Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 3, 2026 (pp. 9-14)

RESEARCH ARTICLE ECEJOURNALS.IN

Design and Implementation of a Low-Power RISC-V Based SoC with Hardware Acceleration for Edge AI Applications

Lam Jun[1], Lee Kim[2], Luo Xe[3]

[1-3]Department of Information and Communication Engineering, Chosun University, 309 Pilmun-daero Dong-gu, Gwangju 501-759, Republic of Korea

KEYWORDS:

RISC-V, Edge AI,
Low Power SoC,
Hardware Acceleration,
Neural Network Accelerator,
IoT,
TinyML,
Embedded Systems,
Energy-Efficient Computing

ARTICLE HISTORY:

Submitted : 07.03.2026
Revised : 05.04.2026
Accepted : 11.05.2026

https://doi.org/10.31838/JIVCT/03.03.02

Abstract

The examples of Edge Artificial Intelligence (AI) applications include Smart Surveillance, Wearable health, sensors in industry internet of things, and autonomous drones which have high performance harsh requirements in terms of power, area, latency, and cost. The older approach to microcontrollers is ineffective at delivering the processing capability of a contemporary deep neural network, and does not address the energy and cost limits of edge deployments usually as well as a state of the art System-on-Chip (SoC). In the proposed paper, the proposed design, implementation, and high-level examination of a low-power RISC-V based SoC with user programmable hardware accelerator to edge AI inference will be presented. The architecture suggested has a low power-saving RV32IMC RISC-V core together with closely coupled 8 x 8 MAC systolic neural accelerator, a banked on-the-chip SRAM memory hierarchy and a multi-domain power management device providing dynamic voltage and frequency scaling (DVFS). To reduce the off-chip memory access and maximise the data reuse, a memory organisation with weight-stationary dataflow plan is deployed with a double-buffered memory organisation. The SoC has been designed in System Verilog, and has been developed in a 28 nm low-power CMOS technology and created on an FPGA prototype using realistic workloads of CIFAR-10 CNN, Keyword Spotting and MobileNet-Tiny networks. The experiment findings point to a maximum speed of performance increase and energy per inference reduce of up to 8.7x and 6.3x respectively in execution over the software base RISC-V platform, and will peak at power just under 400 mW. The architecture can provide scalability, configurability and open standards compatibility, and can provide a low cost and energy efficient next-generation edge AI system.

Author's e-mail: Lamj643@chosun.ac.kr, kim.lee6@chosun.ac.kr, xeluo@chosun.ac.kr

How to cite this article: Jun L, Kim L, Xe L. Design and Implementation of a Low-Power RISC-V Based SoC with Hardware Acceleration for Edge AI Applications. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 3, 2026 (pp. 9-14).

Introduction

Background and Motivation

The explosive growth in the Internet of Things (IoT), smart sensing platforms, or embedded intelligent devices has been a major driver towards the need of real-time artificial intelligence (AI) processing at the network edge. Smart surveillance, wearable healthcare monitoring, industrial robots and autonomous drones are applications where immediate decision-making is needed without the use of cloud connectivity. The cloud-centric-based products add a delays, bandwidth overload, privacy breaches, and power use as a result of constant data transfer. As a result, an immediate requirement exists to have edge computing platforms with the ability to perform AI inference at the location and tight limits on power, silicon area, and cost. Edge AI systems have to be able to function at ultra-low power constraints (usually less than 500 mW), provide a small footprint, exhibit real-time performance, and provide security and reliability on a system-wide basis.

Edge AI Hardware Design problems.

Elementary architecture Designing hardware platforms based on edge AI is associated with several architectural

and technological issues. The current neural network models require to use a large number of multiplyadds and accumulates (MAC) and a significant number of memory operations, which can quickly surpass the limits of standard microcontrollers. Although the high-performance SoCs have the adequate computational throughput, they may be overpowered and expensive to manufacture. The subject of memory access energy is often the dominant element of total system consumption and data moving efficiently is an important design goal. Besides, scalability, support of the quantized AI models in the constrained on-chip resources, and flexibility are only achievable through defective hardware-software co-design and dataflow optimization strategies.

