Real-time Thermal Map Characterization and Analysis for Commercial GPUs with AI Workloads

Jincong Lu[1], Sachin Sachdeva[1], Yuxuan Lin[2] and Sheldon X.-D. Tan[1]
[1]Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521 USA
[2]The Overlake School, Redmond, WA 98053 USA
jincong.lu@email.ucr.edu, ssach008@ucr.edu, dyxlin@outlook.com, stan@ece.ucr.edu

2025 26th International Symposium on Quality Electronic Design (ISQED) | 979-8-3315-0942-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISQED65160.2025.11014443

***Abstract*—AI accelerators such as GPUs are critical for emerging generative AI applications. However, due to the large power consumption of those AI chips, accurate characterization and modeling of spatial thermal map behaviors over practical AI workloads are important for designing better computing resource management and cooling solutions. However, it is challenging to obtain accurate thermal maps of those chips under practical AI workloads due to demanding cooling requirements, which prevents the direct thermal measurement of chips. This paper investigates the spatial thermal maps of the commercial NVIDIA GeForce RTX 4060 GPU chip for the first time. First, we managed to obtain the real-time full-chip thermal maps of the NVIDIA GPU with timed workloads supported by a bottom-cooling thermal IR imaging system. Second, we observed that the thermal hot spots are actually located in the I/O and peripheral areas of the chips, which indicates the intensive data movements in the GPU. The computing areas of GPU, on the other hand, show smoother thermal distributions with a few hotspots, which is quite different than commercial multi-core CPUs. Furthermore, based on real-time information from the NVIDIA System Management Interface (NVIDIA-SMI), we developed a deep neural networks based full-chip thermal map estimation method, called *GPUThermalMap*, for GPUs for the first time, which can be instrumental for dynamic thermal, power, and reliability management. Numerical results highlight the effectiveness of *GPUThermalMap* in achieving highly accurate thermal map predictions, boasting an RMSE of only $0.19^{\circ}\mathrm{C}$ or 0.6% of the full-scale error. It also outperforms the GAN-based method, *ThermalGAN*, by $2.09\mathrm{x}$ in terms of accuracy on average. Besides, the proposed model offers real-time estimation with a rapid speed of $22\mathrm{ms}$ on the target chip.**

I. Introduction

With the ongoing trend of rapid integration and technology scaling, contemporary high-performance multi/many core processors are facing more pronounced thermal limitations than ever before. Research has shown that higher temperatures exponentially degrade the reliability of semiconductor chips [1], making it a significant concern in the industry. This situation becomes even worse for AI chips like commercial GPUs from NVIDIA, where the GPU power consumption can be more than 1000W. For instance, the H100/H200 GPUs can take 700W TDP (thermal design power), while the latest Blackwell GPU B200 has 1200W TDP [2].

To address this trend, runtime power and thermal control schemes are being implemented in most, if not all, new generations of processors. These control schemes play a crucial role in modern processors [3], [4]. However, for these control schemes to be effective, they require accurate real-time thermal information, ideally a spatial thermal map of the entire chip area [5], [6]. On-chip temperature sensors alone cannot provide comprehensive chip-wide temperature information due to their limited number, primarily constrained by the area and power overheads [7].

The work is supported in part by NSF grant under No.CCF-2007135, and in part by NSF grant under No. CCF-2113928.

Several existing methods rely on on-chip temperature sensors. However, the availability of physical sensors is typically limited, and their placement may not accurately capture the true hotspots on the chip. As a result, temperature regulation decisions based solely on these sensors can be misleading [8]. Consequently, a more popular approach is to augment the data from the few on-chip sensors with estimated temperatures of prominent hotspots using thermal models based on estimated power traces [9]. These methods provide higher spatial resolution by enabling real-time monitoring of temperatures for all hotspots on the chip [10]–[13]. However, existing thermal modeling methods still possess certain limitations, such as the difficulty of obtaining functional unit powers and limited accuracy.

Recently machine learning based full-chip thermal maps of commercial multi-core processors and hot spot detection have been proposed by leveraging the universal modeling capability of deep neural networks [14]–[17]. Those methods leverage online real-time utilization and monitoring information such as core frequency, voltage, and various utilization or performance metrics, which are supported by most commercial processors [18]. Software tools such as Intel's Performance Counter Monitor (PCM) [19] and AMD's uProf [20] provide the means to profile these metrics. These methods demonstrate the feasibility of fast online thermal map estimation for the first time. Methods include the Long Short Term Memory (LSTM) based method [14], [15], the Generative Adversarial Network (GAN) based method [16] and recent transformer based models for AMD multi-core CPUs based on uProf utilization metrics [17]. But all those methods have been developed mainly for commercial multi-core CPUs so far.

On the other hand, with the exponential growth of generative AI like ChatGPT [21], thermal and power modeling for commercial AI chips and hardware have become more important due to their massive power consumption and high cost for cooling solutions. However, full-chip thermal modeling for commercial GPUs like GPUs from NVIDIA under AI workloads has yet to be investigated due to several difficulties. One of major challenges is that those chips need heat sinks or active cooling to operate normally. As a result, it is very difficult to obtain real-time thermal images with those heat sinks mounted on the chip from the thermal IR system. Second, there is not study to build machine learning model to map the GPU status and utilization information into the full-chip thermal map, which needs to be investigated.

