Thermal Map Dataset for Commercial Multi/Many Core CPU/GPU/TPU

Jincong Lu and Sheldon X.-D. Tan
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, California, United States
jincong.lu@email.ucr.edu,stan@ece.ucr.edu

ABSTRACT

This paper presents a comprehensive set of datasets containing transient thermal maps of commercial multi/many-core CPU, GPU, and TPU processors. These thermal maps were obtained using a thermal IR imaging system and include data from a range of processors: Intel i5-3337U dual-core CPU, Intel i7-8650U quad-core CPU, AMD Ryzen 7 4800U 8-core CPU, NVIDIA GeForce RTX 4060 GPU, and Google Coral M.2 TPU. Additionally, we review recently proposed thermal map estimation methods developed using various Deep Neural Network (DNN) techniques, including Long Short-Term Memory (LSTM), Generative Adversarial Networks (GAN), and transformer-based models. We provide a detailed discussion and comparison of these models trained on the thermal map datasets. These datasets aim to support and stimulate advanced research in thermal map estimation and modeling within the research community.

ACM Reference Format:
Jincong Lu and Sheldon X.-D. Tan. 2024. Thermal Map Dataset for Commercial Multi/Many Core CPU/GPU/TPU. In *2024 ACM/IEEE International Symposium on Machine Learning for CAD (MLCAD '24), September 9–11, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3670474.3685963

1 INTRODUCTION

With the ongoing trend of rapid integration and technology scaling, contemporary high-performance multi/many-core processors are encountering increasingly severe thermal limitations. Research has demonstrated that elevated temperatures exponentially degrade the reliability of semiconductor chips [1], posing a significant industry concern. This issue is exacerbated for AI chips, such as commercial GPUs from NVIDIA, where power consumption can exceed 1000W. For example, the H100/H200 GPUs have a thermal design power (TDP) of 700W, while the latest Blackwell GPU B200 reaches a TDP of 1200W [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]. As a result, accurate real-time estimation of the full-chip thermal map and identification of some hot spots becomes critical.

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 [14–18]. Those methods leverage online real-time utilization and monitoring information such as core frequency, voltage, and various utilization or performance metrics, which are naturally supported by most commercial processors [19]. Software tools such as Intel's Performance Counter Monitor (PCM) [20] and AMD's uProf [21] 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 the recent transformer-based model for AMD multi-core CPUs based on uProf utilization metrics [18]. For TPUs, thermal map modeling using the hyper-parameters of the Deep Neural Network (DNN) network workload was demonstrated [17].

In this article, we present some measured thermal maps and real-time utilization and monitoring data for five commercial multi/many-core CPU/GPU/TPU processors, which has been released at [22]. The processors comprise an Intel i5-3337U dual-core CPU, an Intel i7-8650U quad-core CPU, an AMD Ryzen 7 4800U 8-core CPU, an NVIDIA GeForce RTX 4060 GPU, and a Google Coral M.2 TPU. Notably, thermal maps from the NVIDIA GeForce RTX 4060 GPU are unveiled here for the first time. Our aim is to disseminate this data to the community, fostering further research on efficient thermal map estimation and hot spot identification for these significant commercial CPUs and emerging AI GPUs and TPUs. Moreover, alongside this data, we provide a summary of existing machine learning based thermal modeling techniques derived from these data, including methods employing LSTM-, GAN-, and transformer-based models.

This article is organized as follows. Section 2 outlines the thermal modeling framework and IR thermography setup employed

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

MLCAD '24, September 9–11, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0699-8/24/09
https://doi.org/10.1145/3670474.3685963

for collecting the thermal and utilization and monitoring data. Section 3 explains the thermal data collected as well as the performance metrics and monitoring data from the mentioned commercial CPU/GPU/TPUs. Section 4 review the machine learning based full-chip thermal map estimation methods based on the real-time utilization and monitoring metrics. Section 5 presents some numerical results and provides comparisons of those machine learning-based models. Section 6 concludes the article.

