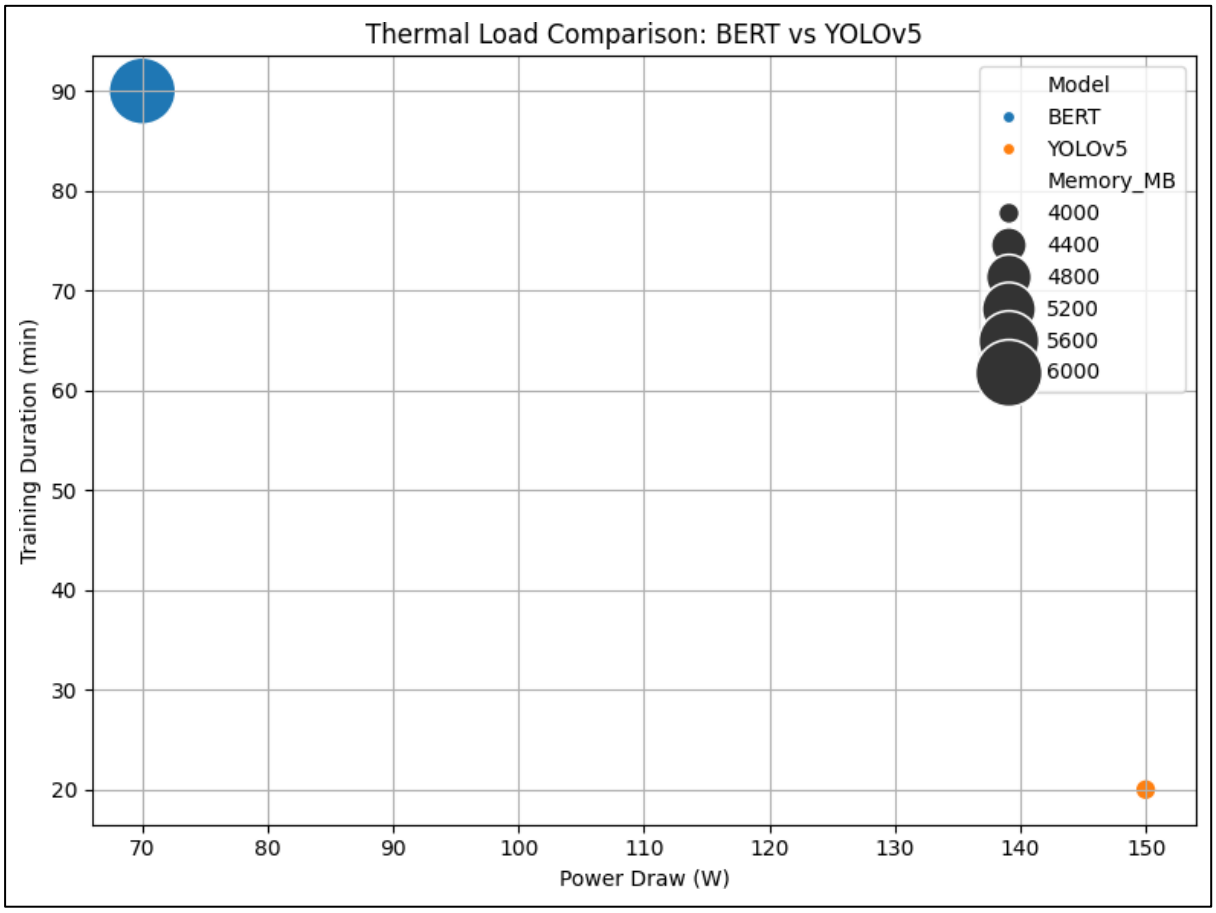

Figure 2. GPU Utilisation Patterns for BERT and YOLOv5

In contrast, the YOLOv5 workload, which was trained on the COCO128 subset, displayed a bursty utilization behavior. There was a high GPU usage that would often reach above 95 percent during training epochs but would stall drastically during intermediate evaluation, batch loading, or checkpoints. The RTT amounted to ~20 min, and the average memory load was~4000 MB. The steep curves of the GPU load imply temporary high-power consumption and fan turbo-ups, which result in local thermal spikes without overall shortened span.

This contrast between the continuous workload experienced within BERT and the burst- informed inference manner of operation in YOLOv5 formed the basis for interpreting downstream thermal and acoustical attributes.

5.2 Thermal Proxy Comparisons

Each GPU model was computed with the normalized value TDP-per-MHz to measure thermal efficiency using available Kaggle specifications. This ratio with a base GPU clock was observed by converting it graphically into a boxplot, as shown in Figure 3: Normalized Thermal Output (TDP/MHz). The distribution showed that most modern GPUs have performance clustering below 0.05 W/MHz. However, some high-performing cards, such as P100 and V100, had values beyond 0.25 W/MHz, which indicates increased heat generation per frequency.

