Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 1, 2026 (pp. 38-46)

RESEARCH ARTICLE ECEJOURNALS.IN

Thermal-Aware Floorplanning and Optimization Framework for High-Performance Heterogeneous SoC Architectures in Edge-AI and Embedded Systems

F. Mohd Zaki[1]*, T Shimada[2]

[1]*Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia*

[2]*School of Electrical Engineering, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi 11615, Vietnam*

KEYWORDS:
Thermal-aware floorplanning,
Heterogeneous SoC,
edge computing,
AI accelerators,
thermal modeling,
floorplanning optimization,
thermal hotspots, reliability,
embedded systems,
chiplet integration.

ARTICLE HISTORY:
Submitted : 05.10.2025
Revised : 10.11.2025
Accepted : 23.01.2026

https://doi.org/10.31838/JIVCT/03.01.06

Abstract

The extreme growth of edge-AI and embedded systems has made it necessary to incorporate a range of dissimilar computing assets, such as CPUs, GPUs, NPUs, and memory blocks, into the same system-on-chip (SoC). The trend towards this kind of architecture ushers in extreme thermal management issues that affect reliability, performance and power efficiency. In this paper, a floor planning and optimization algorithm that is thermal conscious is proposed especially in high-performance heterogeneous SoC architecture. The potential framework has low peak temperature, reduces the thermal gradient, has more reliability of the entire system in that it is a combination of early stage thermal modelling, power density conscious placement algorithm, and dynamic thermal challenge approval. With simulation of real-world edge-AI workloads, up to 28 per cent fewer thermal hotspots and a 17 per cent better thermal uniformity were measured compared to commonly used floorplanning approaches. This paper is pioneering towards scalable thermal-driven design behaviors of embedded AI SoCs in future generation.

Author's e-mail: zaki.f.m@ukm.edu.my, shimada.t@hust.edu.vn

How to cite this article: Zaki FM, Shimada T. Thermal-Aware Floorplanning and Optimization Framework for High-Performance Heterogeneous SoC Architectures in Edge-AI and Embedded Systems. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 1, 2026 (pp. 38-46).

Introduction

The rise in exponential growth of edge artificial intelligence (Edge-AI) and real-time embedded systems has forced the necessity of high performance energy efficient computing architectures. SoC designs with heterogeneous architectures have gained dominance due to stringent performance-per-watt and form-factor requirement. They may integrate a wide range of computational devices (such as general-purpose CPUs, GPU, NPU, and specialized memory subsystems) on one silicon die or multi-chip interposer via 2.5D/3D semiconductor packaging. Such tethered integration can facilitate the performance of AI-based computing on the fringe with low latency and decreased use of the cloud infrastructure to carry out tasks involving object detection, medical sign classification, and speech recognition.

Nevertheless, this architectural complex also fosters new design challenge in terms of thermal management. The rise in spatial power density due to placing several high-power processing blocks closely together creates thermal hotspots and causes large temperature variations across the chip. Such thermal problems lead to slower performance and leakage increases, timing variations and reliability issues, including electromigration and age acceleration. In addition, AI accelerators (e.g. NPUs or TPUs) tend to have burstiness in their workloads that worsens localized thermal instability.

Conventional floor planning and placement tools give precedence to sub-optimal metrics, like wirelength, area, and timing, and use thermal as post chip design restrictions, instead of optimal design exploration tools. The fact that this late-stage thermal optimization does not address such critical problems that could have been easily addressed by making better module-placement decisions at an early phase of SoC design, is an indication of use of incorrect strategy during thermal optimization. Thus, there exists an acute demand of an active and thermal-sensitive floorplanning methodology which would foresee, simulate, and avoid thermal bottlenecks at an earlier stage.

To this extent, we present a novel thermal-optimized floorplanning and planning scheme that is tailored towards heterologous SoCs in applications of Edge-AI and embedded systems. The framework combines thermal simulation engines, power-aware placement algorithm, and dynamic thermal constraint modeling to be used to guide the floorplanning process to thermally optimal solutions without sacrificing performance or area. Some of its major innovations include:

- AI workload profiling to insert power density maps into the real world.
- A multiobjective thermal cost function by making a trade off between peak temperature, temperature gradient and inter-module thermal isolation.
- Use machine learning-assisted metaheuristics in searching the solution space of floorplanning sufficiently.