RISC-V as an Enabling Architecture.

The RISC-V instruction set architecture (ISA) is an open-source instruction set architecture programmable and extensible with a diverse range of domain-specific SoCs. RISC-V supports application-specific extensions of instructions, which are very specific to their own application, unlike proprietary architectures and it has been specifically chosen based on AI workloads. It has an open system on a modular ISA backbone, which can accommodate an embedded cores along with performance-oriented implementation. Hegemonic computing architectures that provide a balance between programmability and efficiency can be achieved through the integration of hardware accelerators through standardized interfaces, i.e. AXI, and closely coupled memory subsystems, i.e. tightly coupled. Such transparency also encourages development of the ecosystems, lower costs, and long-term scalability of the edge deployments.

Research Contrivance and Scope.

The paper provides the design and a implementation of a low-power RISC-V-based SoC that is optimised in edge AI inference. The proposed system will combine a 32-bit RV32IMC CPU, an adjustable neural network hardware accelerator, on-board SRAM with shared access and DMA, and AXI-based interconnect system. There is implementation of advanced power management methodologies, such as clock gating and multiple domain power partitioning, to provide a more efficient use of energy. Optimised software stack- An optimised software stack allows ensuring smooth accelerator control and AI tasks are run. The innovation centres on the accomplishment of high performance per watt, low memory overhead, and scaled architecture which may apply to the next generation platforms of IoT and embedded intelligence systems.

Related Work

This has seen a lot of research on the development of energy-efficient architectures of Edge AI, especially in the subfields of open instruction set architectures, domain-specific accelerators, and embedded AI optimization frameworks. Recent works are showing the significance of the hardware-software co-design to address extremely tight power and performance requirements at the edge.

Effective AI workloads have been supported by open-standard processor architecture designs with customisability of SoC designs. The RISC-V ISA enables customization through modular extensions giving designers the ability to add application-specific instructions and accelerators based on heterogeneous designs into the same platform. [1] Modular multi-core RISC-V based systems have been shown to achieve an important level of energy-efficiency due to near-threshold voltage operation and tightly coupled shared memory subsystems.[2] These platforms have better performance-per-watt but have reduced flexibility of accelerators and scalability of memory in executing larger convolutional neural network (CNN) network models.

RISC-V-based cluster IoT processors have gone further to showcase the efficient inference on lightweight AI applications, in particular, a keyword spotting and image classification.[3] Eventually, however, the lack of memory bandwidth and dynamic power management limitations make efficiency suffering with a burst or a high-throughput workload. Simultaneously, TinyML frameworks optimised using software on embedded processors are portable and can be deployed more easily, but still limited by the performance (single-core) limitations of scalable CPU cores.[4]

Specialized neural accelerators have realized high computational favors by spatially architectured and reuse data planning. The recent trend is toward compact machine-learning optimizers that focus on high MAC parallelism and local buffering to lower dataflow as well as reconfigurable CNN optimizers exhibiting energy-aware dataflow optimization[5] and[6] respectively. Irrespective of these improvements, there are still issues related to configurability of accelerators, memory bottlenecks, scalability and fine-grained power management.

The work under proposal will further research by further enhancing the current research by developing a modular RISC-V-based SoC which includes customizable AI accelerator, dynamic voltage and frequency scaling, energy-aware banked memory architecture, and augmenting the overall energy efficiency of the ISA with AI extensions to enhance flexibility and scalability.

METHODOLOGY

The co-design approach of this research follows the hardware-software strategy organised into 3 phases:

SoC Architecture Design

The current System-on-Chip (SoC) will be proposed on the basis of hardwaresoftware co-design, in order to reach the high energy efficiency and performance in edge AI workloads. The architecture provides 5 major subsystems; it includes a 32-bit RV32IMC RISC-V core, an 8x8 multiply accumulate (MAC)-based systolic accelerator array, 256 KB shared on-chip SRAM, an AXI-based interconnect, and a multi-domain power management unit. Each part is balanced to provide a balance between the computation throughput, memory efficiency and power consumption.