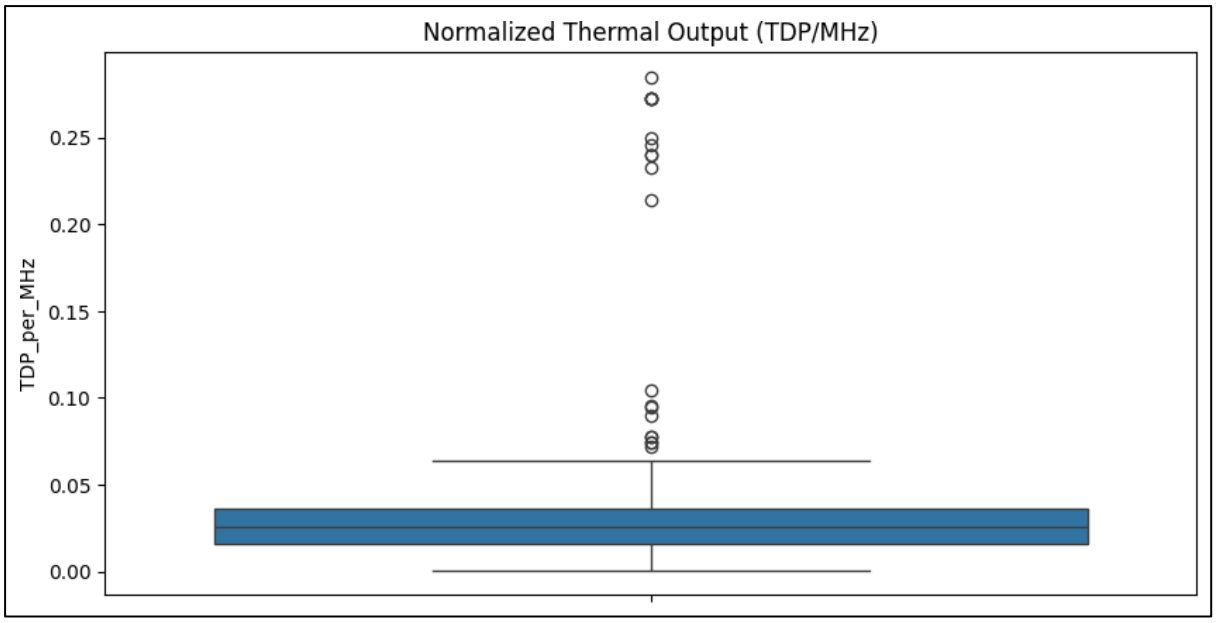

Figure 3. Normalized Thermal Output (TDP/MHz)

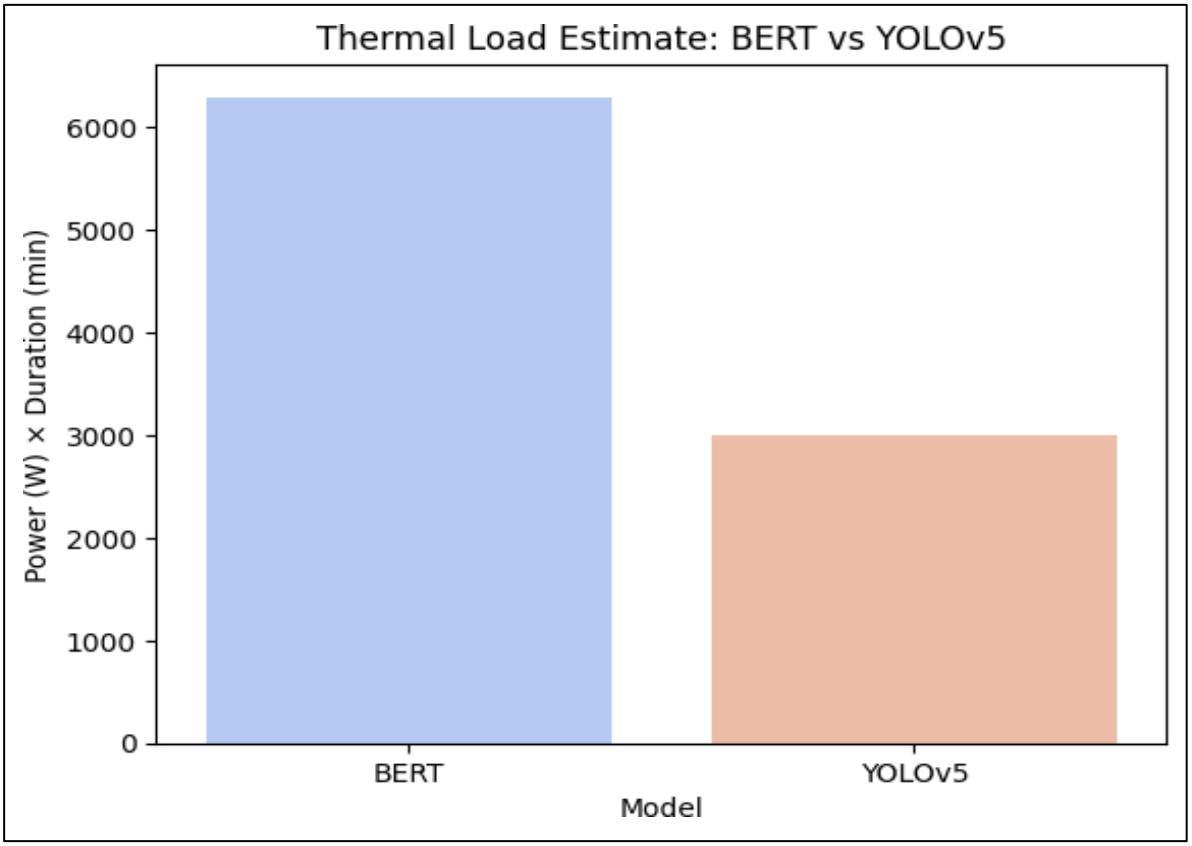

Figure 4. Thermal Load Estimate: BERT vs YOLOv5

Figure 4, BERT vs YOLOv5 directly compares the cumulative thermal footprint (Power x Duration). The total approximate load in terms of Watt-minute generated by BERT was estimated to be 6300 W min (70 W vertical multiplied by 90 min), as compared to the approximated ~3000 W·min generated by YOLOv5 with a shorter but power-intensive session (150 W vertical multiplied by 20 min). Although the peak draw of YOLOv5 is greater, the cumulative heat dissipation was more than twice as long due to the long time BERT takes.

