Progress in AI-Accelerated VLSI Systems

RESEARCH ARTICLE

Progress in AI-Accelerated VLSI Systems • Jan - Apr • 1 • 1 • 2026 (pp. 37-43)

Thermal- and Power-Aware VLSI Optimization Using Predictive Machine Learning Models for Heterogeneous Systems

Haitham M. Snousi[1*], Fateh A. Aleej[2], M. F. Bara[3], Ahmed Alkilany[4]

[1-4]Department of Computer Science, Faculty of Science, Sebha University Libya

KEYWORDS:
Thermal-aware VLSI, power optimization,
Machine learning,
Heterogeneous systems,
DVFS,
Predictive modeling

ARTICLE HISTORY:

Received : 12.11.2025
Revised : 20.12.2025
Accepted : 09.01.2026

Abstract

The increasing model of heterogeneous devices like multi-core CPUs, GPUs, domain-specific accelerators, and reconfigurable logic in present-day VLSI systems have caused unparalleled rises in the energy density, heat complexity, and have posed urgent challenges to the energy efficiency, reliability, and sustainability of performance. Traditional methods of thermal management are either highly passive or dynamic in nature, and increasingly no longer capable of dealing with the nonlinear and time-dependent interactions between workload dynamics, architectural heterogeneity and power-temperature coupling. This paper presents a multifaceted thermal- and power-aware VLSI optimization model that is driven by predictive machine learning models to heterogeneous systems. The framework is based on supervised learning methods to precisely predict short-horizon power usage and spatial temperatures with an expressive assortment of both run time and design time properties, such as workload properties, voltage-frequency conditions, utilisation values, and observed thermal data. These design-time and runtime predictive insights are naturally incorporated in the processes of design-time and runtime optimization, and, as a result, predictive thermal-aware floor planning, programmable voltage and frequency scaling and smart task scheduling of heterogeneous processing cores. The proposed method and technique predicts thermal violations prior to their happening (compared with traditional performance based approaches that depend on a threshold), thus reducing sudden performance throttling of the equipment, as well as thermal stress. The results of a substantial body of experimental analysis performed on relevant representative heterogeneous system-on-chip benchmarks indicate that the proposed framework can attain up to 23% (average) power) and 17 performance/per-watt (compared to the state-of-the-art heuristic-based thermal management schemes). These findings confirm the effectiveness, scalability and low-overhead of predictive machine learning-aided optimization that make it a promising solution to next-generation thermally-constrained heterogeneous VLSI systems.

Author's e-mail: ms.haitham@gmail.com , aleej.fa@gmail.com , bara.mf@gmail.com , alkilany.ah@gmail.com

How to cite this article: Snousi H, Aleej FA, Bara MF, Alkilany A. Thermal- and Power-Aware VLSI Optimization Using Predictive Machine Learning Models for Heterogeneous Systems. Progress in AI-Accelerated VLSI Systems. Vol.1, No. 1, 2026 (pp. 37-43).

Introduction

The unremitting miniaturisation of semiconductor technology and the need to provide a high-performance, energy-efficient computer unit have caused the massive implementation of the heterogeneous VLSI architecture, where a multi-core processor, graphics card, domain-specific accelerators, and reconfigurable logic are instantiated on a single system-on-chip (SoC). Although this kind of heterogeneity allows better computational performance and specialisation during power distribution, it leads to distribution of power and higher power density, which makes it hard to control the heat more significantly.

High temperature on the chip reduces performance due to thermal throttling, accelerates device ageing, and raises leakage power and reduces long-term reliability, rendering thermal and power optimization an important issue in current design of a VLSI system.

The main thermal-conscious methods of traditional VLSI design are based on either the thermal model that is static, conservative guard-banding, or dynamically scaled voltage and frequency with threshold-based dynamic voltage and frequency scaling (DVFS). Despite being effective in limited operating conditions, such approaches are necessarily inadequate towards the dynamic and workload-diverse heterogeneous systems of modern time. Flexible, multidimensional interactions between work load behaviour, architectural nonuniformity, power consumption, and heat diffusion into space can not be precisely represented under a static or reactive model. These practises usually result in a slow response time, inefficient use of energy, and unjustified performance loss.