2 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 chips during its operational load.

2.1 Estimation flow overview

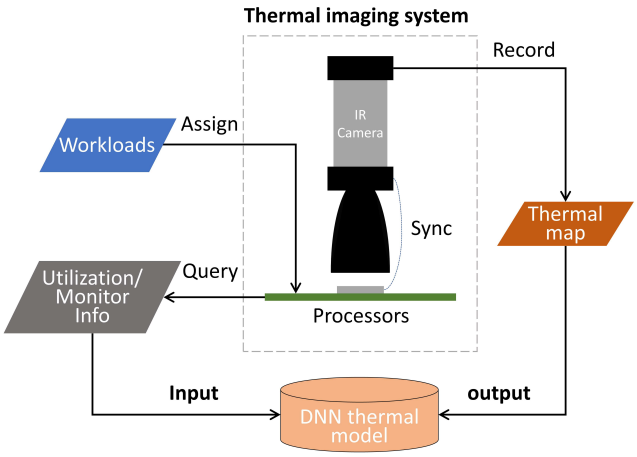


Figure 1: Framework and data acquisition flow

Fig. 1 depicts the framework of the proposed approach. The proposed approach comprises three main components. Firstly, we gather data by logging the chip'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 the online thermal prediction model.

The model training process involves the input and output. The input includes recorded performance 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 thermal prediction.

In the subsequent section, we will discuss the performance metrics and thermal maps data acquisition setup for each chip. Furthermore, in Section 4, we will delve deeper into the DNN thermal models. The rest of the framework is covered in the following subsections.

2.2 Thermal IR imaging system

Accurate measurement of chip 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.

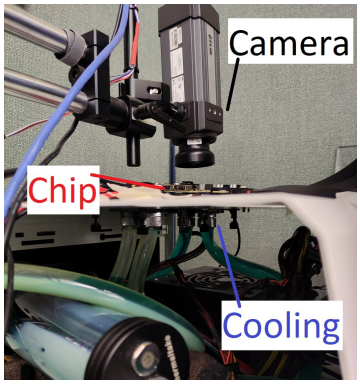


Figure 2: Thermal Imaging system setup

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°C to 120°C and resolves the IR spectral range of 7.5μm to 13μm.

2.3 Operational loads

To better reflect real-world scenarios, the dataset should cover the anticipated usage situations as comprehensively as possible. Therefore, it is necessary to run various workloads on the system during data collection to simulate actual operations. In this work, different workloads were assigned to these chips. For the CPUs, the workloads included daily activities such as idling, word processing, and data compression. Some benchmark suites are also performed.

For task-specific processors such as GPUs and TPUs, we focus more on their performance in highly specialized domains. Consequently, their workloads consist of machine learning models.

3 THERMAL MAP AND UTILIZATION METRICS FOR COMMERCIAL PROCESSORS

3.1 Intel i5-3337U dual-core CPUs

The first multi-core processor we presented is the Intel Core i5-3337U, which is a 2-core / 4-thread CPU processor released in 2012 and used in many laptop computers in the past. Fig. 3 shows the thermal image of this processor.

The primary high-level performance monitoring software supported by Intel is Intel's Performance Counting Monitor (IPCM) [20]. IPCM provides the system-level utilization metrics that we will be utilizing in this work. These provide a comprehensive high-level view of the processor's utilization with system-level metrics such as energy usage, package and core frequency, instruction counts,

cache hit/miss rates, etc., as well as the sensed temperature from the embedded sensors. Table 1 shows the complete list of IPCM performance metrics from both the package and core-wise domains. In total, IPCM provides 86 performance metrics (I_1 to I_{80}) for the Intel i5-3337U.

