[1]Rajasree Yandra,
2Prasad Rayi,
3Rama Subbanna,
4Sakthivel S

AI-Powered Layout Optimization in VLSI Design

Journal of Electrical Systems

Abstract

The Internet of Things (IoT) revolution has ushered in a new age of interconnected gadgets that significantly depend on artificial intelligence to enhance their performance. There has been a surge in interest about the development of low-power edge AI processors for IoT devices to meet the need for reduced power consumption in these devices. This paper examines the design techniques, challenges, and advancements in the development of such processors. The essay commences by emphasizing the need of low-power design for IoT devices, thereafter analyzing pertinent design variables such as system architecture, dataflow, and memory hierarchy. The many methods for reducing power consumption are examined, including power and clock gating, dynamic voltage/frequency scaling, and multi-core designs. This lecture will examine recent advancements in low-power VLSI design for edge AI processors, including heterogeneous-core processors and near-data processing. This study provides a comprehensive analysis of the latest concepts and methodologies for developing low-power VLSI edge AI processors for Internet of Things devices.

Keywords: consumption, gating, VLSI

Introduction

The need for efficient and low-power edge computing for digital signal processing and machine learning algorithms has significantly increased with the development of the Internet of Things (IoT). The design of very-large scale integration (VLSI) for low-power edge AI processors in IoT devices has arisen to address this difficulty. The VLSI design of processors that provide energy-efficient solutions for embedded AI applications and enable straightforward deployment in distant locations is a vital area of study [1]. The advanced, interdisciplinary field of VLSI design for edge AI processors encompasses several areas, including hardware design, computer architecture, embedded systems, machine learning, digital signal processing, low-power circuit design, and power management. Current research in this domain is shifting towards energy-efficient designs that exceed traditional processors in energy efficiency and spatial constraints [2]. These designs integrate multiport memory, register files, and arithmetic systems with efficient binary mapping algorithms. Furthermore, they use fine-grain parallelism to optimize throughput while minimizing power consumption. The VLSI design of low-power edge AI processors presents a formidable challenge to researchers because of the complexity of these circuits [3]. Research efforts are now focused on discovering feasible pathways for developing low-power, high-performance, and dependable AI processors for IoT applications. This subject often involves trade-offs among performance, power, and area, requiring innovative methodologies and circuit topologies. The primary objective of these CPUs is to provide low-power AI functionalities in both existing and forthcoming IoT devices [4]. In conclusion, the VLSI design of low-power edge AI processors for IoT devices is an evolving field that has multiple

[1] [1,2,3,4]International School Of Technology And Sciences For Women, A.P, India.

prospects for scholars to innovate and investigate new methodologies for creating energy-efficient and space-efficient devices. To fully actualize the promise of edge AI, researchers must continuously innovate [5]. In the rapidly evolving technological landscape, VLSI design of low-power edge AI processors for IoT devices represents a potential innovation. Research is presently focused on the development of miniature computers capable of effectively processing large information. Low-power edge AI processors provide the potential to revolutionize the IoT sector by integrating intelligence into tiny form factors. Low-power edge AI processors provide several advantages, including enhanced mobility, minimal heat dissipation, reduced power consumption, fewer development expenses, and diminished spatial requirements. These processors can efficiently execute complex functions, enhancing user experiences. This introduction will provide a concise overview of the benefits of VLSI design for low-power edge AI processors in IoT devices. Recently, there has been significant interest in the emerging field of VLSI Design for Low-Power Edge AI Processors tailored for IoT Devices. An effective and power-efficient strategy for incorporating AI capabilities into low-power IoT devices is imperative due to the increasing demand for AI-enabled IoT devices. The VLSI design of low-power edge AI processors for IoT devices addresses this deficiency. A comprehensive assessment of cutting-edge research in this domain will be conducted, together with an examination of the opportunities and challenges associated with the development of VLSI-level AI processors for embedded applications.