In this work, we aim to address the aforementioned issues and develop a novel way to obtain the thermal images of the commercial NVIDIA GeForce RTX 4060 GPU for the

first time. Based on these measured thermal images and on-chip real-time utilization and monitoring information from the GPU, we successfully developed a transient thermal map estimation method using deep neural networks for the first time. The key contributions of this study are as follows.

- First, we managed to obtain the full-chip thermal maps of NVIDIA GeForce RTX 4060 GPU with machine learning (AI) workloads without heat sink for the first time using the bottom-cooling thermal IR imaging system. To avoid the overheating of the chip, we use timed workload execution to capture timed transient thermal map changes of GPUs, which provides the sufficient data for training of the thermal models. A total of $16,000$ pairs of performance metrics data and thermal maps were collected, with 80% of the data used for training purposes.
- We observed for the first time that for the NVIDIA GeForce RTX 4060 GPU chip, the real hot spots actually are the I/O and peripheral areas of chip, instead of computing areas. This indicates the intensive data movements in and out of GPUs. The computing areas of the chip however shows more smooth thermal distribution than the multi-core commercial processor CPUs as report in the existing literature. Such thermal behavior of GPUs may indicate very efficient small task distributions among GPU stream processors, which is the intrinsic nature of GPU parallel computing. We believe this observation is typical for other NVIDIA GPUs given their similar architectures.
- We develop a deep neural network (DNN) based full-chip thermal map estimation method called *GPUThermalMap* using the transformer architecture for the NVIDIA GeForce RTX 4060 GPU for computing areas for the first time. The new method is trained with practical AI workloads based on real-time information from the NVIDIA System Management Interface (NVIDIA-SMI), which consists of monitoring and utilization information for the GPU.
- Numerical results demonstrate the high accuracy of the thermal map predictions, with a root-mean-square error of only $0.19^{\circ}\mathrm{C}$ or 0.6% of the full-scale error. Also, our *GPUThermalMap* outperforms the GAN-based thermal map estimation method, *ThermGAN*, by $2.09\mathrm{x}$ in terms of accuracy on average, again proving that the transformer-based method is superior to the GAN-based method in the scenario of heat map prediction through time series input. Furthermore, the proposed approach can be deployed on the target chip with a fast inference speed of $22\mathrm{ms}$, making it suitable for real-time estimation.

This article is organized as follows. Section II provides a review of relevant work. Section III outlines the thermal modeling framework and IR thermography setup employed in this study. Section IV discuss some observations about GPU heat map patterns. Section V explains the process of collecting and preparing the training thermal data, as well as the selection of performance metrics features for the proposed method. Section VI describes the architecture of the proposed transformer-based model for thermal map estimation. Section VII presents the experimental results and provides comparisons. Section VIII concludes the article.

II. Related work

Two general strategies are commonly employed to estimate on-chip temperature maps. The first approach involves estimating the full-chip heatmaps using physics-based thermal models and power-related information [12], [13]. These 'bottom-up' numerical methods, such as HotSpot [11], which are based on simplified finite difference methods, finite element methods [22], and equivalent thermal RC networks [23], have been widely used. Recently, top-down behavioral thermal models based on the matrix pencil method [24] and subspace identification method [25], [26] have also been proposed. However, these full-chip thermal analysis methods are computationally expensive and unsuitable for online applications [27].

The second strategy involves employing interpolation-based approaches to estimate full-chip heatmaps from embedded sensor readings [9], [28]. The accuracy of such interpolation-based methods is heavily influenced by the number and placement of sensors. To tackle this challenge, various smart sensor placement algorithms have been proposed to optimize sensor deployment within a specified budget of embedded temperature sensors [28]–[33]. Some studies have utilized Fourier analysis techniques to reconstruct thermal maps [28]. However, the accuracy of such methods is constrained by the non-band-limited nature of temperature signals and the approximations necessitated by non-uniform sensor placement, which is common in heterogeneous multi-core processors. Other approaches aim to minimize the number of thermal sensors and reconstruct thermal maps by interpolating hard sensor information in both frequency and DC domains [29], [30]. Furthermore, the application of eigendecomposition of the interpolation matrix has enhanced sensor placement strategies, leading to nearly optimal sensor numbers and placements [31].

Zhang et al. proposed a statistical method for estimating both power and thermal maps [32], [34]. This method leverages the correlations of power dissipation among different modules of a chip to recover the power map from sensor readings, followed by temperature estimation. However, this estimation relies on power correlation information. In a related approach, Ziabari et al. introduced the power blurring method for rapid 2-D temperature map computation [35]. This method employs the Green's function approach, where the temperature response to unit power impulses is initially computed using finite element thermal analysis. However, the practicality of this method is limited by the availability of accurate thermal models in all cases.

However, the methods mentioned above either require hardware modifications during the design phase, such as inserting or relocating sensors, or they rely on detailed knowledge of the chip's floorplan, power source correlations among functional units, and technology-specific constants that are typically not disclosed by the chip manufacturer. As a result, achieving real-time estimation of the spatial temperature distribution across the entire chip area solely through a post-silicon approach, i.e., estimating the full-chip spatial heatmap $T(x, y)_t$ at time t, remains a significant challenge for existing commercial microprocessors.