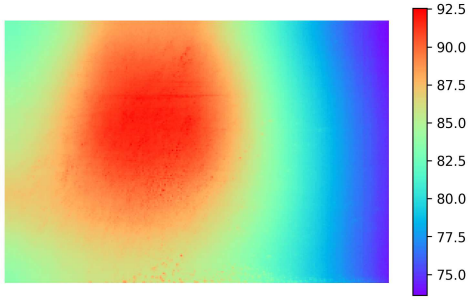


Figure 3: Thermal map of Intel i5-3337U

Table 1: IPCM metrics for the Intel i5-3337U

Pkg.	Pkg.	Core1.1	Core1.2	Core2.1	Core2.2
exec	inst nom	exec	exec	exec	exec
IPC	inst nom%	IPC	IPC	IPC	IPC
freq	C2res%	freq	freq	freq	freq
afreq	C3res%	afreq	afreq	afreq	afreq
L3 miss	C6res%	L3 miss	L3 miss	L3 miss	L3 miss
L2 miss	C7res%	L2 miss	L2 miss	L2 miss	L2 miss
L3 hit	energy (J)	L3 hit	L3 hit	L3 hit	L3 hit
L2 hit	temp	L2 hit	L2 hit	L2 hit	L2 hit
L3 MPI		L3 MPI	L3 MPI	L3 MPI	L3 MPI
L2 MPI		L2 MPI	L2 MPI	L2 MPI	L2 MPI
read rate		C0res%	C0res%	C0res%	C0res%
write rate		C1res%	C1res%	C1res%	C1res%
inst count		C3res%		C3res%	
ACYC		C6res%		C6res%	
physIPC		C7res%		C7res%	
physIPC%		temp		temp	

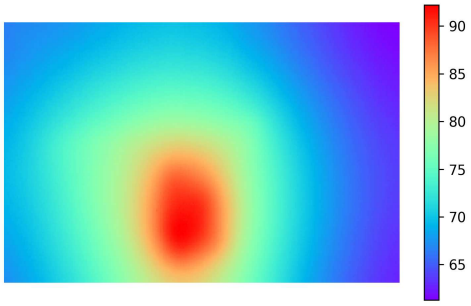
3.2 Intel Intel i7-8650U quad-core CPU

The second multi-core processor we presented is the Intel i7-8650U, which is a 4-core / 8-thread CPU processor released in 2017. Fig. 4 (a) shows the thermal map of the processor with one hot spot (one core is active). Fig. 4 (b) shows the thermal image of this process in which we can clearly see that the temperature at the thermal sensor is quite different than the true hot spot. The complete list of all 170 PCM metrics that we collect and employ for the thermal modeling of Intel i7-8650U is shown in Table 2.

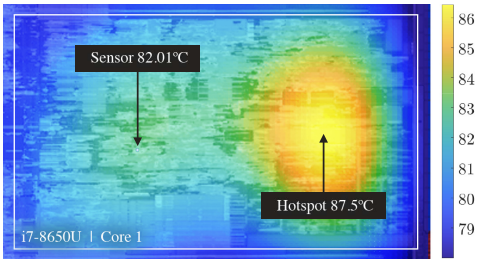
3.3 AMD Ryzen 7 4800U 8-core CPU

The third multi-processor we presented is the AMD Ryzen 7 4800U chip, which has 8 cores and 16 threads and was released in 2020. For this processor, the utilization and monitor metrics are provided by AMD via AMD *uProf 4.0* program [21], which is the performance monitoring software. This software allows us to gather system-level utilization metrics as well as core-wise power characteristics from the chip. By utilizing these metrics, we can gain insights into the chip's performance and thermal behavior.