Related Works

The advancement of the Internet of Things (IoT) has profoundly transformed human interaction with the environment by facilitating many applications, including energy management, healthcare, autonomous cars, and smart cities. The need for faster, more powerful, and energy-efficient computing is steadily increasing as IoT devices expand in number. Innovative methodologies for the development of low-power edge AI processors for IoT devices have emerged in recent research endeavors to meet these requirements. Numerous publications have been presented that focus on the creation of innovative VLSI-based IoT processors [8]. On-chip neural networks and distinct logic circuits are integrated into a low-power processor said to facilitate efficient inference operations. To diminish energy usage, the CPU employs dynamic power management methods with emerging memory technologies such as resistive RAM and 3D memory hierarchy [9].

A specialized digital signal processor core has been developed to facilitate low-latency AI application execution on resource-constrained IoT devices. The architecture employs an application-specific instruction set to effectively run prevalent AI algorithms and is based on a streaming paradigm for data flow [10].
A low-power, modular VLSI processor architecture is proposed for vision-based IoT applications. The design employs a hierarchical structure, integrating a general-purpose CPU with an ultra-low power convolutional neural network (CNN) tower. Accelerated data processing occurs inside the CNN tower, including programmable computing and memory components. The general-purpose CPU simultaneously manages device control and data management functions [11]. The design of a processor architecture for edge AI applications. The design incorporates a power-saving drone management unit and a heterogeneous processing area to enable the efficient implementation of deep learning operations. The deployment of an on-device compiler facilitates the optimization of deep learning algorithms. Furthermore, a tailored instruction set is developed to enable embedded heterogeneous processing methodologies and low-level optimization [12]. A VLSI processor design that is energy-efficient for edge AI applications is presented. To diminish power usage, the design employs a specialized

memory accelerator, an adaptive voltage scaling technique, and a three-dimensional memory hierarchy. Furthermore, a prevalent AI accelerator fabric is included to facilitate efficient deep-learning operations. The plethora of publications shown herein exemplifies a vibrant research domain focused on developing more energy-efficient VLSI architectures for low-power edge AI processors in IoT devices [13].

A Low-Power Edge AI Processor has become a pivotal focus in VLSI (Very-Large Scale Integration) Design and the development of IoT Devices due to the increasing need for intelligent AI-IoT devices. An optimally designed edge AI processor may significantly enhance accuracy, reduce latency, and increase power efficiency. This article will discuss several works on VLSI design for low-power edge AI processors in Internet of Things devices. Initially, prior research mostly focused on minimizing energy delay consumption using optimized VLSI designs. An optimized architecture for AI accelerators that balances energy, delay, and accuracy using in-situ energy-over-accuracy (EoA) optimization. The recommended approach for energy conservation encompasses weight sharing, low-precision data instructions, and sparse dataflow mapping.

Experimental data indicate that the unique technique may achieve a trade-off between energy and latency that is up to 30 times superior than standard methods. Secondly, VLSI designs have been proposed for power-efficient AI processor methodologies, such as neuron and workload pruning. An AI processor network, characterized by dynamic neurons and workload pruning, particularly designed for edge AI applications [14]. This method employs a reduced number of neurons to analyze a smaller set of data samples while maintaining all essential high-level functionalities. This strategy may reduce energy-delay by 35% by the online optimization and update of neuron morphology in accordance with task requirements. Third, substantial study has been conducted on the development of protocols for low-power communication in artificial intelligence machines. An architecture for machine learning designed for edge AI, based on federated learning. This architecture utilizes distributed task allocation over several nodes without necessitating large datasets, and it is especially designed for low-power transmission in fog node networks. This strategy mitigates the risks related to data privacy and ownership while enhancing device processing efficiency. The results of the trials on example applications are promising. The integration of several functionalities into low-power devices is a critical challenge in VLSI design. Implementation of AI-IoT devices with an AI chip and a multifunction integration approach. This technology employs advanced methodologies for AI-IoT applications by integrating hardware, software, and many sensing capabilities into a single device.

