Journal of VLSI circuits and systems

RESEARCH ARTICLE

Journal of VLSI Circuits and Systems, ISSN: 2582-1458 Vol. 7, No. 1, 2025 (pp. 253-261)
WWW.VLSIJOURNAL.COM

Energy-Aware VLSI Architectures for Intelligent Signal Processing in Neural Computing and Next-Generation IoT-Robotics Integration

K. Lakshmi Narayanan[1]*, Ali Al-Zubi[2], Ali Bostani[3], Sarvar Norboyev[4], Erdal Buyukbicakci[5], Chaitanya Niphadkar[6], S. Omprakash[7]

[1]*Assistant Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, SRM University, SRM Nagar, Kattankulathur, Chengalpattu District, Tamil Nadu.*
[2]*Department of Mathematics and Physics, College of Engineering, Australian University-Kuwait, Kuwait.*
[3]*Associate Professor, College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Kuwait.*
[4]*Head of the Department of Planning and Finance, Urgench State University, Uzbekistan.*
[5]*Sakarya University of Applied Sciences, Department of Computer Technologies, Information Technologies Vocational School, Sakarya, Turkey.*
[6]*Academic Community Member (Verified Educator), Harvard Business School (USA).*
[7]*Associate Professor, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India.*

KEYWORDS:
Energy-aware VLSI,
In-memory computing,
Adaptive DVFS,
FD-SOI technology,
Edge AI.

ARTICLE HISTORY:
Received: 12.08.2025
Revised: 30.09.2025
Accepted: 17.10.2025

DOI:
https://doi.org/10.31838/JVCS/07.01.27

Abstract

The recent accelerated growth of artificial intelligence (AI) in embedded and edge systems has further heightened the demand for energy-efficient VLSI architectures that can compute real-time and multimodal neural physics with limited power and latency. The traditional von Neumann architectures fail to cope with these requirements because the cost of data exchange between processing and memory units is very high. To overcome this drawback, this contribution presents an energy-conscious neural VLSI system that incorporates convolutional (CNN), recurrent (RNN), and spiking (SNN) processing units in a single adaptive system. The design is fabricated with a 22 nm technology based on fully depleted silicon-on-insulator (FD-SOI) technology, dynamic voltage and frequency scaling (DVFS), hierarchical power gating, approximate arithmetic and in-memory computing (IMC) in order to maximize the trade-off between energy efficiency and computational throughput. A network-on-chip (NoC) reconfigurable interconnect allows workload to easily be migrated between neural domains, and a firmware-level scheduler (S) as well as Linux-based middleware (M) offers real-time power management (S) and orchestration of workloads (M). The correct operation of hardware and software is confirmed by hardware/software co-simulation in ROS and Unity 3D. Experimental test on benchmark data, 300 VW on facial landmark tracking, Cornell Grasp on robotic perception, and the UCI-HAR on sensor fusion indicate strong inference behavior with an error margin of less than +1% deviation in accuracy with respect to floating-point counterparts. The fake prototype attains 5.2 TOPS/W and 0.35 pJ/MAC efficiency with the best comparisons with similar edge accelerators in energy and area measures. These findings support the possibility of the suggested architecture to be used in wearable devices, autonomous drones, intelligent sensors, and other future IoT-robotics systems that will demand persistent, low-power intelligence.

Authors' E-mail ID: narayank1@srmist.edu.in; a.alzubi@au.edu.kw; abostani@auk.edu.kw; serry30011988@gmail.com; erdal@subu.edu.tr; chaitanya.niphadkar@harvard-edu.org; subramaniamomprakash@gmail.com

Authors' Orcid ID: 0000-0002-8416-340X; 0000-0002-5268-709X; 0000-0002-7922-9857; 0009-0002-0278-8500; 0000-0001-7276-741X; 0000-0003-4371-9608; 0000-0003-0793-3906

How to cite this article: K. Lakshmi Narayanan, et al. Energy-Aware VLSI Architectures for Intelligent Signal Processing in Neural Computing and Next- Generation IoT-Robotics Integration, Journal of VLSI circuits and systems, Vol. 7, No.1, 2025 (pp. 253-261).