AMD *uProf 4.0* offers two types of performance metrics that are relevant to our study. The first type comprises CPU power metrics, including package and core energy usage, core frequency, and



(a) i7-8650U with one hot spot



(b) i7-8650U with temperatures at sensor and true hot spot

Figure 4: Thermal maps of Intel i7-8650U

Table 2: Performance metrics (Intel PCM)

Pkg.	Socket	Socket	Core 1 to 8
INST	EXEC	C6res%	EXEC
ACYC	IPC	C7res%	IPC
TIME	FREQ	C2res%	FREQ
PhysIPC	AFREQ	C3res	AFREQ
PhysIPC%	L3MISS	C6res	L3MISS
INSTnom	L2MISS	C7res	L2MISS
INSTnom%	L3HIT	C8res%	L3HIT
C0res%	L2HIT	C9res%	L2HIT
C2res%	L3MPI	C10res%	L3MPI
C3res%	L2MPI	SKT0	L2MPI
C6res%	READ		C0res%
C7res%	WRITE		C1res%
C8res%	TEMP		C3res%
C9res%	C0res%		C6res%
C10res%	C1res%		C7res%
Energy	C3res%		TEMP

temperature readings from embedded sensors, among others. The second type consists of PMU event counters, which track various events such as instruction counts, cache hits, and branch predictions. In Table 3, we present the list of AMD *uProf 4.0* performance metrics from these two domains. The CPU Power Metrics domain encompasses 42 metrics, while the PMU Events domain includes 58 different types of events. It is important to note that some PMU events have specific unit masks to distinguish different conditions, resulting in further categorization. As a result, we have selected a total of 116 metrics for the PMU Events domain. Combining both domains, we have a total of **158** metrics selected for the AMD Ryzen 7 4800U chip.

3.4 NVIDIA GeForce RTX 4060 GPU

This section presents information for the NVIDIA GeForce RTX 4060 GPU (3072 CUDA cores, 8 GB GDDR6 Memory, released in

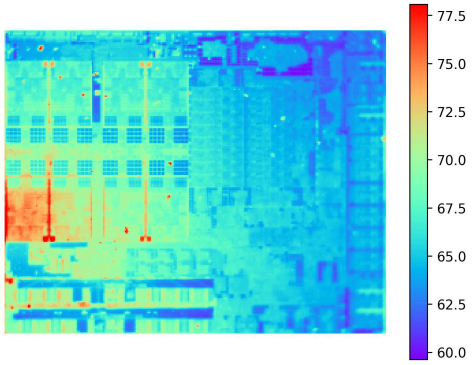


Figure 5: Thermal map of AMD R7 4800U

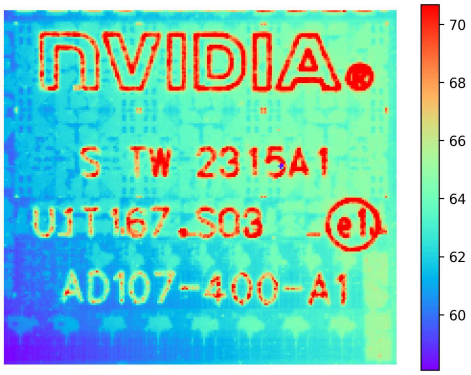


Figure 6: Thermal map of NVIDIA RTX 4060

Table 3: Selected Performance Metrics (AMD R7 4800U)

CPU power metrics		
ThreadEffectiveFrequency 0-15	ThreadPerformanceState 0-15	Socket0Power
CorePower 0-7	Socket0Temperature	
PMU events		
FpRetSseAvxOps	FpRetiredSerOps	FpDispFaults
LsBadStatus2	LsLocks	LsRetClClush
LsRetCpuid	LsDispatch	LsSmiRx
LsIntTaken	LsSTLF	LsStCommitCancel2
LsMabAlloc	LsRefillsFromSys	LsL1DTlbMiss
LsMisalLoads	LsPrefInstrDisp	LsInefSwPref
LsSwPfDcFills	LsHwPfDcFills	LsAllocMabCount
LsNotHaltedCyc	IcCacheFillL2	IcCacheFillSys
BpL1TlbMissL2TlbHit	BpL1TlbMissL2TlbMiss	BpL1BTBCorrect
BpL2BTBCorrect	BpDynIndPred	BpDeReDirect
BpL1TlbFetchHit	DeDisUopQueueEmpty	DeDisUopsFromDecoder
DeDisDispatchTokenStalls1	DeDisDispatchTokenStalls0	ExRetInstr
ExRetCops	ExRetBrn	ExRetBrnMisp
ExRetBrnTkn	ExRetBrnTknMisp	ExRetBrnFar
ExRetNearRet	ExRetNearRetMispred	ExRetBrnIndMisp
ExRetMmxFpInstr	ExRetCond	ExDivBusy
ExDivCount	ExRetMsprdBrnchInstrDirMsmtch	ExTaggedIbsOps
ExRetFusBrnchInst	L2RequestG1	L2RequestG2
L2CacheReqStat	L2PfHitL2	L2PfMissL2HitL2
L2PfMissL2L3		