On the other hand, machine learning-based approaches offer new perspectives on real-time full-chip thermal map estimation methods for commercial multi-core CPUs. These methods utilize real-time on-chip utilization and monitoring information as inputs for generating thermal and power maps. Such ideas have been explored recently by DNN-based approaches leveraging real-time performance metrics [14]–[17]. Sadiqbatcha et al. first proposed an LSTM-based approach called Realmaps, which utilizes Intel PCM metrics for estimating full-chip thermal maps in commercial off-the-shelf multi-core processors [14], [15]. This approach has shown promising results in terms of accuracy and speed of inference for real-time applications. On top of this, an improvement was made by employing an image-friendly CNN model based on the GAN architecture. This approach, known as ThermGAN, demonstrates better results than the LSTM-based methods in terms of accuracy [16]. Recently, a transformer-based DNN method has been proposed to estimate the full-chip thermal map of AMD multicore chips using uProf utilization metrics [17]. This method exhibits superiority over the GAN-based and LSTM-based methods due to its powerful modeling capability for time series data via the attention mechanism. Furthermore, the GAN model was extended to model Google TPU chips for the first time [36]. This study shows for the first time that, for TPUs that do not have many online utilization and monitoring metrics, thermal models can still be built successfully using only the hyper-parameters of DNN models running on TPUs.

However, none of the existing methods have investigated the transient thermal modeling of commercial GPUs, especially with practical AI workloads. There are several challenges, as mentioned earlier. However, given the exponential growth of generative AI applications and their heavy dependency on GPU computing for model training and inference, it becomes increasingly important to fully understand the thermal behaviors of commercial GPUs. This understanding is crucial for designing better online computing resource management methods and improved cooling solutions.

III. Thermal Map Estimation Framework

In this section, we will provide a concise overview of the proposed approach, accompanied by a description of the thermal camera setup employed to gather essential data from the NVIDIA GeForce RTX4060 GPU chip during its operational load.

A. Estimation flow overview

The proposed approach comprises three main components. Firstly, we gather data by logging the GPU's metrics during workload execution. Concurrently, we capture comprehensive thermal map measurements of the entire chip using a thermal infrared (IR) imaging system. Subsequently, we utilize the collected data to train a transformer-based online thermal prediction model.

The model training process involves the input and output. The input includes recorded GPU metrics, which act as indicators for generating predictions. The output consists of the offline measured thermal maps, serving as the model's training targets. When the model converges, it can be employed for online GPU thermal prediction.

Fig. 1 depicts the framework of the proposed approach. In the subsequent section, we will discuss the data acquisition process, covering each step outlined in the figure. Furthermore, in Section VI, we will delve deeper into the machine learning model.

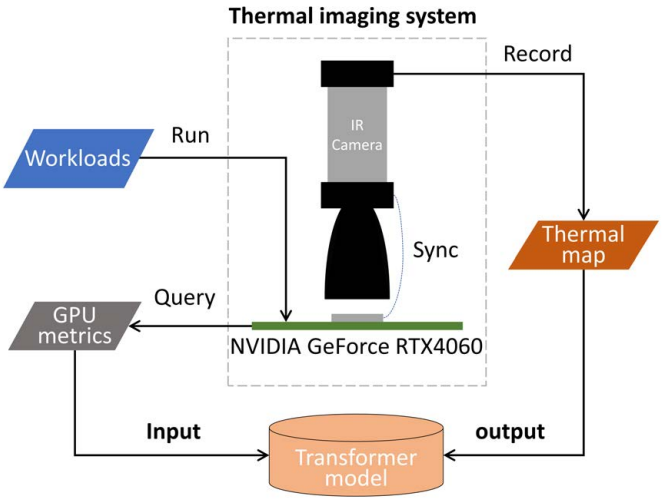


Fig. 1. Framework and data acquisition flow

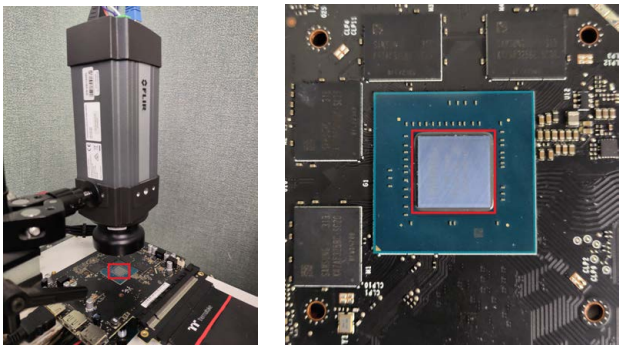
B. Thermal IR imaging system

Accurate measurement of GPU surface temperature maps is a prerequisite for the success of machine learning methods. To achieve this, an advanced infrared thermal imaging system, as shown in Fig. 2, is deployed for measurement. However, the top surface of the core module is usually obscured by the heat sink. To obtain the thermal map, a bottom-side liquid cooling system [8] is adopted instead of the traditional top-side heat dissipation method. As heat dissipation from the bottom side requires passing through the PCB board, the efficiency is significantly reduced compared to top-side heat dissipation. Therefore, a thermoelectric (Peltier) device is installed on the PCB beneath the processor module to improve efficiency. As a result, the front side of the processor is fully exposed to the infrared camera without any possible interference from intervening materials.

