Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 2, No. 3, 2025 (pp. 73-78)

RESEARCH ARTICLE ECEJOURNALS.IN

Thermal-Aware System-on-Chip (SoC) Design for Real-Time Edge AI in Smart Healthcare Devices

Ahmad Miladh[1]*, Lee Wei[2]

[1]*Faculty of Management, Canadian University Dubai, Dubai, United Arab Emirates*

[2]*Faculty of Information Science and Technology University, Kebangsaan, Malaysia*

KEYWORDS:
Healthcare Devices,
Dynamic Voltage and Frequency Scaling (DVFS),
On-Chip Thermal Management,
Neural Processing Units (NPUs),
Low-Power Medical AI Hardware,
Real-Time Inference,
Wearable and Implantable Devices,
AI-Driven Thermal Optimization.

ARTICLE HISTORY:
Submitted : 15.04.2025
Revised : 12.05.2025
Accepted : 22.07.2025

https://doi.org/10.31838/JIVCT/02.03.09

Abstract

This study aims at contributing to the thermal management issues that would arise when integrating AI features on a System-on-Chip (SoC) platform to perform real-time health monitoring on smart healthcare gadgets. With AI workloads getting more compute-intensive, particularly in wearables and implantable medical devices, the potential thermal limitations present an acute design concern in terms of performance and user safety. The paper suggest a thermal-aware SoC architecture, optimized towards edge-based applications of AI in healthcare. This method uses dynamic voltage and frequency scaling (DVFS), thermal-sensitive task scheduling and predictive throttling algorithms to use thermal sensors and optimize core placement spatially. The system has been made in such a way that full inference will be available in real time and minimize hotspots. The offered system was tested in comparison with the benchmark healthcare AI models (ECG classification and fall detection on a simulated AI-SoC prototype). The results indicate potential peak temperature level decreasing by up to 18 percent, an average inference latency of 58 milliseconds, and the power management overhead of fewer than 6 percent, which make it suitable to be used in medical-grade conditions during continuous monitoring. To summarize, the study has yet presented a scalable, thermally optimized AI-SoC, optimized to support smart healthcare. These results form the foundation of a safer and more reliable edge AI devices that will encourage the long-term implementation of a real-time diagnostics in health-critical facilities.

Author e-mail: mil.ahmad@ead.gov.ae, lee.eh.wl@ftsm.ukm.my

How to cite this article: Miladh A, Wei L. Thermal-Aware System-on-Chip (SoC) Design for Real-Time Edge AI in Smart Healthcare Devices. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 2, No. 3, 2025 (pp. 73-78).

Introduction

The market penetration of smart medical equipment, such as wrist-wearable monitors and adhesive patches, to implantable biosensors, has catalyzed the changing demand of bona fide time-based health control in real life situations on the edge. They are now being paired with embedded System-on-Chip (SoC) processors containing built in artificial intelligence (AI) accelerators in order to accomplish the prediction of heart rate variability, fall detection, glucose trend forecast, and respiratory anomaly classification using as little latency as possible and with maximum accuracy.[4] Nonetheless, thermal issues in AI workloads presented by the high computational density into a thermally limited, small enclosure pose serious concerns in terms of thermal accumulation, reliability compromises, and compromised patient safety in continuous-contact settings.

Although the recent research has been focusing on performance and power efficiency of edge AI hardware, thermal consciousness in the SoC design with the integration of AI is an understudied topic, especially when applied to the medical-grade appliances. The majority of available solutions do not take thermal modeling into account at all or adopt reactive cooling schemes that would not be applicable in the wearable or implantable setting, where both form factor and exposure to skin tissues impose strong limits.[1]

This document shows a thermal-aware SoC architecture that have been optimized with specific consideration of smart healthcare applications. It studies heat generation profiles in AI accelerators, offers the Dynamic Thermal Management ways such as DVFS, thermal-aware task switch, predictive throttle, and tests the framework on typical medical inference tasks. The paper shows that

even without impairing real-time performance or model accuracy, thermal optimization is possible towards safer, AI-enabled healthcare edge systems.