Table 4: 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

2023). NVIDIA provides a command-line management and monitoring tool for its GPUs, known as the NVIDIA System Management Interface (NVIDIA-SMI). Through this utility, we can query the device state and gather GPU metrics to obtain information about the GPU's performance and thermal behavior.

NVIDIA-SMI supports several types of GPU metrics, each type consisting of several items. Table 4 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.

3.5 Google Coral M.2 TPU

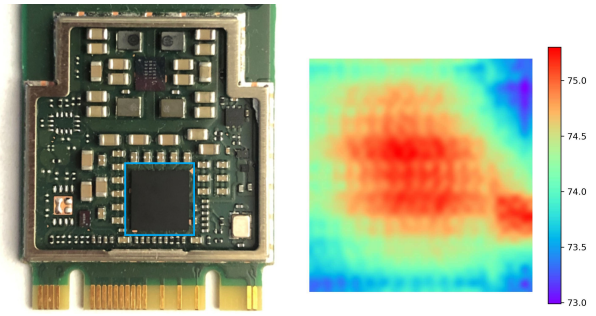
The last processor we provided is Google Coral M.2 TPU, which was released in 2020. The M.2 TPU board has one or two Edge TPU coprocessors, as shown in Fig. 7 (a). Each Edge TPU coprocessor is capable of performing 4 trillion operations per second (4 TOPS), using 2 watts of power. But different than other commercial CPUs and GPUs, as we presented earlier, Google TPUs do not have real-time utilization and monitoring metrics information.

To mitigate this issue, we instead use the hyperparameters of the AI workloads as the inputs for training the thermal map models. The rationale is that the TPU hardware resources that the network demands are tightly related to the network model architecture, size, operations, etc. Hence, we are able to characterize the TPU's power from the workloads' hyperparameters, such as operation type, count, workload size, etc.

Table 5: Selected Workload Features (Coral M.2 TPU, Google Edge)

Overall			
image_shape	pooling_mode	onchip_mem_rem	num_op_tpu
width_multiplier	model_size	offchip_mem_used	num_op_cpu
depth_multiplier	onchip_mem_used	total_op_cnt	infer_time
Operational Statistics			
add	full_connect	pad	reduce_max
avg_pool_2d	l2_norm	quant	relu
concat	max_pool_2d	reshape	strslc
conv2d	mean	sft_max	hard_swish
depconv2d	mul	sub	

We divide the network's hyperparameters as features into two groups, called the *overall features* and *operational statistics*. Neural network models that are coded to run on CPU need to be compiled to a TPU readable version. In our study, EdgeTPU Compiler [23] is employed to transform a CPU version network model to a TPU version. On the one hand, the overall features, such as model size and memory usage, are recorded through the process. On the other hand, we collect statistical information for the type and count of operations indicated by the network in the meantime. We mark that for different TPUs, different tools may be involved. However, network information should always be reachable. Today's world has a vast number of neural networks and hundreds of kinds of operations. To find the most popular operational features of network models, we explored a number of the most popular and widely used open-source deep neural network models from TensorFlow [24]. Table 5 shows the 31 features selected for the network workloads.