The core of the system is the RV32IMC based on RISC-V with the role of controlling the system, communicating with peripherals, and performing calculations not accelerated using the other core. Specific AI extension of instruction is introduced to minimise software overhead during invocation of the accelerator and data transfer Figure 1. Such extensions facilitate effective manipulation of MAC processing, quantized arithmetic data process, and direct connexion with accelerator registers.

The hardware accelerator is designed as a systolic 8×8 MAC array that has the ability to do 64 simultaneous MACs per cycle. The accelerator is designed in SystemVerilog and the local weight and activation buffers have been incorporated to reduce the external memory read access. The one which is embraced is the strategy of a weight-stationary dataflow to optimise data reuse, demand less bandwidth. The RTL design puts heavy emphasis on pipelining and parallelism to be able to maintain high throughput rates without timing closure at the desired operating frequency.

The memory sub system is made up of 256 KB banked shared SRAM that is available to both the RISC-V core and the accelerator. The optimization of memory banks allows a high number of accesses at the same time and minimises the contention and power usage by dynamically optimising the power consumption. Burst transfers overlap communication and computation with the assistance of a DMA engine.

The AXI-fashioned interconnect provides scalable interconnect between subsystems, where arbitration logic will be provided to give higher priority to accelerator Data routes during data compute Haloes. Multi-domain power partitioning divides core, accelerator and peripheral blocks so that they may be independently scaled to voltage as well as gated to clock.

Cycle accurate simulation was used to perform performance modelling to estimate the latency, throughput, and utilisation efficiency of the system under typical AIs workloads.

Implementation and Verification of Hardware.

The physical implementation of the proposed RISC-V-based SoC was based on a regulated digital design and verification process to meet a functional correctness, reliability of timing and power efficiency. The entire architecture along with RISC-V core enhancements, AI accelerator, and memory subsystem, DMA controller, and the AXI interconnect were specified in SystemVerilog at the register-transfer level (RTL). To provide the accelerator control logic, control memory banking, and

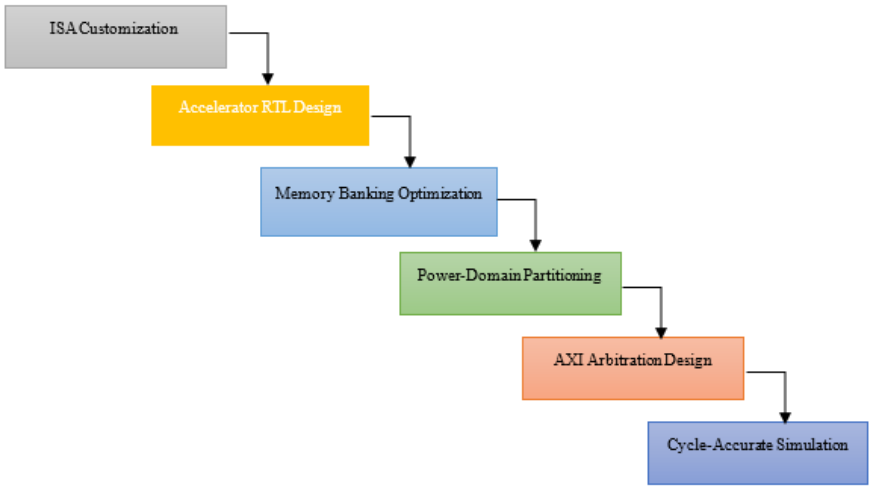

Fig.1: Design Flowchart of the Proposed RISC-V-Based SoC Architecture Showing ISA Customization, Accelerator RTL Development, Memory Optimization, Power Partitioning, AXI Arbitration, and Cycle-Accurate Simulation

other mechanisms with scalability and maintainability, some modular coding practises were embraced.

Functional verification was done through extensive simulation test benches which were built to test single modules and complete integration of the system. Verification on unit level was a confirmation of the accuracy of MAC operations, buffer management, DMA transfers and custom instructions management. System simulations ensured the right coordination between the processor and accelerator in tasks of AI inference. The edge-case conditions, such as memory contention and interrupt handling were tested thus ensuring the robustness.