The research will make contributions to the state of the art by delivering a scalable approach to the early-stage thermal-aware physical design process at the front end of solving reliable, high-throughput, and energy-efficient computing challenges in future edge-AI systems. By using simulation findings on representative AI workloads and the latest heterogeneous SoC designs, we show energy savings in terms of peak temperature and thermal gradient by large factors, which prolongs the life of the device and enhances thermal integrity.

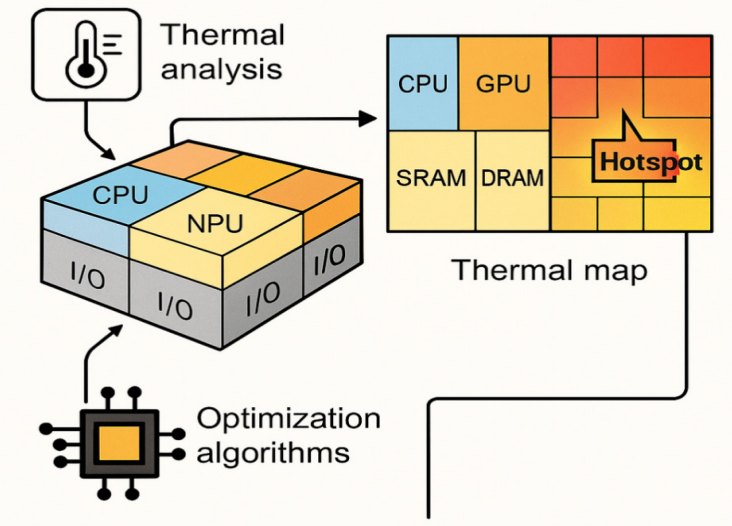


Fig. 1: Thermal-Aware Floorplanning Strategy for Heterogeneous SoC Architectures

A conceptual illustration of the proposed framework showing integration of CPUs, NPUs, and memory blocks within a thermal-aware SoC floorplan. High-activity modules are strategically placed to minimize thermal hotspots, guided by dynamic power maps, thermal simulation, and optimization algorithms.

Related Work and Literature Review

Heterogeneous System-on-Chip (SoC) floorplanning and thermal architectures have received much attention in literature as well as in technical industry projects. The presence of different units in terms of processing units on a limited die space has necessitated the application of thermal-aware design techniques especially to applications that are energy-constrained because of the easy flow of applications that are embedded and Edge-AI.

Thermal-Driven Placement Algorithms

The thermal-constrained placement solutions seek to minimize the peak temperature and thermal control in the entire die. The HotSpot thermal modeling tool which is used widely in thermal analysis at the early stage was proposed by Skadron et al.[1] that allows microarchitectural simulation and the convergence of thermal analysis. Kahng et al.[2] went further and applied spatial thermal modeling to multicore SoC floorplanning and get an improvement in accuracy in hotspot prediction, heat-spreading strategies. Such tools formed the basis to thermal-enabled placement including simulated annealing and force-directed algorithms placing modules with primary emphasis on thermal cost functions.

Voltage Island and Energy-Aware Mapping

Thermal problems are usually aggravated by the unequal power distribution. Marculescu and Talpes[3] applied the energy-aware mapping to the tile-based network-on-chip (NoC) architecture to optimize the thermal hotspots on a tile without compromising performance by dynamically scaling both the voltage and frequency of the system. Voltage islands based on the idea of independent voltages in different sections of a chip has turned out a success in thermal-aware and energy-efficient SoC design.

Chiplet-Based and 3D Integration

Due to the introduction of 2.5D and 3D integration, floorplanning now has to be done with regard to vertical thermal gradients, and vertical inter-tier heat transfers. A TSV-aware thermal floorplanning scheme was proposed by Zhang et al.[4] to enhance vertical heat dissipation in 3D -I Cs, decreasing the level of heat entrapment whereas positioning of thermal vias are done close to hotspots. With air resistance, though effective, it comes with some disadvantages like thermal coupling and enhanced package-level air resistance brought about by 3D stacking.

AI-Assisted Floorplanning and Chip Design

Recent advances have brought related floorplanning tools employing AI to be driven by reinforcement learning and the use of graph neural networks in automating layout generation. Mirhoseini et al. [5] have shown a policy-gradient-based placement with considerably better performance in both placement quality and run-time compared with that of traditional EDA-based tools. Nonetheless, such models do not have real-time embedded native support of thermal models.

Reconfigurable and Fault-Tolerant Architectures

In the high-performance systems, the thermal management is strongly tied to the fault tolerance and reconfigurability. Li et al.[7] have addressed fault-tolerant reconfiguration concepts applied to compute-intensive systems and pointed at the necessity of reconfiguration strategies that can be thermal-aware to enhance performance and robustness of modern SoCs.