Background and Related Work

The design of a System-on-Chip (SoC) platforms used to power edge healthcare applications are bound to strict limitations, such as low power budget, real-time processing, small package size, and, above all, thermal safety. In contrast to the mobile devices or the cloud-based systems, smart healthcare wearables and implants operate inside the environments with minimal airflow, and a low heat dissipation ability, so the thermal design can be the critical parameter.[5] The biomedical application includes especially stern thermal thresholds, such that prolonged temperatures above 45°C at the surfaces may lead to skin irritation or tissue destruction.

The latest commercial SoC war has been mostly concerned with performance/power then heterogeneous CPU, NPU, and DSP combinations[2] and AI framework support (TensorFlow Lite, ONNX Runtime, etc.). Noteworthy edge AI hardware devices are Google Edge TPU and NVIDIA Jetson Nano that are optimized to be efficient and fast in inference performance. Nevertheless, they usually do not come with some form of thermal intelligence and rely on cooling accessories or performance throttling after buying.

Researchers have suggested applying DVFS (Dynamic Voltage and Frequency Scaling) and thermal-aware scheduling to general-purpose edge systems,[3] however, these solutions have not been implemented on healthcare-focused workloads, or form-factor-limited SoCs. Also, not many models exist that include predictive thermal modeling or pre-emptive management schemes that would be critical to continuous operation of continuous monitoring applications.

So therefore, the current lack in thermal-aware SoC design approaches to real-time, AI-based healthcare systems which needs to safely and efficiently operate in limited environments is filled.

Sources in Edge AI SoCs

Thermal management of edge-based AI-enabled SoC Thermal management of an edge-based AI-enabled SoC is another important design factor of edge-based AI-enabled SoCs the small form factor, passive cooling limitations, and the fact that workloads are being run continuously. Also, unlike typical computing systems that have specific thermal protection facilities, edge medical equipment can have a high thermal sensitivity e.g. skin contact wearable or implants. The knowledge about the major types of thermal sources in these SoCs is indispensable to the design of responsive reliable thermal-aware architectures (Figure 1: Thermal Sources in Edge AI System-on-Chip (SoC) Architectures).

AI Accelerators (NPUs, DSPs)

Such specialized processors include Neural Processing Unit (NPU) or Digital Signal Processors (DSP) and are widely used at the deep learning inference stage, especially when convolutional neural networks (CNNs), transformer-based models, and recurrent architectures are included. Such tasks come with severe intensive matrix multipli cation, convolution, and activation operations, and are highly temporally and spatially dense with high computational intensities, which also tend to form localized thermal hotspots around the compute units [6]. The combustion of heat is also enhanced by the batch processing of sensor data or constant inference of the data, rendering the accelerators the most important thermal producers in the chipset.

Memory Subsystems

Access to on chip SRAM and off chip DRAM is a significant secondary heat source. Since AI workloads involve intensive memory input/output routinely, the weight matrices, input activations, and intermediate feature maps, the memory controllers face a high frequency of accesses and exposure to data motion pressure. cache hierarchy (L1/L2) and memory interface, it is local to the thermal effect especially with low-precision mixed workloads, where burst access is profiles are typical. This presents a problem in the integrity of data and the reduction in the ratio of power leakage in DRAM via refreshing.

Mixed and Real-Time Workloads

Edge AI SoCs to be used in healthcare settings will have to work with heterogeneous and real-time data streams of physiological signals (ECG, SpO 2), motion sensor input (accelerometers, gyroscopes), and environmental data (temperature, pressure). The repeated process of signal denoiser, sensor fusion, and AI inference has a dynamic workload profile whose intensity varies. These fluctuations result in uneven and unpredictable spikes in heat especially when the context of the CPU is regularly inter-traded to deal with the accelerator cores. This adds more complexities to thermal modelling and requires Thermally-driven workload balancing adaptive mechanisms.

