Runtime Thermal Profiling and Management for NPU using Distributed Triple-RO IVT Sensors and AI-assisted Profile Prediction and Task Scheduling

Zhantong Zhu*, Xueping Liu*, Mingxuan Li, Le Ye, Ru Huang, and Tianyu Jia†
School of Integrated Circuits, Peking University, Beijing, China
*Equally credited authors, †Corresponding author email: tianyuj@pku.edu.cn

***Abstract*—Thermal management is crucial for AI computing chips, where high power density leads to large spatial and temporal hotspot variations. In this work, we present an AI-assisted thermal profiling and management solution for NPUs, featuring distributed all-digital triple-RO IVT sensors, ML-based thermal profile prediction, and reinforcement-learning-based task scheduling. Implemented in a 22nm NPU design, the system deploys 14 distributed sensor nodes across compute and memory regions for fine-grained runtime temperature monitoring. A lightweight MLP predicts short-horizon thermal profiles, and a floorplan-aware task scheduler uses the predicted state to dynamically dispatch tasks before severe hotspot buildup. The triple-RO IVT sensor achieves 2.3°C, 8.0mV, and 8.1% relative current inaccuracy. Our AI-assisted thermal management achieves a promising 3.2°C-8.2°C peak temperature reduction and 2.5%-18.3% equivalent performance improvement for representative NPU workloads.**

***Index Terms*—thermal management, AI-assisted, IVT sensor**

I. Introduction

Thermal management remains a persistent and critical challenge in the era of AI computing. Driven by escalating power density from AI computation and technology scaling, on-chip thermal density and hotspot variation continuously increase, exacerbating transistor aging and degrading energy efficiency [1]. As shown in Fig. 1, existing thermal management schemes typically employ one thermal sensor (T-sensor) per CPU core for temperature monitoring and hotspot mitigation [2], [3]. However, such approaches encounter limitations in both sensing accuracy and management granularity for designs with higher power and thermal density, e.g., AI NPUs. This limitation becomes more severe in AI accelerators, where the heterogeneous compute and memory blocks create non-uniform hotspots during runtime.

Digital ring oscillator (RO)-based T-sensors enable localized hotspot detection [4], which can be extensively utilized for fine-grained thermal profiling [5]. Moreover, sensing physical properties such as voltage and current together with temperature can provide richer runtime information for better thermal management. Although significant progress has been made at the circuit level for thermal sensing, further exploration is needed to leverage these distributed sensing circuits for complete thermal profiling and management in practical NPU design cases.

In this work, we present a runtime thermal profiling and management solution for NPUs with a distributed sensor network and AI-assisted thermal profile prediction and task scheduling. Our work has the following key features: 1) distributed thermal monitoring utilizing full-digital triple-RO current-voltage-temperature (IVT) sensors, 2) runtime AI-assisted thermal profile prediction (TPP) with a lightweight ML model, and 3) reinforcement learning (RL)-based floorplan-aware task scheduling for peak temperature mitigation. By combining fine-grained sensing, short-horizon thermal prediction, and adaptive task dispatch in a closed loop, our solution supports proactive thermal management rather than reacting only after hotspot formation. Our triple-RO IVT sensor achieves 2.3°C, 8.0mV, and 8.1% relative current inaccuracy, and the on-chip prediction model attains a mean absolute error (MAE) of 1.5°C. Collectively, our RL-based task scheduling enables 3.2-8.2°C peak temperature reduction, equivalently bringing 2.5-18.3% performance improvement for representative NPU workloads.

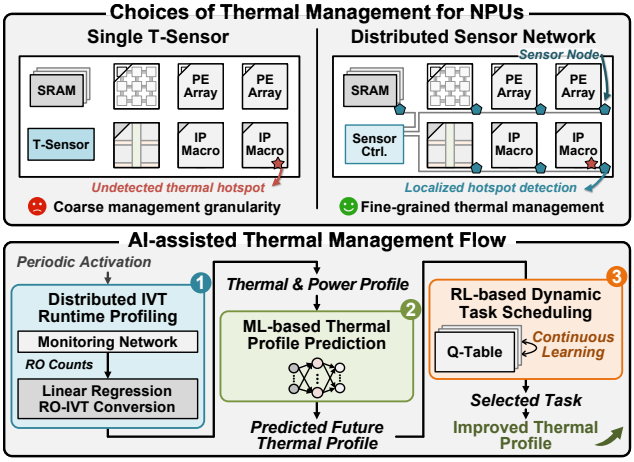


Fig. 1. Thermal management solutions for NPUs and our proposed AI-assisted thermal management scheme.