(a) Coral M.2 TPU module (b) The Coral TPU thermal map

Figure 7: Google Coral M.2 TPU module (a) and its thermal map (b)

4 DNN-BASED FULL-CHIP THERMAL MAP ESTIMATOR

Recently, several methods have been proposed to leverage real-time on-chip utilization and monitoring information for generating thermal and power maps. These methods predominantly employ DNN techniques utilizing real-time performance metrics. We summarize these methods in this section, and their performance comparison will be discussed in the next section.

4.1 LSTM-based method

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed to capture long-term dependencies and handle vanishing gradient problems. LSTMs are particularly suited for holding time series tasks.

Sadiqbatcha et al. proposed an LSTM-based approach named *Realmaps*, which utilizes Intel PCM metrics to estimate full-chip thermal maps in commercial off-the-shelf multi-core processors [14, 15]. *Realmaps* has demonstrated promising results regarding accuracy and inference speed for real-time applications.

However, LSTM often struggles to achieve high accuracy due to its inherent limitations in capturing complex temporal patterns over extended sequences.

4.2 GAN-based method

Generative Adversarial Networks (GAN) consist of two neural networks, a generator and a discriminator, which compete against each other in a zero-sum game. The generator creates synthetic data samples, while the discriminator attempts to distinguish between real and generated samples. GAN has shown impressive results in generating images.

Therefore, to further improve the thermal prediction accuracy, an enhanced method using a Convolutional Neural Network (CNN) model based on the GAN architecture was introduced, known as *ThermGAN*. This approach has shown superior accuracy compared to LSTM-based methods [16].

Furthermore, the GAN model was extended to model Google TPU chips for the first time [17]. This study demonstrated that, even for TPUs with limited online utilization and monitoring metrics, thermal models could be successfully developed using only the hyper-parameters of DNN models running on TPUs.

However, the training process of GAN can be highly unstable, requiring careful tuning of hyperparameters and architecture. Additionally, GAN is not inherently designed to handle time series data, making it challenging to apply them effectively to tasks that involve sequential inputs, like what we have here.

4.3 Transformer-based method

Transformer is a type of neural network architecture that has revolutionized the field of natural language processing and is increasingly applied to various sequential data tasks. Unlike traditional RNNs and LSTMs, Transformers do not process data in a sequential manner. Instead, they utilize a self-attention mechanism that allows them to attend to all positions in the input sequence simultaneously. This capability makes Transformers highly effective at capturing long-range dependencies and complex patterns in time series data. Consequently, Transformers can achieve higher accuracy in tasks that involve sequential inputs, as they are better equipped to handle the intricacies of temporal dynamics. Their ability to process inputs in parallel also makes them more efficient and scalable compared to traditional sequential models.

More recently, a transformer-based DNN method called *ThermTransformer* was proposed to estimate the full-chip thermal map of AMD multicore chips using uProf utilization metrics [18]. This method outperforms both GAN-based and LSTM-based approaches due to the transformer's powerful modeling capability for time-series data through the attention mechanism.

5 NUMERICAL RESULTS AND DISCUSSIONS

In this section, we present some numerical results for full-chip thermal map estimation methods using the dataset provided. As

Table 6: Accuracy comparison among different methods

Dataset	Accuracy Term	ThermTransformer [18]	ThermGAN [16]	Realmaps [15]
AMD R7 4800U	Average RMSE	0.360	0.596	2.190
	Maximum RMSE	2.776	10.880	19.762
	RMSE deviation	0.234	0.958	2.684
NVIDIA RTX 4060	Average RMSE	0.187	0.381	2.075
	Maximum RMSE	0.831	2.290	9.885
	RMSE deviation	0.134	0.249	1.892