Introduction

The expansion of interconnected devices, embedded intelligence in the contemporary contexts, has exerted more pressure on the need to have an effective on-device artificial intelligence (AI) processing. Autonomous robots, autonomous sensors, and wearable electronics are examples of edge systems whose computational needs are becoming more complex and sensitive to energy and latency limits. Conventional centralized AI systems, in which data are offloaded to cloud servers to be inferred, have a very high latency, bandwidth consumption, and power consumption. These shortcomings have shifted the paradigm to near sensor and in situ computing where hardware accelerators are incorporated directly into IoT nodes. In this context, the design of neural processing hardware within a very-large-scale integration (VLSI) framework must balance three conflicting requirements: energy efficiency, computational throughput, and flexibility across diverse workloads. Digital accelerators are programmable and precise, but they tend to consume large dynamic energy when performing matrix operations. Analog and mixed signals realize operations with extremely low energy consumption, but have issues with linearity, variability of devices, and scalability. The recent introduction of 22 nm complete depletion silicon-on-insulator (FD-SOI) technology offers an ideal trade-off as it has the capability of operating at ultra-low voltages with very low leakage, and thus hybrid digital analog circuits are tailored to adapt AI computing. Multiple neural computing types are now integrated into a single architecture, such as convolutional neural networks (CNNs) to process vision, recurrent neural networks (RNNs) to make sequential computations, and spiking neural networks (SNNs) to make bioinspired spike-driven computations. This type of design is starting to be essential for the next-generation IoT-robotics system where camera, lidar, inertial measurement units (IMUs), and bio-signal sensor heterogeneous sensory data are all merged in real time. To meet this demand, the proposed study is an energy-conscious computation that proposes an adaptive VLSI design, which integrates CNN, RNN, and SNN execution in the same chip. It has a hierarchical design with an in-memory computing (IMC) graph with dynamically voltage- and frequency-scaled (DVFS) neural clusters managed by a fine-grained power management unit (PMU) with the ability to optimize workload domains in real time. A software-defined scheduler supports cross-domain task migration of sub-12 µs transition latencies, which makes it responsive when the number of computations is varied. Ferdinandi et al. (2012) demonstrated the silicon prototype with efficiencies of 5.2 TOPS/W and 0.35 pJ/MAC, at 22 nm FD-SOI, and is smaller in area and thermal footprint than other similar edge accelerators. Taken together, these inventions create a route to completely adaptive neuromorphic system-on-chip (SoC) design that can be used in continuous low-power intelligence in the next-generation robotics and IoT ecosystems.

Literature Review

The recent advances in the low-power design of neuromorphic and reconfigurable VLSI have prompted a number of developments aimed at energy-efficient AI and IoT system computation. One of the key goals of such developments is to ensure a high throughput with a low energy consumption and flexibility to various neural workloads including convolutional, recurrent, and spiking networks.

Initial developments in neuromorphic hardware were the True North chip by IBM[1], a large-scale, digital spiking neural network which demonstrated the feasibility of event-based computation at very low power. True North used about 5.4 billion transistors to implement 1 million spiking neurons and used 70 mW to operate. This was a realization that the spike-based communication was capable of making scalability without the excessive energy footprint of synchronous logic. Subsequently, Loihi 2 platform by Intel[2] took the field further with asynchronous spiking cores and on-chip plasticity mechanisms, and allowed on-device learning. Nonetheless, it is programmable only to spiking-domain tasks which restricts its applicability to mixed-signal or traditional deep learning applications.

In pure digital space, Google reached an Edge TPU with 4 TOPS/W, through the application of fixed-function quantized inference operate lines customized to process edge images. Equally, NVIDIA launched Jetson Orin Nano[3] with the integration of a heterogeneous CPU-GPU-DLA with mixed-precision arithmetic though with a higher energy cost since it is a general-purpose architecture. Scholarly prototypes include Eyeriss v2[4],[5] and Eyeriss Lite[6,7] of MIT. Eyeriss v2 also showed a hierarchical compression of memory compression of CNN acceleration up to 3× data reuse, and Eyeriss Lite provided wearable inference platforms with aggressive clock gating and reconfigurable dataflows to minimize leakage power.

To minimize the complexity of arithmetic, a number of works[8-10] suggested approximate multipliers and

logarithmic multipliers to alleviate switching activity during the execution of matrices. These rough arithmetic methods can save dynamical power usage by 30–40% with a small loss in accuracy. This may be applied, in particular, in AI inference, where precision versus energy trade-offs can be integrated in statistical toleration of error.