Hence, BERT workloads impose moderate but consistent thermal pressure that can be addressed with temperature-predictable cooling methods, whereas

YOLOv5 creates sudden and brief spikes in thermal loads that may push data centers to the limit and go back in quick succession.

5.3 Acoustic Classification

An acoustic prediction procedure was carried out based on the two-category characterization: GPUs with TDP greater than 150 W received the label HN (>45 dBA), and those that were less than or equal to 150 W became M (<45 dBA). Figure 5: Estimated Acoustic Level Based on

TDP reveals that most of the GPUs landed in the moderate group, but above that, a significant number, roughly equal to V100, P100, etc., registered more than a high noise level.

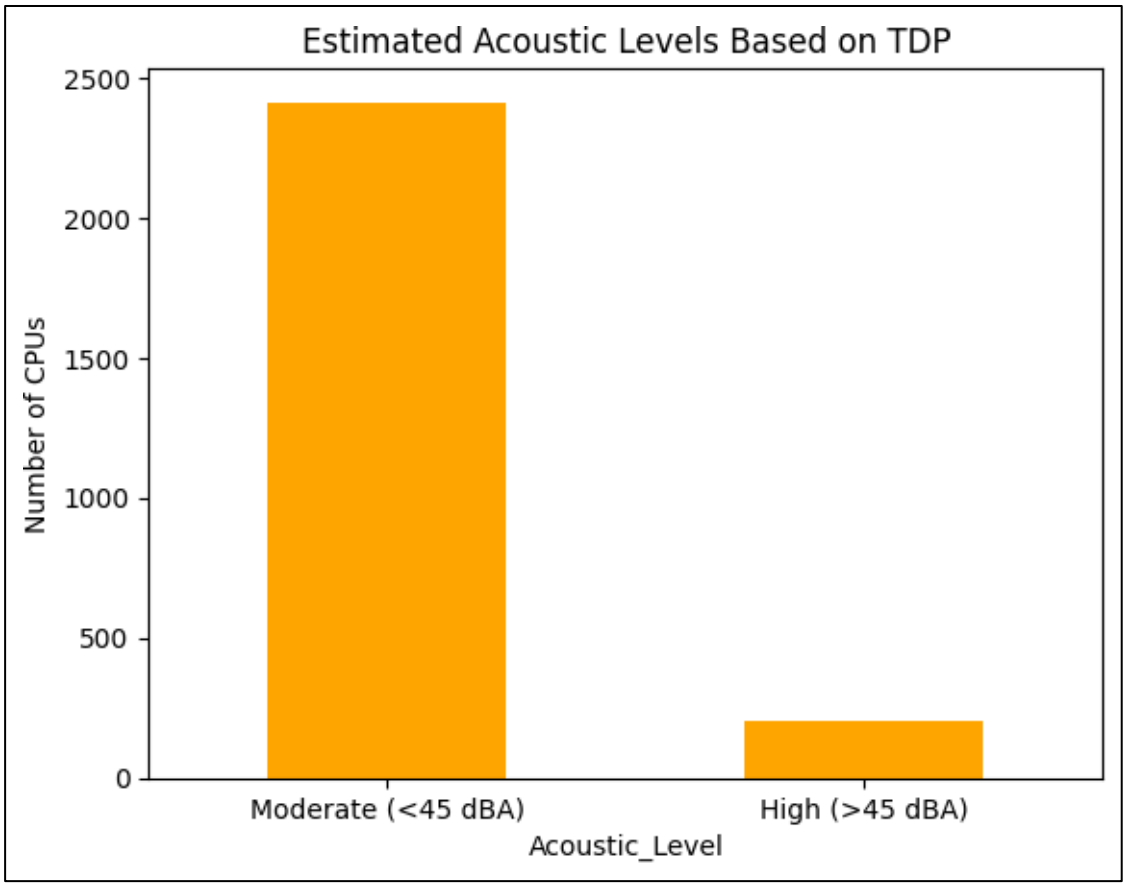

Figure 5. Estimated Acoustic Levels Based on TDP

This is vital in deploying workloads. The fact is that BERT was largely implemented on the NVIDIA T4 (TDP 70 W), so it was always linked to a minimal acoustic footprint. YOLOv5, however, using large-TDP GPUs, measured in the high-noise category, demonstrated that inference runs may cause rather unreasonably high acoustic stress with the ratio. Such results indicate that high-draw and bursty workloads such as YOLOv5 must be directed to racks with better acoustic isolation or redundancy cooling.

The various scatter and box plots investigated the overall architectural characteristics of GPUs. Figure 6: GPU Clock Speed vs Memory Size uses the graph to illustrate the poor relationship between the two variables, GPU frequency and onboard memory, using raw data from Kaggle. Outliers, long-clock speed-low-memory-footprint GPUs were also present, which signaled specialised or outdated designs. Most new AI GPUs operated at a range of 1200 1800 MHz with memory capacity varying between 16 and 32 GB.

