Energy-Efficient Computer Systems: RISC-V Extensions for Machine Learning Inference at IoT's Edge Computing

Ben Sujin[1], M.Sangeetha[2]
[1,2]Lecturer, Computer Engineering Department, University of Technology and Applied Sciences,
Nizwa, Sultanate of Oman

Article Info

Article History:

Received Jun 13, 2025
Revised Jul 10, 2025
Accepted Aug 16, 2025

Keywords:

Energy-Efficient Computer Systems
Instruction-level parallelism
RISC-V
Internet of Things
Machine learning (ML)

ABSTRACT

Over the past few decades, there have been numerous turning points in the massive transformation of computing systems. The limits of instruction-level parallelism (ILP) and the end of Dennard's scaling pushed the semiconductor sector toward multi-core devices, notwithstanding Moore's law, which directed the industry to pack more and more transistors and logic into the exact same volume. The era of domain-specific architectures (DSA) and processors for novel workloads like machine learning (ML) and artificial intelligence (AI) has recently begun. In addition to the difficulties brought on by tighter integration, extreme form factors, and increasingly varied workloads, these trends—possibly with additional limitations—further complicate the architecture, design, implementation, and power consumption optimization of systems. Nowadays, across the board, energy efficiency is a first-order design constraint and parameter for computing equipment. The creation of energy-efficient computer systems using RISC-V architecture modifications specifically suited for machine learning inference is investigated in this study. While preserving inference speed and accuracy, the suggested extensions seek to decrease energy consumption, minimize instruction overhead, and maximize hardware utilization. Through the integration of domain-specific accelerators, memory-access optimizations, and lightweight vector operations, the extended RISC-V platform exhibits notable gains in performance per watt when compared to traditional architectures. Results from experiments and benchmark assessments demonstrate how well-suited the method is for real-time Internet of Things applications including industrial automation, smart healthcare, and environmental monitoring. AI-enabled IoT systems that are low-power, scalable, and sustainable are being advanced by this work.

Corresponding Author:

M.Sangeetha,
Lecturer, Computer Engineering Department,
University of Technology and Applied Sciences, Nizwa, Sultanate of Oman
Email: sangeetha.mani@utas.edu.om

1. INTRODUCTION

In recent years, we have contributed to the exponential expansion of Internet-of-things (IoT) connected devices that are present in a variety of application areas, including surveillance of structural health, agricultural, and health tracking [1]. In this situation, the Internet of Things end-nodes must use signal processing techniques to gather data from low-power sensors and transmit it wirelessly over the network. In addition to giving IoT nodes intelligent capabilities and expanding IoT applications with DL-enhanced tasks (like self-governing nanodrones), machine learning (ML) algorithms, particularly cutting-edge Deep Learning (DL), offer "information distillation" solutions that allow for the extraction of useful information from the raw data collected by sensors. They can wirelessly transmit a small amount of condensed information by "squeezing" raw data into a much more grammatically dense format [2] (e.g., determining classes, feature levels, and characters).

This reduces traffic on the Internet of Things network and lessens reliability and security issues, which are now made worse by the large increase in raw data flowing through the network. With the aim of implementing DL capability at the very edge of the IoT, a broad study area has been drawn to the obvious advantages of embedding knowledge on IoT nodes. This endeavor must contend with the high memory and computational demands of popular deep learning techniques, which conflict with the typically limited memory and computing capabilities of deeply embedded equipment that are fueled by harvesting power or battery packs.