II. System Overview

Fig. 2 shows the overall chip architecture, which includes an NPU design with heterogeneous compute units, an AI-assisted thermal management unit (TMU), a host RISC-V CPU, a memory system with 256 KB SRAM, and peripheral circuits. The NPU performs diverse Transformer workload [6], [7] acceleration using 4 PE arrays (PE_0 - PE_3) and 7 custom computation IP macros (IP_0 - IP_6). A total of 14 triple-RO-based IVT sensors are distributed in the NPU, with one IVT sensor collocated with each PE array or IP macro, and 3 other sensors are placed near SRAM buffers. This sensor placement allows the system to capture both compute hotspots and thermal spreading toward nearby memory regions.

The TMU contains a sensor processing unit (SPU), a thermal profile prediction (TPP) module, and an RL-based NPU task scheduler. The SPU collects RO counts (OSC_0 -

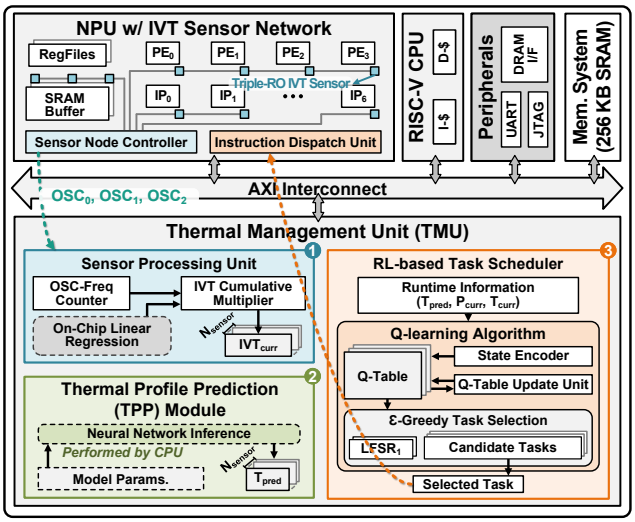


Fig. 2. Top-level system architecture of our chip.

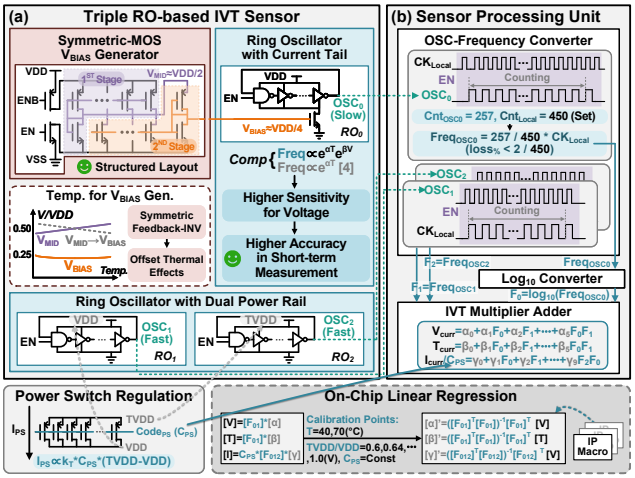


Fig. 3. (a) Design details of the triple-RO IVT sensor, (b) Sensor processing unit for on-chip linear regression.

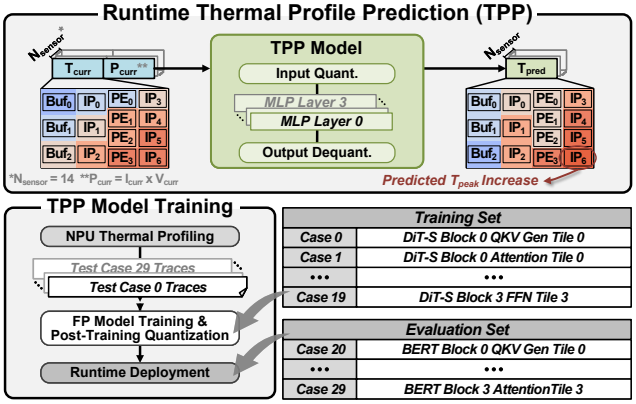


Fig. 4. ML-based thermal profile prediction (TPP).