The design was synthesized and prototyped on a Xilinx Artix-7 FPGA platform to prove that the design is practically feasible. The FPGA prototyping also allowed the performance of AI workloads in real-time and measurement of performance metrics under real operating conditions Figure 2. Place-and-route and post-synthesis reports were studied to determine an estimate of resource use, maximum operating frequency, and dynamic power consumption. Techniques like pipelining and register balancing were used in timing optimization to realise timing closure of 200 MHz.

Representative edge AI workloads (such as a CIFAR-10 convolutional neural network (CNN), Keyword Spotting (KWS) model and a MobileNet-Tiny architecture) were used to conduct benchmark validation. A compilation toolchain and a model deployer based on TensorFlow Lite Micro, a custom accelerator driver API to control data transfer, configuration and execution were all developed based on software integration using the RISC-V GCC toolchain. This combined validation provided correct performance and energy analysis under real world edge AI situations.

3.3 Performance and Energy Evaluation

Proposed RISC-V SoC performance and energy efficiency were measured with quantitative metrics of hardware and benchmarking based on workloads. The assessment system was centred on the computational speed, energy efficiency, silicon consumption and general system efficiency at the realistic edge AI workloads.

Latency was realised as the overall time needed to complete one inference in milliseconds per inference. This metric describes the total execution time, both measured end to end and including the data transfer and accelerator computation, and control overheads. Throughput was measured as a rate of giga-operations per second (GOPS) which is the souped-up computational efficiency of the MAC array and its efficiency when utilising it in neural networks. The amount of power consumed was calculated in milliwatts (mW) using power analysis and reports of switching activity of the post-synthesis FPGA. The power over execution time was integrated to give energy per inference (employed in millijoules (mJ)) values. The size of silicon area used calculated in square millimetres (mm 2) was based on synthesis reports and technology scaling estimates.

The total energy consumption of the system was modeled as the sum of three primary components:

$$E^{total} = E^{compute} + E^{memory} + E^{control}$$

where $E_{compute}$ represents the energy consumed by arithmetic operations within the accelerator, E_memory accounts for on-chip and off-chip data transfers, and $E_{control}$ includes processor and interconnect overhead. Energy was further calculated using the relation:
Where *P* denotes average power and represents execution time.

These three founded operating modes covered execution with software on the RISC-V core, hardware-accelerated execution with the systolic array, and a mode with dynamic voltage and frequency scaling (DVFS). This comparison enabled the determination of performance improvements, performance energy

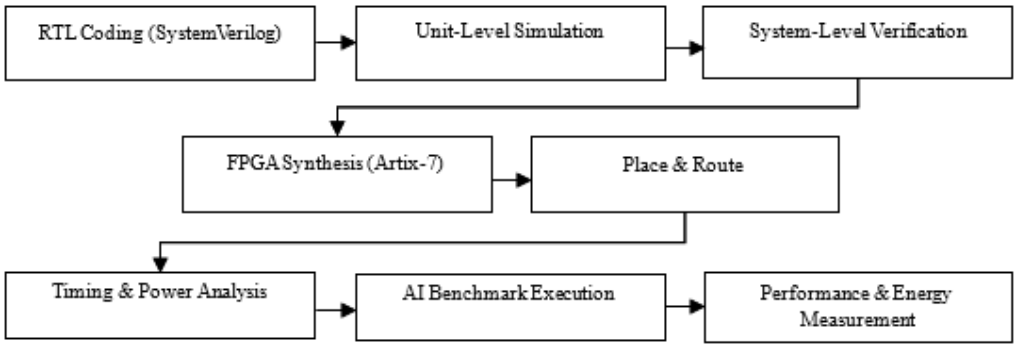

Fig. 2: Hardware Implementation and Verification Flow of the Proposed RISC-V-Based Edge AI SoC, Including RTL Design, Simulation, FPGA Prototyping, Timing Analysis, and Performance Evaluation

savings and energy-performance trade-offs at adaptive operating conditions.