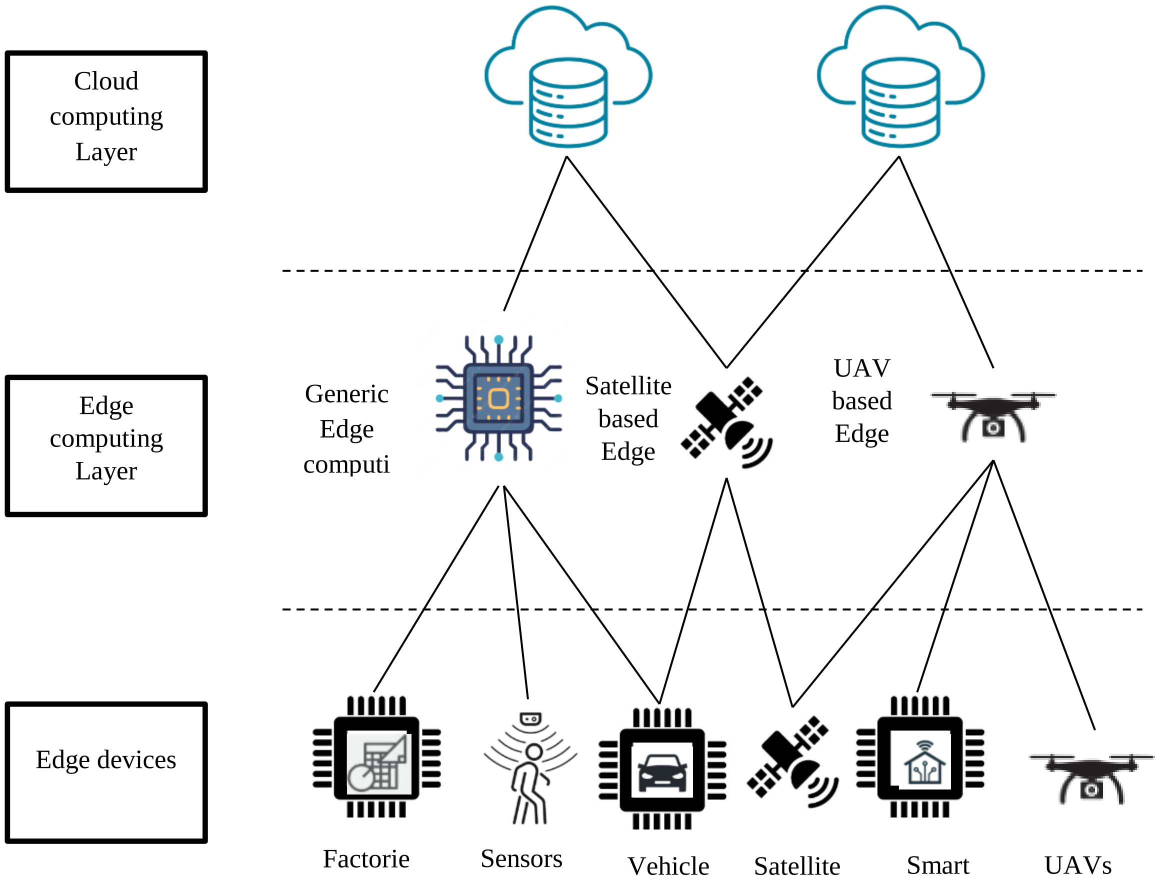

Figure 1. An example of an architecture for edge computing

The energy conservation system's block diagram shows how the hardware and software components are arranged methodically to allow for energy-efficient machine learning inference on RISC-V based architectures at the edge of the Internet of Things. In order to guarantee optimal performance per watt and accuracy in inference tasks, the system is primarily built to track energy use, optimize computational tasks, and dynamically distribute resources in Figure 1. The input data sources, which usually comprise sensor data, image/video streams, or biological signals recorded by Internet of Things devices, are shown first in the diagram. These inputs serve as the system's cornerstone since they are the unprocessed data that needs to be processed using machine learning models. The data goes through a feature extraction and preprocessing unit after the input step. This block is in charge of filtering noise, lowering the dimensionality of the incoming data, and carrying out quick calculations that get the data ready for more in-depth analysis. By avoiding heavy or unnecessary computations, preprocessing at the edge reduces the stress on the phases that follow, saving energy. Preprocessing, for instance, might entail filtering heart-rate signals in a medical wearable device before sending them to the primary inference engine.

The RISC-V core combined with Deep Learning Accelerator (DLA) extensions is the next essential part. This block serves as the energy conservation system's computational foundation. Because of its open-source flexibility and adaptability, the RISC-V instruction set architecture (ISA) enables the addition of specific extensions that are tailored for deep learning applications. Lightweight vector instructions, multiply-accumulate (MAC) units, and domain-specific accelerators for matrix operations—all crucial components of convolutional and recurrent neural networks—are frequently included in these additions. The DLA significantly lowers energy consumption and increases throughput by ensuring that most energy-intensive calculations are transferred from the general-purpose processor to specialized hardware.

The energy monitoring and power management block is connected to the computational core. In order to balance performance and energy efficiency, this subsystem dynamically adjusts the frequency, voltage, and resource allocation while continuously monitoring the system's power consumption. To reduce idle power dissipation, strategies including clock gating, power gating, and Dynamic Voltage and Frequency Scaling (DVFS) are used. For example, the system lowers operational frequency and turns off unused modules when the workload is light or sporadic, resulting in substantial savings in energy without sacrificing real-time performance. The memory subsystem, which consists of external memory units and on-chip cache, is another crucial piece in the diagram. The energy conservation system minimizes energy-hungry data transfers by introducing optimizations like data reuse buffers, compression algorithms, and memory-access scheduling, as memory access is frequently a bottleneck in machine learning workloads. The system makes sure that the amount of energy used for memory operations is maintained to a minimum by employing clever caching techniques and putting frequently accessed data closer to the CPU.

In IoT edge scenarios where devices need to communicate with cloud servers or other edge devices, the communication interface block is essential. Here, energy efficiency is attained by the use of edge-level decision-making to cut down on needless communication overhead, data compression before to transmission, and lightweight communication protocols. For example, the system uses the DLA to analyze data locally and only sends the pertinent inference findings, saving bandwidth and energy, rather than transmitting raw video feeds to the cloud. The application layer and output stage, the last block in the design, is where the user or higher-level decision-making systems get the outcomes of machine learning inference. These applications span a variety of industries, including environmental monitoring, smart cities, healthcare, and industrial automation. When anomalous activity is detected, for instance, the output of a smart surveillance program might be an alert; in agriculture, it might be a recommendation for irrigation based on the analysis of sensor data. Crucially, by guaranteeing that energy-efficient computation results in useful, real-world consequences, the application layer completes the energy conservation system's cycle.

Overall, the block diagram shows a well-integrated system with energy-saving contributions from every module. Input data capture, preprocessing, RISC-V DLA computation, power management, memory optimization, communication efficiency, and, at the end, application-specific outputs provide a comprehensive approach to sustainable computing platform design [3]. The system maintains scalability and low power consumption while adapting to a variety of IoT use cases by utilizing RISC-V's extensibility.

This block diagram's importance arises from both its modular design and its capacity to emphasize the interdependencies among various subsystems. Preprocessing improvements, for example, immediately lower computational energy, whilst power management unit improvements increase device lifetime. Memory optimizations also lower communication costs by lowering data transmission overhead. This interdependence demonstrates that energy conservation is the outcome of cooperative optimization throughout the entire architecture rather than the responsibility of a single block.

2. RELATED WORKS

The constraints of the conventional von Neumann architecture, which uses distinct hardware for memory and computation, are revealed by the "memory wall" problem in computer architecture. In memory-intensive applications like artificial intelligence, the memory bus bandwidth limits system performance due to the division between the compute blocks used for logic processing and the memory blocks used for data and program storage. There have been a number of earlier suggestions in the literature to lessen or eliminate the memory wall by bringing the memory and calculation units closer together [4]. The IRAM design suggested using the DRAM main memory chip to fabricate a compute logic processor. However, because of the constraints of fabrication technology at the time, this early concept for the PiM model was not widely embraced. AI Processing Units (APU) has been implemented in the base logical die of a 3D layered main memory in numerous earlier works.