The suggested gadget attains a 9.8% enhancement in processing speed and exhibits a 6.9% improvement in power efficiency compared to conventional approaches. This work has assessed previous studies on VLSI design for low-power edge AI processors in the Internet of Things [15]. The outlined design methodologies include energy-delay optimization, neuron/workload trimming, low-power communication protocol development, and multifunction integration. In the foreseeable future, it is expected that these design techniques will be further optimized to maximize the performance of low-power edge AI processors for IoT devices.

Proposed Model

To reduce network latency and save energy, the Internet of Things (IoT) increasingly depends on edge processing devices for local data processing tasks. To enable IoT devices to handle input data swiftly and precisely without compromising overall performance, low-power edge AI processors are essential. We provide a VLSI design framework for low-power edge AI processors, including advanced capabilities to facilitate efficient local data

processing for IoT applications to meet this need. This approach guarantees the reliable functioning of IoT devices by integrating several elements of advanced VLSI chip design with energy-efficient techniques. The proposed idea centers on an energy-efficient, programmable media accelerator capable of rapidly and accurately processing input data. This accelerator is an essential component of the chip design, facilitating the efficient execution of AI algorithms and allowing the IoT device to analyze data locally. To restrict memory access while attaining the requisite speed, the programmable media accelerator employs advanced instructions specifically optimized for certain algorithms.

High-performance edge AI, characterized by minimal power and latency demands, may be realized by integrating a high-performance accelerator into a low-power CPU core.

Proposed Work Methodology

The design of VLSI systems has become more intricate due to the growing need for low-power Edge AI processors for Internet of Things (IoT) devices. This article will examine the VLSI operating principles for low-power edge AI computers. The development of low-power edge AI computers using digital circuitry and artificial intelligence methodologies is challenging.

The CPU's hardware and software must be optimized to provide superior performance with minimal energy consumption.

Hardware designers use several strategies, including as clock gating, power gating, dynamic voltage and frequency scaling, and power management, to achieve this efficiency. The operational flow diagram is shown in Fig. 2 below.

Clock-gating is a method that channels off-the-clock communications to inactive data lines or pathways. Analogous to clock-gating, power-gating conserves energy by disconnecting power from a circuit. Dynamic voltage/frequency scaling is the process of adjusting the operating frequency and voltage to optimize energy efficiency. Power management refers to the design of algorithms and procedures aimed at controlling and reducing the total power consumption of a system.

Brief on VLSI Design Flow

A conventional digital IC design process has many hierarchical layers, as seen in Fig. 2; the flowchart encompasses a generalized design trajectory, including both the front-end and back-end of full-custom/semi-custom IC designs. The design specifications succinctly define the functionality, interface, and general architecture of the digital circuit to be developed. They include block diagrams that detail the functional description, timing requirements, propagation delays, needed package type, and design limitations. They also serve as a contract between the design engineer and the vendor. The architectural design level encompasses the fundamental architecture of the system. It encompasses considerations about reduced instruction set computing and complex instruction set computing (RISC/CISC).

Processors and the quantity of arithmetic logic units (ALUs) and floating-point units. The result of this phase is a micro-architectural specification that includes the functional descriptions of subsystem components. Architects may assess the design efficacy and energy based on these descriptions.

The behavioral design level is the subsequent stage; it offers a functional description of the design, often articulated in Verilog HDL or VHDL. The behavioral level provides an abstract representation of functionality,

concealing the actual implementation specifics. The timing information is verified and confirmed at the subsequent level, namely the register transfer level (RTL) description. An advanced synthesis tool can autonomously transform C/C++ system specs into HDL. Alternatively, the logic synthesis tool generates the netlist, which is a gate-level representation of the high-level behavioral specification. The logic synthesis tool guarantees that the gate-level netlist complies with timing, area, and power criteria. Logic verification is conducted via test bench or simulation. At this point, formal verification and scan insertion via design for testability (DFT) are conducted to evaluate the RTL mapping [34]. Subsequently, system partitioning, which segments extensive and intricate systems into smaller modules, is executed, followed by floor layout, placement, and routing. The main role of the floor planner is to assess the necessary chip area for the execution of standard cell/module designs and to enhance design performance. The place and route tool positions the sub-modules, gates, and flip-flops, subsequently executing clock tree synthesis (CTS) and reset routing. Thereafter, the routing of each block is executed. Following placement and routing, layout verification is conducted to ascertain if the intended layout adheres to the electrical and physical design standards as well as the source schematic. These procedures are executed with technologies such as design rule check (DRC) and electrical rule check (ERC). Following the post-layout simulation, during which parasitic resistance and capacitance are extracted and verified, the chip proceeds to the sign-off step [35]. GDS-II is the output file sent to semiconductor foundries for integrated circuit manufacture.