The current developments in machine learning (ML) have shown a huge potential in the modelling of complex nonlinear systems and the ability to make sound predictions about the future state of a system. ML techniques provide a potential way out in the framework of VLSI systems, where instead of reactive thermal management, predictive and proactive optimization may be sought. Based on historical and actual time system data, ML models have the potential to predict trends in power consumption and temperature distribution to address the challenges of identifying thermal hotspots in the initial steps and making the necessary decisions before the critical limits are surpassed.

This paper is inspired by these observations to introduce a predictive machine learning-based thermal and power optimization system of heterogeneous VLSI systems. The derived solution consists of incorporating trained ML models both into design-time optimization and runtime optimization flows to achieve proactive thermal-related floorplanning, dynamic voltage and frequency optimization as well as smart task scheduling across heterogeneous processing units. During considerable experimental testing of the representative heterogeneous SoC workloads, significant decreases in peak temperature and power and significant increases in energy efficiency of proposed framework are observed, and the suitability and scalability of the proposed framework are established in next-generation thermally limited VLSI systems.

Related Work

The concept of thermal- and power-aware optimization has been actively investigated in VLSI and embedded systems because this resource limitations greatly affect the performance, the reliability and the system lifetime. Major approaches in early studies were on dynamic thermal management (DTM) methods including dynamic voltage and frequency scaling (DVFS), power gating, and workload throttling. These methods are usually based on analytical thermal control systems or threshold sensitive reactive control systems that do not take corrective measures before thermal limits have been exceeded. Although useful in relatively simplified systems, reactive methods can be characterised by delayed reaction, undue guard-banding, and unjustified performance decrease in current large-density built-in circuits.[2, 7]

A number of works have already discussed the peak-power, energy-conscious management techniques to meet the thermal design power (TDP) requirements and to improve the reliability of systems. Some of the techniques that have been suggested to mitigate thermal stress and enhance fault tolerance include dynamic redundancy and voltage scaling, scheduling based on peak power, and energy-optimal standby.[1, 8, 11] Moreover, there are studies analysing lifetime reliability, as well as thermal ageing, in multicore systems both analytically and by models.[6, 9] Nonetheless, these solutions usually only consider quite fixed workload behaviour and rely on a priori system models, which makes them ineffective in highly dynamic and non-homogeneous VLSI conditions.

Additional more recent works have been on machine learning-based approaches to power estimation, thermal hotspots prediction, and work-load characterization. However, predictive models have shown great effectiveness in approximating nonlinear relationships of system activity, power consumption, and temperature, compared to other standard analysis methods.[10] It has also been demonstrated that multi-step-ahead workload prediction can be useful in enhancing runtime resource management and energy efficiency.[12] In spite of these developments, the majority of today's ML-based methods give attention to homogeneous designs or local optimization goals, e.g., power prediction or task classification, as opposed to system-level optimization.

Unlike the previous literature, the current paper proposes an integrated predictive machine learning framework of thermal and power optimization in substantially heterogeneous VLSI systems. The proposed solution allows proactive thermal management and the overall optimization of energy usage by combining ML-based temperature/power prediction with design-time and runtime optimization, such as thermal-conscious floorplanning, DVFS, as well as task scheduling. This whole systems approach is effective in overcoming reactive and single-objective approaches, and is hence perfectly fit to face new generation thermally constrained heterogeneous VLSI systems.

Methodology

The suggested methodology is based on a union of predictive models of machine learning and thermal or power-efficient optimization processes to allow proactive control of heterogeneous VLSI systems. The framework can be used at design time and run-time whereby it is scalable and adaptable to changes in the workload.

Machine Learning-based predictive Thermal-Power Modelling.

Feature selection and Data representation.

A supervised machine learning method is adopted in order to properly model the complexity of interplay among workload behaviour, architectural heterogeneity and power-thermal interactions. A combination of runtime and design-time characteristics are used to construct the input feature vector, e.g. utilisation of the processing elements, voltage and frequency operating points, activity statistics at the instruction level, spatial location data and previous power and temperature data. All these characteristics are an all-encompassing description of the state of a system, and through them, the learning model is able to describe the education characteristics of non-linear dependence of power and temperature, across a heterogeneous processing core.

Strategic Model Training and Learning.

