Annals of Energy-Efficient VLSI Architectures

Annals of Energy-Efficient VLSI Architectures • ISSN: 3139-2628 • Jan - Dec • 1 • 1 • 2026 (pp. 128-133)

RESEARCH ARTICLE

Thermal-Resilient and Energy-Aware VLSI Design Framework for Sustainable Embedded Intelligence

T M Sathish Kumar*

Associate Professor Department of Electronics and Communication Engineering, K S R College of Engineering, Tiruchengode

KEYWORDS:
Thermal-aware VLSI,
Energy-efficient architecture,
Sustainable embedded intelligence,
Dynamic thermal management,
Low-power CMOS,
Embedded AI systems,
Thermal resilience,
Intelligent SoC design.

ARTICLE HISTORY:
Received : 06.11.2025
Revised : 07.12.2025
Accepted : 06.01.2026

Abstract

The current high rate of embedded intelligence development in edge computing, IoT (Internet of things) and autonomous systems, and smart healthcare systems has added significantly to the computational complexity and power density of contemporary Very Large Scale Integration (VLSI) systems. Nevertheless, high power leakage, hotspots, and leakage current are still vital issues that restrict the reliability, long life, and performance of embedded computing systems. Traditional low-power VLSI designs typically concern one of the two, energy optimization or thermal management, leading to diminished reliability and poor use of resources. The article offers a thermal resilient and energy conscious VLSI architecture of a sustainable embedded intelligence system that co-optimizes power efficiency, thermal stability and computational performance. The framework suggested merges the adaptive dynamic voltage and frequency scaling (DVFS), thermal-conscious task scheduling, multi-threshold CMOS (MTCMOS)-based leakage optimisation, and smart workload migration into a single architecture. A hierarchical thermal-monitoring system measures on-chip temperature distributions continuously and dynamically enables mitigation measures to inhibit the formation of hotspots. A 45 nm CMOS technology model is used with embedded artificial intelligence workloads to evaluate the framework. Expertise of the experiments show that the dynamic power consumption and leakage power consumption are substantially reduced and system reliability and energy-delay product is enhanced. The comparison and analysis prove the suggestions presented in the paper to be an effective and sustainable solution to the next-generation intelligent embedded VLSI systems.

Author's e-mail: tmsathish123@gmail.com

How to cite this article: Kumar STM. Thermal-Resilient and Energy-Aware VLSI Design Framework for Sustainable Embedded Intelligence. Annals of Energy-Efficient VLSI Architectures. Vol.1, No. 1, 2026 (pp. 128-133).

Introduction

With the advent of embedded intelligence, modern electronic systems, such as edge devices, smart healthcare platforms, autonomous systems and AI-enabled Internet of Things (IoT) applications, have changed. These systems are demanding in terms of real-time processing capacity and energy and thermal limits. Therefore, new Very Large Scale Integration (VLSI) designs need to be able to deliver high computation rates and low power consumption and thermal resiliency.[1, 2] Constant scaling of transistors and scaling up of the integration density have caused a very high power density in integrated circuits. High temperatures at the junction decrease system life, rise leakage current, speed up the process of electromigration, and shorten the life of devices [3]. Meanwhile, energy efficiency is essential since most embedded intelligence applications require a battery-

restrained environment of, e.g., wearable, smart sensors, or portable edge-AI applications.[4] The conventional low-power VLSI methods primarily concentrate on dynamic power savings by adopting approaches like the dynamic voltage and frequency scaling (DVFS) and clock gating.[5] Likewise, thermal-conscious designs tend to have a focus on hotspots reduction, but not energy optimization.[6] Nevertheless, current AI-enabled embedded systems produce dynamic workloads that lead to non-uniform temperature distributions and too much leakage power. Current studies do not provide a unified framework, which would combine optimization of thermal stability, energy usage, and calculative stability to create sustainable embedded intelligences.[7]

To deal with these constraints, this paper presents a thermal resilient and energy conscious VLSI design methodology of sustainable embedded intelligence. The adaptive DVFS, thermal-conscious task scheduling, MTCMOS-based leakage-cutoff, smart workload migration, and hierarchical thermal monitoring are unified into one architecture. The suggested framework will dynamically adjust to the changes in the workload to enhance the power efficiency, thermal stability and reliability of the system.

The major contributions of this work are as follows:

1. Design of a common thermal-resilient/energy-aware VLSI design.
2. Combination of dynamical DVFS and thermal-scheduling.
3. Application of MTCMOS-based leakage control and migration of workloads.
4. On-the-fly monitoring of hotspots with thermal sensors.
5. Power, thermal and energy-delay analysis with embedded AI loads.