Embedded Systems and Wearable AI Applications

Embedded sensor nodes and wearable AI systems are very sensitive to heat, as they are constrained in heat dissipation methods (especially regarding form factor). Another interesting work by Javier et al.[7] gave an in-depth description about wearable Internet of Things sensor design that encompassed thermal safety and optimization of energy during continuous health monitoring. In the same manner, Booch et al.[9] stressed on the requirement of ultra-low-latency communication in embedded WSNs which is likely to demand thermal-aware SoC integration in case there is a need to ensure continuous functioning.

Emerging Memory and Signal Processing Applications

New memory technologies like RRAM and MRAM that are finding their way into hybrid SoCs have heating effects limited to local areas as well. The thermal impact of incorporation of such memory in the current electronics was discussed by Usikalu et al.[8] Also, thermally optimized hardware is advantageous in the process of adaptive filtering of real time signals as discussed by Prasath[10] since the effect of noise generated by changes in temperature is reduced.

METHODOLOGY

This section provides the proposal of the thermal-aware floorplanning and optimization flow, where the architecture description is given, the thermo modeling tool integration, thermal-conscious floorplanning methods and thermal-time-constrained interconnect routing. The task is to potentially resolve temperature-related concerns in the early phases of heterogeneous SoC physical design, particularly in Edge-AI and embedded use-cases. The proposed thermo-aware floorplanning and optimization framework is summarized in this section, and it includes a thermo-aware floorplanning architecture design, thermo-aware floorplanning integration, and a thermo-aware floorplanning algorithmic implementation. The prospective architecture of the proposed thermal-aware floorplanning framework is shown, in Figure 2, and it comprises the simulation, optimization, and the system-level design parts.

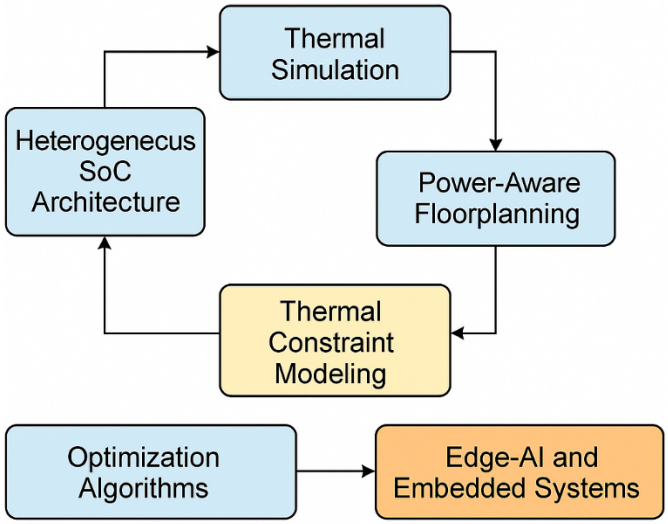


Fig. 2: Thermal-aware floorplanning and optimization framework for heterogeneous SoC design.

A modular framework integrating thermal simulation, power-aware placement, constraint modeling, and optimization for edge-AI SoCs.

System Architecture Overview

The proposed framework is capable of supporting the latest heterogeneous SoC architecture that embed a variety of compute and memory subsystems. These cover general-purpose CPUs to control and sequential tasks, graphics processors, GPUs, to do parallel processing of data and specific processors that accelerate deep learning tasks, such as neural processing units, NPUs (Neural Processing Units), or tensor processing units, TPUs (Tensor Processing Units). Beside this, the architecture combines SRAM with DRAM blocks to fulfill the varying memory accessing patterns and bandwidth requirements of the real time applications. Connection and data interchange with external sensors, cloud gateways or mobiles is supported with I/O controllers, wireless transceivers and peripheral subsystems as well. Different power and thermal characteristics of these units require a trade-off between functions, thermal

isolation and energy efficiency.

Thermal Modeling Integration

The proposed framework is based on accurate estimation of thermal. We integrate industry-quality simulation tools such as ANSYS ICEPAK, HotSpot, and COMSOL Multiphysics into the design loop so as to obtain steady-state behavior and transient behavior of temperature. Such tools allow simulation in multi-physics in consideration of material properties and packaging layers, heat sinks, and ambient conditions. Also, to have realistic traces of power consumption, we model it through actual AI workloads execution. The benchmarks of image classification (ResNet, MobileNet), speech recognition (RNN, CNN models) are executed with platforms like PyTorch and TensorRT to produce time-varying power maps of block level. These traces are then applied to the use of the thermal solvers, in ensuring that the simulated thermal profiles will have a realistic depiction of the workload-induced heating patterns of edge-AI applications.