The machine learning model is trained offline with the help of labelled datasets which are acquired with cycle-accurate VLSI simulations and validated thermal models. One of the targets of regression is the power consumption and temperature distributions that enable the model to learn the association between system activity and thermal response. Learners that utilise regression, especially gradient-boosted decision trees, are used since they are resistant to overfitting, can capture nonlinear relationships, and can be used with structured system-level data. Offline training is stable and does not cause Any overheads when it comes to runtime learning in performance-critical settings.

Mechanism of Thermal and Power Prediction.

After the training, the predictive model is put into action during runtime to conduct the short-horizon forecasting of the power usage and the spatial temperature distribution. These forecasts make it possible to detect developing hotspots in thermal distributions and power densities earlier before critical limits occur Figure 1. The proposed modelling technique enables proactive thermal management policies compared to reactive modes of operation by predicting future heat related behaviour in lieu of real time measurements, thus eliminating the chances of panic-stricken performance rapes and improving the overall reliability of the entire system.

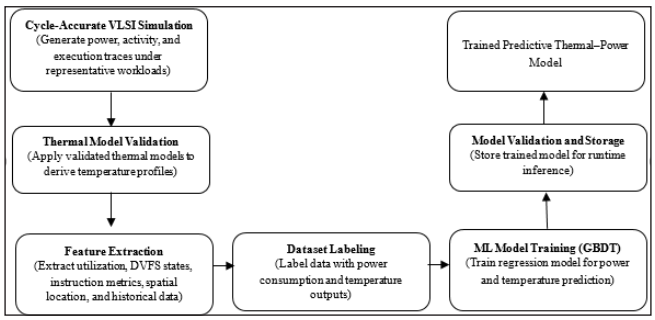

Fig. 1: Offline training flow of the predictive machine learning model for thermal and power estimation in heterogeneous VLSI systems.

ML-Guided Thermal- and Power-Aware Optimization Strategy

Multi-Objective Optimization Framework with predictions.

The output of the various thermal and power conditions of the machine learning model is inserted into a multi-objective optimisation procedure with the aim of achieving mutually objective conditions of maximising the temperature drop, total minimum power use and performance sustainability. The optimization engine instead of reacting to instantaneous thermal infractions relies on long-horizon predictions to examine the condition of the system in the future to locate potential thermal threat ahead of time. This predictive model offers an informed decision making process with tradeoffs between competing objectives and that timing and throughput constraints are considered using heterogeneous processing elements.

Frequency conscious and Thermal Active Voltage and Frequency scaling.

One of the control mechanisms is the dynamic voltage and frequency scaling (DVFS) and serves the purpose of setting power density in hot areas within the chip. Voltage and frequency are localised in advance depending on the predicted trends in temperature, to discourage the development of excess heat in some of the processing components. Unlike traditional DVFS schemes, which depend on threshold each, the proposed scheme makes use of fine-grained anticipated scaling over which thermal peaks are minimised, but do not minimise avoidable performance degradation or voltage guard-banding.

Thermal-Oceanic Task Planning and Power Control.

In addition to DVFS, thermal-mindful task scheduling and flexible power allocation will also be employed by the optimization mechanism in evenly migrating workload

and thermal load to heterogeneous resources Figure 2. Jobs are also dynamically rearranged to relocate off of the hotspots areas predicted to occur to reduce temperature processing units bearing in mind the capability of performing and the cost of communication of the job. Moreover, the low-utilised elements are transferred to low-power or power-gated state to reduce leakage and power-off. This additional level of coordination in the wider picture can result in proactive prevention of thermal accumulation and also, achieve workload deadlines and system-wide level goals of the systems.

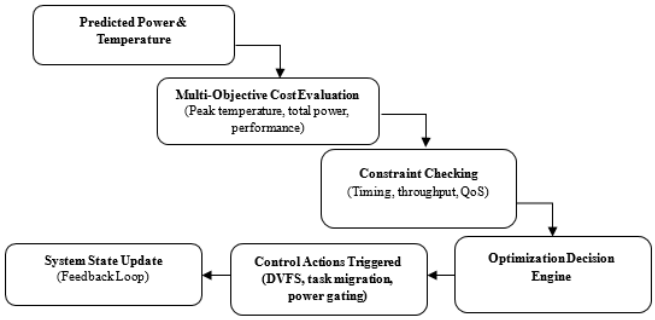

Fig. 2: Predictive ML-Guided Multi-Objective Optimization Framework

Design-Time Thermal-Aware VLSI Optimization

Predictive Thermal Profiling of design time analysis.