The most fundamental and ultimate need in an energy-saving system is the integration of microcontrollers with various components, including sensors, relays [5], and business appliances like fans, lights, etc. Devices, users, and the system can be linked in a variety of ways, depending on the system, range, corporate office size, and user convenience. The appliances must all interface with the system if a user wants to control every corporate appliance via it. Only those appliances that the user wishes to control must interface with the system; all other devices can be operated manually. The fundamental idea behind this project is to automatically turn on and off business gadgets like fans and lights. In this case, the relay acts as a switch by transmitting data between the appliances and the microcontroller. A microcontroller is coupled to a PIR sensor and a proximity sensor. The microcontroller is also connected to control devices. Human detection provides information to the microcontroller.

RISC-V has served as the foundation for a number of open-source projects that have tried to close this gap. With an emphasis on energy-efficient near-sensor processing, PULPino created an ultra-low-power microcontroller platform. GAP8 [6] expanded this architecture by adding an 8-core cluster that was tuned for parallel DSP and CNN workloads. Although these architectures demonstrated RISC-V's potential in edge-constrained contexts, their efficacy on control-intensive or serial algorithms was limited due to their heavy reliance on parallelism and compiler-assisted task decomposition. Furthermore, they lacked dedicated memory co-optimization and flexible SIMD execution engines, both of which are essential for maximizing the use of contemporary quantized deep learning algorithms.

Over time, there has been a notable increase in the demand for low-cost and low-power accelerators. In particular, RISC-V-based ISA designs have aided in the global emergence of open-source hardware solutions during the global disruption of the semiconductor supply chain brought on by the COVID-19 pandemic and further intensified by Russia's invasion of that country in early 2022 [7]. In order to remain as inclusive as possible, we have also had difficulties keeping up with the quick rise in the number of open-source hardware designs that are produced. It was also challenging to gather specifications and more detailed information about commercially available CPU cores and SoCs because some of them were only accessible with a private license.

There are more benefits to integrating BNN inference into a RISC-V core than just domain-specific hardware acceleration. An FPGA-deployed RISC-V processor may perform BNN inference while concurrently managing other instructions, obviating the need for an additional processing unit, in contrast to conventional FPGA-based deep learning accelerator that only use specialized processing units [8]. It has been demonstrated that this hybrid capacity improves real-time AI systems, especially in situations when deep learning inference and dynamic task execution are required.

3. METHODS AND MATERIALS

3.1 RISC-V VP Based on System C

A potential RISC-based processor design, RISC-V is an open ISS (Instruction Set Structure) that was developed at U.C. Berkeley and is kept up to date by the open foundations. Processor chips based on RISC-V Core IP are already on the market, and RISC-V Core has already been produced and published as Core IP [9]. Nonetheless, RISC-V is commonly employed in research based on FGPA or RTL because it is still in its infancy. The FPGA or RTL level design has also been used in the research and development stages of the RISC-V-based CNN Architecture; however, this method has the drawback of requiring a comparatively long period for system verification, analysis, and optimization.

System C is the foundation for the recently established RISC-V Virtual Platform, which is effective for system testing in a comparatively short amount of time. The Virtual Platform is a highly adaptable and extendable system that may include other TLM-connected module for checking unique functions in the RISC-V VP context. It was developed and tested with a modular bus system using TLM 2.0 for the RISC-V RV32IM core [10]. The bus delivers the initiator's activity to the destination port by routing it according to the memory-mapped address. The CPU may also manage foreign or local interruptions. The PLIC-based IC (Interrupt Controller) handles the external interrupt, while the CLINT (Core Local Interrupt Controllers) handles the local interrupt. Software that has been compiled using the RISC-V cross-compiler can run on this virtual platform. After being cross-compiled, the SW is produced as an executable file in the ELF format, which is then put into the main memory and run as firmware [11]. The primary memories module serves as the software's memory region and is additionally linked to the bus.

By incorporating the CNN DLA modules into the virtual platform, we created a DLA prototype system built on the RISC-V VP platform. The CNN DLA module is assigned to a portion of the RISC-V CPU core's address range and is linked to the TLM 2.0 Bus via the target port [12], as seen in Figure 2. The internal module that carries out CNN functions including convolution, initialization, and pools through read/write over the TLM bus makes up the CNN DLA module. After being loaded into main memory, the DNN data is moved to the DLA module via DMA techniques.

