Journal of Integrated VLSI and Signal Processing • ISSN 3139-2369 (O) • Jan - Apr • 1 • 1 • 2026 (pp. 18-25)

RESEARCH ARTICLE

Low-Power Embedded VLSI Architectures for AI-Driven Signal and Image Processing Systems

Ashu Nayak*

Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India.

KEYWORDS:
Low-power VLSI architectures;
Embedded AI;
Signal and image processing;
Energy-efficient design;
AI accelerators;
Edge computing

ARTICLE HISTORY:

Received : 09.11.2025
Revised : 12.12.2025
Accepted : 18.01.2026

Abstract

The heavy increase in artificial intelligence (AI)-driven signal and image processing on embedded systems and edge-based systems has further increased the urge to utilise energy-efficient very-large-scale integration (VLSI) architecture. Traditional processor based solutions are too slow in moving data and energy proportional which corresponds to an inappropriate solution in power-constrained systems. The present paper discusses a low-power embedded VLSI architecture that is optimised to be used in AI-based signal and image processing applications with a focus on architecture-algorithm co-design and energy-efficient optimization. A common system model is established to account dynamic and leakage power terms and to measure energy efficiency in regards to energy / inference. The presented architecture is a combination of specialised AI acceleration and programmable signal processing data streams, with localised hierarchy of shared memory to minimise the overhead of data transfer. Some energy-efficient methods, such as adaptive dataflow management, clock and power gating, dynamic voltage and frequency scaling, and workload-aware precision adaptation are added. Compared to the conventional architecture, experimental analysis of the proposed architecture is able to establish similar power consumption and energy per operation, and simultaneous competitive throughput, which confirms the use of the proposed architecture in low-power embedded AI systems.

Author's e-mail: ku.ashunayak@kalingauniversity.ac.in

How to cite this article: Nayak A. Low-Power Embedded VLSI Architectures for AI-Driven Signal and Image Processing Systems. Journal of Integrated VLSI and Signal Processing. Vol.1, No. 1, 2026 (pp. 18-25).

Introduction

The ongoing blistering development of artificial intelligence (AI) has radically changed the system of signal and image processing and made possible advanced solutions, i.e. recognition of objects, speech analysis, medical signal interpretation, intelligent sensor, etc.[8–10] More and more, such AI workloads are being implemented on embedded and at the edge to achieve a high-level of requirements regarding latency, p rivacy, and bandwidth efficiency.[4, 7] Nonetheless, it is particularly difficult to run computationally intensive algorithms of AI on embedded systems that have resource constraints, in particular regarding the power consumption and energy efficiency.[2,5] Consequently, architectural design of low-power very-large-scale integration (VLSI) systems has emerged as a particularly important research topic in the design of next-generation signal and image processing systems.[1, 4] Traditional general purpose CPU based and embedded graphics architectures such as general-purpose CPUs and embedded GPUs tend to be poorly matched to embedded AI workloads because of their poor energy proportionality and excessive data transfer latency .[4, 5] With these architectures, memory access power often overshadows software power particularly in data-intensive signal and image processing applications.[2, 3] This inefficiency is further compounded by the fact that the modern workloads are heterogeneous (they consist of classical signal processing kernels and deep learning inference)[10, 11] in the first place. This make a pure power consumption optimization at the algorithm level or circuit level an inadequate that holistic architecture-algorithm co-designs be undertaken to experience significant energy-saving at the system level.[2, 6] More recent studies have examined the application of domain-specific accelerators and neural processing units in order to enhance performance per watt of AI workloads.[3,7,11] Although these solutions are beneficial in terms of offering