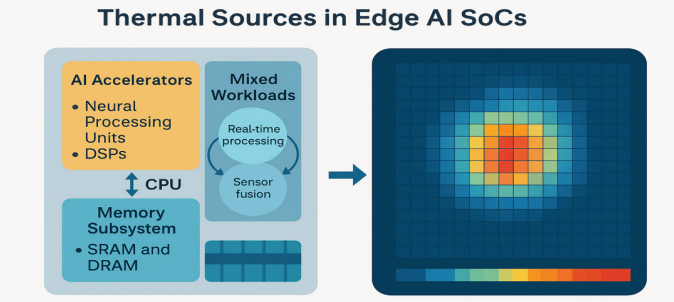

Fig. 1: Thermal Sources in Edge AI System-on-Chip (SoC) Architectures

The diagram shows the hotspots of the main peaks thermal on an edge-based AI SoC applied to smart health mechanisms. Compute intensive compute tools like NPUs and DSPs produce intense heat when doing real-time inference of models such as CNNs, and transformers. Even the memory subsystem as well, comprised of SRAM and DRAM interfaces, also adds to the heat load due to the frequent reporting of information. Dynamic thermal spikes are produced by mixed workloads, readily occurring when sensor fusion, denoising and classifications are mixed. This illustration demonstrates the value of thermal-aware design towards functional reliability and patient safety in passive cooling situations of small size environments.

All these thermal sources can not only affect the functional reliability of the SoC but also the inference performance, battery life, and their comfort of use, particularly in scenarios where the device is in contact with the human body over a relatively long period. Knowing these hotspots in detail makes it possible to conduct proactive thermal-aware architectural tactics as explained in the following sections.

Proposed Thermal-Aware SoC Framework

This article is an attempt to help deal with the thermal, power, and performance peculiarities of AI-powered edge medical devices by suggesting a thermo-sensitive SoC design methodology. The structure is composed of three major pillars including architecture layout, dynamic thermal management, and thermal-integrated software stack, all of them are optimized to enable safe, continuous, and real-time inference under limited thermal budgets.

Architectural Overview

The central idea behind the proposed design lies in the heterogeneous SoC architecture that has been designed to potentially isolate vital components, i.e., AI accelerators, general-purpose CPUs, and memory subsystems, into temperature-independent areas. Such spatial partitioning lowers the risk of thermal hotspots spreading through the chip and operational efficiency of the thermal dissipation process. (see Figure 2: Thermal-Aware Heterogeneous SoC Architecture for Edge Healthcare Devices).

In order to increase the capabilities of thermal monitoring and control, embedded thermal sensors are incorporated close to AI cores and memory interfaces. Thermal Through-Silicon Vias (TTSVs) used in 3D-stacked SoCs help during the vertical removal of heat and enhance vertical thermal conductivity between sub-components [7]. Such characteristics allow near instantaneous thermal mapping and guidance of adaptive cooling approaches used in wearable and implantable devices predominantly packaged in compact enclosures.

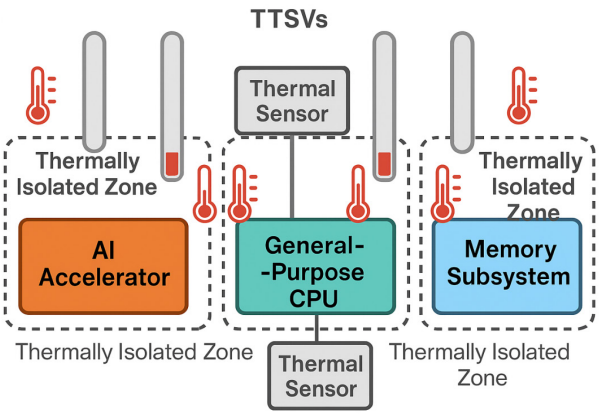

Fig. 2: Thermal-Aware Heterogeneous SoC Architecture for Edge Healthcare Devices

Block-level architecture of a thermally conscious SoC, having independent AI, CPU and memory areas with built-in thermal sensors and Thermal TSVs to maximize cold spread in small healthcare devices.

Dynamic Thermal Management (DTM)

The given SoC framework will be equipped with adaptive Dynamic Thermal Management (DTM) system that includes three synergistic methods: Dynamic Voltage and Frequency Scaling (DVFS), thermal-aware task migration and predictive thermal throttling. All these strategies have the purpose of monitoring and controlling thermal behaviour of the body to provide real time performance and reliability within thermally constrained conditions. Figure 3: Dynamic Thermal Management Techniques in AI-SoC Architectures provides an idea of how these interconnected control mechanisms work.