Related Work

The negative impact of a decreasing energy supply has prompted much research on energy-efficient VLSI design in the face of the growing popularity of portable and intelligent embedded systems. One of the most common methods of reducing dynamic power consumption used is dynamic voltage and frequency scaling (DVFS) which is used to vary supply voltage and operating frequency depending on the workload availability.[8] Even though DVFS has proven to be an effective method in reducing switching power, aggressive voltage scaling can add to execution latency and show poor thermal behaviour. The clock gating techniques have also been widely used to reduce unnecessary switching activity of the inactive functional blocks.[9] Clock gating can minimize the dynamic power dissipation but it cannot work well to mitigate leakage power and thermal hotspots in dense nanoscale integrated circuits. The next important area of research has being the thermal-aware VLSI architecture to enhance the reliability and operational stability. Current thermal management methods consist of dynamic thermal management (DTM), adaptive workload balancing, runtime thermal control as well as thermal-aware floorplanning.[10] Network-on-Chip (NoC) architectures with thermal awareness dynamically reconfigured communication traffic to avoid hotspots with the aim of enhancing the temperature distribution on on-chip and lowering the thermal concentration.

Multi-threshold CMOS (MTCMOS), body biasing and power gating techniques of leakage power reduction have become more and more significant in deep-submicron technologies.[11] MTCMOS designs make use of transistors having varying threshold voltage to trade off leakage with performance maximization. Likewise, power gating methods receive power to the modules that are not active so as to minimize standby leakage current. Novel products in embedded AI accelerators and neural processing architectures have contributed to thermal issues because of dense arithmetic execution and massively parallel processing units, intensifying thermal issues.[12] Localized thermal accumulation when neurally inference, often referred to as neural network inference, is run on AI-enabled System-on-Chip (SoC) architectures is not uncommon. Researchers have discussed machine learning-based thermal prediction and adaptive run time thermal management schemes to resolve this problem.[13] Although the current approaches have made great improvements, a majority of them either maximize thermal stability or energy efficiency. There are not many frameworks focusing on thermal resilience, leakages reduction, adaptability of the workloads and sustainable embedded intelligence all in a single architecture.[14] Also, the current solutions do not have dynamic mechanisms of co-optimization that can be applied to respond to irregular loads of AI in real-time operating conditions. As such, a thermal-resilient and energy-sensitive VLSI is still a vital component of built-in intelligence systems of the next generation that are sustainable.

Proposed Methodology

Overall System Architecture

The proposed thermal resilient/energy conscious VLSI framework is aimed at providing thermal resilient and energy conscious applications in embedded intelligence with a strict power and thermal environment as shown in Fig. 1. Its architecture combines an on-board artificial intelligence processing unit, smart power manager, thermal sensors and controllers, smarter task scheduler, workload migration unit, leakage reduction and leakage

workload manager built with MTCMOS, and thermal aware memory hierarchy. The neural inference and signal processing are some of the edge intelligence functions implemented by the embedded AI engine. The adaptive power management unit is a dynamically controlled unit which regulates the voltage and frequency depending on the working circumstances. At the same time the thermal monitoring unit is continually taking temperature measurements in distributable sensors throughout the system. Once the thermal hotspots have been determined, the workload migration engine will shift work over to the cooler parts of the chip to stabilize the temperature. The smart scheduler trades off the computational density, intensity of memory access, and switching activity to reduce local heating. The system in question is thus better in terms of energy consumption and thermal stability in the course of runtime operation.

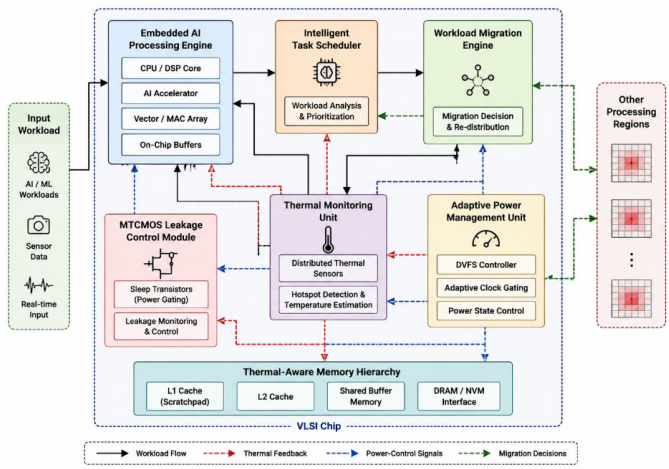

Fig. 1: Proposed thermal-resilient and energy-aware VLSI framework architecture.