huge gains to a particular application, most of them lack flexibility, and their integration with the classic signal processing pipelines, or they are simply too costly to use in embedded systems.[11, 12] Still, techniques of aggressive reduction of power like the voltage scaling and power gating, when utilised without any architectural knowledge, may negatively impact on the performance robustness and the system reliability.[1, 5] Such constraints indicate the requirement of integrated embedded VLSI designs, which can collectively concentrate on computation, data movement, and power management in the same way.[2, 4] This paper suggests an embedded VLSI architecture with a low power consumption related to an AI-based signal and image processing system. The suggested solution stresses the close coupling between acceleration of AI and programmable signal processing data streams, with the aid of shared and locality-sensitive memory hierarchy.[11, 12] A combination of both architectural optimization techniques and adaptive power management techniques such as dataflow-aware execution, fine-grained clock and power gating, dynamic voltage and frequency scaling, and workload-aware precision adaptation is used to create energy efficiency.[2, 5, 6] A system model that is unified is developed to measure power use and energy efficiency to allow systematic assessment and compare it with traditional embedded baselines.[3, 7] The rest of this paper is structured in the following way. Section 2 provides a review of related work and determines the essential gaps in research. Part 3 includes the proposed methodology consisting of the system modelling, architectural design, and optimization methods of energy-saving. The implementation details and the experimental setup are explained in Section 4. Section 5 examines the energy efficiency and performance of the architecture in question. Section 6 deals with the implications and results, and the paper ends with an 7 which gives directions of a future research.

Associated Work and Research gap.

Signal and image processing systems that are energy efficient in VLSI architectures have been ample in the reaction to the mounting computational load of embedded and edge systems.[1, 4, 5] The first architectural design included low-power digital signal processing (DSP) designs, with the most common techniques to minimise dynamic power consumption: fixed-function accelerators, pipelined data paths, and coefficient optimization.[5, 10] Although such strategies resulted in significant energy efficiency of classical signal processing operations like finite impulse response (FIR) filtering and fast fourier transform (FFT), they are not flexible enough to handle modern AI-driven operations with high data parallelism and non-uniform access patterns.[2, 8] As the deep learning-based image and signal processing began to appear, the focus of research has moved to more specialised AI accelerators and neural processing units (NPUs).[2, 3, 7] These architectures leverage the idea of massive parallelism and reusing a data to enhance performance-per-watt of a convolutional and matrix-based operation.[3, 11] Other designs use the reduced precision arithmetic, systolic array structure to reduce energy of the computation.[3, 6] Nonetheless, currently numerous AI-oriented accelerators are specific to a particular neural network model and are not efficient to execute hybrid workloads that mix conventional signal processing and AI inference.[11, 12] Besides, there is inadequate architectural flexibility and inadequate underuse and overheads in terms of energy consumption when the architecture is applied in embedded systems.[4, 7]

Recent work has examined embedded and edge AI platform that uses AI accelerators with general-purpose processors or DSP cores.[7, 12] These heterogeneous mechanisms provide better flexibility, yet they often have great data moving overheads as related to disjointed memory hierarchies and inefficient interconnects.[2, 5] Consequently, memory access power emerges as a significant cause of overall power that can be seen when processing imag e related data and tasks that are data intensive.[2, 3] Also, aggressive power optimization methods, including voltage scaling and coarse-grained power gating are widely used without enough knowledge of workload behaviour to achieve optimal energy and performance and may disrupt activity performance.[1, 5] Table 1 contains a comparative summary of representative low-power VLSI architectures of signal and image processing. The table indicates areas of application, technology nodes, reported power or energy features and major constraints of current designs.[2, 3, 11, 12] As illustrated in the table, the past architectures are generally most efficient in signal processing or AI inference, but never in united and energy-aware models.[4, 7, 11]

Although such advances have occurred, there are still a number of research gaps. First, existing architectures to a significant extent assume that AI inference and conventional signal processing are different execution tasks, leading to the redundancy of hardware and the inefficient movement of data. Second, the majority of designs are concerned with computational efficiency and are underestimating the effect of memory hierarchy and dataflow on total energy consumption. Third, the power management plans tend to be primarily static and do not dynamically improve in response to different workload patterns which are prevalent with embedded systems. The latter restrictions impetrate the necessity of a unification of low-power embedded VLSI architecture

Table 1: Comparative Analysis of Existing VLSI Architectures for Signal and Image Processing

Architecture	Application	Tech Node	Power / Energy	Key Limitation
Low-Power DSP Accelerator	FIR, FFT	65 nm	Low dynamic power	Limited support for AI workloads
CNN Accelerator	Image classification	45 nm	High energy efficiency for CNNs	Poor flexibility, model-specific
Embedded GPU	Vision processing	28 nm	High throughput, high power	Excessive energy consumption
Heterogeneous SoC (CPU+NPU)	Edge AI	22 nm	Moderate energy efficiency	High data movement overhead
FPGA-Based Accelerator	Signal & image processing	28 nm	Reconfigurable, moderate power	Area and configuration overhead

tightly combining AI acceleration with signal processing data stream and utilising energy optimization methods that focus on workloads. The filling of these gaps is the main theme of the proposed methodology that is discussed below.