OSC_2) from each sensor node and converts them into IVT codes (I_{curr}, V_{curr}, T_{curr}) using calibrated coefficients. The IVT codes from all sensor nodes form 14-element vectors to represent NPU power ($P_{curr} = I_{curr} \times V_{curr}$) and thermal profile ($T_{curr}$). Using T_{curr} and P_{curr}, the future thermal profile states (T_{pred}) are predicted using a lightweight quantized ML model. This short-horizon prediction enables the controller to anticipate hotspot buildup before temperature peaks fully develop. The runtime parameters T_{pred}, P_{curr}, and T_{curr} form the environmental state input for the task scheduler, which utilizes a Q-table to track the policy performance of different state-action pairs and employs an ε-greedy strategy to select the upcoming task(s). In this way, the TMU forms a runtime loop of sensing, prediction, and thermal-aware scheduling. The selected tasks are then sent to the NPU instruction dispatch unit for scheduling.

III. Distributed IVT Runtime Profiling

Fig. 3 (a) illustrates the circuit details of the triple-RO IVT sensor. Each IVT sensor includes three ring oscillators, i.e., RO_0 to RO_2. RO_0 is designed for temperature monitoring with a sub-threshold tail current source, which is driven by a symmetric-MOS V_{BIAS} generator. In contrast to the 3-T voltage reference [4], we employ a structured two-stage inverter for digital-friendly V_{BIAS} generation, with each stage connecting its input to the output. The first-stage inverter features a larger number of PMOS fingers, enabling its output voltage to increase with temperature. In contrast, the second stage has a larger NMOS finger count to neutralize the voltage-thermal effects. The generated V_{BIAS} is independent of temperature changes while being proportional to voltage, resulting in $Freq_{OSC0}$ exhibiting an exponential relationship with both temperature and voltage. This design also mitigates higher-order supply voltage influence, reducing measurement errors during linear regression of IVT results. The remaining ring oscillators, RO_1 and RO_2, are designed for voltage and current monitoring under VDD and TVDD voltage domains (TVDD$\geq$VDD), connected by power switches (PS).

Each IVT sensor forwards its three RO outputs to the SPU for sampling and counting by an on-chip reference CK_{Local}, as shown in Fig. 3 (b). The counting process terminates upon reaching predefined values, and $Freq_{OSC0,1,2}$ are calculated proportionally to CK_{Local}. $Freq_{OSC0}$ undergoes an additional logarithmic converter to eliminate the exponential base. The three frequencies are processed by regression coefficient matrices $[\alpha][\beta][\gamma]$ in the cumulative multiplier to produce I_{curr} [15:0], V_{curr} [9:0], and T_{curr} [9:0]. To derive regression coefficients, output frequencies from 14 sensors are measured at 22 (2 calibration T $\times$ 11 VDD) testing points.

IV. AI-assisted Thermal Management

Fig. 4 shows our ML-based runtime thermal profile prediction (TPP). Utilizing T_{curr} and P_{curr} vectors from 14 sensor nodes, the CPU performs online ML inference to predict thermal profile (T_{pred}) over a time horizon of tens of microseconds. The TPP model is trained in full precision and subsequently quantized, using a training set from 20 NPU cases from Diffusion Transformer (DiT) inference [7]. For runtime deployment, we use 10 separate cases from BERT inference [6]. The TPP adopts 4-layer MLP model architecture with W8A8 quantization, achieving 1.5°C mean absolute error (MAE) with negligible performance overhead.

Each NPU test case comprises several computation tasks, such as matrix multiplication, which can be dispatched to PE array or IP macro considering TPP results by the RL-based floorplan-aware task scheduler (Fig. 5). Compared to round-robin scheduling that may consecutively issue computation-intensive tasks to the same compute unit, e.g., IP_6, our task scheduler employs Q-learning policy [8] and utilizes T_{curr},

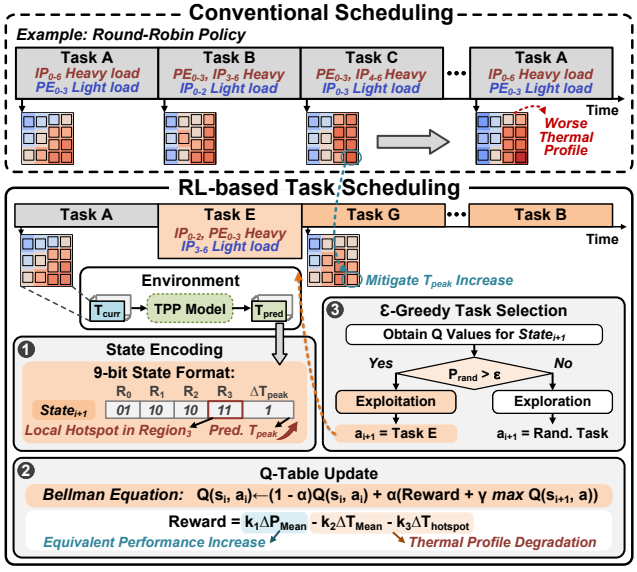
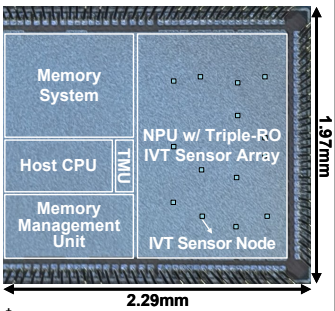