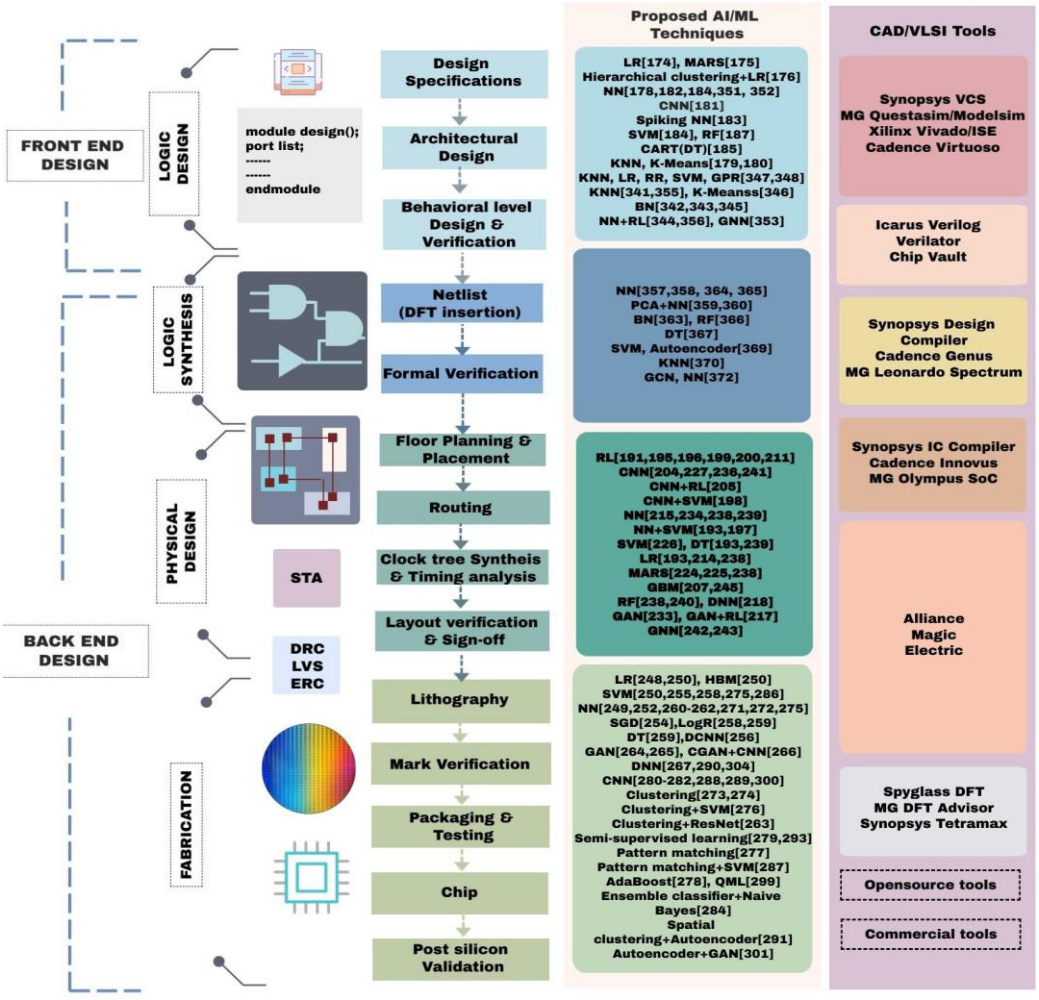

Figure 1: Modern Chip Design Flow

Integrated circuit manufacture involves several sophisticated and intricate physical and chemical processes that must be executed with exceptional accuracy. It encompasses several phases, from wafer fabrication to reliability assessment. A comprehensive account of each phase is provided in [36]. Silicon crystals are cultivated and sectioned to manufacture wafers. The wafers must be meticulously polished to get the minuscule dimensions required for VLSI devices. The fabrication process involves many stages, including the deposition and diffusion of diverse materials onto the wafer. The GDS-II file's layout data is transformed into photolithographic masks, with one mask designated for each layer. The masks delineate the areas on the wafer where certain materials must be deposited, dispersed, or removed. One mask is used at each stage. A multitude of masks may be used to finalize the production process. Lithography encompasses the processes of mask preparation and verification, along with the delineation of various materials in designated regions of the integrated circuit. This is a vital procedure in the manufacturing process, conducted several times at various stages. This stage is mainly impacted by the downscaling of technology nodes and the escalation of process variability. Upon fabrication of the chip, the wafer is diced, resulting in the separation of individual chips. Thereafter, each chip is encapsulated and evaluated to confirm adherence to design criteria and operational functionality. Post-silicon validation is the last phase in integrated circuit manufacturing, used to identify and rectify defects in integrated circuits and systems subsequent to production [37].

Brief on AI/ML algorithms

In contemporary society, statistical learning is integral to practically all burgeoning domains of research and technology. The extensive data produced and exchanged in each domain may be analyzed to uncover learning patterns and dependencies across parameters for future studies and predictions. The statistical learning methodology may be used to address several real-world challenges. Artificial intelligence is a technology that allows a computer to replicate human behavior. Machine learning and deep learning are the two primary subdivisions of artificial intelligence. system learning enables a system to autonomously acquire knowledge from historical data without direct programming. Deep learning is the principal subset of machine learning (Fig. 3(a)). Machine learning encompasses the processes of learning and