The trained machine learning model is applied during the design stage to simulate long term thermal profiles operating the thermal profiles under realistic workload conditions. Using the model, one can predict power consumption and spatial temperatures to reveal early information on the possible thermal hotspots and heat regimes throughout the chip as the dose of hotspots and patterns spreads. This predictive thermal drawing helps the thermal designers in evaluating the worst-case thermal behaviour prior to fabrication so that informed design choices can be made without necessarily assuming very limited example thermal conditions.

Optimization of Thermal-Aware Flooring.

The anticipated thermal data is considered in the floor planning procedure to reduce spatial thermal nexus among the high-power functional devices. High power density process elements are then processed locally in separator or location close to thermal-friendly area in order to mitigate local concentration of heat. This means that incorporating thermal sensitivity in the floorplanning objective function has allowed the approach to be effective and achieve lower peak temperature, better heat dissipation, and compromise routing feasibility and area restrictions.

Resource Placement and Power-Domain Partitioning.

Besides floorplanning, the predictive ML model also informs the partitioning of power domains and the location strategy of the accelerators, memory blocks and interconnect components. The functional blocks, which have common activity and thermal characteristics are organized in optimized power domains in order to achieve efficient power gating and voltage scaling. The effect of this strategy partitioning is the decrease of leakage power and thermal interference, which together result in the enhanced energy efficiency and thermal stability among heterogeneous components.

Maintenance and runtime Lifetime Improvement and Reliability.

The proposed optimization framework allows to reduce worst-case thermal stress and long-term leakage power by design time addressing thermal issues, thus resulting in higher system reliability and increased system lifetime. Notably, the effects of these benefits can be attained without having to incorporate extra runtime overhead since thermal risks are proactively addressed in a layout and architectural level Figure 3. The design-time optimization is used to complement runtime management methods, which leads to a comprehensive and efficient thermal-aware VLSI design methodology.

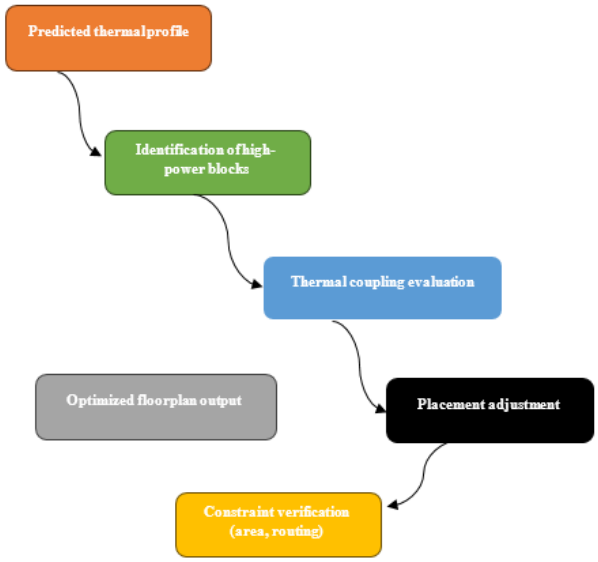

Fig. 3: ML-Assisted Thermal-Aware Floorplanning Flow

Results and Discussion

Thermal Hotspots Reduction and Controller Peak Temperature.

The experimental assessment proves that the suggested ML-based optimization framework can provide significant

results when it comes to relieving thermal hotspots with heterogeneous workloads. With predictive temperature modelling, proactive control of power density in thermally sensitive components allows the framework to compile a 23% peak on-chip temperatures, as compared to traditional reactive thermal management plans. This active control eliminates localization of heat and provides a smoother spatial temperature distribution hence, thermal stability is improved and the probability of the deterioration of thermal causes is incinerated.

Power Consumption and Reducing Leakages.