Fig. 5. Reinforcement learning-based NPU task scheduler.

P_{curr}, and T_{pred} to determine upcoming tasks. RL-based task migration has also been explored for 3D NoC systems [9]. In this work, we integrate RL-based scheduling with distributed IVT sensing and ML-based TPP in an NPU design. First, T_{pred} is encoded into a 9-bit $State_{i+1}$ to represent predicted regional and global peak temperatures (T_{peak}). The 14 sensors are divided into 4 regions based on floorplan, with regional T_{peak} converted to 2-bit encoding, plus an additional bit to indicate the temperature-change direction. This compact state encoding keeps the Q-table small enough for practical runtime deployment while preserving thermal information most relevant to scheduling decisions. Next, Q-values for the previous $State_i$ are updated using the Bellman equation with a reward function that incorporates both performance improvement and thermal profile degradation feedback. Finally, task selection employs an ε-greedy policy with exploration/exploitation transition. In the exploration phase, a random action is selected with probability ε, while in the exploitation phase, the best action is chosen with a probability of $(1 - \varepsilon)$. As the example shown in Fig. 5, the selection of Task E allows $IP_{3\text{-}6}$ to run a lighter workload. This floorplan-aware scheduling mitigates potential T_{peak} increase, enabling equivalent performance improvement.

V. Measurement Results

Fig. 6 presents the chip micrograph and specifications. We implement our thermal profiling and management system in a 22nm NPU design that achieves 3TOPS@INT8 at 0.8V and 500MHz. This NPU serves as the validation platform for both sensor characterization and closed-loop thermal management evaluation. Fig. 7 shows the measured IVT error across 11 temperatures (0-100°C) and 11 voltages (0.6-1V) for 14 sensors. Maximum σ errors are only 0.75°C at (60°C, 0.88V), 2.12mV at (100°C, 0.64V), and 1.12% relative current at (100°C, 1.0V). We also show the worst-case per-sensor errors in the line plots, where the $\pm 3\sigma$ errors of the 14 sensors are plotted as dashed lines. This small spread across sensors confirms good matching and calibration consistency for distributed deployment.



†The 22nm test chip uses heat-isolation packaging for testing purpose.

Process	22nm†
Supply Voltage	0.6V - 1.0V
Frequency	50MHz - 800MHz
NPU Components	4 PE Arrays + 7 IP Macros
NPU Performance*	3 TOPS
Sensor Node Area	40μm² (x14)
Measurement Time	0.3μs - 100μs
Scheduling Intvl.	3ms - 56ms
Thermal Management Scheme	AI-assisted Thermal Profile Prediction & Task Scheduling
Peak Temp. Reduction	3.2°C - 8.2°C
NPU Perf. Improvement	2.5% - 18.3%

*500MHz, 0.8V, sparsity = 50%, 1MAC = 2OPs.

Fig. 6. Chip micrograph and specifications.

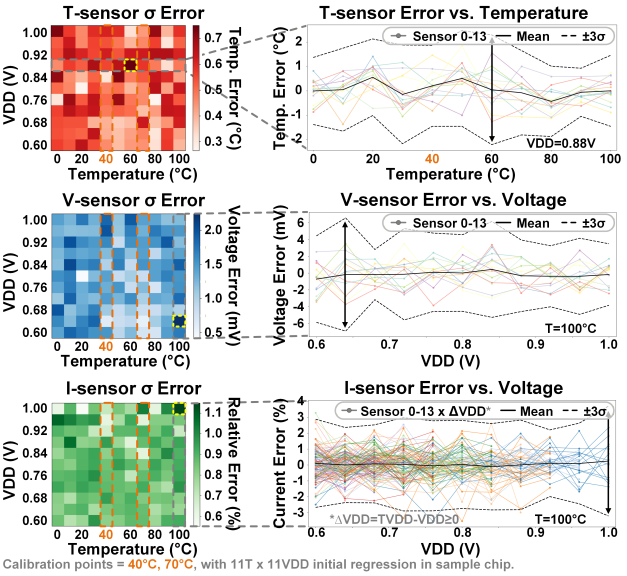


Fig. 7. IVT sensor measurement results.