Methodology

In this part, the development of a proposal of the proposed approach in designing a low power embedded VLSI architecture to implement AI-driven signal and image processing systems is given. The methodology plays with the concept of system level modelling, system architecture and energy-efficient optimization element into a single system to conquer the problem of embedded execution under power constraints. This section is aimed at making the system model more formal, defining the optimization goals, and offering the analytical background of energy-efficiency assessment.

System Model and Problem Formulation

The target system available in the present paper is an embedded processing platform with heterogeneous workloads of classical signal processing kernels (e.g., FIR filtering and FFT) and AI-based image and signal inference tasks (e.g., convolutional neural networks). The common features of these workloads are high data reuse, varied intensity of computation, and high frequency of memory accessibility with power consumption and reduction of data flow being significant design factors. In order to allow the systematic analysis and optimization, an integrated system model is developed, so as to reflect the main contributors of power and energy consumption. The embedded VLSI system is represented as total power consumption, which represents the dynamic and leakage power consumption. Dynamic power comes with switching activity in calculations and through the movement of data, whereas leakage power comes about as a result of sub threshold and gate leakage currents in scaled CMOS technologies. The overall power usage is represented as

$$P_{total} = P^{dyn} + P_{leak} \quad (1)$$

Where P_{dyn} denotes the dynamic power and P_leak represents the leakage power. Switching activity, capacitance, supply voltage and operating frequency are the major influence factors of dynamic power. As can be estimated with the popular CMOS power model, it is approximated as.

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

Where α is the activity factor, C is the effective switched capacitance, V is the supply voltage, and f is the clock frequency. This equation eventually emphasizes the high quadratic vulnerability of power consumption on supply voltage and thus voltage and frequency scaling are promoted as a method to low-power design. Energy per inference is used as an important measure to analyse the energy efficiency of AI-based signal and image processing tasks. This measure is a combination of the computational and memory access energy and is especially well suited to the task of comparing heterogeneous architecture. The amount of energy required to inference is defined as.

$$E_{inf} = \frac{P_{total}}{T} \quad (3)$$

Where T denotes the effective throughput, measured as the number of inferences or processing operations completed per unit time. Minimizing E_{inf} forms the primary optimization objective of the proposed architecture, subject to constraints on throughput and functional correctness. According to the above formulation, the design objective of this work will be to minimise the overall energy usage with the acceptable performance of the mixed signal and image processing workloads. This goal is achieved by means of architectural integration, dataflow optimization and adaptive power management techniques that are covered in the following subsections.

Low-Power Proposed embedded VLSI architecture.

The suggested low-power embedded VLSI architecture is meant to serve efficiently the heterogeneous workloads

that involve classical signal processing tasks and AI-inspired signal and image inference with rigorous energy requirements. The major design consideration is to reduce movements of data and power as well as ensure adequate computational throughput to embedded application. To do this the architecture closely combines AI acceleration, customisable signal processing data paths, a hierarchical shared memory, and energy conscious control mechanisms in a single framework. The block diagram of the proposed embedded VLSI architecture is provided in figure 1. As the figure shows, the initial input in the system is the raw signal and image data on sensor and image based inputs, which is acquired using the input interface. An important signal conditioning processing like filtering, framing and spectral transformation is carried out in a pre-processing unit to prepare the data to go through other processing processes. Such operations save unnecessary computation followed by increasing locality of data, which will save on energy usage in subsequent stages.