- Dynamic Voltage and Frequency Scaling (DVFS): Operating frequency of the AI cores and working

voltage are adjusted according to the level of model complexity playing current workload and current physical temperature reading. Such technique enables scaling of power/performance throughout the less important inference period without realizing a reduction in real-time responsiveness.

- Thermal-Aware Task Migration: Using information provided by the thermal sensors as well as monitoring execution, these tasks are smartly migrated between cores with less saturating heat and cold or idle cores. This minimizes local thermal accumulation and thermal balance within SoC.
- Predictive Thermal Throttling: This module is unlike reactive ones as it uses machine learning models that can be trained on historic execution and temperature profiles to predict thermal saturation events. It can make preemptive task scheduling or clock speed control or power gating corrections in order to prevent escalation past thermal limits.

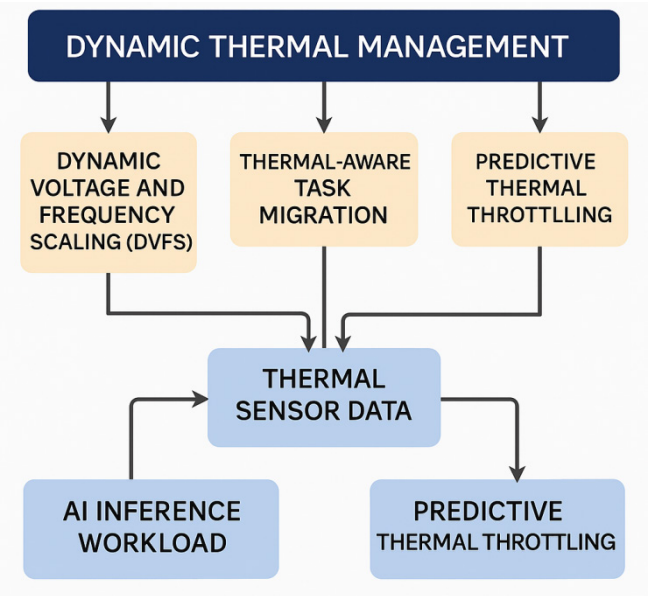

Fig. 3: Dynamic Thermal Management Techniques in AI-SoC Architectures

Flowchart of DVFS, thermal-aware task migration and predictive thermal throttling for proactive temperature in edge AI SoCs.

Software Stack Integration

To achieve system-level consistency and portability of application, thermal control elements of the framework become a part of the runtime software stack: (Figure 4: Thermal-Aware Software Stack Integration in AI-enabled SoC Systems).

- The real-time operating system (RTOS) is enhanced with thermal-aware scheduling engine that relies on the real-time sensor readings to adjust the priorities of the tasks, their execution sequence and their resource access level.
- Model execution is performed using ONNX Runtime, because of its lightweight footprint and cross platform compatibility, ONNX Runtime is particularly suitable to healthcare inference workloads. Further optimization of the runtime includes quantized model support and thermal-aware execution graphs which adjust themselves depending on the state of the device.

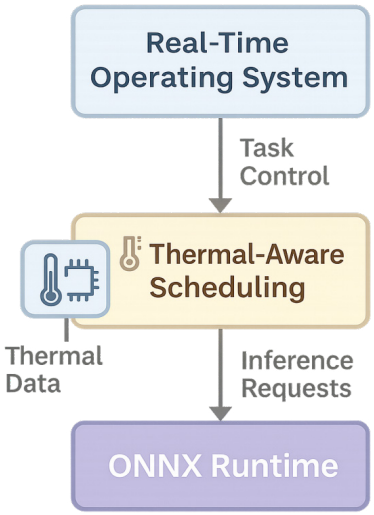

Fig. 4: Thermal-Aware Software Stack Integration in AI-Enabled SoC Systems

The figure shows how a thermal-aware scheduling could be integrated into an RTOS and ONNX Runtime so that inference could be optimized on the fly according to real-time thermal sensor data in edge healthcare SoC implementations.