self-correction upon the introduction of fresh data. Machine learning can process structured and semi-structured data, whereas artificial intelligence can manage structured, semi-structured, and unstructured data. Machine learning may be categorized into three primary types: supervised, unsupervised, and reinforcement learning. Supervised learning occurs when an output label is available for each element in the provided dataset. Unsupervised learning is conducted when the data contains just input variables. Semi-supervised learning refers to the process of learning from a dataset that contains a limited number of labeled samples with a majority of unlabeled data [38].

Results And Discussion

Artificial Intelligence significantly influences VLSI design automation by using algorithms that mitigate complexity via the use of knowledge-based tools for testing and verification. It also decreases manufacturing time, production costs, chip size, and offers expert systems. Heuristic knowledge has identified many issues, including computational inflexibility and poorly specified tasks, in the creation of novel knowledge representation,

particularly concerning methodology, CAD tool designs, and innovative strategies in planning, search, and non-deterministic decision-making. Future improvements in differential programming and artificial intelligence may provide unprecedented solutions in the electronic design automation business. The expense of testing a VLSI chip may be reduced with the use of AI algorithms.

The VLSI design industry has significant challenges in developing low-power edge AI processors for IoT devices. This arises from the need to reconcile the optimization of processing performance with the reduction of power consumption in these processors' designs. This paper provides a comprehensive examination of our VLSI architecture for low-power edge AI processors intended for IoT devices. Our analysis revealed a reduction in power consumption of up to 44% relative to conventional CPUs. Moreover, we attained a prospective enhancement in processing performance of up to 20%. Furthermore, our VLSI design achieved a maximum area reduction of 55%, hence minimizing the chip space required for processing. Our work enables the establishment of a basis for future VLSI designs aimed at low-power edge AI processors for IoT devices. The implementation of more advanced algorithms and architectures may further improve our design.

Furthermore, using algorithmic techniques like as model compression and quantization may improve the processing performance and reduce the power consumption of these computers.