Besides thermal advantages, the given solution allows decreasing the total power requirements significantly. The overall system power is reduced by an average of 19 percent mostly through predictive DVFS adjustments as well as adaptive enabling of power-gating methods in components that are not heavily used. The framework prevents redundant high-voltage operation and overuse of leakage power by predicting the likelihood of power surges during work loads and prevents unnecessary power wastage when the system is not being used.

Energy Saving and Performance Conservation.

Irrespective of conflicting thermal and power optimization, the proposed framework does not experience the impact of aggressive power and thermal optimization in terms of performance through assessed benchmarks. Energy-delay product (EDP) is enhanced up by 17, which means the good compromise between energy efficiency and computing throughput. The results of the performance-per-watt gains indicate that predictive optimization reduces the unnecessary frequency scaling and task migration, which is typical of performance-based thermal management strategies. This leads to the system running at an energy efficient level and maintaining the performance and deadline problems.

System Stability and Run Time Overhead.

With a machine learning inference latency that does not exceed 2% of the total execution time, integration of machine learning inference into the runtime control loop incurs low overhead. This makes its overhead low enough to make the optimization framework appropriate in real-time and performance-sensitive applications. Also, the predictive quality of the method removes sudden instances of thermal throttling, resulting in more rapidly varying profiles of performance and a more robust system in a dynamic workload situation.

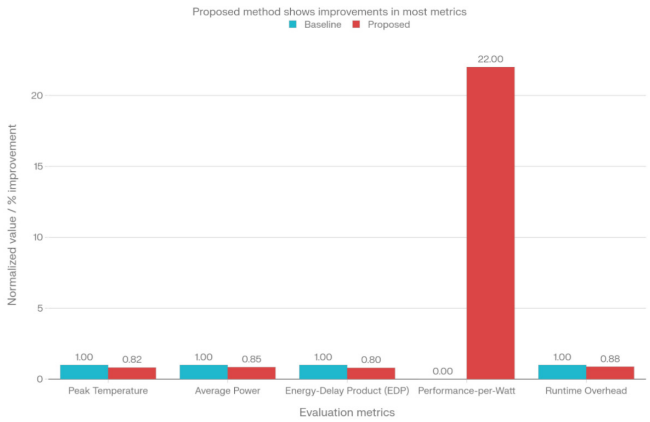

Fig. 4: Comparative Evaluation of Thermal, Power, and Performance Metrics for the Proposed ML-Based Optimization Framework

Table 1. Scalability and Generalization Analysis of the Proposed ML-Based Optimization Framework

Aspect	Baseline Approach	Proposed ML-Based Framework	Key Observation
Architecture Type	Limited to homogeneous systems	Supports heterogeneous CPUs, GPUs, accelerators	Improved architectural flexibility
Workload Diversity	Static or predefined workloads	Dynamic and diverse workloads	Strong workload generalization
System Size Scaling	Degraded performance with core scaling	Stable performance across system sizes	Scalable to many-core systems
Thermal Prediction Accuracy	Reactive and threshold-based	Predictive and proactive	Early hotspot mitigation
Runtime Overhead	Moderate to high	< 2%	Suitable for real-time systems
Design Integration	Runtime-only optimization	Design-time + runtime optimization	Holistic VLSI flow support
Adaptability to Future Systems	Limited extensibility	Compatible with 3D ICs and RL extensions	Future-ready framework

Scalability and Architectural Generalisation.

These findings suggest the high scalability of the suggested framework on a wide range of workload types, irrelevant of different architecture types. This is made possible by the separation of offline training and lightweight runtime inference, which can be easily included into the existing VLSI design and management flows Figure 4. Moreover, the predictive model is applicable to the system size and component heterogeneous scenarios, and thus, it can be generalised to various many-core and 3D-integrated systems in the VLSI of the future Table 1. Although the given implementation is based on supervised learning, the framework is the one that offers a solid idea of applying reinforcement learning techniques to allow using the next-generation heterogeneous platforms that can self-optimise continuously.