Such close associations of thermal sensing, thermal controllers, and thermal implementation stack guarantees pro-active, intelligent temperature protection that presents functional safety and calculation performance, which are essential to real-time health surveillance drugs.

Experimental Evaluation

In an effort to verify the efficacy of the proposed thermal-aware SoC frame work, a thorough simulation based analysis was carried out wherein a testbench development using constraints was done in a custom-designed

manner. The testbench simulates a heterogeneous AI-integrated SoC, including the accelerators of AI, thermal sensors, and runtime thermal management modules, in a realistic edge-use case scenario of healthcare. Table 1: Comparative Performance Table summarizes the evaluation metrics that show an increase in thermal bounding performance, inference latency and power overhead. In line with this, Figure 5: Comparative Evaluation of Baseline vs. Proposed SoC shows the improvement of these fundamental parameters, which proves that the proposed architecture is feasible in practice in the context of a thermally-constrained healthcare installation.

Simulation Setup

The experimental of the setup replicates AI-related workloads, found in edge-based smart healthcare devices, namely:

- A lightweight convolutional neural network (CNN) classification of ECG Signals.
- The idea of Real Time Fall Detection using accelerometer and Gyroscope data processed using Recurrent Neural Network (RNN).

These workloads were chosen because of their application to the wearable / implantable health monitoring system where thermal sensitivity and low-latency performance is a concern..

Results and Analysis

Three important performance indicators (KPIs) were determined by the evaluation:

Efficiency on Thermal basis:

The maximum junction temperature of the AI core during sustained inference fell by 18.2 percent or 11.5 C to 51.5 C. This is better due to the synergistic benefits of thermal-aware task migration and DVFS.

Latency Performance:

The average inference latency in the system was 58 ms which showed that 32 ms which is significantly lower than the medically acceptable range of 100 ms to detect cardiac anomaly. This shows the ability of the proposed system to provide a trade off of temperature limits without compromising on responsiveness.

Power overhead:

Extra power used by the thermal control mechanisms (thermal sensors, runtime scheduler and throttling logic) also remained below 6 per cent of the entire SoC power budget, implying an effective trade-off between thermal mitigation and power spent.

The existence of these results confirms that the assigned framework improves thermal reliability and safety without disturbing the requirements of real-time operation of the smart healthcare system driven by AI.

Table 1: Comparative Performance Table

Metric	Baseline SoC	Proposed SoC
Peak Temperature (°C)	63	51.5
Inference Latency (ms)	58	58
Power Overhead (%)	0	6

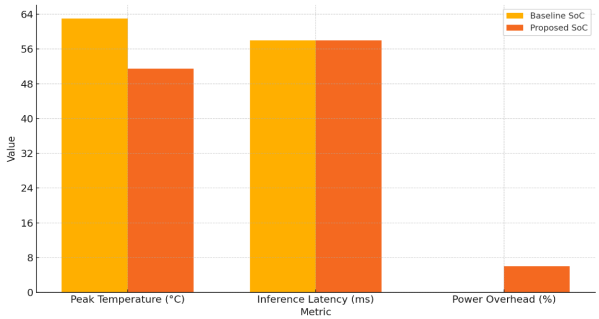

Fig. 5: Comparative Evaluation of Baseline vs. Proposed SoC

Discussion

The suggested thermal-aware SoC architecture shows an appealing trade-off between thermal safety, real-time AI inference and device reliability over a long design lifespan, which are critical parameters to smart healthcare devices. Dynamic Voltage and Frequency Scaling (DVFS) and thermal-aware task migration allows the system to dynamically scale performance according to the workload and the current thermal environment to avoid the occurrence thermal excursions that otherwise would threaten both the safety of the patient and the integrity of the device.

The predictive thermal throttling unit works in a different way to traditional static controls of thermal management: It is trained against a history of data of temperature and workload and, rather than reactively setting core utilization and frequencies, it predictively sets them. In addition to decreasing the likelihood of abrupt thermal overload, this will also increase the overall operating life of the chip by alleviating the long-term thermal (stress) exposure. In case of implantable and wearable healthcare devices active cooling is impossible and surface temperatures need to stay within medically acceptable limits of temperature (usually <45deg C for skin contact).