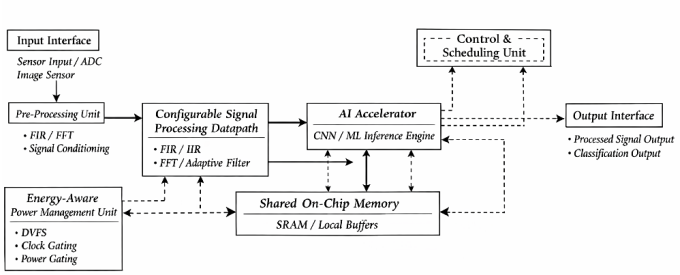

Fig. 1: Block Diagram of the Proposed Embedded VLSI Architecture

Two tightly coupled processing subsystems are used in the core computation an adjustable signal processing data path and a specific AI accelerator. The data path takes care of the usual operations of DSPs such as FIR/IIR filtering, FFT and adaptive filtering and therefore they are able to run classical signal processing kernels. Simultaneously, the AI accelerator is configured to be machine learning inference optimised, with convolution and activation functions, and implemented with a parallel multiplyaccumulate (MAC) array and adjustable precision to achieve better energy efficiency. The two subsystems are closely interconnected, enabling yet another simple execution of hybrid tasks, which are signal preprocessing coupled with AI-based inference. An on-chip common on-chip memory hierarchy is of primary importance in lowering the overhead of data movement. Figure 1 shows on-chip SRAM and local buffers are shared between signal processing data path and AI accelerator which allow intermediate data and weights to be reused efficiently. It is important to note that this shared memory concept helps to minimise expensive accesses to external memory which are known to consume most of the energy in the embedded AI systems. The efficiency of energy is also increased using an energy-conscious power management unit (PMU) and a centralised control and scheduling unit. Clock gating, power gating, and dynamic voltage and frequency scaling (DVFS) are all dynamically used by the PMU basing on the characteristics of the workload and the activity of the system. The control and scheduling unit orchestrates the execution of tasks, controls dataflow and movements between processing units and balances the performance and energy consumption by adjusting the precision and operating mode. The output interface eventually provides out information on processed signals and inference results. The building arrangement and the engineering functions of the key elements can be outlined in Table 2. The table points out the role that each field plays in this proposed design in ensuring efficient functioning and low-power activities (Table 2).

Optimization and Power Management Techniques should be energy-saving.

The concept of energy efficiency of AI-based signal and image processing systems cannot be asserted to be possible by merely having fixed architecture development; it needs the ability to dynamically adjust to workload

Table 2: Architectural Parameters of the Proposed Design

Component	Configuration	Function
Input Interface	Sensor / Image input support	Acquires raw signal and image data
Pre-Processing Unit	FIR, FFT, signal conditioning	Prepares data and improves locality
Signal Processing Data path	FIR/IIR, FFT, adaptive filter	Executes classical DSP operations
AI Accelerator	Parallel MAC array, configurable precision	Performs AI-based inference
Shared On-Chip Memory	SRAM, local buffers	Reduces data movement and energy
Control & Scheduling Unit	Centralized controller	Manages task mapping and execution
Power Management Unit	DVFS, clock and power gating	Enables energy-aware operation
Output Interface	Digital output interface	Provides processed results

specifics, data flow, and performance demands. To deal with it, the proposed architecture takes into account an all-inclusive suite of energy-efficient optimization and power management strategies that are utilised in a co-ordinated fashion within the computation, memory, and control layers. These methods are systemized in an energy conscious dataflow and power control system, as shown in Figure 2. Based on the Figure 2, an energy-conscious dataflow controller controls workload execution and dynamically schedules AI inference and signal processing tasks according to their computational intensity and potential reuse of data. The controller can easily optimise data locality by choosing suitable tiling, buffering and execution order and minimise redundant accesses to memory as they are a prime source of energy consumption in embedded systems. This dataflow-aware implementation model can be used effectively to achieve efficient use of the signal processing data path and the AI accelerator in different workload scenarios.