5.4 Cross-Hardware Performance Patterns

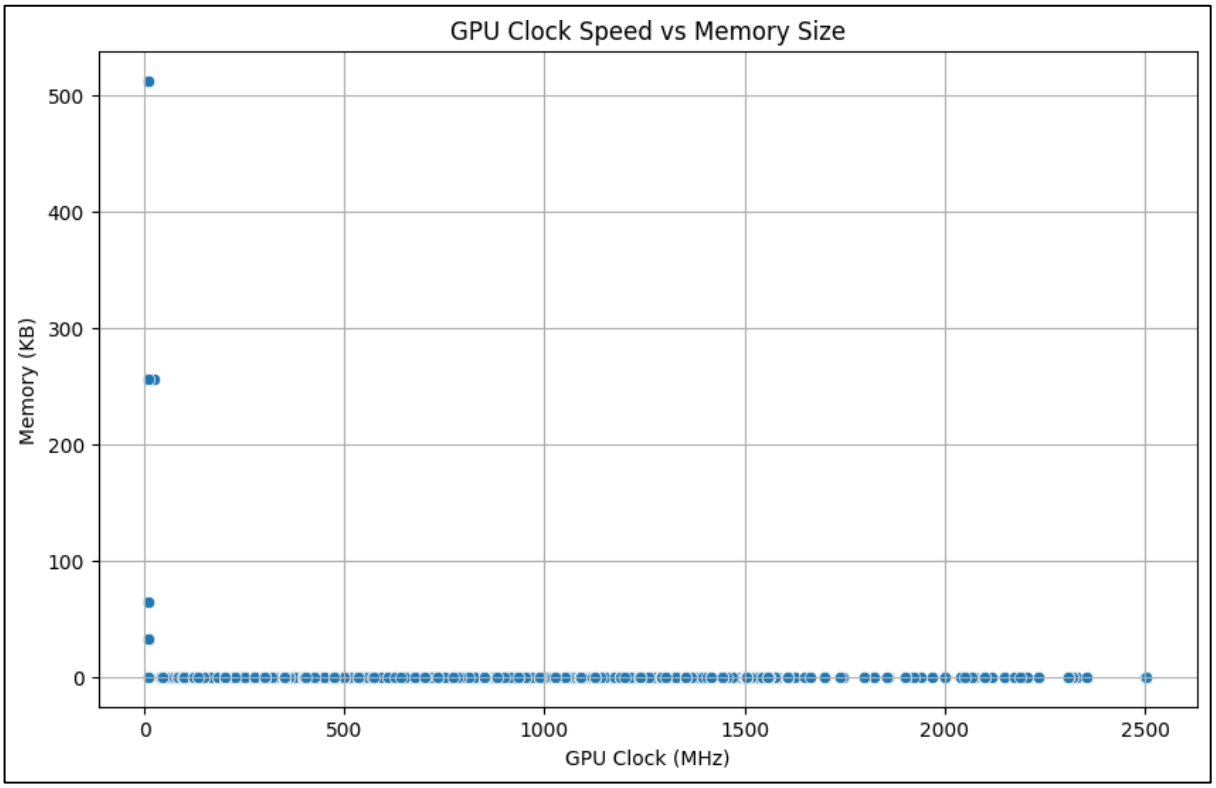

Figure 6. GPU Clock Speed vs Memory Size

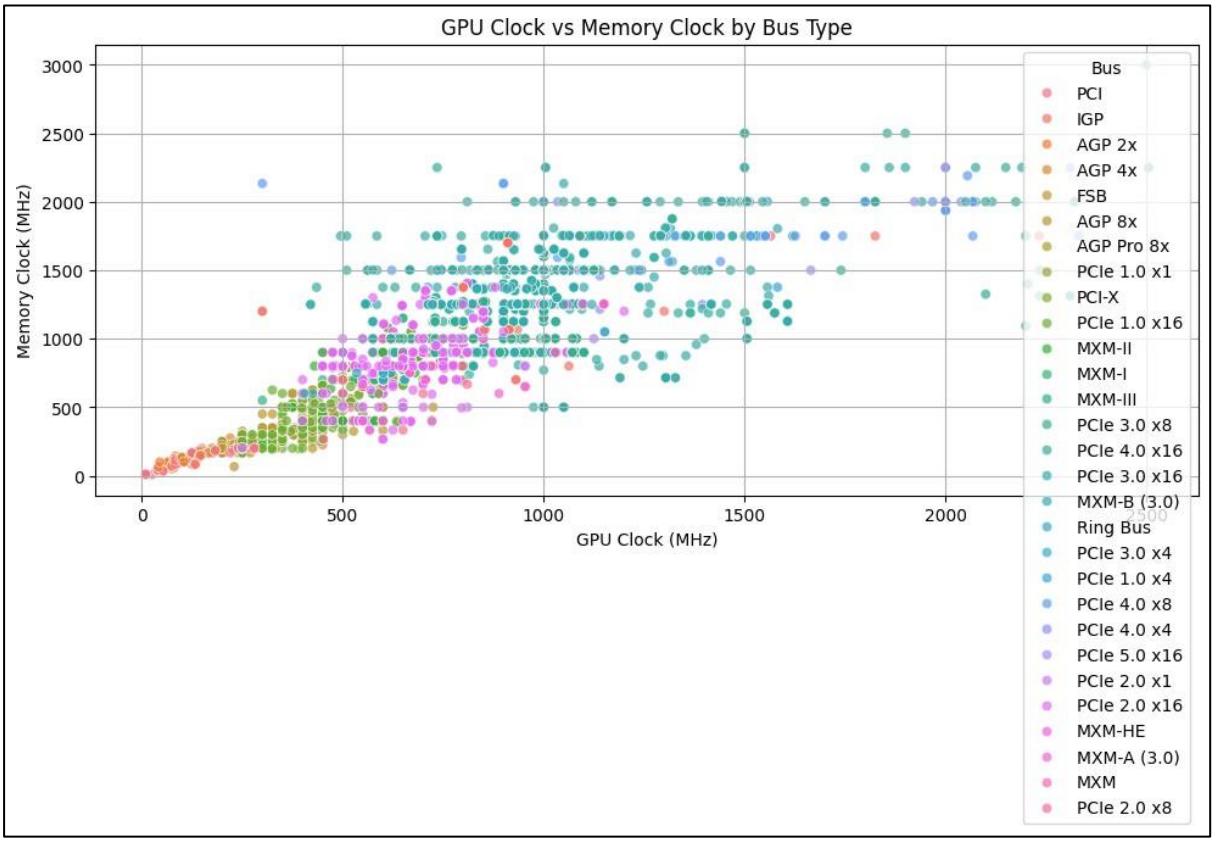

Figure 7. GPU Clock vs Memory Clock by Bus Type

Figure 7 indicates the bus interface's significance in memory performance. PCIe 4.0 and NVLink-based cards went further with memory clocks centered even higher than 1500 MHz, while older buses (AGP, PCI) plateaued at much lower frequencies. As the memory bandwidth not only influences the throughput but also results in thermal accumulation, this further supports the idea that the current generation of GPUs is simply better prepared to handle the thermal spikes of huge AI models.

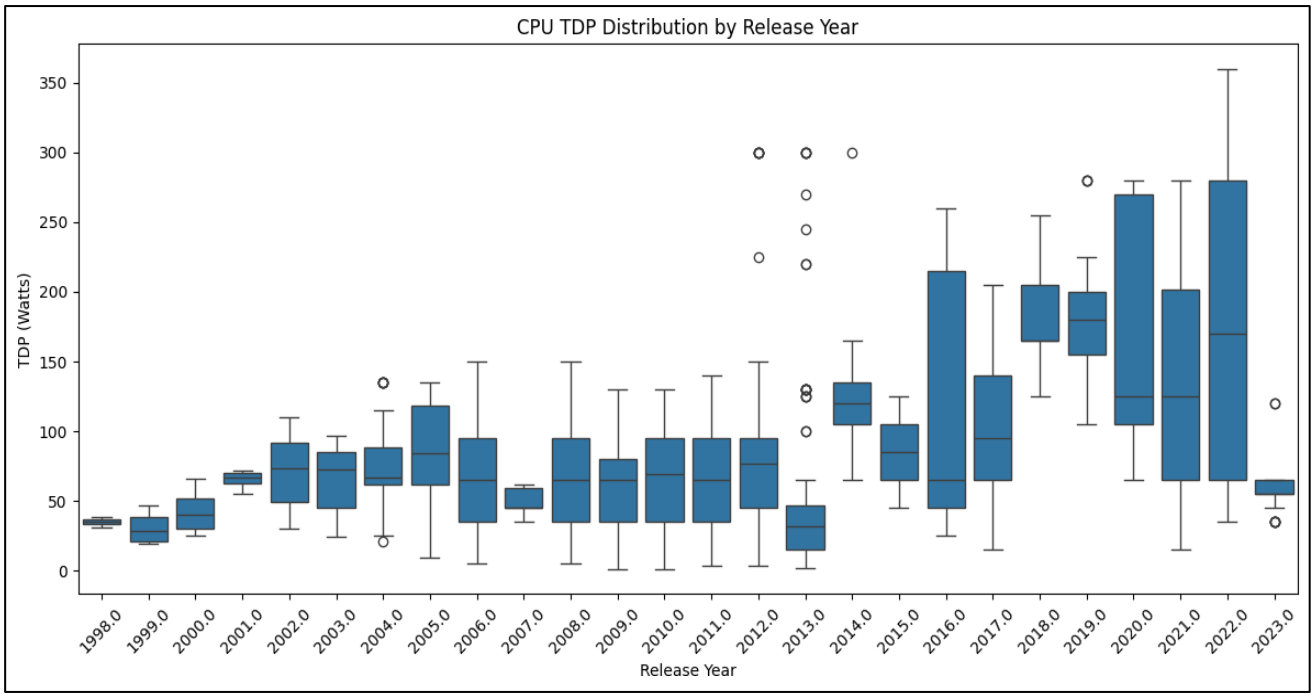

Figure 8. CPU TDP Distribution by Release Year

There has been a generational change to thermal design philosophy, as illustrated in Figure 8. The TDP figures scarcely ranged above 100 W between 1998 and 2010. Since 2015, there has been a sharp rise in the median TDP, with several 2020-2023 GPUs going above 250 W. This indicates the development of architecture and the increased need for AI-optimized silicon.

6. DISCUSSION

6.1 Workload-Specific Thermal Dynamics