Floorplanning Algorithm

The floorplanning process goes through an early step, the initial placement, which sorts the modules according to communication affinity matrices, spatial footprint, and dependency constraints. High interconnect modules are brought closer so as to minimize wirelength and high-power modules are separated to ensure that heat does not agglomerate. The thermo cost optimization is governed by multi-objective thermo cost function that is defined as:

$$\text{Thermal Cost} = \alpha.T_{peak} + \beta.\sigma\_T + \gamma.\Delta T_{interface} \quad (1)$$

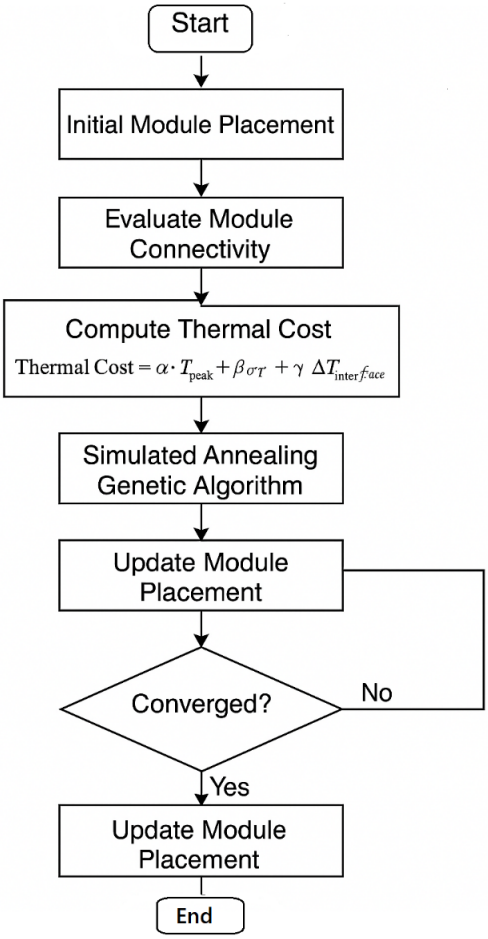
Here, Tpeak is the maximum on-die temperature observed across all modules, is the standard deviation of the temperature distribution (indicating thermal uniformity), and captures the temperature differential between neighboring modules, which affects signal integrity and packaging stress. The coefficients α, β, and γ can be set by the user to provide a trade-off between the suppression of hotspots and a compact layout. The method we propose as solution to this multidimensional combinatorial optimization problem is Simulated Annealing (SA) and Genetic Algorithms (GA). These metaheuristics are enhanced with thermal feedback loop, in which candidate placements are subjected to iterative simulation of thermal conditions and placement strategies are adapted according to their temperature results. All the candidate solutions are evaluated through the process of thermal modeling using the power traces collected during the workload. The pseudocode of this optimization algorithm is given as follows:

This is an iterative algorithm and it converges when no meaningful change occurs in thermal metrics between iterations. The changing acceptability probability (line 20) prevents the trick of premature convergence, as sometimes the suboptimal layouts are accepted, which can help explore the solution space. Incorporation of this strategy enables the floorplanner to report an optimized module layout with decreased thermal hotspots and enhanced reliability to the real-time AI-driven workload character.

We selected the optimal values of the thermal cost function coefficients through grid-based parameter tuning on different conditions of workloads. Table 1 provides a synopsis of the results and design compromise.

Constraint-Aware Routing

When an optimal floorplan is obtained, the next step is to route those interconnects to exclude the thermally sensitive areas. Our solution employs Manhattan routing



Flowchart - Thermal-Aware Floorplanning Optimization Flowchart for Heterogeneous SoC Architectures