The model of the IR camera is FLIR A325sc. It can capture thermal images with a maximum resolution of 240×320 pixels (px) at a maximum frequency of 60Hz. The factory-calibrated IR sensor ensures accuracy within a temperature range of $-20°\mathrm{C}$ to $120°\mathrm{C}$ and resolves the IR spectral range of $7.5\mu m$ to $13\mu m$.

IV. GPU Thermal Patterns and Unique Features

In this section, we present observations regarding the heat map patterns and some unique features of the NVIDIA GeForce RTX 4060 GPU. We chose this GPU because it is a relatively low-power NVIDIA GPU compared to high-performance GPUs like the H100 or H200. Consequently, we can run the GPU without heat sinks to obtain transient thermal maps using a back cooling setup in our IR thermal imaging



(a) (b)

Fig. 2. (a) Thermal Imaging system setup (b) GPU chip under-test, NVIDIA GeForce RTX4060. Core module is shown in the red box.

system. For high performance GPUs, more aggressive back cooling techniques will be investigated in the future.

Fig. 3. The thermal image example of NVIDIA RTX4060

Fig. 3 shows a thermal map example of the NVIDIA GeForce RTX 4060 GPU chip. One notable feature is the presence of the NVIDIA logo and various chip design and manufacturing information directly printed on the chip, which are clearly visible thermally. These printed details alter the thickness and reflectivity of the surface material, leading to inaccuracies in the temperature readings by the thermal camera for these areas (as they are no longer covered by aluminum). Therefore, in this study, the areas covered by text and their corresponding pixels will be treated as missing information or blocked-out areas. We may explore thermal image recovery techniques in future research.

The second unique feature of GPU thermal maps is the observation that the hot spots are mainly located at the edge of the chip, as depicted in Fig. 3, where many bright areas are seen on the top, right, and some on the bottom sides of the chip. These areas primarily correspond to the input and output interfaces of the chip. Specifically, we find that the interfaces between the processors and the GPU memory module are located above and to the right, while the PCI-4 interface to the motherboard is situated at the bottom right. Additionally, the interface to the display is on the bottom left, while the main computing stream processors are in the middle area. This observation indicates that the movement of input and output data plays a crucial role in the GPU's heat production process and often dominates the spatial distribution of the temperature. It is noteworthy that this finding is somewhat surprising, as GPUs are typically considered to be computing-dominant hardware, rather than I/O or data movement dominant devices.

On the other hand, the processor or main computing regions located in the middle exhibit relatively smooth thermal characteristics. They typically do not experience continuous migration of hotspots during computations, which is quite surprising given that hot spots are not sensitive to the workloads for GPUs. The temperature maps locally reflect the shapes of internal components while overall conforming to a skewed temperature gradient. This is in stark contrast to the temperature patterns observed in multi-core CPUs [14]–[17], where significant temperature differences between active and inactive cores are commonly observed, along with hotspot migrations under different workloads. This observation may indicate very efficient small task or thread distributions among GPU stream processors, reflecting the intrinsic nature of GPU parallel computing and task allocation and dynamic management.

We note that although we only analyzed one NVIDIA GPU, We believe the other NVIDIA GPUs will have similar observations given their similar architectures.

V. Data preparation and GPU metrics selection

Since we aim to predict thermal maps from the real-time GPU metrics through supervised learning, having a high-quality dataset is crucial. To collect experimental data, we need to run various AI workloads as environments, then systematically record GPU metrics and temperature maps at regular intervals. When processing the data, by stacking the GPU metrics from several previous timestamps at each timestamp, we obtain a historical performance metric time series. This time series, along with the thermal map measured at each timestamp, forms a single data point. As the system runtime increases, we can accumulate a comprehensive dataset conducive to analysis. This section will provide a detailed exposition of the data collection methodology.

A. AI workloads used for training

In practical applications, the target chip may need to perform highly specialized tasks, and experiments can be designed to obtain the required data. In this study, we trained several well-known image recognition models, including MobileNet and InceptionResNet, configured with different architectural hyperparameters to encompass various scenarios. A total of 100 AI workloads were tested, which we consider sufficient for our training purpose. We believe the selected AI workloads are representative of many practical AI tasks. Future work will include testing additional workloads, such as transformer-based language models.

As described in Section III-B, we employed bottom-side liquid cooling to obtain the thermal images of the chip's front side. However, even entry-level GPUs have high power consumption, making it difficult to achieve completely effective cooling through the PCB board. Continuous operation of neural network workloads can cause the GPU's heat generation to exceed its cooling capacity, eventually triggering overheat protection. To address this, we limit the runtime of each workload cycle to around around ten to twenty seconds, then stop the workload to allow the GPU to cool down adequately before running the next cycle.

TABLE I
SELECTED GPU METRICS (NVIDIA GEFORCE RTX 4060)

Temperature		
core GPU temperature		
GPU Operation Mode		
performance state	fan speed	current GOM
pending GOM		
Memory		
installed memory	reserved memory	allocated memory
free memory	protected memory	allocated protected memory
free protected memory	compute mode	compute capability
Utilization		
gpu utilization	memory utilization	encoder utilization
decoder utilization	jpeg utilization	ofa utilization
Frequency		
graphics clock	SM clock	memory clock
video encoder/decoder clock		
Encoder Stats		
session count	average fps	average latency
Mode		
current ECC mode	pending current ECC mode	current MIG mode
pending MIG mode	current GSP firmware mode	
Error Counter (corrected & uncorrected)		
device memory	DRAM	register file memory
L1 cache	L2 cache	texture memory
CBU	SRAMs	entire chip
Retired Pages		
single bit ECC	double bit ECC	pending retirement

B. Thermal map acquisition

In this study, we primarily investigate the thermal spatial distribution within the processor region. Due to factors such as the focal length range of the camera, it is challenging to have the target area occupy the entire 240×320 px field of view of the camera. To address this issue, the thermal map of the target area used for training is cropped and reduced to 183×205 px. Additionally, to ensure synchronization with the collection of GPU metrics, the camera records at a speed of 10Hz.