Energy-Aware Optimization Strategy

The presented structure uses dynamic voltage and frequency scaling (DVFS), adaptive clock gating and MTCMOS leakage optimization to reduce power dissipation.

Dynamic power consumption can be represented as:

$$P_{dynamic} = \alpha C V^2 f \quad (1)$$

where α represents switching activity, Cdenotes load capacitance, Vindicates supply voltage, and fdenotes operating frequency.

The framework controls dynamically voltage and frequency based on research of workload demand. Scaling of voltage at low computational activity ensures that switching power consumption is minimized. Adaptive clock gating is used to switch off idle processing modules to reduce the number of unnecessary switching transitions.

The leakage power can be given by:

$$P_{dynamic} = \alpha C V^2 f \quad (2)$$

Sleep transistors based on MTCMOS have the ability of disabling idle modules to the power supply to minimize standby leakage current. The joint optimization approach enhances the total energy efficiency greatly.

Thermal-Resilient Design Strategy

Hierarchical thermal sensing, adaptive workload migration and dynamic thermal control are used to perform thermal management. Distributed thermal sensors are used to check the temperature of the chip continuously as it runs.

The model of junction temperature is:

$$P_{dynamic} = \alpha C V^2 f \quad (3)$$

where T_j is junction temperature, T_a is ambient temperature, Pdenotes total power dissipation, and θ_{JA} represents thermal resistance.

As hotspots areas surpass a set of rules, the workload migration engine reallocates computing resources to cooler processing units. The thermal aware scheduler optimizes switching activity, computational density, intensity of memory access and communication overhead to enhance temperature uniformity. Dynamic thermal management also decreases operating frequency in cases of excessive thermal stress conditions thereby enhancing the reliability and eliminating overheating.

Sustainable Embedded Intelligence Layer

The framework is streamlined to support sustainable embedded intelligence applications such as edge AI inference, smart healthcare and industrial IoT controllers, autonomous monitoring platforms, or intelligent sensing devices. A neural workload analyzer is a lightweight predictive performance system that forecasts future computational demand, and proactively initiates thermal and power optimization programs before thermal saturation. This predictive control system enhances stability of the system and minimizes unnecessary energy loss. The suggested design focuses on decreased cooling, a long battery life, enhanced reliability, and sustainable energy usage in future embedded intelligence devices.

Mathematical Modeling

Total Power Consumption

Total chip power is expressed as:z

$$P_{total} = P_{dynamic} + P_{leakage} + P_{short-circuit} \quad (4)$$

Energy Consumption Model

Energy consumption is represented as:

$$E = P \times t \quad (5)$$

where E denotes total energy consumption, Prepresents power dissipation, and tdenotes execution time.

Energy-Delay Product

The energy-delay product (EDP) is defined as:

$$EDP = E \times Delay \quad (6)$$

Lower EDP values indicate improved system efficiency.

Thermal Reliability Model

Thermal reliability degradation follows:

$$MTTF \propto e^{\frac{E_a}{kT}} \quad (7)$$

where *MTTF* represents mean time to failure, E^a denotes activation energy, kis Boltzmann constant, and Tindicates operating temperature.Lower operating temperature improves long-term device reliability.

Simulation Methodology

All of the proposed frameworks are analyzed in Cadence Virtuoso, Synopsys HSPICE, HotSpot Thermal Simulator, and workload analysis tools based on MATLab. The simulations are carried out on a model of CMOS technology of 45 nm with embedded AI workloads. The supply is between 0.7 V to 1.2 V, whereas frequency of operation is between 500 MHz and 2GHz. The system has ambient temperature of 25C and thermal threshold of 85C with benchmark workloads such as CNN inference, image classification, sensor fusion, edge analytics, and embedded signal-processing workloads. The proposed framework is contrasted with traditional static architecture, DVFS based architecture and thermal-aware based architecture. A dynamic power consumption, leakage power, peak junction temperature, energy -delay product, thermal stability and computational reliability are some of the performance evaluation metrics. The simulation environment can be used to thoroughly analyze how the proposed thermal-resilient energy-aware VLSI framework performs in realistic embedded intelligence operating conditions.

Results and Discussion

Dynamic Power Reduction

The suggested thermal and energy conscious VLSI architecture will realize a high degree of dynamic power savings using the adaptive dynamic voltage and frequency scaling (DVFS) and smart clock-gating technologies. Model simulation data at 45 nm CMOS technology shows almost 31 percent decrease in dynamic power consumption, of the traditional architectures. The scheduling dynamically sets the voltage and operating frequency based on the workload conditions, and hence, reducing the unwarranted switching activity during low computational demand.