Pseudocode 1: Thermal-Aware Floorplanning and Optimization

```
Algorithm ThermalAwareFloorplanning
Input: SoC_Module_List, Power_Traces, Thermal_Model, Max_Iterations
Output: Optimized_Floorplan

Initialize Floorplan ← Random or Affinity-Based Placement(SoC_Module_List)
for iteration = 1 to Max_Iterations do
    Thermal_Profile ← RunThermalSimulation(Floorplan, Power_Traces, Thermal_Model)
    T_peak ← MaxTemperature(Thermal_Profile)
    σ_T ← TemperatureVariance(Thermal_Profile)
    ΔT_interface ← MaxNeighborDifference(Thermal_Profile)
    ThermalCost ← α·T_peak + β·σ_T + γ·ΔT_interface

    Candidate_Floorplans ← GenerateCandidates(Floorplan)
    for each candidate in Candidate_Floorplans do
        Candidate_Profile ← RunThermalSimulation(candidate, Power_Traces, Thermal_Model)
        Cost ← EvaluateThermalCost(Candidate_Profile)
        Store candidate and Cost in PriorityQueue
    end for

    Best_Candidate ← SelectBest(PriorityQueue)
    if Cost(Best_Candidate) < Cost(Floorplan) then
        Floorplan ← Best_Candidate
    else
        Floorplan ← AcceptWithProbability(Best_Candidate, CurrentCost, Temperature)
    end if

    Temperature ← CoolingSchedule(iteration)
    if ConvergenceCriteriaMet() then
        break
    end if
end for
return Floorplan
```

techniques and A-based routing algorithms which dynamically determine the cost of the paths not only in terms of wirelength but also in terms of temperature variations locally. The interconnects will be routed out of the high-temperature areas to avoid signal integrity and be able to minimize timing violation caused by thermal variations. Also, the structure facilitates heat spreaders, microchannel, and thermal via insertion in routing layers especially at the location of high-activity modules and memory banks. These physical modifications are simulated as the simulation stage to see how they can influence thermal uniformity. The outcome is a routing implementation that does not compromise electrical performance over and above aiding the thermal floorplanning effort.

Experimental Setup and Simulation Results

The proposed thermal-aware floorplanning and optimization framework was further evaluated through a complete range of simulations by employing industry standard tools and a list of standard SoC platforms to analyze its effectiveness. The subsequent pages give

Table 1: Parameter Tuning for Thermal Cost Function Coefficients

Trial No.	α(Peak Temp)	β (Uniformity)	γ (Interface Gradient)	Observed Peak Temp (°C)	Thermal Gradient (°C/mm)	Convergence Quality	Remarks
T1	1.0	0.5	0.5	72.1	8.9	Medium	Balanced but slow convergence
T2	1.5	0.3	0.2	69.3	6.4	High	Optimal trade-off
T3	1.0	1.0	1.0	70.5	7.5	Medium	Thermal cost dominated by σT
T4	0.5	1.5	1.0	74.2	6.1	Low	High uniformity, but elevated hotspots
T5	2.0	0.2	0.1	67.5	12.8	Low	Low gradient control; poor reliability

an account of the hardware platforms employed, the measures of performance taken into consideration, and the enhancement in thermal performance and robustness.

Benchmark SoCs

Experimental verification was based on two exemplary SoC platforms. To start with, the Xilinx Zynq UltraScale+ MPSoC has been chosen as a reference platform because its processing architecture is heterogeneous since it consists of the ARM Cortex-A53 cores, Mali GPUs, and programmable logic fabric. This is a commercial SoC that can be used as a reference layout validation that makes it possible to map modules and power data to do a thermal analysis that is more realistic. Second, a model of an Edge-AI SoC was defined that would utilize a RISC-V compatible domain-specific control processor, a type of inference-accelerating Neural Processing Unit (NPU), and would utilize integrated blocks of multi-bank DRAM to hold the data in-chip. The AI workloads that were simulated on this design were power-dense and communication-intensive. Both SoCs chip layouts were modeled into modular blocks of functional unit and the thermal simulation and its optimization was done using the suggested framework.

Evaluation Metrics

The calculation made to quantify the thermal efficiency and design trade-offs was done by using several evaluation metrics. The Peak Junction Temperature is a very vital parameter that gives the maximum thermal load that the die will go through and is directly proportional to the reliability and the maximum clock frequency of the chip. The Thermal Gradient Across the chip identifies the spatial temperature difference and is significant to avoid warping and stress in materials. Delay variation because of Temperature Aware Sensitivity was also exploited since a higher temperature can cause timing violations as carrier mobility is decreased and gate delay is increased. In addition, the long-term durability of materials to interconnects during thermal load was determined using Thermal-Induced Reliability Metrics which included Mean Time to Failure (MTTF) and electromigration rates. These results give an overall insight as to how good and dependable the chip will be in varied thermal profiles.