C. GPU metrics acquisition

An NVIDIA GeForce RTX 4060 chip (3072 CUDA cores, 8 GB GDDR6 Memory, released in 2023) is studied in this work. NVIDIA provides a command line management and monitoring tool for their GPUs, known as NVIDIA System Management Interface (NVIDIA-SMI). Through this utility, we can query the device state and gather GPU metrics to obtain information about the performance and thermal behavior of the GPU.

NVIDIA-SMI supports several types of GPU metrics, each type consisting of several items. Table I provides a list of the metrics we selected, including the readings from sensors such as GPU core temperature, as well as metrics on current GPU memory usage, frequency, operating mode, and more. It's worth noting that the Error Counter category includes counts for both corrected and uncorrected errors, effectively doubling the number of counters. In total, we have 53 metrics for the NVIDIA GeForce RTX 4060 GPU chip.

We use the historical time series of metric vectors from the previous 10 frames as input for thermal prediction, making the processing part of this similar to that of natural language tasks. Longer data histories can increase the information available to the model and improve prediction accuracy, but they also impose a heavier burden on the model. The selection of this parameter can be traded off based on the actual application scenario.

VI. GPUTHERMALMAP FOR FULL-CHIP THERMAL ESTIMATION FOR GPUS

A. Review of the transformer DNN architecture

Overall, our task is to generate images from time series data. Previous research on CPU heat map prediction has demonstrated that the transformer-based method is superior to other methods (like LSTM-based or GAN-based methods) when time series processing is needed [17]. Transformer employs self-attention mechanisms, enabling the model to access all historical data simultaneously in a highly parallelized manner, addressing the issue of long-range dependencies. Below, we briefly review the transformer architecture and self-attention mechanism.

Given a sequence of input vector $\{x_i\}$, the attention unit first embeds each vector x_i into three vectors: query vector q_i, key vector k_i, and value vector v_i, which are computed as

$$q_i = x_i W^Q, k_i = x_i W^K, v_i = x_i W^V$$

Here q and k share the same dimension d_{qk}. Then a weighted average will be performed among $\{v_i\}$. The ith weight comes from the dot production between $\{q_i\}$ and $\{k_i\}$, divided by $\sqrt{d_{qk}}$ and then passed through a softmax for the normalization. The weighted average is the attention vector $Attention$ $(\{q_i\}, \{k_i\}, \{v_i\})$. If we write the sequence of vectors in a way of matrices (i.e. Q, K, V for the matrices where the i-th rows are q_i, k_i, v_i), then the attention can be represented in a more concise form.

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{qk}}}\right) V \qquad (1)$$

where softmax is along the rows. Now the input vectors sequence $\{x_i\}$ becomes a single vector output. In general, we may want to process the data with multiple attention interests. Then we can have multiple $\left(W^Q, W^K, W^V\right)$ (which is called the head), concatenate multiple attention vectors together into a matrix, and then project it into a final output matrix as shown below:

$$\begin{aligned}&\text{MultiheadedAttention}\,(Q, K, V) = \\ &\qquad \text{Concat}\left(\text{Attention}\left(\text{QW}_\text{i}^\text{Q}, \text{KW}_\text{i}^\text{K}, \text{VW}_\text{i}^\text{V}\right)\right) \text{W}^\text{O}\end{aligned} \qquad (2)$$

where $\left({W_i}^Q, {W_i}^K, {W_i}^V\right)$ are the mulitheads, and W^O is the final projection matrix. By repeating multiple attention layers, we have the typical encoder or decoder structure in the original transformer model [37].

B. Proposed thermal estimation framework

Fig. 4 illustrates the framework of our model, named *GPUThermalMap*. Since our objective is to generate image outputs, we have not adopted the traditional encoder-decoder architecture but only employ an encoder structure. Subsequently, we use the multi-layer perceptron (MLP) to generate

TABLE II
ACCURACY COMPARISON BETWEEN GPUTHERMALMAP AND THERMGAN

	GPUThermalMap	ThermGAN [16]	ThermGAN/GPUThermalMap ratio
Average RMSE	0.190°C	0.397°C	2.089
Maximum RMSE	0.812°C	2.323°C	2.861
RMSE deviation	0.132°C	0.252°C	1.909

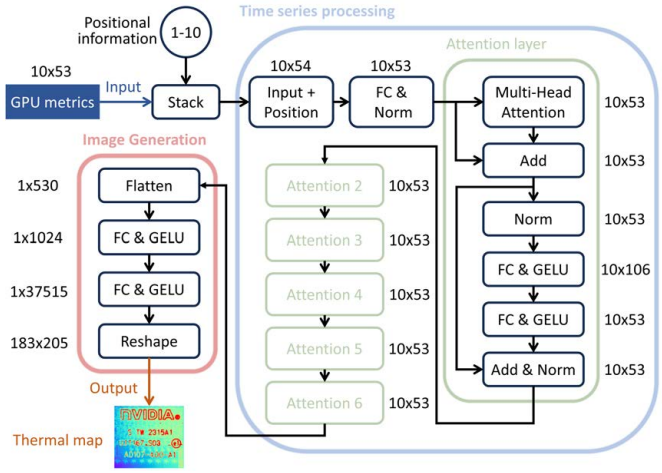


Fig. 4. Architecture of proposed *GPUThermalMap*

the output and reshape it into a 183×205 image. This is somewhat akin to GPT-1 [38] only using a decoder structure to predict the next word.