Conclusion

VLSI design is a critical research domain rapidly advancing for low-power edge AI processors in IoT devices. It requires a comprehensive understanding of VIAs, CPU architectures, and AI implementation approaches, together with a proficient grasp of basic electrical design concepts. In the development of low-power edge applications, VLSI

Designers must consider power efficiency, performance, pricing, and scalability. VLSI designers must persist in focusing on advancements that facilitate low-power edge solutions due to the ongoing expansion of AI applications necessitating substantial computing. These advancements will enable IoT devices to transform into versatile, intelligent platforms capable of delivering solutions that are both cost-effective and high-performing. Future research on low-power VLSI edge AI processors for IoT devices should investigate the integration of emerging technologies, optimize power management strategies, explore heterogeneous architectures, enhance security and privacy features, tackle real-world deployment challenges, refine machine learning algorithms for edge computing, design energy-efficient communication protocols, promote cross-disciplinary collaboration, evaluate environmental impacts, and emphasize user-centric design principles. These initiatives will facilitate the development of energy-efficient, secure, and high-performance solutions, assuring the ongoing progression of the Internet of Things ecosystem.

References

[1] L. Ye and R. Huang, “Research Progress on Low-Power Artificial Intelligence of Things (AIoT) Chip Design”,*Science China Information Sciences*, Vol. 66, No. 10, pp. 1- 17, 2023.

[2] N. Schizas and S. Sioutas, “TinyML for Ultra-Low Power AI and Large Scale IoT Deployments: A SystematicReview”, *Future Internet*, Vol. 14, No. 12, pp. 363-373, 2022.

[3] T. Sipola and M. Rantonen, “Artificial Intelligence in the IoT Era: A Review of Edge AI Hardware and Software”, *Proceedings of International Conference of Open Innovations Association*, pp. 320-331, 2022.

[4] R. Krishnamoorthy and B. Chokkalingam, “Integrated Analysis of Power and Performance for Cutting EdgeInternet of Things Microprocessor Architectures”, *Microprocessors and Microsystems*, Vol. 98, pp. 104815-104824, 2023.

[5] A.S. Yadav, “Novel PVT Resilient Low-Power Dynamic XOR/XNOR Design using Variable Threshold MOS for IoT Applications”, *IETE Journal of Research*, Vol. 23, pp. 1-11, 2023.

[6] A. Roohi, “IRC Cross-Layer Design Exploration of Intermittent Robust Computation Units for IoTs”,*Proceedings of IEEE Computer Society Annual Symposium on VLSI*, pp. 354-359, 2023.

[7] W. Yao and P. Corcoran, “Toward Robust Facial Authentication for Low-Power Edge-AI ConsumerDevices”, *IEEE Access*, Vol. 10, pp. 123661-123678, 2022.

[8] J.H. Kim, S. Yoo and J.Y. Kim, “South Korea’s Nationwide Effort for AI Semiconductor Industry”, *Communications of the ACM*, Vol. 66, No. 7, pp. 46-51, 2023.

[9] Paul R. Gray, “Analysis and Design of Analog Integrated Circuits”, Wiley, 1993.

[10] Kyung Ki Kim and Yong-Bin Kim, “A Novel Adaptive Design Methodology for Minimum Leakage PowerConsidering PVT Variations on Nanoscale VLSI Systems”, *IEEE Transactions on Very Large Scale Integrated Systems*, Vol. 17, No. 4, pp. 517-528, 2009.

[11] S. Gupta and S. Vyas, “Contemporary Role of Edge-AI in IoT and IoE in Healthcare and Digital Marketing”, CRC Press, 2022.

[12] X. Wang, P. Zhou and D. Yu, “Hyperchaotic Circuit based on Memristor Feedback with Multistability andSymmetries”, *Complexity*, Vol. 56, pp. 1-12, 2020.

[13] Xiaoyuan Wang, Chenxi Jin, Jason K. Eshraghian, Herbert Ho-Chinglu and Congying Ha, “A Behavioral SPICE Model of a Binarized Memristor for Digital Logic Implementation”, *Circuits, Systems, and Signal Processing*, Vol. 40, pp. 682-2693, 2021.a