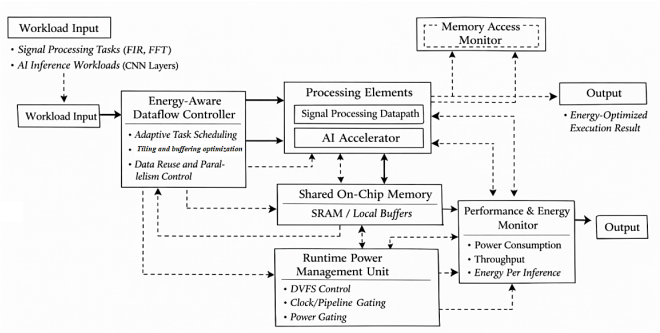

Fig. 2: Energy-Aware Dataflow and Power Management Framework

The model also incorporates a runtime power management unit (PMU) which implements adaptive power-saving policies in the reaction to real-time system activity. Dynamic voltage and frequency scaling (DVFS) is used to change operating point according to throughput needs, as well as small-scale clock gating can selectively cross-fuelling of idle functional units and minimise the unnecessary switching activity. Moreover, the power gating is employed to entirely disable idle blocks when low utilisation states occur hence limiting leakage power. These systems work under the control of the PMU that is constantly fed with system monitors. The activity and performance monitoring of memories is imperative to close-loop optimization of energy. Figure 2 shows that a memory access monitor monitors the intensity of read/writes and the use of buffers, whereas a performance and energy monitor measures the level of power, throughput, and energy per inference. The resulting metrics are recapitulated to both the dataflow controller and the PMU with regards to the possibility of adjusting the execution and power states of the system in real-time. This is through closed loop feedback where energy efficiency is not compromised to performance associated requirements. The proposed framework through deep learning of workload-sensitive dataflow control and adjustable power management attains significant energy savings in both dynamic and leakage power and still maintains the speed to complete AI-based signal and image processing workloads. These techniques are quantitatively measured in results section where the effectiveness of these methods is compared.

Experimental Flow and Execution.

In this section, the description of the experimental set up and the methodology of implementation to test the proposed low-power embedded VLSI architecture has been described. The experimental evaluation aims at confirming the efficiency of the architectural design and energy-conscious optimization strategies described in the Section 3.2 and 3.3, and shown in Figure 1 and Figure 2, when put under the typical AI-driven signal and image processing workloads. The proposed architecture was executed based on a hardware-oriented design flow, which lets one approximate power and performance as well as the efficiency of the energy use. Figure 1 represents the architectural elements of the building whereby the signal processing data path was presented as configurable, the AI accelerator, shared on-chip memory, and power management, were all modelled and built into a single simulation environment. The dataflow and runtime power management framework shown in Figure 2 was energy-aware to change the state of execution as well as power dynamically at runtime as a control layer.

Implementation Platform and Tools.

The design was tested on a representative embedded hardware platform and design tools of industry standard. The AI accelerator/ signal processing modules operated with settings that accommodate mixed workloads and was thus able to run classical DSP kernels as well as neural network differencing. The estimation of power was done by factoring in activity based switching models and leakage estimates at scale CMOS technologies. There was also the enabling of dynamic voltage and frequency scaling, clock gating, and power gating mechanisms to evaluate their effect on the energy consumption.

Standards and Measures of Evaluation.

A series of representative signal and image processing workloads were chosen to represent the realistic embedded AI workloads. Signal processing can use finite impulse response (FIR) filtering and fast fourier transform (FFT) and image processing loads are founded

on convolutional neural network (CNN) inference layers. The metrics applied to performance and energy efficiency were power consumption, throughput, and energy per inference, which aligns with the system model stated in Section 3.1.The experimental set up and benchmark measurements could be found in Table 3 that presents the central parameters needed during the performance and energy efficiency evaluation.

Table 3. Experimental Setup and Benchmark Configuration

Parameter	Value
Implementation Platform	Embedded hardware simulation / FPGA prototype
Technology Node	28 nm CMOS
Operating Frequency	100–300 MHz
Supply Voltage	0.7–1.0 V (DVFS enabled)
Signal Processing Benchmarks	FIR filtering, FFT
AI Benchmarks	CNN inference layers
Memory Configuration	On-chip SRAM with local buffers
Power Optimization Techniques	DVFS, clock gating, power gating
Evaluation Metrics	Power, throughput, energy per inference

This controlled experiment will give a balanced and thorough analysis of the proposed architecture and energy-efficient optimization methods. The findings derived in this system are discussed and reported in the next section.

Performance Evaluation and Results.