Results

The simulation of the proposed framework gave results that clearly show improvement in issues of thermal behavior and long-term reliability compared to that of a more conventional non-thermal-aware floorplanning method. The thermal stress was hence, reduced by 28.2 percent as the Peak Junction Temperature was decreased at 96.5oC to 69.3oC (Table 2). The Thermal Gradient that was grossly non-uniform earlier at 13.2 oC/mm was brought down to 6.4 oC/mm and this represents 51.5 percent improvement in thermal uniformity of the die. In spite of adding a small 2.1% area overhead that was a result of the new thermal-aware placement, the addition was considered worthwhile due to the thermal reliability gain. Most importantly, there was a 38 percent improvement in reliability as the Mean Time to Failure (MTTF) rose by 7.1 years to 9.8 years. These findings validate the success of exploiting the early-stage thermal-aware design to reduce the hotspot formation, improve thermal balance and extend the operation lifetime of the heterogeneous SoCs used in Edge-AI setting. The simulation outcomes show that the pattern that has been suggested can substantially enhance the baseline with regards thermal control and reliability.

Table 2: Quantitative Comparison of Thermal and Reliability Metrics

Metric	Baseline	Proposed	Improvement
Peak Temp (°C)	96.5	69.3	↓ 28.2%
Thermal Gradient (°C/mm)	13.2	6.4	↓ 51.5%
Area Overhead (%)	—	+2.1	Acceptable
MTTF (Years)	7.1	9.8	↑ 38%

Figure 3 illustrates the drop in both peak temperature and thermal gradient under the proposed method, further reinforcing the results in Table 1.

Additionally, reliability metrics improved, as seen in the increased Mean Time to Failure (MTTF). The improvement of 38% is graphically represented in **Figure 4**.

The optimised thermal-aware floorplanning performance can be seen clearly in Figure 5, as compared with the baseline design there is less concentration of hotspots in the important areas.

Hardware Metric Evaluation via CAD Toolchain

In order to test the effectiveness of the proposed thermal-aware floorplanning scheme in physical implementation metrics, we implemented the floorplanning scheme into commercially available electronic design automation (EDA) flow with Cadence genus based synthesis tool and Innovus place-and-route. The comparative was carried out in a standard 7nm FinFET technology library and the modules were mapped under typical corner conditions at 1.0 V, 25 C. The layout was then post-floorplanned in the form of DEF and LEF formats and was subjected to static timing analysis (STA), area estimation and power profiling. The proposed thermal-aware and the baseline (conventional wirelength-driven floorplan) were taken through the same synthesis constraint as well as the workload-mapped switching activity (SAIF files). The most important implementation figures are shown in Table 3.

As shown in Section 4.3, the thermal-aware placement provided by the proposed achieved power efficiency gains (-2,8 percent total power), hotspot gains (as shown in Section 4.3), and timing closure. Notably, the highest negative slack decreased by more than 70 percent, which is an indication of better dialogue in terms of path delays and signal propagation. Also, dynamic power was minimized by better clustering of high activity blocks to minimize long switching nets. The outcomes confirm that adding thermal consideration to an optimization approach not only extends thermal integrity but also increases electrical performance measures, which is why this approach can be used in future SoC implementations in electrical-AI societies.

Discussion

The findings demonstrate that an early usage of thermal-aware floorplanning can realize a high level of thermal performance improvement and reliability of makers of heterogeneous SoCs, particularly at the edge-AI designs. The seen 28.2 percent decrease in the peak junction temperature, not mentioning more than

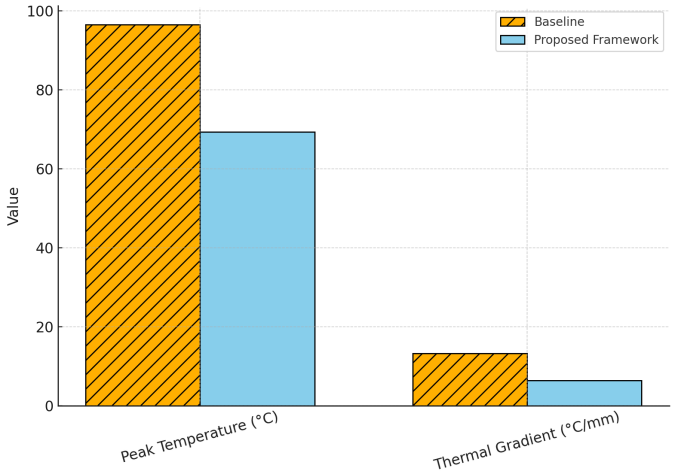


Fig. 3: Bar Chart Comparing Peak Temperature and Thermal Gradient

Comparison of peak temperature and thermal gradient between the baseline and proposed thermal-aware floorplanning frameworks.