The first part is to process the time series input, which is included in the blue box. The input $\{x_i\}$ can be viewed as 10 vectors, each with a length of 53. Since the transformer do not inherently capture sequential structure, incorporating positional information as prior knowledge can aid the model's learning. To achieve this, we first augment the length of each vector by 1 and append the corresponding position 1-10 to each vector. Subsequently, we pass each vector through a fully-connected (FC) layer to revert it to a vector of length 53. This completes the positional encoding.

Next are six attention layers with identical structures, highlighted in green boxes. The structure of the first attention layer is depicted in the figure. It begins with a multi-head attention structure mentioned in the section VI-A. Then, the data is added with a skip connection and normalized by layer normalization (denoted by "Norm"). Then we have two fully connected layers operating on the columns, using the Gaussian Error Linear Unit (GELU) [39] as the activation function. The result is added to the output before layer normalization, and then passed through another layer normalization before being forwarded to the next attention layer.

After the attention layers, the processing of the time series concludes and transitions into the image generation process, as indicated by the brown box. The time series data is flattened and then passed through two FC layers to match the final image size before being reshaped into the final 183×205 thermal image output.

Last not least, for the estimation, we minimize the L2 loss between $F(x)$ and ground truth y. The loss function is:

$$loss_G = \mathbb{E}_{(x,y)}[||y - F(x)||^2] \tag{3}$$

As mentioned in section IV, we cannot obtain the real temperature of the parts covered by text. Therefore, in this equation y and $F(x)$ only represent the uncovered parts.

VII. EXPERIMENTAL RESULTS AND DISCUSSIONS

We implemented *GPUThermalMap* with Python 3.8 and TensorFlow (version 2.11.0) [40], a widely adopted open-source machine learning library. The model was trained on a Linux server equipped with a Xeon E5-2699v4 2.20GHz processor and an NVIDIA Titan RTX GPU. The dataset ultimately comprised 16,000 data points, of which 12,800 were allocated for training and 3,200 for testing purposes.

A. Thermal map estimation accuracy

We used the Root-Mean-Square Error (RMSE) between the predicted thermal map and the ground truth across all pixels except those covered by text to measure accuracy. On the test set, the average RMSE is $0.190°\mathrm{C}$, with a standard deviation of only $0.132°\mathrm{C}$, across a temperature range of 42.24 to $76.56°\mathrm{C}$. Fig. 5 illustrates the estimated and measured thermal maps, showcasing examples from the test set. Each column in the figure represents the comparison results at a specific time step. We display the results for three time steps. Clearly, *GPUThermalMap* has achieved very high precision and the learning of the contour is also very successful.

Besides, we compared *GPUThermalMap* with the GAN-based full-chip thermal map estimation methods, *ThermGAN* [16], in terms of prediction accuracy on the same dataset. The training is stopped when there is no obvious improvement for both models. We ended up training each model for 150 epoches. The results are shown in Table II. The last column is the ratio of the error of *ThermGAN* to that of *GPUThermalMap*. We can see that *GPUThermalMap* is about 2.09x more accurate than *ThermGAN* in average. In terms of maximum RMSE, *GPUThermalMap* is about 2.86x more accurate. This is also the similar case for the RMSE deviation. In summary, the experimental results confirm that in scenarios where the input is a time series, the transformer-based approach indeed exhibits greater advantages in terms of accuracy compared to the GAN-based method.

B. Computational efficiency

On the server mentioned at the beginning of this section, training the model until convergence takes from several hours to tens of hours. Once the model training is completed, it can be deployed on the target system and the average inference time measured is $22\mathrm{ms}$. This speed ensures that the model can keep up with real-time thermal map predictions at a frequency