The present section contains a quantitative validation of the proposed low-power embedded VLSI to the proposed experimental setup described in Section 4. The analysis is based on the major performance indicators, such as power consumption, energy per operation, throughput and area overhead, to determine the efficiency of the architecture suggested and energy-conscience optimization methods in the case of mixed signal and image processing workloads. The first analysis is of power consumption and energy efficiency due to the need to determine the appropriateness of the proposed design to energy-limited embedded systems. Figure 3 takes into contrast the power consumption and energy per operation in the proposed architecture with a representative base embedded architecture in signal processing and AI-based image processor workloads. The proposed architecture uses less power at all of the workloads considered, as demonstrated in Figure 3. The major cause of such reduction is the data shared on-chip memory hierarchy that cuts the expensive data traffic, and an execution aware of workloads that enhances the use of resources. Moreover, the energy per call is greatly minimised, especially when technologies based on AI are used to handle image processing functions, where memory accessibility energy usually contributes to overall consumption. The use of dataflow-efficient execution, clock gating, power gating, and dynamic voltage and frequency scaling are effectively used to prevent dynamic and leakage power.

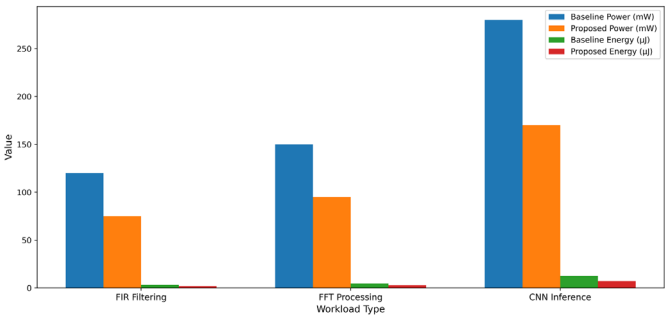

Fig. 3: Power and Energy Consumption

Queuing during energy constrained performance is tested by examining the relationship between throughput and energy efficiency. The throughput versus the energy efficiency trade-off of the proposed architecture compared with the baseline design is shown in Figure 4. The proposed architecture in comparison to baseline architecture is expected to exhibit energy proportionality at a broad range of throughput values as illustrated in Figure 4. Notably, the gains in energy efficiency do not cause throughput to drop and this means that performance scaling does not result in a proportional increase in energy consumption. This usage demonstrates the usefulness of the energy-conscious power management framework of dataflow/runtime balancing performance and energy usage.

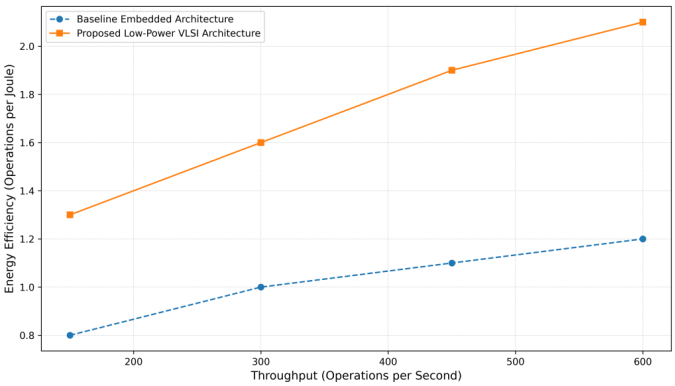

Fig. 4: Throughput vs Energy Efficiency Trade-off

Along with the measures of power and performance, one can also look at the area overhead in order to evaluate

the implementation feasibility. A small increment to silicon area is presented by the implementation of the AI accelerator, configurable signal processing data path and power management logic over baseline designs. Nevertheless, such overhead can be removed through the removal of unnecessary processing units, as well as through the adoption of common on-chip memory resources. On the whole, the designed architecture delivers a desirable ratio between power consumption, energy efficiency, throughput, and area, which is why the proposed architecture is a good fit in terms of embedded AI-driven signal and image processing applications.

Discussion

As it has been shown in the experimental results, the suggested low-power embedded VLSI architecture is capable of homogeneous tradeoff in terms of power efficiency and energy proportionality between the heterogeneous signal and image processing workloads. The power consumption decrease and energy per operation reduction in Figure 3 proves the point that architectural integration and optimization of data movements have a decisive role on exhibiting better energy efficiency on embedded AI systems. Specifically, the on-chip memory hierarchy, in which the common shared on-chip memory is used very heavily, greatly alleviates the energy premium commonly linked to the frequent off-chip memory accesses that is often the bottleneck in embedded systems in traditional architectures. The trends of throughput versus energy efficiency exhibited in Figure 4 helps to understand the scalability of the proposed design even better. The proposed architecture has a desirable energy-performance balance, unlike that of baseline architectures that show a decline in energy efficiency with increased throughput over a large operating range. This observation underscores the success of dataflow control that is energy conscious as well as adaptive power management in facilitating operation that is energy proportional. The architecture prevents over-provisioning resources when workload intensity and system activity are low as voltage, frequency, and power states are dynamically adjusted to meet these demands, and when larger throughputs are needed, the architecture scales to higher throughputs.