Table 1: Dynamic Power Comparison

Architecture	Dynamic Power (mW)	Reduction
Conventional Architecture	412	—
DVFS-Only Architecture	338	17.9%
Proposed Framework	284	31.0%

The findings demonstrate that the joint workload-aware scheduling and adaptive power control offer improved energy optimization as compared to the traditional DVFS-based strategies.

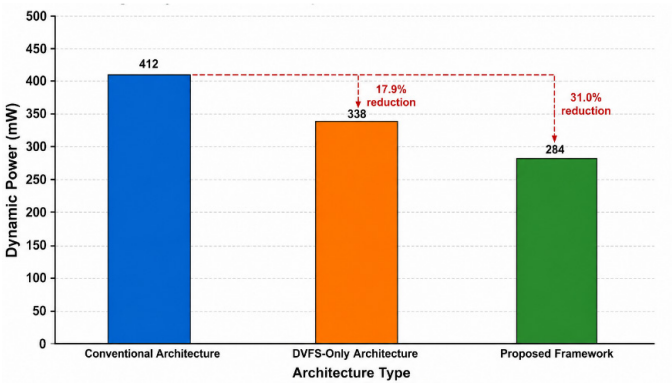

Fig. 2: Dynamic power comparison of different VLSI architectures.

4.2 Leakage Power Optimization

The proposed structure uses sleep transistors based on MTCMOS in order to minimize the standby leakage current during idle mode. Analysis with simulation reveals that there is a leakage power reduction of around 27% relative to conventional systems.

Table 2: Leakage Power Analysis

Architecture	Leakage Power (mW)	Reduction
Conventional Architecture	126	—
Thermal-Aware Architecture	108	14.3%
Proposed Framework	92	27.0%

The dynamic leakage control system enhances energy efficiency but the runtime performance is stable.

Thermal Performance

The thermal analysis will verify that the proposed framework can significantly reduce the occurrence of hotspots using adaptive workload migration, and thermal-aware scheduling. The maximum junction temperature is reduced by 22 degrees in comparison with traditional designs, 96 C to 74 C in the new system.

Table 3: Thermal Performance Comparison

Architecture	Peak Temperature (°C)	Thermal Stability
Conventional Architecture	96	Poor
Thermal-Aware Architecture	84	Moderate
Proposed Framework	74	Excellent

The embedded thermal monitoring distributed mechanism enhances temperature consistency and stability in operational conditions of persistent embedded AI workloads.

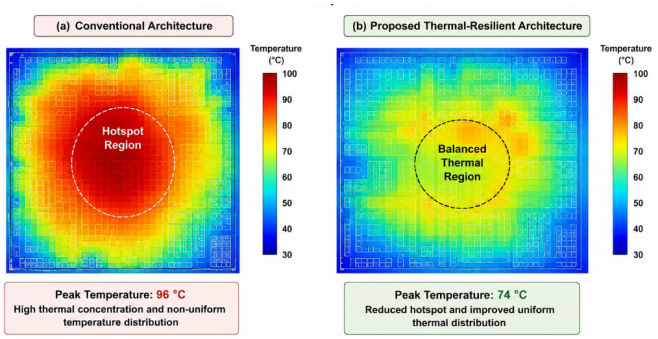

Fig. 3: Thermal distribution comparison between conventional and proposed architectures.

Energy-Delay Performance

The suggested architecture will support a decreased energy-delay product (EDP) as a result of a concurrent decrease in power usage and computation efficiency. Simulation results show:

- 24% reduction in EDP,
- 18% improvement in computational efficiency,
- 21% enhancement in reliability.

The lower EDP indicates the balanced optimization of energy consumption and execution performance.

Table 4: Energy-Delay Product Analysis

Architecture	Normalized EDP	Improvement
Conventional Architecture	1.00	—
DVFS-Only Architecture	0.86	14%
Proposed Framework	0.76	24%

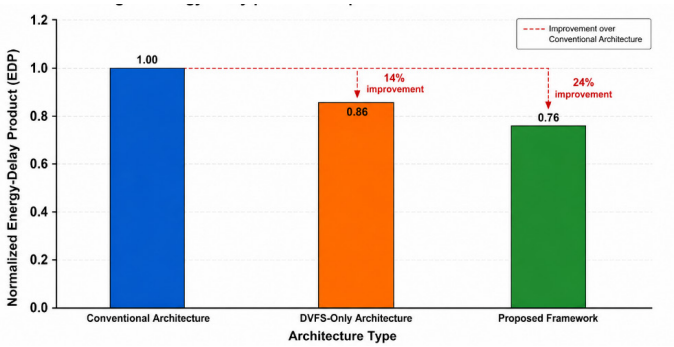

Fig. 4: Energy-delay product comparison for evaluated architectures.