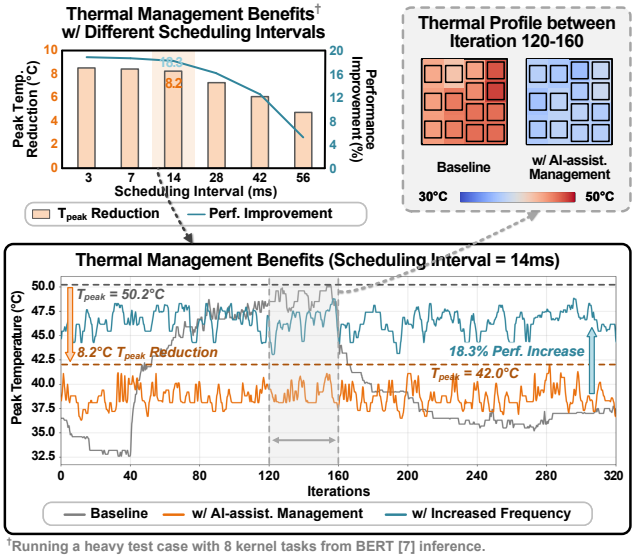
Fig. 8 shows the comparison with prior digital sensors. While prior digital sensors can only monitor current [10], temperature [5], or VT [4], our sensor is developed using a full-digital flow for simultaneous detection of three parameters (I/V/T). Our IVT sensor achieves the lowest area cost with $40\mu m^2$ and has a 65% reduction in energy per conversion compared to prior VT-sensor [4], while also providing additional current sensing capability with better accuracy than [10]. Compared to prior T-sensor [5], our sensor can detect VT with a short measurement time of $0.3\mu s$. These characteristics make the proposed sensor suitable for dense on-chip distribution and frequent runtime sampling.

Fig. 9 illustrates the benefits of our AI-assisted thermal management across scheduling intervals ranging from 3ms to 56ms. We use one heavy test case with 8 kernel tasks from BERT inference. With the scheduling interval set to 14ms, the baseline round-robin policy causes a maximum T_{peak} of 50.2°C. In contrast, our RL-based scheduler dynamically adjusts the execution order while maintaining the same total workload, reducing peak temperature by 8.2°C (orange vs. gray line). Because the workload and functionality are unchanged, the measured benefit directly reflects the advantage of thermally aware task reordering. Thermal profile comparison further shows alleviated floorplan hotspots with AI-assisted management. This T_{peak} reduction allows us to further boost the operating frequency to achieve an 18.3%

	[9]	[5]	[4]	This Work			
Process (nm)	4	16	65	**22**			
Sensor Type	I	T	VT	**IVT**			
Implementation	Fully Digital	Digital	Digital	**Fully Digital**			
Circuit Design	RO	RO	Dual RO	**Triple RO & PS**			
Sensor Area (μm²)	679.6*	350	67	**40**			
Measure. Time (μs)	1~100	10	30	0.3	**5**	30	100
Power (μW)	Not reported	11-18	26.6	96.3	**55.6**	53.4	53.1
Energy/Conv (nJ)	Not reported	Not reported	0.8	0.03	**0.28**	1.6	5.3
T Inaccuracy (°C$_{pp}$)	/	1.5	1.5	3.6	**2.3**	1.9	1.7
V Inaccuracy (mV$_{pp}$)	/	/	5.4	9.6	**8.0**	6.3	5.8
I Inaccuracy (%$_{pp}$)	8.7	/	/	8.7	**8.1**	7.8	7.7

*Including sensor processing and conversion logic.

Fig. 8. Comparison with prior digital sensors.



†Running a heavy test case with 8 kernel tasks from BERT [7] inference.

Fig. 9. Benefits of AI-assisted thermal management.

performance improvement (blue vs. orange line).