The other notable observation is diverse response of the proposed techniques on signal processing and AI-based workloads. Although the two classes involve decreased power consumption, the energy savings are increased in the case of image processing tasks facilitated by AI. The result can be explained by an increase in data reuse and computational density of neural network inference, which was successfully utilised by the built-in AI accelerator and dataflow-conscious implementation plan. By comparison, the workloads of classical signal processing, typically either more regular (e.g. DSP) or a streaming behaviour, are well suited to clock gating and lower memory access overhead. The trade-off presented by the simple area overhead added by adding AI acceleration and power management logic is a viable compromise. Whereas the suggested design proposes extra expenditure on devices than undergirding base architectures, this disadvantage is substituted by the extenses in both energy efficiency and performance growing. System level wise, the enhanced energy efficiency can also be expressed into longer battery life, diminished thermal strain, and reliability, that are vital factors to embedded and edge computing platforms. Altogether, the discussion provides insight that addressing energy efficiency in embedded AI systems is most effective with the help of the holistic perspective that coordinates the architecture, dataflow, and the runtime power management. The given design shows that this kind of integration can potentially result in significant energy efficiency without compromising on performance, which gives a reasonable basis to next-generation AI-based signal and image processing systems.

Conclusion and Future Work

The article described a low-power embedded VLSI architecture to design signal and image processing systems based on AI, to meet the increasing need to execute advanced computer computations on resource-constrained embedded systems. Through a close combination of an AI accelerator, a customizable signal processing data path, and a common on-chip memory hierarchy, the architecture proposed ignores data flow by ensuring minimal data mobility and also enhances the use of resources. Besides, an energy-sensitive dataflow and power management system was also proposed to dynamically adjust the state of execution and power using the nature of the workload and system activity. Extensive experimental results have shown that the suggested architecture is capable of recording high power savings and energy per operation in addition to competitive throughput in diverse workloads of heterogeneous processors. The comparisons in power and energy and the trade-off analysis of throughput and energy efficiency prove the proposed design provides better energy proportionality at the expense of performance. The architecture is suitable to embedded and edge AI applications despite having a small overhead of energy usage as the resulting advantages in energy efficiency and scalability are promising. The upcoming research will also consider the expansion of the proposed architecture to a fully ASIC implementation and silicon validation in further determination of benefits in terms of energy and area. Other directions mentioned are to use reliability

and thermal-conscious optimization, enable dynamically reconfigurable AI, and focus on closer integration with new memory technologies. The extensions will also extend the usability of the suggested framework to the next-generation low-power intelligent embedded systems.

References

1. Chen, Y. H., Yang, T. J., Emer, J., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9*(2), 292–308. https://doi.org/10.1109/JETCAS.2019.2910237
2. Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing, 7*(3–4), 197–387. https://doi.org/10.1561/2000000039
3. Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., & Dally, W. J. (2016). EIE: Efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News, 44*(3), 243–254. https://doi.org/10.1145/3007787.3001163
4. Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. *Communications of the ACM, 62*(2), 48–60. https://doi.org/10.1145/3282307
5. Horowitz, M. (2014). Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (pp. 10–14). IEEE. https://doi.org/10.1109/ISSCC.2014.6757323
6. Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... Yoon, D. H. (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture* (pp. 1–12). https://doi.org/10.1145/3079856.3080246
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.
8. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
9. Rieger, M. L. (2019). Retrospective on VLSI value scaling and lithography. *Journal of Micro/Nanolithography, MEMS, and MOEMS, 18*(4), 040902. https://doi.org/10.1117/1.JMM.18.4.040902
10. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE, 105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740
11. Yang, T. J., Chen, Y. H., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5687–5695). https://doi.org/10.1109/CVPR.2017.599
12. Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., & Cong, J. (2015). Optimizing FPGA-based accelerator design for deep convolutional neural networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (pp. 161–170). https://doi.org/10.1145/2684746.2689060