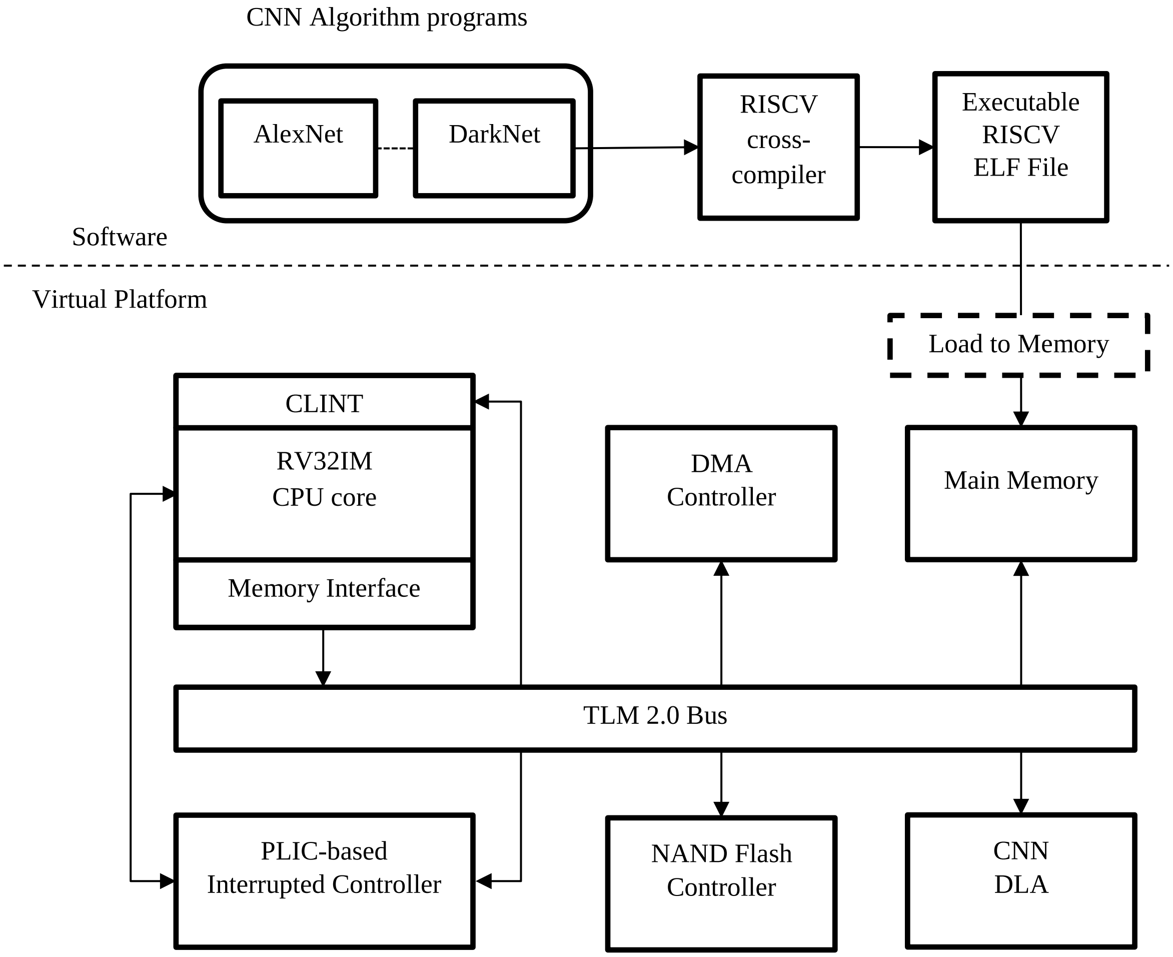

Figure 2. Overview of the RISC-V Virtual Platform Architecture

Registers assigned to the CNN addressing range are written to and read from by the CPU core, which manages the CNN module. Through the execution of DNN applications like Darknet, a virtual platform that includes CNN DLA can carry out deep learning interpretation. The DNN applications are loaded into primary memory as executable files in the ELF format after being compiled by a RISC-V cross processor.

3.2 Applications of the RISC-V DLA System

The RISC-V Deep Learning Accelerator (DLA) system has emerged as a promising architecture for deploying machine learning inference in energy-constrained environments. By combining the open-source flexibility of RISC-V with hardware acceleration for deep learning tasks, the DLA system enables efficient and scalable deployment of AI at the edge. Its adaptability and low-power design make it suitable for a broad spectrum of applications, ranging from healthcare and robotics to industrial automation [13], smart cities, and consumer electronics.

One of the most impactful applications of the RISC-V DLA system is in healthcare and medical technologies. The demand for continuous patient monitoring and real-time diagnostic assistance has driven the adoption of wearable and portable devices that integrate machine learning at the edge. With RISC-V DLA, biomedical signals such as electrocardiograms (ECG), electroencephalograms (EEG), and blood glucose levels can be processed directly on wearable devices, minimizing reliance on cloud servers. This not only reduces latency but also preserves patient privacy, as sensitive data remains local to the device. Moreover, imaging applications such as ultrasound and X-ray interpretation can be enhanced through deep learning inference on edge-based diagnostic machines, making advanced healthcare accessible even in remote or resource-limited settings. Such capabilities align well with the broader goal of personalized and preventive medicine, where rapid decision-making is critical.

Another major application domain is autonomous systems and robotics, where real-time data processing and decision-making are essential. Autonomous vehicles, drones, and service robots rely heavily on deep learning models for tasks such as object detection, obstacle avoidance, and path planning. The RISC-V DLA system provides a platform capable of executing these computations with high efficiency while keeping energy consumption to a minimum, thereby extending operational time for battery-powered systems. For example, drones equipped with a RISC-V DLA can perform aerial surveillance, package delivery, or agricultural monitoring without requiring frequent recharging. Similarly, industrial robots deployed on factory floors can use the accelerator to interpret sensory inputs in real time, enabling safe human-robot collaboration and enhancing productivity in smart manufacturing.

The deployment of RISC-V DLA in smart cities and surveillance systems highlights its role in large-scale societal applications. Video surveillance cameras embedded with deep learning accelerators can perform local inference for facial recognition, anomaly detection, and crowd monitoring, reducing the need to transmit vast amounts of raw data to centralized servers. This approach not only decreases communication bandwidth requirements but also strengthens data privacy and security. Additionally, smart traffic systems can leverage edge-based analytics powered by the RISC-V DLA to dynamically control traffic lights, monitor congestion, and enhance pedestrian safety. Environmental monitoring in urban areas, such as analyzing air quality and noise levels through distributed IoT nodes, further illustrates the potential of energy-efficient deep learning inference for sustainable city management.

In the context of Industrial IoT (IIoT), the RISC-V DLA system enables predictive maintenance and process optimization. By embedding machine learning inference capabilities directly into industrial machines, the system allows continuous monitoring of vibrations, temperature, and pressure to detect early signs of equipment failure. This predictive maintenance reduces downtime and minimizes operational costs. Furthermore, edge-based inference reduces reliance on cloud connectivity, ensuring that time-sensitive operations can proceed without delays. The flexibility of the RISC-V instruction set architecture allows domain-specific customization, enabling accelerators to be optimized for particular industrial workloads. In smart grids and energy distribution systems, DLA-enabled devices can balance power consumption, detect anomalies, and ensure reliable energy supply in real time.