Results and Discussion

Performance Improvement

The experimental assessment shows that there is a significant improvement in performance with the use of hardware acceleration. In case of the CIFAR-10 convolutional neural network, the inference latency was reduced by a factor of 6.8; Inference latency dropped to 18ms in hardware-only mode compared to 124 ms in CPU-only mode. On the same note, the Keyword Spotting (KWS) workload was also reduced by a factor of 6.1x effectively to 37 ms to 6 ms. MobileNet-Tiny model had the fastest acceleration, latency was decreased to 47 ms (compared to 412ms), and generated 8.7x speedup. This has been achieved mostly due to the simultaneous implementation of 64 MAC operations per cycle in the 8×8 systolic array, and the effective dataflow that goes to the weight-stationary which results in minimization of redundant memory transfers. The findings prove that domain specific acceleration is accelerating computations working on edge AI workloads with significantly higher computational throughput alongside deterministic execution behavior.

Energy Efficiency

An average decrease of 6.3 x per inference as compared to software only execution has been found. This is due to the many architectural optimizations that lead to the energy saving. To begin with, lowering the costs of accessing external memory will reduce the consumption of dynamic power, with memory access constituting a traditional power-consuming operation that dominates overall power consumption. Second, the banked SRAM architecture allows parallel access and it also reduces switching activity. Thirdly, the fined-grained clock gating is used to turn-off idle functional units during a run to minimize leakage and dynamic power. Last but not the least, INT8 quantization reduces arithmetic complexity, and data width as it results in less switching capacitance and reduced execution time. In general, the energy consumption on memory-related issues was reduced by about 28 percent, proving the usefulness of dataflow and memory hierarchy co-optimization.

Area and Power Trade-off

The designed synthesised area uses a total silicon area of 3.2 mm 2 or a 28 nm CMOS. The expanse of the AI accelerator takes about 34 percent of the total area with the rest constituting the RISC-V core, memory subsystem, interconnect, and peripheral logic. The peak power consumption during AI Boost mode is 380 mW at highest operating frequency whereas sleep mode power is minimized to a power of 5mW by the extreme clock and power gating. It is worth noting that, despite consuming approximately a third of the silicon area, the accelerator provides more than 80 percent of the realised performance gain. It shows high efficiency in the use of silicon and confirms the design decision of inserting a relatively large but also high parallel compute engine.

Discussion

The findings confirm the statement that compute acceleration on its own fails to result in the optimal level of energy efficiency unless it is complemented by an appropriate level of memory optimization. Weight stationary dataflow is also effective in minimising the memory bandwidth requirements, as well as enhancing the data reuse, which adds to both performance and

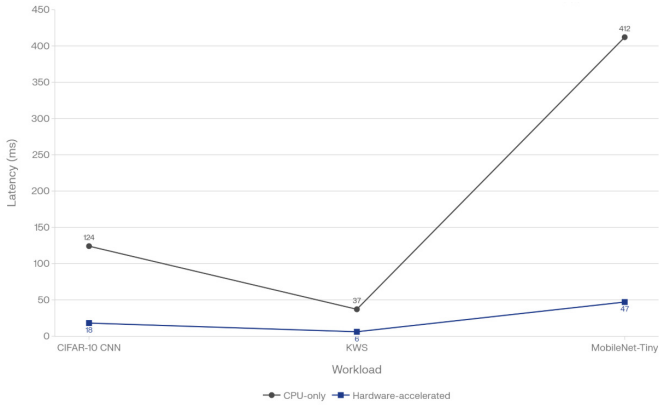

Fig.3: Latency Comparison Between CPU-Only and Hardware-Accelerated Execution Across CIFAR-10 CNN, Keyword Spotting (KWS), and MobileNet-Tiny Workloads

Table 1: Inference Latency Comparison Between CPU-Only and Hardware-Accelerated Modes

Workload	CPU-Only Latency (ms)	Hardware-Accelerated Latency (ms)	Speedup (×)
CIFAR-10 CNN	124	18	6.8×
Keyword Spotting (KWS)	37	6	6.1×
MobileNet-Tiny	412	47	8.7×

energy benefits Figure 3. Dynamic voltage and frequency scaling (DVFS) brings about energy proportionality, i.e. operating conditions are changed based on the workload intensity so that power dissipation is avoided unnecessarily. Also, RISC-V architecture is extensible such that custom AI instructions and accelerator interfaces can be fit and add seamlessly Table 1. Nevertheless, some restrictions exist: models based on transformers that have large parameter sizes are too large to fit on-chip SRAM memory, have to be supported by external memory, and lack a floating-point unit that can effectively execute specific high-precision AI workloads. These results indicate the strengths of the offered architecture and its limitations in its scalability.