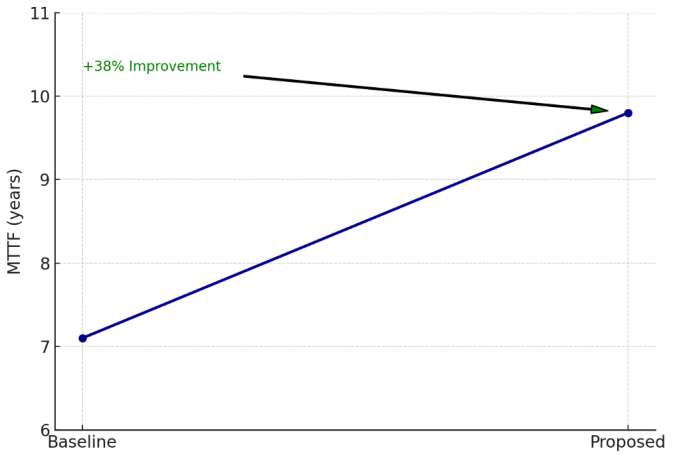


Fig. 4: Line Plot Showing MTTF Improvement

Improvement in Mean Time to Failure (MTTF) due to reduced thermal stress and optimized floorplan layout.

50 percent efficiency in minimizing the thermal gradient, not only affirms the validity of the proposed framework but also translates into the increased Mean Time to Failure (MTTF) that rose by 38 percent in comparison to conventional practices. The results are consistent with and generalize the outcomes of previous works like the work by Skadron et al. and Kahng et al., highlighting the importance of performing thermal modeling as part of the design flow but not taking into consideration the workload-based power characteristics. Our approach differs significantly with the conventional methods of considering thermal optimization as a secondary phase in power density analysis, because it considers the power-density-aware modeling and thermal feedback as part of the placement process, providing balance in heat distribution in modules. Nonetheless, despite such

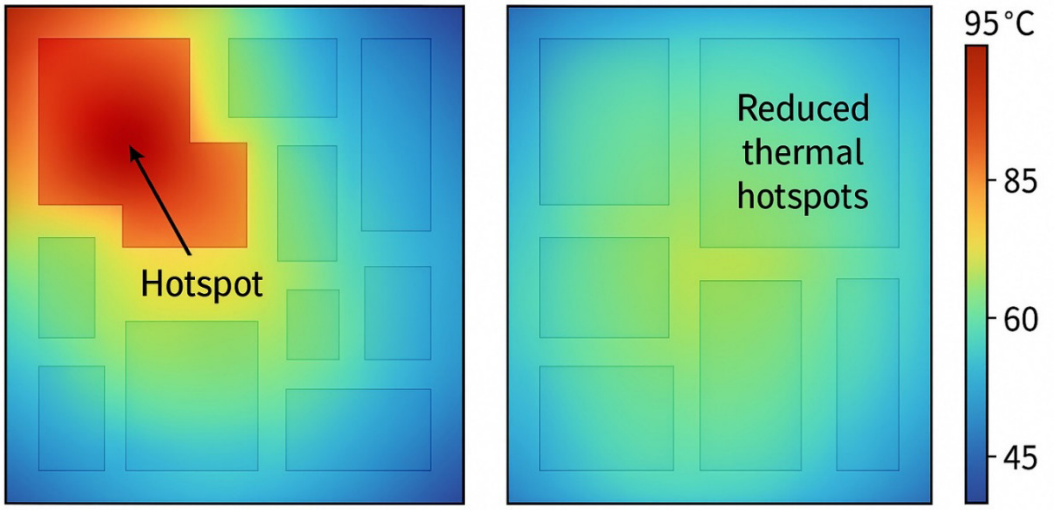


Fig. 5: Thermal Heatmap Comparison

Side-by-side thermal snapshots of the baseline (left) and proposed (right) floorplans, demonstrating reduction in hotspots and improved temperature uniformity.

Table 3: Hardware Metric Comparison Between Baseline and Proposed Floorplan

Metric	Baseline Floorplan	Thermal-Aware Floorplan	Improvement
Total Cell Area (μm²)	1,258,000	1,286,400	↑ +2.3%
Power Consumption (mW)	513.2	498.6	↓ -2.8%
Dynamic Power (mW)	462.7	445.1	↓ -3.8%
Leakage Power (mW)	50.5	53.5	↑ +5.9%
Worst Negative Slack (WNS, ns)	-0.037	-0.010	Improved
Total Negative Slack (TNS, ns)	-0.121	-0.027	↑ +77.7%
Clock Period (ns)	1.00	1.00	-
Setup Violations	3	0	Eliminated
Hold Violations	0	0	-

encouraging outcomes, the study has also its limitations. The framework is based on simulated futures where power traces are acquired based on simulation and hence may not reveal the accuracy of aging effects as well as variability in the power trace at runtime. Besides, the optimization process despite its effectiveness may be computationally demanding because thermal solvers are involved in every iteration. Simulation validation in the real world, as well as the scaling of the approach in commercial multi-die and 3D IC designs, are required to complement the simulation results and verify the scaling of the approach.