Consumer electronics and wearable devices represent another rapidly expanding application area for the RISC-V DLA system. Modern devices such as smartphones, augmented/virtual reality (AR/VR) headsets, and smart-watches increasingly depend on deep learning models for enhanced user experiences. The DLA system allows these devices to run applications such as natural language processing (NLP) for voice assistants, gesture recognition for intuitive control, and real-time image classification for AR/VR interactions, all while conserving battery life. Wearable’s benefit particularly from the system’s low power consumption, enabling continuous monitoring of user activity, sleep cycles, and fitness levels with AI-driven insights delivered directly on-device.

Beyond consumer-focused domains, environmental monitoring and agriculture provide additional contexts where the RISC-V DLA system has transformative potential. In precision agriculture, cameras and sensors equipped with DLA accelerators can identify plant diseases, monitor crop growth, and optimize irrigation patterns in real time, thereby improving yield and reducing resource usage. In environmental applications, distributed IoT sensor networks can track wildlife populations, detect forest fires, and predict weather anomalies with edge-based machine learning models. These applications benefit from the energy efficiency of the DLA system, which enables long-term operation in remote or resource-limited environments powered by renewable sources such as solar energy.

Finally, the defense and aerospace sectors are also potential beneficiaries of RISC-V DLA integration. In defense, deep learning accelerators can support target recognition, surveillance, and situational awareness, all of which demand real-time inference under strict power and resource constraints. In aerospace, satellites equipped with DLA-enabled processors can process high-resolution images and scientific data on-board, reducing the need to transmit large datasets to ground stations and thus improving efficiency and responsiveness.

Taken together, these applications demonstrate the versatility and scalability of the RISC-V DLA system in addressing the challenges of modern AI workloads. By providing an open, customizable, and energy-efficient platform, the system supports innovation across critical domains, from healthcare and industry to consumer electronics and environmental sustainability. Its capacity to deliver real-time deep learning inference at the edge positions it as a cornerstone for the future of intelligent, connected, and energy-aware computing systems.

3.3 The energy conservation system

Microcontroller 1 is linked to a PIR sensor, a proximity sensor, and loads via a relay, a power supply, and an LCD display in Figure 3 [14]. The PIR and proximity sensors determine whether or not there is anyone moving in a certain region. The microcontroller then receives this data.

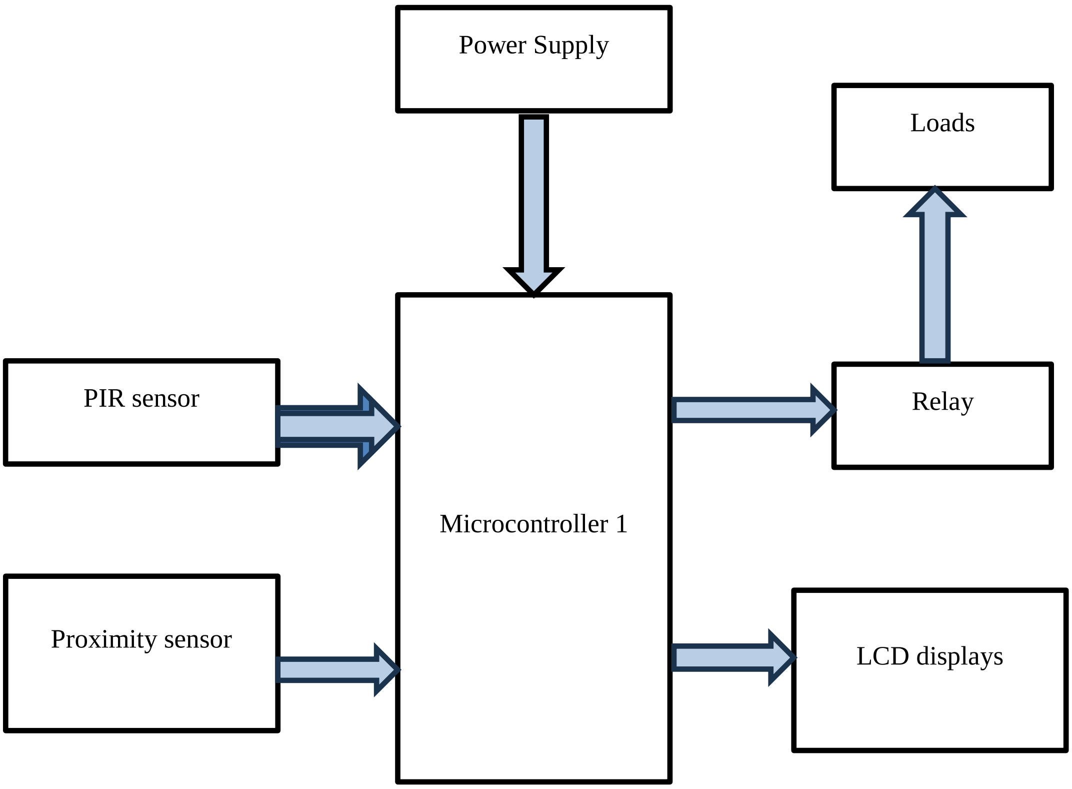

Figure 3. The energy conservation system's block diagram

The energy conservation system is a structured framework designed to minimize power consumption while maintaining acceptable levels of computational performance in modern computing platforms. Within the context of IoT and edge computing, where devices are often resource-constrained and battery-powered, energy conservation becomes a critical design objective. The system integrates a combination of hardware-level optimizations, software strategies, and architectural extensions, with the goal of achieving higher performance-per-watt in executing machine learning workloads. At its core, the system ensures that energy resources are utilized efficiently, unnecessary computations are eliminated, and power-hungry operations are replaced with optimized alternatives.