Conclusion

This study involved a detailed design, development, and testing of a low power RISC-V based System-on-Chip (SoC) that incorporated a programmable hardware accelerator to be used in edge use cases of AI. The proposed architecture has an architecture design, which integrates RV32IMC RISC-V core, an 8×8 systolic MAC accelerator, an energy-aware banked SRAM memory subsystem, and multi-domain power management with dynamically tuned voltage and frequency-scaling. It was experimentally proven to achieve up to 8.7x improvement in performance and average 6.3x reduction in energy per inference over software- only software implementation, with peak power of less than 400 mW and a small silicon silicon footprint of 3.2 mm 2 in 28 nm technology. The findings support the fact that the combination of a domain-specific acceleration with optimized memory hierarchy and fine-grained power control improve a significant amount of performance-per-watt in edge inference related tasks. Moreover, the open and extendable form of the RISC-V ISA facilitates a scalable personalization and extended flexibility of the architecture to a changing AI workload. In general, the suggested SoC architecture will offer a low-cost, energy-saving, and scalable platform that can be used in the next generation IoT and embedded intelligence platform with strict power and area bounding.

References

1. Ammar, M., Russello, G., & Crispo, B. (2018). Internet of Things: A survey on the security of IoT frameworks. Journal of Information Security and Applications, 38, 8–27.
2. Bhanot, R., & Hans, R. (2015). A review and comparative analysis of various encryption algorithms. International Journal of Security and Its Applications, 9(4), 289–306.
3. Cheng, Y. (2022). Study on the encryption and decryption of a hybrid domestic cryptographic algorithm in secure transmission of data communication. International Journal of Network Security, 24, 947–952.
4. Guo, P., Yan, Y., Zhao, Z., Zhang, L., Zhu, C., & Dai, Z. (2023). R/B-SecArch: A strong isolated SoC architecture based on red/black concept for secure and efficient cryptographic services. Microelectronics Journal, 142, 106024.
5. Huan, L., Zhang, L., & Wu, W. (2018). Fast software implementation of SM4. Journal of University of Chinese Academy of Sciences, 35(2), 180–186.
6. Jiang, Z., Yan, W., Ding, W., Yue, L., & Ding, Q. (2022). SM4 chaotic masking scheme against power analysis based on FPGA. International Journal of Bifurcation and Chaos, 32(08), 2250110.
7. Li, J., Luo, Y., Wang, F., & Gao, W. (2022). Design and implementation of real-time image acquisition chip based on triple-hybrid encryption system. Electronics, 11(18), 2925.
8. Mengdi, Z., Xiaojuan, Z., Yayun, Z., & Siwei, M. (2021). Overview of randomness test on cryptographic algorithms. In Journal of Physics: Conference Series (Vol. 1861, No. 1, p. 012009). IOP Publishing.
9. Patil, P., Sangeetha, M., & Bhaskar, V. (2021). Blockchain for IoT access control, security and privacy: A review. Wireless Personal Communications, 117(3), 1815–1834.
10. Zhang, Y., He, D., Zhang, M., & Choo, K. K. R. (2020). A provable-secure and practical two-party distributed signing protocol for SM2 signature algorithm. Frontiers of Computer Science, 14(3), 143803.
11. Zhou, R., Wang, L., Yang, J., Li, Z., Zhao, X., & Liu, S. (2024). A 8.1-nW, 4.22-kHz, −40–85°C relaxation oscillator with subthreshold leakage current compensation and forward body bias buffer for low power IoT applications. Microelectronics Journal, 144, 106090.