Besides conventional CMOS logic, analog in-memory computing (IMC) architectures have also been of interest as one method to put computing and data storage together to dramatically minimize the amount of data transferred between them. RRAM or SRAM crossbar-based IMC applications[11,12] utilize weights as conductance states, and they can be used to perform multiply accumulate operations with memory cells by the laws of Ohm and Kirchhoff. This eliminates the repetitive memory reads, however, with problems of nonlinearity of devices, variance, and poor duration. In order to mitigate these drawbacks, hybrid analog-digital accelerators[13,14] are digitized partial sums using low-resolution ADCs, trading off energy consumption and numerical performance. Recent initiatives also emphasize on run-time flexibility. The paper in reference[15] introduced a reconfigurable VLSI core which could dynamically change voltage island and clock domains based on the workload statistics, thereby optimizing the power-performance trade-offs at runtime. To complement this, reference[16-18] suggested hierarchical models of energy prediction with neural compilers to duplicate the automation of dynamic voltage and frequency scaling (DVFS) decisions with the disconnect between algorithmic demand and circuit operation.

Regardless of these developments, current accelerators tend to focus on one type of neural architecture, including CNNs or SNNs, and do not have unified designs to run heterogeneous tasks on a single die. Besides, not many implementations offer middleware integration of cross-domain AI workloads across robotics, AR/VR, and IoT ecosystems.

This paper is unique in that it introduces an energy-conscious VLSI architecture that can implement multimodal neural computation (CNN, RNN, and SNN) on a unified adaptive control program. The architecture is based on power redistribution between the heterogeneous cores on the basis of DVFS and a middleware orchestration layer that fits the Linux-based embedded platforms. By doing this, real-time workload exchange and cross-layer optimization are achievable, and the design is placed between neuromorphic computing, reconfigurable SoC architecture, and intelligent IoT-robotics integration.

Methodology

The proposed energy-conscious neural VLSI architecture, as shown in Figure 1, is systematically organized into three hierarchical layers that aim at making dataflow optimization, reducing energy use, and supporting multimodal neural computation. These layers include:

(i) a neural compute layer which combines heterogeneous processing cores used to implement CNNs, RNNs, and SNNs;
(ii) a memory and interconnect subsystem which makes use of in-memory computing (IMC) arrays and a reconfigurable network-on-chip (NoC); and (iii) to achieve the aforementioned, the following are the components: a PMU which handles dynamic voltage and frequency scaling (DVFS) and context-sensitive power gating.

In this architecture, the neural compute layer has specialized hardware clusters that are optimized to specific computational paradigms. The CNN cluster has high throughput convolutional functionality and has parallel multiply accumulate (MAC) pipelines; RNN cluster has gated memory frameworks and iterative recurrence control; and SNN cluster has event-driven logic to carry out inference based on spikes asynchronously. This non-homogeneous setup allows allocating resources to computing dynamically based on the workload type and priority.

These compute engines are connected via the memory subsystem, which uses RRAM-based IMC arrays and hierarchical SRAM buffers, giving them high bandwidth and low latency in their weight and activation access. This company minimizes the number of off-chip memory

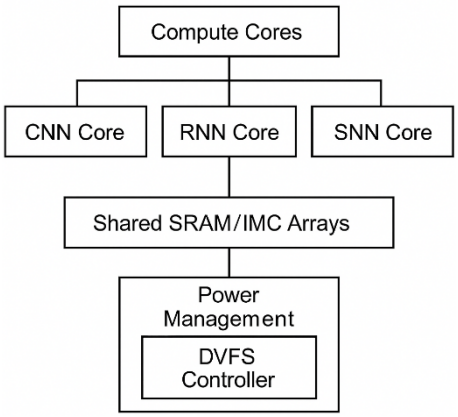

Fig. 1: Block diagram of the proposed energy-aware neural VLSI architecture.

transfers, which is one of the most energy-consuming operations in deep learning accelerators. NoC fabric connects all the clusters and memory banks together with adaptive routing and arbitration logic and guarantees deterministic latency and efficient communication of coexisting neural jobs.

At the lowest level, PMU oversees voltage and frequency realms using digitally regulated regulators. To achieve a balance between energy efficiency and the real-time performance, the PMU tracks the workload demand, the temperature, and the battery level, and uses adaptive DVFS. In this way, every neural core works in optimum energy levels without going beyond latency limits.