The system starts with energy monitoring and control systems that continuously measure how much electricity is used by different modules. These monitoring units provide feedback to the central controller, enabling dynamic adjustments such as lowering clock frequencies, scaling voltages, or shutting down inactive components. Techniques such as Dynamic Voltage and Frequency Scaling (DVFS) and clock gating are fundamental here, as they allow the processor and accelerators to adapt energy usage to the computational demand in real time. This ensures that no excess energy is consumed when workloads are light or intermittent, thereby extending device lifetime. The RISC-V processor with specific enhancements for deep learning inference is a vital component of the energy-saving system. Unlike general-purpose processors, which consume high power for repetitive matrix and tensor operations, the RISC-V architecture can be extended with domain-specific instructions and accelerators tailored for machine learning tasks. By introducing lightweight vector instructions, multiply-accumulate (MAC) units [15], and custom accelerators for convolutional operations, the system is able to execute inference workloads with significantly reduced energy overhead. Offloading these operations from the general-purpose core to specialized hardware blocks further improves efficiency, enabling edge devices to perform real-time analytics without draining power reserves.

Another essential component of energy conservation is the memory subsystem. One of the most energy-intensive processes in computer systems is the transmission of data between the processor and memory. To mitigate this, the energy conservation system incorporates strategies such as data reuse buffers, intelligent caching, and compression techniques. These reduce the number of expensive memory accesses by reusing locally stored data whenever possible. By minimizing redundant data movement, the system ensures that energy is conserved while maintaining high throughput.

Communication efficiency is another dimension of the energy conservation system, particularly relevant in IoT scenarios. It takes a lot of energy and bandwidth to send raw sensor data or video feeds to cloud servers. This is addressed by the system's emphasis on edge-level inference, in which only crucial results are sent after local preprocessing and analysis of the data.This not only reduces the volume of data transmitted but also ensures faster response times while conserving energy. Lightweight communication protocols, adaptive transmission scheduling, and selective reporting further contribute to efficient energy utilization.

From an application perspective, the energy conservation system demonstrates its importance across multiple domains. In healthcare, wearable devices use these techniques to continuously monitor patient signals without frequent recharging. In industrial IoT, predictive maintenance systems leverage energy-aware computation to operate continuously in harsh environments. In smart cities, edge devices optimize surveillance and traffic management while operating under strict energy budgets. These examples highlight that the system is not just a technical framework but also a practical enabler of sustainable computing across sectors.

Edge Computing's Difficulties

Edge computing has potential, but a number of obstacles prevent its broad use. Because edge devices are frequently limited by low-power processors and batteries, one of the main problems is limited computational and energy resources, which makes it challenging to complete complicated machine learning workloads. Since edge networks are made up of several devices with different software, hardware, and networking capabilities, scalability and heterogeneity present additional difficulties that make system setup and standardization more difficult. Since sensitive data is processed locally and needs strong safeguards against breaches and illegal access, data security and privacy continue to be major concerns.

System performance can also be impacted by network latency and dependability, especially in situations when decisions must be made instantly. Lastly, there are major operational problems in managing and maintaining distributed edge devices, including software upgrades, fault tolerance, and resource allocation. To fully profit from edge computing in IoT and AI-driven applications, these problems must be resolved.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Here are the findings from the case study experiments. Different instruction extensions that were introduced to the original RISC-V ISA have been used to group the results.

4.1 Utilizing SIMD MAC instructions for performance

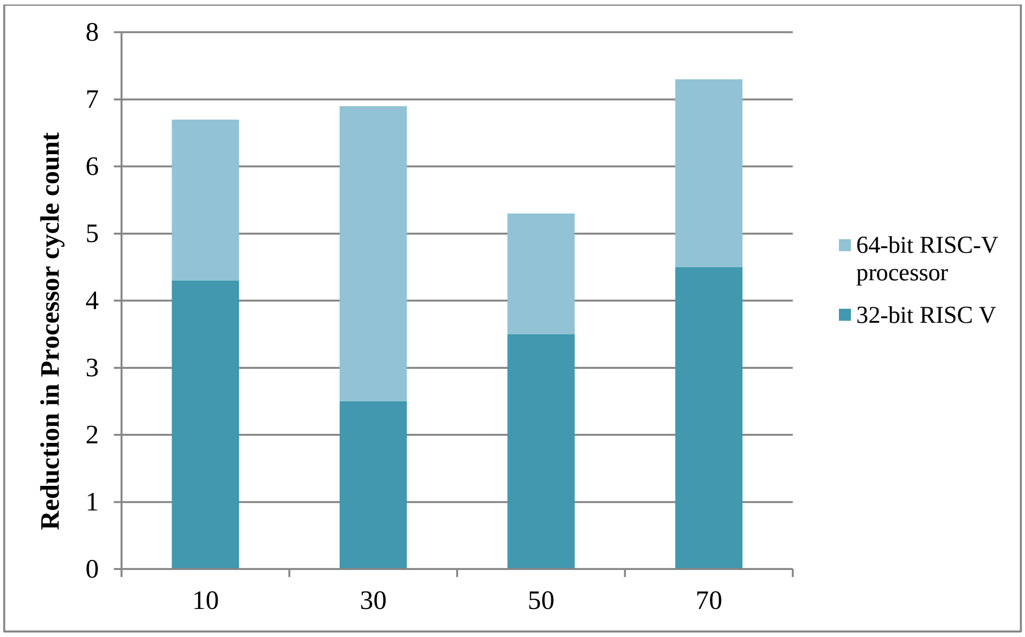

Figure 4. Enhancement of cycle count performance using SIMD MAC instruction

Given the abundance of data level parallelism found in neural networks, EXTREM EDGE was used to simulate a packed SIMD implementation of vectorized MAC instruction with four parallel vector lanes. This experiment produced an overall performance boost of 32% in 32-bit RISC-V processor designs and 45% in 64-bit RISC-V processor implementations, as illustrated in Figure 4 [16]. Using an actual scalar MAC instruction, the SIMD MAC experiment's results clearly demonstrate the benefits of the EXTREM-EDGE methodology, which was used to quickly and easily simulate SIMD MAC instructions in order to explore instruction design space. Without actually implementing the large vector "V" RISC-V ISA extension, EXTREM-EDGE users can study sophisticated vector ISA instructions by utilizing such ISA simulation techniques.