Thermal patterns in BERT and YOLOv5 workloads exhibit opposite trends, which suggests significant details of workload-specific stresses on GPUs. BERT is a transformer-based NLP model with a long, stable pattern of use with large memory occupation and a steady pattern of GPU usage [28]. Previous studies showed that transformer-based NLP models have enduring energy demands during fine-tuning operations, particularly on big datasets such as the SQuAD v2 [17]. In our experiments, this resulted in a very high cumulative thermal load given a relatively moderate power draw, translating into the bar plot array comparing Thermal Load.

Conversely, a convolutional object detection model (YOLOv5) tuned to real-time usage had a hostile usage pattern with large power surges and occasional idle periods on the GPU [29]. These spikes in the behavior of our GPU utilization logs confirm previous findings that computer vision models are characterized by short electro-thermal spikes of immense demand, which overtax their steady-state cooling capabilities unless handled appropriately.

These behavioral differences are essential to infrastructure management regarding their thermal implications. BERT's thermal footprint is predictable, encouraging it to fit into a consistent cooling range, whereas a burst pattern common to YOLOv5 could cause spikes of overheating or fan speed if the burst has not been scheduled with such thermal padding. Such dynamics call for mitigating the significance of matching workload types with suitable equipment and cooling approaches, especially in mixed-use settings.