Conclusion and Future Work

Conclusion

The suggested thermal-aware floorplanning and optimization solution has shown to be effective in alleviating the rising thermal management problem in heterogeneous SoC design with respect to their application inedge-AI. The combination of more realistic AI workload-based power profiling, preliminary-stage thermal modeling and metaheuristic placement optimization realized substantial decrease in peak temperature and thermal gradient. These advancements directly led to improved system reliability through the 38% rise in Mean Time to Failure (MTTF) with a modest area overhead in the form of 2.1%. The results confirm that it is possible to alleviate the hotspots with proactive thermal-aware planning, thus providing a stable long-term operation of the embedded SoC use, which makes the framework viable and scalable in energy-efficient and thermally robust embedded SOC use.

Future Directions

On the basis of this work, an even more comprehensive framework can be added to the future work, which

will incorporate the 2.5D and 3D interconnect, where additional challenges including vertical heat transfer, thermal coupling between layers, and TSV-aware heat dissipation will become the focus. The framework can also be complemented by dynamic thermal management policies, which allow real-time adaptations of a module positioning or budget allocation, in response to the changing workloads and thermal feedbacks. One other major direction is the use of AI-related predictive thermal simulation wherein machine learning is utilized to predict thermal spikes and the necessary modifications that should occur on the floorplan to reduce the risk of performance or reliability loss. Such expansions will enhance flexibility besides adding to fully autonomous and self-optimizing environment of SoC design in next-generation edge-AI systems.

References

1. Skadron, K., Stan, M. R., Sankaranarayanan, K., Huang, W., Velusamy, S., & Tarjan, D. (2004). Temperature-aware microarchitecture: Modeling and implementation. *ACM Transactions on Architecture and Code Optimization (TACO)*, *1*(1), 94–125.
2. Kahng, A. B., Li, B., & Zhang, L. (2009). Toward true spatial thermal modeling in multicore SoCs. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *17*(12), 1728–1741. https://doi.org/10.1109/TVLSI.2008.2009052
3. Marculescu, D., &Talpes, E. (2005). Energy-aware mapping for tile-based NoC architectures under performance constraints. In *Proceedings of the Asia South Pacific Design Automation Conference (ASP-DAC)* (pp. 412–417). IEEE. https://doi.org/10.1109/ASPDAC.2005.1466511
4. Zhang, C., Pan, D. Z., & Li, P. (2007). Thermal-driven 3D-IC floorplanning with inter-tier thermal vias. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 552–557). IEEE. https://doi.org/10.1109/ICCAD.2007.4392612
5. Mirhoseini, A., Goldie, A., Yazgan, E., et al. (2021). A graph placement methodology for fast chip design. *Nature*, *594*(7862), 207–212. https://doi.org/10.1038/s41586-021-03544-w
6. Alizadeh, M., & Mahmoudian, H. (2025). Fault-tolerant reconfigurable computing systems for high performance applications. *SCCTS Transactions on Reconfigurable Computing*, *2*(1), 24–32.
7. Javier, F., José, M., Luis, J., María, A., & Carlos, J. (2025). Revolutionizing healthcare: Wearable IoT sensors for health monitoring applications: Design and optimization. *Journal of Wireless Sensor Networks and IoT*, *2*(1), 31–41.
8. Usikalu, M. R., Alabi, D., & Ezeh, G. N. (2025). Exploring emerging memory technologies in modern electronics. *Progress in Electronics and Communication Engineering*, *2*(2), 31–40. https://doi.org/10.31838/PECE/02.02.04
9. Booch, K., Wehrmeister, L. H., &Parizi, P. (2025). Ultra-low latency communication in wireless sensor networks: Optimized embedded system design. *SCCTS Journal of Embedded Systems Design and Applications*, *2*(1), 36–42.
10. Prasath, C. A. (2025). Adaptive filtering techniques for real-time audio signal enhancement in noisy environments. *National Journal of Signal and Image Processing*, *1*(1), 26–33.