4.2 Effectiveness of in-memory VMM training

In end-to-end AI applications, a 17x gain on a small VMM kernel may not always correspond to an equivalent improvement. Therefore, we ran further tests to assess how in-memory VMM instruction affected the reference neural network model ResNet-8. The ResNet-8 benchmark neural network model saw an overall speedup of 4.41x thanks to in-memory VMM operations (including the overhead of VMM memory load/store sequences), as seen in Figure 5.

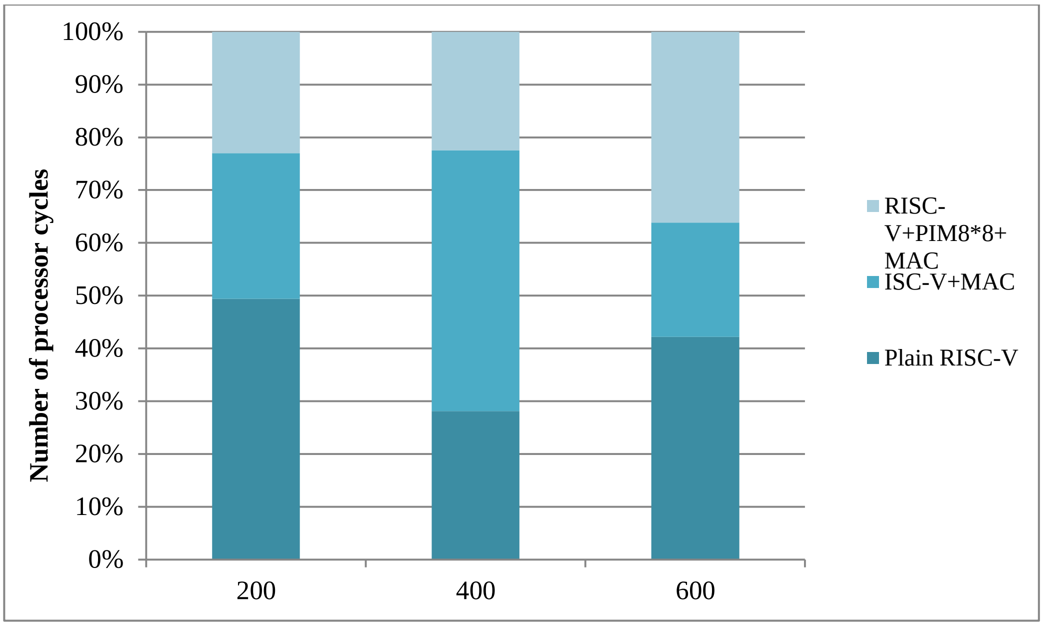

Figure 5. Comparing the ResNet-8 model's performance speedup due to MAC instructions and in-memory VMM instruction to that of a baseline RISC-V microprocessor

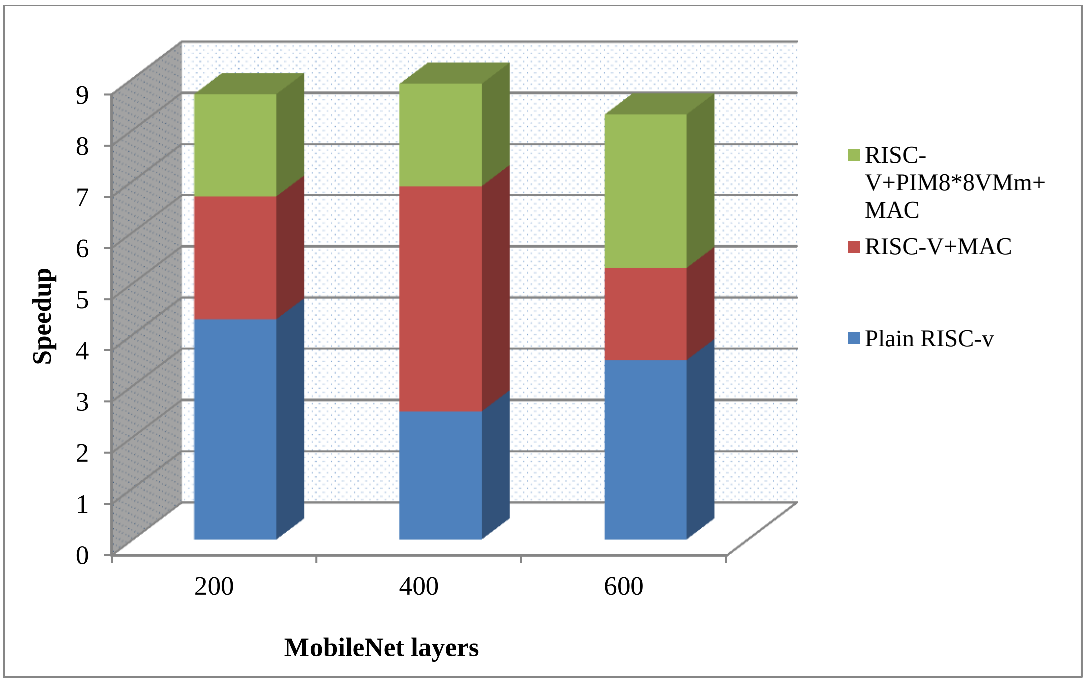

Figure 6. Comparing the MobileNet device's speedup due to MAC instructions and in-memory VMM instructions to that of a baseline RISC-V CPU

To better examine the effects of in-memory VMM in instruction, we assessed the results using the MLPerf Tiny benchmark's MobileNet V1 neural network model. The ResNet-8 model's typical 2D convolution layers are not the same as the depth-separable convolution layers seen in the MobileNet framework. For processing the entire MobileNet neural network model, EXTREM-EDGE improved performance by 1.35x with MAC procedures and 2.15x with the inclusion of in-memory VMM instructions to the RISC-V ISA, as shown in Figure 6.