6.2 Acoustic Engineering Insights

The acoustic classification through proxy performed in this study gives the initial information about the trends of GPU-dependent noise in cloud and institutional data centers. We could classify workload based on indirect power-based heuristics by defining a conservative value of TDP of 150 W as a boundary between moderate and high levels of acoustics. The findings indicated that the BERT, usually performed on T4 GPUs, with

comparatively low TDP (70 W), was clearly inside the zone labeled Moderate Acoustic Load. YOLOv5 sessions, in turn, often use V100 or P100 GPUs with a TDP of 250~300 W, which fall in the “High Noise” category.

These measurements are confirmed by system-level measurements in which the authors point to high-performance workloads on GPUs greater than 200 W TDP persistently causing fan speeds in excess of 5000 RPM and resulting in acoustic emissions reaching more than 45 dBA [29]. Although it was impossible to measure absolute values of dBA because of the Colab limitation in this study, our classification gives a convenient rough estimate of the acoustic disturbances caused by fans.

As a practical implication, YOLOv5 and all other short-duration but high-performance scalers are best suited to be slotted on thermally decoupled GPU servers or racks with better noise suppression. In some university labs and cloud-native server clusters, inference-heavy GPU workloads are already co-located in acoustically shielded areas. Our proxy analysis affirms this design philosophy, which means that the inherent need to map the behavior of AI models to physical infrastructure attributes should be reinforced.

6.3 Sustainability in Shared Environments

Sustainability-wise, such trade-offs open prospects for deploying long-duration batch jobs (e.g., BERT) and bursty inference jobs (e.g., YOLOv5) in multi-tenant cloud GPU infrastructures. Long-term jobs can be thermally friendly because their maximum load is small. Still, their total energy demand is significant, well beyond 6000 Wmin in our experiments, which casts doubt on cooling expenses and total power demand.

In contrast, bursty jobs might only take a short time to complete but can cause rapid swings in system temperature, promoting more intense fan cycling and immediate energy spikes [30]. This corresponds to previous findings, which proved that uneven workloads within heterogeneous systems contribute to oscillating power consumption and disrupt dynamic cooling control [18].

According to our study, schedulers should concentrate on sustainability-related constraints to account for the peak and total thermal loads when scheduling workloads to common GPUs. For example, high-burst jobs can be coupled to thermal buffer periods, and long jobs can be alternated with low-load background tasks to more equally share the thermal stress. In addition, the metrics that we use to achieve our method are interpretable and could be implemented into cloud orchestration systems to embrace green AI frameworks.

6.4 Accuracy and Limitations of Proxy Approach

Interpretability is one of the suggested framework's strengths. Compared to more complicated simulations or closed-loop monitoring systems, our proxy-based approach only uses moderately available runtime logs and already published GPU specifications. This makes it compatible with reproducibility between platforms and scalable where telemetry APIs are absent or constrained [31]. The method helps monitor the lightweight infrastructure with minimum required equipment by translating the observed utilization, power consumption, and TDP values into normalized thermal and acoustic metrics.

However, it has some significant limitations. No direct measure of temperature or fan RPM values available, which restricts checks of proxy estimates against actual ground truth measurements. Although our thresholds and formulas were oriented on vendor documentation and available benchmarks, we are generalizing these values to all the conditions in the data center, and it is an approximate process. For example, selecting the rack airflow design, ambient room temperature, or type of material used in the heat sink can influence the thermal behavior independent of any aspect related to the workload.

Moreover, the binary definition of acoustic impact (only TDP thresholds are considered) can be too simplistic in practice. The fan speed curves are not linear and are usually controlled by vendor-specific firmware algorithms, which could differ between vendors and across different releases of the BIOS. Therefore, future work can be expected to consider integrating controlled laboratory measurements of GPU acoustics under benchmark workloads to refine proxy calibration.

6.5 Comparison with Known GPU Specs

The final phase of our work was dedicated to analyzing the correspondence between noticed model behavior and official GPU specifications. The bar plot shows that BERT workloads consumed more total thermal loads overall despite using lower-TDP hardware, owing to the

longer runtime. Conversely, YOLOv5 also produced a greater instantaneous load on greater- TDP GPUs, but completed faster, leading to less dissipated energy. The bubble plot also confirms this finding, as in the lower-right quadrant (low power, high duration), we have BERT, and in the upper-left quadrant (high power, low duration), we have YOLOv5. Another factor that supports the distinction in resource allocation patterns between the two models is the size of the bubbles, which would equate to the usage of GPU memory.

These correlations are consistent with observed GPU performance tables published by NVIDIA and performance evaluations of Anzt et al. (2021), indicating that V100 and P100 GPUs are optimized to run a high-throughput burst workload. In contrast, the T4s are better suited to a sustained, latency-tolerant workload. Such differences in design are empirically confirmed with our results in workload testing in the real world of Colab GPUs.

Therefore, this work has effectively shown that even open-sourced GPU specifications can be used to forecast the behavior of AI models via straightforward, albeit effective, proxy techniques. The techniques help gain knowledge on the suitability of the workloads, thermal profiling, and acoustic prediction in restricted conditions where access to hardware telemetry is not available directly.

7. CONCLUSION

This paper presented and confirmed a proxy model for assessing GPU clouds' thermal and acoustic efficiency in training AI tasks. Even in platforms such as Google Colab Pro, where the direct telemetry of such variables as GPU temperature or fan speed is difficult to acquire, the framework proved capable of high-quality derivation of interpretable and actionable insights based on GPU usage, power consumption, memory use, and known characteristics (such as TDP).