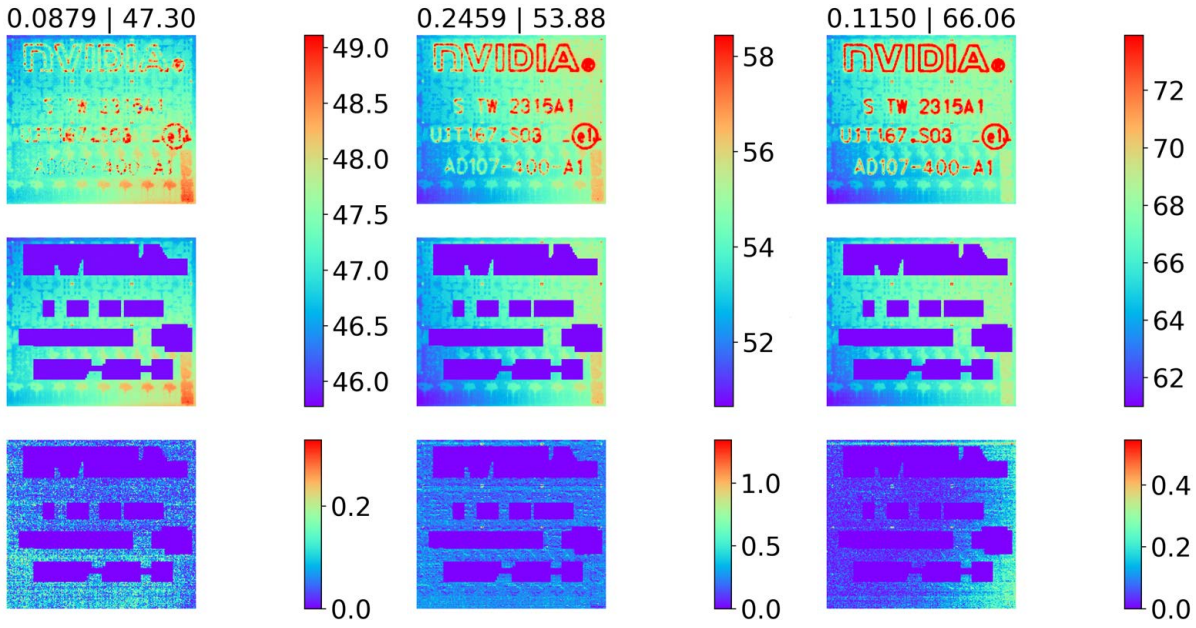


Fig. 5. Measured thermal maps (row #1), estimated thermal maps (row #2), and error maps (row #3). The numbers above the maps indicate the *Temperature RMSE* | *Average Temperature* (unit: °C). Each column indicate result at one particular time step.

of 10Hz or with 100ms intervals, just like our experimental setup.

VIII. Conclusion

In this article, we obtained the full-chip thermal maps of commercial NVIDIA GeForce RTX 4060 GPU with AI workloads. The unique features of GPU thermal patterns has been discussed, which is different from commercial CPU thermal patterns. On this basis, we have proposed a machine learning based approach, named *GPUThermalMap*, for real-time estimation of full-chip thermal maps for the RTX 4060 GPU. This method leverages NVIDIA-SMI metrics as input for the machine learning models. To construct the dynamic thermal map model, we have developed a modified self-attention architecture to model thermal maps. Numerical results highlight the effectiveness of *GPUThermalMap* in achieving highly accurate thermal map predictions, boasting an RMSE of only 0.19°C or 0.6% of the full-scale error. Additionally, it outperforms the GAN-based method, *ThermalGAN*, by 2.09x in terms of accuracy on average. Furthermore, the proposed model offers real-time estimation with a rapid speed of 22ms on the target chip.

References

[1] "Critical Reliability Challenges for The International Technology Roadmap for Semiconductors (ITRS)," 2003, in International Sematech Technology Transfer Document 03024377A-TR, 2003.

[2] "Nvidia blackwell b200 gpu thermal design power," https://www.nvidia.com/en-us/data-center/dgx-b200/.

[3] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *Micro, IEEE*, vol. 32, no. 3, pp. 122–134, May 2012.

[4] M. Taylor, "A landscape of the new dark silicon design regime," *IEEE/ACM International Symposium on Microarchitecture*, vol. 33, no. 5, pp. 8–19, October 2013.

[5] K.Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *International Symposium on Computer Architecture*, 2003, pp. 2–13.

[6] J. Kong, S. W. Chung, and K. Skadron, "Recent thermal management techniques for microprocessors," *ACM Comput. Surv.*, vol. 44, no. 3, pp. 13:1–13:42, jun 2012. [Online]. Available: http://doi.acm.org/10.1145/2187671.2187675

[7] S. Sadiqbatcha, J. Zhang, H. Zhao, H. Amrouch, J. Henkel, and S. X.-D. Tan, "Post-silicon heat-source identification and machine-learning-based thermal modeling using infrared thermal imaging," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2020.

[8] H. Amrouch and J. Henkel, "Lucid infrared thermography of thermally-constrained processors," in *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, July 2015, pp. 347–352.

[9] F. Beneventi, A. Bartolini, P. Vivet, and L. Benini, "Thermal analysis and interpolation techniques for a logic + wideio stacked dram test chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 4, pp. 623–636, April 2016.

[10] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in VLSI circuits: Principles and methods," *Proc. of the IEEE*, vol. 94, no. 8, pp. 1487–1501, Aug. 2006.

[11] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, May 2006.

[12] Y. Yang, Z. P. Gu, C. Zhu, R. P. Dick, and L. Shang, "ISAC: Integrated space and time adaptive chip-package thermal analysis," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 16, no. 1, pp. 86–99, 2007.

[13] H. Wang, S. X.-D. Tan, G. Liao, R. Quintanilla, and A. Gupta, "Full-chip runtime error-tolerant thermal estimation and prediction for practical thermal management," in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, Nov. 2011.

[14] S. Sadiqbatcha, Y. Zhao, J. Zhang, H. Amrouch, J. Henkel, and S. X. D. Tan, "Machine learning based online full-chip heatmap estimation," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020, pp. 229–234.