Fig. 10 shows comparisons with prior thermal management solutions. In contrast to prior processor solutions [2], [3], our design utilizes a distributed sensor network to enable fine-grained thermal monitoring and management with only 1.8% area overhead. Compared to [4], [5], which only focused on thermal sensing without management, our work demonstrates a complete solution from thermal monitoring and profile prediction to task scheduling. Our design is also the only solution deploying on-chip ML for AI-assisted TPP. Our thermal management supports scheduling iteration intervals ranging from 3ms to 56ms. Representative test cases from the BERT model are executed with a scheduling interval of 14ms, and our solution achieves 3.2-8.2°C temperature reduction and 2.5-18.3% equivalent performance improvement for representative NPU workloads [6], [7], [11]. These results verify that fine-grained sensing combined with predictive scheduling can translate on-chip thermal awareness to meaningful system-level performance gains.

VI. Conclusion

In this work, we demonstrate a complete runtime thermal profiling and management solution for NPUs. The proposed system combines a distributed monitoring network of triple-RO IVT sensors, ML-based runtime TPP, and RL-based floorplan-aware task scheduler to dynamically manage workloads. We evaluate our solution on a 22nm 3TOPS NPU design with heterogeneous compute units. The IVT sensor occupies only $40\mu m^2$ and achieves measurement inaccuracies of just 2.3°C, 8.0mV, and 8.1% for temperature, voltage and current, respectively. With this fine-grained runtime observability, the task scheduler reduces peak temperature by 3.2–8.2°C, corresponding to 2.5–18.3% equivalent performance improvement. These results show that coupling distributed multi-parameter sensing with short-horizon prediction and floorplan-aware scheduling is a practical path toward scalable thermal management in future AI accelerators.

	ISSCC'21 [3]	ISSCC'23 [2]	ISSCC'25 [5]	ISSCC'25 [4]	This Work
Process (nm)	7	4	16	65	**22**
Load Design	DSP	Mobile SoC	Synth. Load	Synth. Load	**NPU**
Total Area	Not reported	108.6 mm²	2.04 mm²	Not reported	**4.51 mm²**
Distributed Sensing	/	/	√	√	√
Sensor Type	IT	I	T	VT	**IVT**
# Sensors	3 for T	1/Core	204	12	**14**
Sensing Area (μm²)	Not reported	Not reported	81600	94800*	**31400****
AI-assisted Thermal Profile Prediction	/	/	/	/	√
Thermal Management Scheme	Task Scheduling	Task Scheduling /DVFS	/	/	**RL-based Task Scheduling**
Measure. Time	Not reported	Not reported	10μs	30μs	**0.3-100μs**
Sched. Interval	Not reported	Not reported	/	/	**3-56ms**
Temperature Improvement	Not reported	10°C	/	/	**3.2-8.2°C**††
Performance Improvement	11%†	Not reported	/	/	**2.5-18.3%**††

*Including SFP of 94000μm² **Including SPU (18400μm² for I, 13000μm² for V/T)
†Avg. perf. improvement measured across different load conditions.
††Across representative test cases w/ different intensities, at sched. interval of 14ms.

Fig. 10. Comparison with prior thermal management solutions.

Acknowledgment

This work is supported by NSFC Grant No. 92464202.

References

[1] H. Amrouch *et al.*, "On the Efficiency of Voltage Overscaling under Temperature and Aging Effects," *IEEE Transactions on Computers*, vol. 68, no. 11, pp. 1647–1662, 2019.

[2] B.-J. Huang *et al.*, "2.2 a 5G Mobile Gaming-Centric SoC with High-Performance Thermal Management in 4nm FinFET," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 40–42, 2023.

[3] V. K. Kalyanam *et al.*, "35.3 Thread-Level Power Management for a Current- and Temperature-Limiting System in a 7nm Hexagon™ Processor," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, pp. 494–496, 2021.

[4] H. S. Ozturk *et al.*, "8.7 a Dual VDD-Temperature Sensor Employing Sensor Fusion with 2.4°C, 9mV (±3σ) Inaccuracy in 65nm CMOS," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, pp. 170–172, 2025.

[5] Z. Ahmed *et al.*, "8.8 Fine-Grained Spatial and Temporal Thermal Profiling of a 16nm CMOS Buck Converter and SoC Load-Current Emulator Using Low-Voltage Micron-Scale Thermal Sensors," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, pp. 172–174, 2025.

[6] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[7] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," 2023.

[8] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998.

[9] J. Tang and J. Hong, "Reinforcement Learning-driven Task Migration for Effective Temperature Management in 3D NoC Systems," *Scientific Reports*, vol. 15, p. 11933, Apr. 2025.

[10] C.-Y. Lu *et al.*, "14.4 a Fully Digital Current Sensor Offering Per-Core Runtime Power for System Budgeting in a 4nm-Plus Octa-Core CPU," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67, pp. 260–262, 2024.

[11] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021.