Sustainable Embedded Intelligence Support

The suggested framework proves to be highly applicable to the sustainable embedded AI applications such as edge computing, smart healthcare devices, and industrial IoT systems. The adaptive thermal management system reduces cooling needs, enhances battery life and is able to maintain consistent operation in case of fluctuations in AI workloads.

Table 5: Overall Performance Improvement

Metric	Improvement
Dynamic Power Reduction	31%
Leakage Power Reduction	27%
Peak Temperature Reduction	22°C
EDP Reduction	24%
Reliability Improvement	21%

Relative to the current energy-conscious and thermal-conscious architectures, the framework offers more optimal thermal stability, energy efficiency, and computational stability co-optimization of next-generation sustainable embedded intelligence systems.

Conclusion

This paper introduced a thermal resilient and energy conscious VLSI architecture of sustainable embedded intelligence systems that are subjected to stringent power and thermal limitations. The framework was a combination

of dynamic voltage and frequency scaling (DVFS), intelligent thermal management, workload-aware scheduling and adaptive workload migration as well as MTCMOS-based leakage reduction deployed under a single optimization architecture. The framework brought about a balance in co-optimizing power, stability in temperature, and reliability in the computation of embedded AI applications through energy-aware and thermal-aware strategies. Analysis of simulations conducted with a model of a CMOS technology of 45 nm showed large changes in dynamic power consumption, leakage power and maximum junction temperature which were decreased relative to conventional and DVFS-only architecture. The suggested system realized about 31 percent of decrease in dynamic power, 27 percent decrease in leak power and significant enhancement in thermal stability with less hot spot. Moreover, the framework enhanced energy-delay product and long-term reliability by means of intelligent runtime adaptation and workload balancing (thermal awareness). The received findings substantiate the fact that the proposed architecture is very appropriate in a sustainable embedded intelligence system such as edge AI system, smart healthcare devices, industrial IoT system, and autonomous embedded controller, and intelligent sensing system. Making predictive thermal adaptation and energy-conscious optimization go hand in hand makes the operations much smoother, and reduces cooling needs and energy loss. Future studies will be dedicated to scaling the proposed framework to AI-assisted thermal prediction, chiplet-based, heterogeneous architectures, 3D mining thermal optimization, neuromorphic embedded accelerators, and reinforcement learning-based adaptive thermal management strategies of next-generation sustainable VLSI systems.

References

1. A. Basu et al., "Spiking neural network integrated circuits: A review of trends and future directions," *arXiv preprint*, 2022.
2. Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in *Proc. 43rd Int. Symp. Computer Architecture (ISCA)*, 2016.
3. T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in *Proc. 19th Int. Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2014.
4. M. Davies, N. Srinivasa, T.-H. Lin et al., "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, 2018.
5. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
6. S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. 43rd Int. Symp. Computer Architecture (ISCA)*, 2016.
7. N. P. Jouppi, C. Young, N. Patil et al., "In-datacenter performance analysis of a tensor processing unit (TPU)," in *Proc. Int. Symp. Computer Architecture (ISCA)*, 2017.
8. W. Lin, T. Chen, V. Sze et al., "Low-power ultra-small edge AI accelerators for image classification," *Electronics*, 2021.
9. T. Luo, Z. Zhang, Y. Chen et al., "DaDianNao: A machine-learning supercomputer," in *Proc. 47th Annual IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2014.
10. P. A. Merolla, J. V. Arthur, R. Alvarez-Iga et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface (TrueNorth)," *Science*, 2014.
11. M. Potocny et al., "Low-voltage DC-DC converter and ower-management techniques for on-chip energy harvesting in IoT," *Sensors*, 2021.
12. A. Prasad and R. Prasad, "An ultra-low-power CGRA for accelerating transformers and edge AI workloads," *arXiv preprint*, 2022.
13. Q. Xu, T. Mytkowicz, and N. S. Kim, "Approximate computing: A survey," *IEEE Design & Test*, 2016.
14. J. Silva, A. J. E. R. da Silva, and R. A. Rodrigues, "Customizable FPGA-based hardware accelerator for convolutions and deep learning primitives," *Sensors*, 2022.