[15] S. Sadiqbatcha, J. Zhang, H. Amrouch, and S. X.-D. Tan, "Real-time full-chip thermal tracking: A post-silicon, machine learning perspective," *IEEE Transactions on Computers*, 2021.

[16] W. Jin, S. Sadiqbatcha, J. Zhang, and S. X.-D. Tan, "Full-chip thermal map estimation for commercial multi-core cpus with generative adversarial learning," in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*. New York, NY, USA: ACM, Nov. 2020, pp. 1–9.

[17] J. Lu, J. Zhang, and S. X.-D. Tan, "Real-time thermal map estimation for AMD multi-core CPUs using transformer," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–7.

[18] K. Zhang, A. Guliani, S. Ogrenci-Memik, G. Memik, K. Yoshii, R. Sankaran, and P. Beckman, "Machine learning-based temperature prediction for runtime thermal management across system components," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 2, pp. 405–419, Feb 2018.

[19] Intel, "Intel Performance Counter Monitor (PCM)," https://software.intel.com/en-us/articles/intel-performance-counter-monitor.

[20] AMD, "AMD uProf software profiling tool," https://developer.amd.com/amd-uprof/.

[21] "Chatgpt from openai." [Online]. Available: https://chat.openai.com

[22] S. P. Gurrum, Y. K. Joshi, W. P. King, K. Ramakrishna, and M. Gall, "A compact approach to on-chip interconnect heat conduction modeling using the finite element method," *Journal of Electronic Packaging*, vol. 130, pp. 031 001.1–031 001.8, September 2008.

[23] Y. C. Gerstenmaier and G. Wachutka, "Rigorous model and network for transient thermal problems," *Microelectronics Journal*, vol. 33, pp. 719–725, September 2002.

[24] D. Li, S. X.-D. Tan, E. H. Pacheco, and M. Tirumala, "Parameterized architecture-level dynamic thermal models for multicore microprocessors," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 15, no. 2, pp. 1–22, 2010.

[25] T. Eguia, S. X.-D. Tan, R. Shen, D. Li, E. H. Pacheco, M. Tirumala, and L. Wang, "General parameterized thermal modeling for high-performance microprocessor design," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 2011.

[26] Z. Liu, S. X.-D. Tan, H. Wang, Y. Hua, and A. Gupta, "Compact thermal modeling for packaged microprocessor design with practical power maps," *Integration, the VLSI Journal*, vol. 47, no. 1, January 2014, in press, online access: http://www.sciencedirect.com/science/article/pii/S0167926013000412.

[27] Y.-K. Cheng, C.-H. Tsai, C.-C. Teng, and S.-M. Kang, *Electrothermal Analysis of VLSI Systems*. Kluwer Academic Publishers, 2000.

[28] R. Cochran and S. Reda, "Spectral techniques for high-resolution thermal characterization with limited sensor data," in *Proc. Design Automation Conf. (DAC)*, 2009, pp. 478–483.

[29] A. Nowroz, R. Cochran, and S. Reda, "Thermal monitoring of real processors: Techniques for sensor allocation and full characterization," in *Proc. Design Automation Conf. (DAC)*, 2010.

[30] S. Reda, R. Cochran, and A. N. Nowroz, "Improved thermal tracking for processors using hard and soft sensor allocation techniques," *IEEE Transactions on Computers*, vol. 60, no. 6, pp. 841–851, June 2011.

[31] J. Ranieri, A. Vincenzi, A. Chebira, D. Atienza, and M. Vetterli, "Eigenmaps: Algorithms for optimal thermal maps extraction and sensor placement on multicore processors," in *Proceedings of the 49th Annual Design Automation Conference*, ser. DAC '12. New York, NY, USA: ACM, 2012, pp. 636–641. [Online]. Available: http://doi.acm.org/10.1145/2228360.2228475

[32] Y. Zhang, B. Shi, and A. Srivastava, "Statistical framework for designing on-chip thermal sensing infrastructure in nanoscale systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 2, pp. 270–279, 2014.

[33] X. Li, X. Li, W. Jiang, and W. Zhou, "Optimising thermal sensor placement and thermal maps reconstruction for microprocessors using simulated annealing algorithm based on pca," *IET Circuits, Devices Systems*, vol. 10, no. 6, pp. 463–472, 2016.

[34] Yufu Zhang, A. Srivastava, and M. Zahran, "Chip level thermal profile estimation using on-chip temperature sensors," in *2008 IEEE International Conference on Computer Design*, 2008, pp. 432–437.

[35] A. Ziabari, J. Park, E. K. Ardestani, J. Renau, S. Kang, and A. Shakouri, "Power blurring: Fast static and transient thermal analysis method for packaged integrated circuits and power devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 11, pp. 2366–2379, 2014.

[36] J. Lu, J. Zhang, W. Jin, S. Sachdeva, and S. X.-D. Tan, "Learning based spatial power characterization and full-chip power estimation for commercial tpus," in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, ser. ASPDAC '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 98–103. [Online]. Available: https://doi.org/10.1145/3566097.3568347

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017.

[38] A. Radford and I. Sutskever, "Improving language understanding by generative pre-training," in *arxiv*, 2018.

[39] K. G. Dan Hendrycks, "Gaussian Error Linear Units (GELUs)," *arXiv e-prints*, p. arXiv:1606.08415v4, June 2016.

[40] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/