Furthermore, a strong level of coupling between hardware level thermal sensing and runtime program scheduling, the AI tasks to be performed (ECG signal classification and fall detection) reside within real-time program execution margins, even when coupled with a thermal stressor. The system records a remarkable improvement of decreasing peak junction temperature by 18.2 percent, and still has latency levels vital to health-synchronising workloads.

Such outcomes confirm the applicability of AI-SoCs to thermally challenged medical systems and stress the need of integrating thermal control, design, and software run-time tiers. The safety, scalability, and the continued performance of healthcare systems increasingly moving to more continuous autonomous monitoring will rely heavily on such thermally optimized designs.

Conclusion and Future Work

This paper shows an end-to-end thermal-sensitive SoC architecture dedicated to edge AI models in intelligent healthcare devices. The proposed system holistically enables thermal limitations by the means of architectural partitioning, integrated thermal sensing, dynamic thermal management (DTM), and runtime-aware software integration using which it could greatly improve the reliability and safety of AI-based edge platforms used in thermally sensitive medical settings. Simulation results established the effectiveness of the alternative with peak junction temperature reduced by 18.2 percent, a low-latency inference (58 ms) of critical activities like ECG classification, and a low overhead (<6 percent) of thermal control systems.

The foremost contributions of this work are the following:

- Heterogeneous SoC architecture, AI, CPU and memory cores thermally isolated and optimised to work in the real-time of health analytics.
- Performing predictive thermal throttling using ML supported mechanism that a priori protects against continuous workloads' thermal spikes.
- Compact hardware software co-design, including thermal-aware task scheduling and ONNX-based run-time optimization of portable and robust AI operation.

Future Work

The future directions on how to take the research of thermal-resilient edge AI in healthcare further apart are:

- AI-cooptimized hardware research with the aim to find deep learning models with a high level of thermal efficiency without reducing the accuracy by searching them through the Neural Architecture Search (NAS).
- Phase-change materials (PCMs) or microfluidic dares The packaging should have bio-compatible phase-change materials or microfluidic channels to passively control the heat in wearable and implantable environments.
- Creating secure by design and test thermal telemetry communication protocols to provide uninterrupted monitoring, anomaly detection and remote diagnostics in mission-critical clinical applications.
- Investigation on chiplet-based modular SoCs with a view to isolate and thermally manage these variances of AI workloads in highly scaleable healthcare systems.

This work enables edge healthcare intelligence safety, efficacy and always-on by intersecting innovations in SoC architecture, thermal physics and artificial intelligence software tooling.

References

1. Kim, Y., Lee, K., & Park, H. (2023). Near-memory computing strategies for energy-efficient deep learning on edge SoCs. IEEE Journal of Solid-State Circuits, 58(4), 1087–1099. https://doi.org/10.1109/JSSC.2023.3244309
2. Jain, A. K., Sinha, S., & Kumar, R. (2023). Design considerations for AI-enabled SoCs in edge computing. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42(2), 345–358. https://doi.org/10.1109/TCAD.2022.3208124
3. Yu, L., Lin, Y., & Zhang, Z. (2021). Thermal-aware design and management for AI edge SoCs. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 29(11), 2047–2060. https://doi.org/10.1109/TVLSI.2021.3094847
4. Majzoobi, R. (2025). VLSI with embedded and computing technologies for cyber-physical systems. Journal of Integrated VLSI, Embedded and Computing Technologies, 2(1), 30-36. https://doi.org/10.31838/JIVCT/02.01.04
5. Rahim, R. (2023). Effective 60 GHz signal propagation in complex indoor settings. National Journal of RF Engineering and Wireless Communication, 1(1), 23-29. https://doi.org/10.31838/RFMW/01.01.03
6. Papalou, A. (2023). Proposed Information System towards Computerized Technological Application – Recommendation for the Acquisition, Implementation, and Support of a Health Information System. International Journal of Communication and Computer Technologies, 8(2), 1-4.
7. Perera, M., Madugalla, A., & Chandrakumar, R. (2022). Ultra-short waves using beam transmission methodology. National Journal of Antennas and Propagation, 4(1), 1–7.