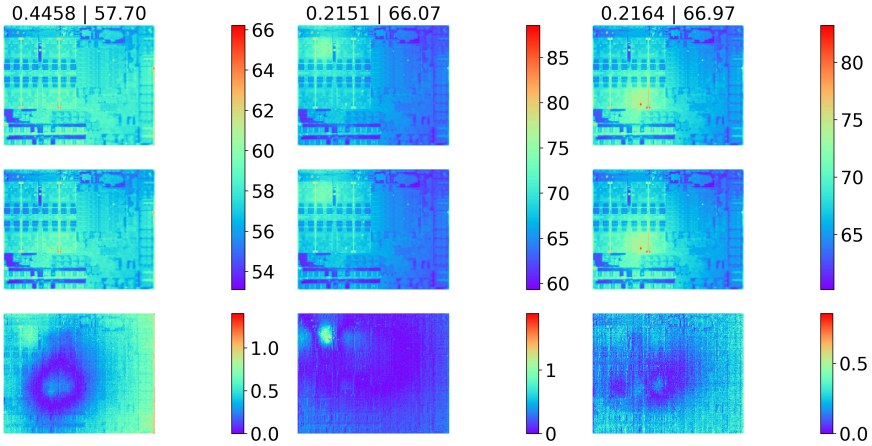


Figure 8: Measured thermal maps (row #1), estimated thermal maps (row #2), and error maps (row #3). The numbers above the maps indicate the *Temperature RMSE* | *Average Temperature* (unit: °C). Each column indicates the result at one particular time step.

mentioned in the previous section, we have *Realmaps*, *ThermGAN*, and *ThermTransformer*. We evaluate all three methods on two commercial processors: (1) AMD Ryzen 7 4800U 8-core microprocessor, which features the Zen 2 (Renoir) architecture and is commonly used in thin and light laptop computers. (2) NVIDIA GeForce RTX 4060, which has 3072 CUDA cores, 8 GB GDDR6 Memory, and was released in 2023.

The implementation of the proposed models is based on Python 3.8 and utilizes TensorFlow (2.11.0) [24], a widely adopted open-source machine learning library. The model is trained on a Linux server equipped with an Xeon E5-2699v4 2.20GHz processor and an NVIDIA Titan RTX GPU. All the DNN models are trained on the same dataset, and the training process terminates when there is no significant improvement in performance. The training procedure typically takes a few to several hours to complete.

We begin by examining the accuracy of predicting thermal maps using the proposed method on the given dataset. To evaluate the accuracy, we calculate the Root-Mean-Square Error (RMSE) between the generated thermal map and the measured thermal map across all pixels we are interested in. The results are presented in Table 6.

In the AMD Ryzen 7 4800U dataset, the temperature ranges from 44.11 to 90.08°C. The average RMSE of the *ThermTransformer* on the test set is 0.360°C, with a standard deviation of only 0.234°C. These results are remarkably accurate considering the range of the data. *ThermTransformer* demonstrates approximately 1.66x higher accuracy than *ThermGAN* on average and 6.09x higher accuracy than *RealMaps*. In terms of the maximum RMSE, *ThermTransformer* is about 3.92x and 7.12x more accurate than *ThermGAN* and *RealMaps*, respectively. This trend also applies to the NVIDIA RTX 4060 dataset.

Fig. 8 illustrates the estimated and measured thermal maps using *ThermTransformer*, showcasing examples from the AMD CPU dataset. Each column in the figure represents the comparison results at a specific time step. We display the results in three-time steps. It is evident that the model has successfully learned the contour of the real thermal map.

6 CONCLUSION

This paper provides a comprehensive set of datasets featuring transient thermal maps for a variety of commercial multi/many-core CPU, GPU, and TPU processors, captured through thermal IR imaging. The datasets encompass processors such as the Intel i5-3337U dual-core CPU, Intel i7-8650U quad-core CPU, AMD Ryzen 7 4800U 8-core CPU, NVIDIA GeForce RTX 4060 GPU, and Google Coral M.2 TPU. We hope that these datasets will foster further research and advancements in thermal map estimation and modeling within the research community.

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, 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

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

[19] 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.

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

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

[22] "Commerical Multi/Many Cores Thermal Map Dataset," https://github.com/sheldonucr/commercial_thermal_map_dataset.

[23] "Edge TPU Compiler," available from coral.ai/docs/edgetpu/compiler. [Online]. Available: https://cloud.google.com/edge-tpu

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