5. CONCLUSION

The design of energy-efficient computer systems has become a necessity in the era of the Internet of Things, where billions of devices must process and transmit data under severe power and performance constraints. This study has shown that integrating RISC-V extensions with deep learning accelerators offers a highly effective pathway toward achieving energy-efficient machine learning inference at the IoT edge. By leveraging the modularity of the RISC-V instruction set, specialized vector operations, and domain-specific accelerators, the system significantly reduces computational overhead, optimizes memory usage, and lowers energy consumption while preserving inference accuracy and speed.

The exploration of applications across healthcare, smart cities, industrial IoT, and consumer electronics further demonstrates the practical relevance and scalability of this approach. At the same time, the incorporation of energy monitoring, dynamic power control, and communication-efficient strategies ensures that the architecture aligns with the sustainability requirements of large-scale IoT deployments. Ultimately, RISC-V based energy-efficient computer systems represent not just a technical advancement, but also a step toward creating a sustainable digital ecosystem. The insertion of AI functional units (AFU) to the processor pipeline by EXTREM-EDGE follows a strict integration strategy, enabling the execution of both AI and non-AI operations on the same processor. EXTREM-EDGE designs have demonstrated up to 4.41x speedup utilizing a set of 3 (vmm.ld, vmm.sd, and vmm) unique instructions on the ResNet-8 neural network simulator from the MLPerf Tiny test, and up to 45% performance improvement with a single MAC instruction added to RISC-V ISA.

REFERENCES

[1] Challa, N. (2021). Investigating the Potential of Enterprise Service Bus as a Fundamental Facilitator for Future Information Technology Infrastructure. Journal of Economics & Management Research. SRC/JESMR-275. DOI: doi. org/10.47363/JESMR/2021 (2), 208, 2-3.

[2] Javadi, M., Raeisi, Z., & Latifian, A. (2025). Enhancing Production Strategies Using Service-Oriented Architecture and Enterprise Service Bus in Manufacturing Companies. Journal of Business and Management Studies, 7(3), 318-332.

[3] Garofalo, A., Tagliavini, G., Conti, F., Benini, L., & Rossi, D. (2021). XpulpNN: Enabling energy efficient and flexible inference of quantized neural networks on RISC-V based IoT end nodes. IEEE Transactions on Emerging Topics in Computing, 9(3), 1489-1505.

[4] Liu, Q., & Amiri, S. (2025). Optimised Extension of an Ultra-Low-Power RISC-V Processor to Support Lightweight Neural Network Models. Chips, 4(2), 13.

[5] Radford, C. (2025). Design and Optimization of Low-Power RISC-V Processors for Edge AI Applications. Journal of Computer Technology and Software, 4(7).

[6] Garofalo, A., Tortorella, Y., Perotti, M., Valente, L., Nadalini, A., Benini, L., ... & Conti, F. (2022). DARKSIDE: A heterogeneous RISC-V compute cluster for extreme-edge on-chip DNN inference and training. IEEE Open Journal of the Solid-State Circuits Society, 2, 231-243.

[7] Garofalo, A. (2022). Flexible Computing Systems for AI Acceleration at the Extreme Edge of the IoT.

[8] Rutishauser, G., Mihali, J., Scherer, M., & Bonini, L. (2024, July). xtern: Energy-efficient ternary neural network inference on risc-v-based edge systems. In 2024 IEEE 35th International Conference on Application-specific Systems, Architectures and Processors (ASAP) (pp. 206-213). IEEE.

[9] Sordillo, S., Cheikh, A., Mastrandrea, A., Menichelli, F., & Olivieri, M. (2021). Customizable vector acceleration in extreme-edge computing: a RISC-V software/hardware architecture study on VGG-16 implementation. Electronics, 10(4), 518.

[10] Garofalo, A., & Benini, L. (2025). Leveraging RISC-V for HW/SW Co-Design of Flexible and Efficient TinyML SoCs. IEEE Design & Test.

[11] Taheri, F., Bayat-Sarmadi, S., & Hadayeghparast, S. (2022). RISC-HD: Lightweight RISC-V processor for efficient hyperdimensional computing inference. IEEE Internet of Things Journal, 9(23), 24030-24037.

[12] Angeline A, A. Risc-V Based Processor Design for Machine Learning Application on Edge. Sasipriya and R, Haridhrakajan, Risc-V Based Processor Design for Machine Learning Application on Edge.

[13] El Zarif, N., Hemmat, M. A., Dupuis, T., David, J. P., & Savaria, Y. (2024). Polara-Keras2c: Supporting Vectorized AI Models on RISC-V Edge Devices. IEEE Access.

[14] Hussain, T., Tahir, M. W., Mushtaq, M., & Khalid, S. (2025, February). Design and benchmarking of a low-cost RISC-V-based high-performance computing cluster for edge computing. In International Conference on Energy, Power, Environment, Control and Computing (ICEPECC 2025) (Vol. 2025, pp. 579-586). IET.

[15] Christofas, V., Amanatidis, P., Karampatzakis, D., Lagkas, T., Goudos, S. K., Psannis, K. E., & Sarigiannidis, P. (2023, September). Comparative Evaluation between Accelerated RISC-V and ARM AI Inference Machines. In 2023 6th World Symposium on Communication Engineering (WSCE) (pp. 108-113). IEEE.

[16] Tabanelli, E., Tagliavini, G., & Benini, L. (2022). Optimizing random forest-based inference on RISC-V MCUs at the extreme edge. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41(11), 4516-4526.