We identified the opposite tendencies of GPU behavior based on similar benchmarking studies of two sample AI workloads, BERT (NLP) and YOLOv5 (computer vision). BERT showed an inferred longer duration, sustained feature use with a high cumulative thermal loading but low peak power, whereas YOLOv5 had blistering, multiplicative features that used the GPU with a high-power draw but lower cumulative thermal load. The range of these distinctions was measured based on thermal loading estimation and acoustic classification with the help of TDP- based averages and standardized comparisons such as TDP-per-MHz.

The results were backed up by visualizations using bubble plots, boxplots, scatterplots, and bar charts, providing evidence of workload-specific GPU strain and efficiency profiles. The acoustic proxy showed that workloads in YOLOv5 regularly caused GPUs to enter high-noise states, but with BERT, execution on low-TDP GPUs such as the T4 kept it in moderate sounds. Hence, the results highlight that a lightweight, reproducible thermal and acoustic evaluation in the limited cloud areas is viable. This allows more efficient scheduling of workload, energy- sensitive computation, and infrastructure design without the necessity of invasive sensors or access to proprietary telemetry. The fact that the framework concurred with the publicly accessible datasets guarantees its broad applicability in research programs in universities and industries relevant to sustainable implementations of AI systems.

8. RECOMMENDATIONS AND FUTURE WORK

The findings of the current work present a few significant suggestions to system designers, data scientists, and infrastructure engineers involved in the AI implementation:

Thermal-Aware Scheduling: Workloads must also be scheduled based on the profile of thermal load rather than the GPU availability. The steady heat-producing BERT-like models are better applied in conditions of constant cooling capacity. Burst compatible models, such as YOLOv5 and others, can benefit thermally insulated nodes by avoiding overheating or fan surge.

Acoustic Zoning in Data Centers: Depending on the correlation between TDP and noise levels identified, high-TDP GPUs must be assigned to acoustically isolated racks, particularly in academic labs or server rooms located at points of consumption where background noise is

essential. Stress benchmarking on acoustic output should be considered in future AI-capable hardware.

Integrated Proxy Monitoring Tools: The methodology can be automated to be used in dashboarding schemes on GPU usage, which calculate and present utilization trends and approximate thermal/acoustic load in real time. It allows active policies to be actively tuned to cool

and gives real-time alerting without any changes to the hardware.

Reproducibility via Public Datasets: This study's reproducibility is proven by the fact that the Kaggle GPU/CPU datasets and the runtime logs (acquired via nvidia-smi) can be integrated into the qualitative analysis. These proxies should be standardized in developing AI benchmarking tools and academic pipelines.

Future research will involve identifying the limitations of the present-day approach. Firstly, value addition to the telemetry, such as Google Colab Pro+ or AWS EC2 telemetry (in case API access is available), would allow correlating proxy values with real temperature/fan signals, resulting in higher calibration accuracy. Second, using physical acoustic sensors in an experiment should lead to a ground truth measure to optimize the dBA classification. Third, the framework's application can be generalized to multi-GPU workloads and hybrid CPU-GPU systems (TPUs).

Finally, combining such thermal-acoustic profiling with a model of energy costs and environmental quantities (e.g., carbon intensity of power consumption) would aid in green AI efforts. This would enable developers to make intelligent choices regarding performance, accuracy, and the sustainability of their AI compute workflow.

ACKNOWLEDGMENTS

The authors would like to acknowledge the help of Google Colab Pro, which allowed us to run GPU-based workloads in a real-world cloud environment with constraints on most dimensions. And we are also grateful to the Kaggle open hardware datasets contributors who not only gathered and updated free, publicly available, comprehensive GPU and CPU specifications, but also because of which the reproducibility of this work was made possible. Their work made a large proxy-based thermal and acoustic study possible without physical access to hardware. Furthermore, we would like to embrace the general open-source community (Matplotlib, Seaborn, and pandas developers), whose visualization and analysis tools played a significant role in understanding and presenting the findings.

References

1. Artificial Intelligence, Machine Learning, and Deep Learning for Advanced Business Strategies: A Review | Partners Universal International Innovation Journal [Internet]. [cited 2025 Jun 28]. Available from: https://puiij.com/index.php/research/article/view/143
2. High Performance Computing for Understanding Natural Language: Computer Science & IT Book Chapter | IGI Global Scientific Publishing [Internet]. [cited 2025 Jun 28]. Available from: https://www.igi-global.com/chapter/high-performance-computing-for-understanding- natural-language/273400
3. Mustafa F, Gilbert A. Scalable Data Architectures for Generative AI: A Comparison of AWS and Google Cloud Solutions [Internet]. Unpublished; 2024 [cited 2025 Jun 28]. Available from: https://rgdoi.net/10.13140/RG.2.2.26378.07364
4. Rafsanjani H, Marwazi A, Sitompul D. Thermal Management and Power Optimization in Modern CPU and GPU Architectures.
5. Katal A, Dahiya S, Choudhury T. Energy efficiency in cloud computing data center: a survey on hardware technologies. Cluster Comput. 2022 Feb 1;25(1):675–705.
6. CPUs Versus GPUs | SpringerLink [Internet]. [cited 2025 Jun 28]. Available from: https://link.springer.com/chapter/10.1007/978-981-97-9251-1_9
7. Thermal intelligence: exploring AI's role in optimizing thermal systems – a review | Interactions [Internet]. [cited 2025 Jun 28]. Available from: https://link.springer.com/article/10.1007/s10751-024-02122-6
8. A Survey of Cloud-Based GPU Threats and Their Impact on AI, HPC, and Cloud Computing.
9. A systematic review of scheduling approaches on multi-tenancy cloud platforms - ScienceDirect [Internet]. [cited 2025 Jun 28]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0950584920302214
10. Ajayi R. Integrating IoT and cloud computing for continuous process optimization in real time systems. Int J Res Publ Rev. 2025 Jan;6(1):2540–58