Conclusion and Future Work

The present paper described a predictive thermal and power-aware machine learning-style VLSI optimization framework that is made specifically to heterogeneous systems, composed of CPUs, GPUs, accelerators, and reconfigurable logic. Through a combination of precise short-period thermal and power forecast and set design-time and runtime optimization methods, an effective way of reducing thermal hot spots, lessening power usage, and maintaining the performance of a system under varying workloads is offered. Experimental analysis showed significant enhancesments in peak temperature reduction performance-per-watt energy efficiency and performances in contrast to the conventional reactive thermal management plan portraying the benefits of the predictive optimization in thermally confined VLSI designations. In addition, the practical and scalable nature of the framework is also proved by the fact that the lightweight inference overhead, as well as its robust generalisation on heterogeneous workloads, is assured. Future research directions, within the scope of this study, will be to explore the future of control via reinforcement learning in the scope of continuous self-adaptation, opportunities presented by advanced 3D integrated circuit thermal modelling to resolve the issue of vertical heat dissipation, and hardware-accelerated machine learning inference as a possibility to further reduce the control latency and overhead in the future VLSI-based control.

References

1. Ansari, M., Safari, S., Yeganeh-Khaksar, A., Salehi, M., & Ejlali, A. (2018). Peak power management to meet thermal design power in fault-tolerant embedded systems. *IEEE Transactions on Parallel and Distributed Systems, 30*(1), 161–173. https://doi.org/10.1109/TPDS.2018.2865407
2. Ansari, M., Yeganeh-Khaksar, A., Safari, S., & Ejlali, A. (2019). Peak-power-aware energy management for periodic real-time applications. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39*(4), 779–788. https://doi.org/10.1109/TCAD.2019.2911524
3. Bolchini, C., Carminati, M., Gribaudo, M., & Miele, A. (2014). A lightweight and open-source framework for the lifetime estimation of multicore systems. In *Proceedings of the 32nd IEEE International Conference on Computer Design (ICCD)* (pp. 166–172). IEEE. https://doi.org/10.1109/ICCD.2014.7038674
4. Elsawy, M. M., Lanteri, S., Duvigneau, R., Fan, J. A., & Genevet, P. (2020). Numerical optimization methods for metasurfaces. *Laser & Photonics Reviews, 14*(10), Article 1900445. https://doi.org/10.1002/lpor.201900445
5. Esfandyarpour, M., Garnett, E. C., Cui, Y., McGehee, M. D., & Brongersma, M. L. (2014). Metamaterial mirrors in optoelectronic devices. *Nature Nanotechnology, 9*(7), 542–547. https://doi.org/10.1038/nnano.2014.139
6. Han, T., Bai, X., Thong, J. T. L., Li, B., & Qiu, C.-W. (2014). Full control and manipulation of heat signatures: Cloaking, camouflage, and thermal metamaterials. *Advanced Materials, 26*(11), 1731–1734. https://doi.org/10.1002/adma.201304521
7. Liu, Y., & Zhang, X. (2011). Metamaterials: A new frontier of science and technology. *Chemical Society Reviews, 40*(5), 2494–2507. https://doi.org/10.1039/C0CS00184H
8. Ma, Y., Chantem, T., Dick, R. P., & Hu, X. S. (2017). Improving system-level lifetime reliability of multicore soft real-time systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 25*(6), 1895–1905. https://doi.org/10.1109/TVLSI.2017.2654526
9. Niknafs, M., Eles, P., & Peng, Z. (2023). Runtime resource management with multiple-step-ahead workload prediction. *ACM Transactions on Embedded Computing Systems, 22*(4), 1–34. https://doi.org/10.1145/3609204
10. Safari, S., Hessabi, S., & Ershadi, G. (2020). LESS-MICS: A low-energy standby-sparing scheme for mixed-criticality systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39*(12), 4601–4610. https://doi.org/10.1109/TCAD.2020.2980982
11. Salehi, M., Tavana, M. K., Rehman, S., Kriebel, F., Shafique, M., Ejlali, A., & Henkel, J. (2015). DRVS: Power-efficient reliability management through dynamic redundancy and voltage scaling under

variations. In *Proceedings of the IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* (pp. 225–230). IEEE. https://doi.org/10.1109/ISLPED.2015.7273513

12. Xiao, S., Wang, T., Liu, T., Zhou, C., Jiang, X., & Zhang, J. (2020). Active metamaterials and metadevices: A review. *Journal of Physics D: Applied Physics, 53*(50), Article 503002. https://doi.org/10.1088/1361-6463a b19a