11. Real-Time Thermal Map Characterization and Analysis for Commercial GPUs with AI Workloads | IEEE Conference Publication | IEEE Xplore [Internet]. [cited 2025 Jun 28]. Available from: https://ieeexplore.ieee.org/abstract/document/11014443
12. An Overview of Thermal and Mechanical Design, Control, and Testing of the World's Most Powerful and Fastest Supercomputer | J. Electron. Packag. | ASME Digital Collection [Internet]. [cited 2025 Jun 28]. Available from: https://asmedigitalcollection.asme.org/electronicpackaging/article- abstract/143/1/011005/1082291/An-Overview-of-Thermal-and-Mechanical-Design
13. Design, Operation and Maintenance of Direct and Indirect Evaporative Cooling Systems in Data Center Thermal Management - ProQuest [Internet]. [cited 2025 Jun 28]. Available from: https://www.proquest.com/openview/eed632faea3362b05f921b8213a5b9ab/1?pq-origsite=gscholar&cbl=18750&diss=y
14. Harnessing Machine Learning in Dynamic Thermal Management in Embedded CPU-GPU Platforms | ACM Transactions on Design Automation of Electronic Systems [Internet]. [cited 2025 Jun 28]. Available from: https://dl.acm.org/doi/full/10.1145/3708890
15. Bagai R. Comparative Analysis of AWS Model Deployment Services. IJCTT. 2024 May 30;72(5):102–10.
16. Zhang J, Zhang W, Xu J. Bandwidth-efficient multi-task AI inference with dynamic task importance for the Internet of Things in edge computing. Computer Networks. 2022 Oct 24;216:109262.
17. Ansar W, Goswami S, Chakrabarti A. A Survey on Transformers in NLP with Focus on Efficiency [Internet]. arXiv; 2024 [cited 2025 Jun 28]. Available from: http://arxiv.org/abs/2406.16893
18. A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization - ScienceDirect [Internet]. [cited 2025 Jun 28]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1383762121001739
19. Knebel FP. Designing and implementing digital twins with cloud and edge computing: challenges and opportunities. Projetando e implementando Gêmeos Digitais com computação em nuvem e de borda: desafios e oportunidades [Internet]. 2024 [cited 2025 Jun 28]; Available from: https://lume.ufrgs.br/handle/10183/276593
20. Advances in Numerical Modeling for Heat Transfer and Thermal Management: A Review of Computational Approaches and Environmental Impacts [Internet]. [cited 2025 Jun 28]. Available from: https://www.mdpi.com/1996-1073/18/5/1302
21. SOUND AND NOISE: MEASUREMENT AND DESIGN GUIDANCE – HANDBOOK OF HUMAN FACTORS AND ERGONOMICS - Wiley Online Library [Internet]. [cited 2025 Jun28].Availablefrom: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119636113.ch18
22. GPU Devices for Safety-Critical Systems: A Survey | ACM Computing Surveys [Internet]. [cited 2025 Jun 28]. Available from: https://dl.acm.org/doi/abs/10.1145/3549526
23. Chowdhury U, Rodriguez J, Tradat M, Soud Q, Wallace S, O"Brien D, et al. Acoustics Analysis of Air and Hybrid Cooled Data Center. In: 2024 23rd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm) [Internet]. 2024 [cited 2025 Jun 28]. p. 1–11. Available from: https://ieeexplore.ieee.org/abstract/document/10709368
24. Distributed Cloud Computing Infrastructure Management [Internet]. [cited 2025 Jun 28]. Available from: https://www.scirp.org/journal/paperinformation?paperid=143462
25. TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms | Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 [Internet]. [cited 2025 Jun 28]. Available from: https://dl.acm.org/doi/abs/10.1145/3676641.3716025
26. Enhancing Reservoir Modeling and Simulation Through Artificial Intelligence and Machine Learning: A Smart Proxy Modeling Approach - ProQuest [Internet]. [cited 2025 Jun 28]. Availablefrom: https://www.proquest.com/openview/f3886ad9ad2f556a032981db194fcfa4/1?pq-

origsite=gscholar&cbl=18750&diss=y

27. Ganesh P, Chen Y, Lou X, Khan MA, Yang Y, Sajjad H, et al. Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. Transactions of the Association for Computational Linguistics. 2021 Sep 21;9:1061–80

28. Synchronizing Object Detection: Applications, Advancements and Existing Challenges | IEEE Journals & Magazine | IEEE Xplore [Internet]. [cited 2025 Jun 28]. Available from: https://ieeexplore.ieee.org/abstract/document/10499817

29. Aalborg University, Liu J. Automatic Analysis of People in Thermal Imagery [Internet] [Ph.d]. Aalborg University; 2022 [cited 2025 Jun 28]. Available from: https://vbn.aau.dk/en/publications/automatic-analysis-of-people-in-thermal-imagery

30. AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications | IEEE Conference Publication | IEEE Xplore [Internet]. [cited 2025 Jun 28]. Available from: https://ieeexplore.ieee.org/abstract/document/9773216/

31. Tapis: An API Platform for Reproducible, Distributed Computational Research | SpringerLink [Internet]. [cited 2025 Jun 28]. Available from: https://link.springer.com/chapter/10.1007/978-3-030-73100-7_61