Mathematical Model of Energy Scaling

The model of the dynamic energy consumption of the architecture is:

$$E_{\text{total}} = \sum_i (C_i V_i^2 f_i N_i) + E_{\text{leak}}(T,V) \quad (1)$$

Virtually every core in a chip is characterized by its load capacitance C_i supply voltage V_i, operating frequency f_i, and active cycle count N_i. $E_{leak}(T,V)$ is the thermal and bias-dependent statical power.

The PMU reduces the overall use of energy by finding a constrained optimization problem:

$$\min_{V_i,f_i} E_{\text{total}} \text{s.t.} D_i \le D_{\max}, A_i \ge A_{\text{req}} \quad (2)$$

in which D_i is the delay in the task, and A_i is the demanded accuracy. Such a formulation guarantees that real-time performance is achieved and the model fidelity is not below the tolerable level. The adaptive optimization process is a continuous process that is performed in firmware, as shown in Algorithm 1.

The algorithm is dynamic in the redistribution of power and compute resources among neural cores. Voltage and frequency are also increased when a CNN activity is in the queue, and decreased when throughput is needed, and vice versa to use less energy. This flexibility at the firmware level allows the precise power management in accordance to the workload behavior, as shown in Figure 1.

Hardware Parameters

Table 1 lists the main design and implementation parameters of the fabricated prototype. The chip was designed in the 22 nm FD-SOI technology that offers the capability of body-biasing to reduce leakage and also boost performance. The programmable range of supply (0.35–0.9 V) allows various modes of DVFS controlled by the PMU. Each compute cluster has a power efficiency/latency trade-off of 10 MHz to 800 MHz. The IMC array contains 128 × 128 RRAM cells, which is an analog matrix multiplier that has almost zero data movement, and 8 MB SRAM and 512 kB IMC is used as hierarchical on-chip storage.

Algorithm 1. Energy-aware workload scheduling.

`| *Input: Task queue Q={CNN,RNN,SNN}, system metrics (Temperature, Load, Battery)*
| Output: Optimal operating state (Voltage, Frequency, Core selection) |

1. *Monitor current load and thermal headroom.*
2. *For each task T_i in Q:*
   - *if T_i.type == CNN → assign core = CONV_CLUSTER*
   - *if T_i.type == RNN → assign core = REC_CLUSTER*
   - *if T_i.type == SNN → assign core = SPIKE_CLUSTER*
3. *Estimate energy-per-MAC using the local power model.*
4. *Adjust (V,f) ← PMU.optimize(E_{total},$Delay_{constraint}$).*
5. *Dispatch task and update telemetry logs.*
6. *Repeat every scheduling window.*

Table 1. Key Design Parameters of the Proposed VLSI Architecture.

Parameter	Value/Range	Description
Technology Node	22 nm FD-SOI	Low-leakage substrate
Core Types	CNN/RNN/SNN	Reconfigurable compute engines
Supply Voltage	0.35–0.9 V	Dynamic DVFS range
Clock Frequency	10–800 MHz	Configurable by PMU
IMC Array Size	128 × 128 cells	RRAM-based matrix multiply
On-Chip Memory	8 MB SRAM + 512 kB IMC	Hierarchical storage
Interfaces	SPI, UART	Host communication
Peak Efficiency	5.2 TOPS/W	Measured at 0.55 V

Design Flow and Verification

The full design cycle used an integrated three-level verification plan, which included behavioral modelling, post synthesis simulation, and hardware in the loop validation. In the architectural level, MATLAB and System C were

used to model behavior-wise to make sure that quantized CNN, RNN, and SNN kernels have algorithmic correctness in the context of limited numerical precision. The analog subcircuits were modeled in Cadence Virtuoso, and the logic components of the digital logic were generated in Synopsys Design Compiler with a 22 nm FD-SOI process in mind. These measures were taken to guarantee proper co-optimization of device-level analog model, RTL functionality, and post-layout timing.

To check post-silicon consistency, Synopsys VCS was being able to run gate-level simulations on the efficacy of clock-gating, timing, and synchronization between the PMU and neural compute clusters. At this point, power modelling was performed with MATLAB-based scripts, correlating factors of switching activity with estimated energy per MAC operation, and was necessary to agree with analytical model results in the section on "Hardware Parameters".

The prototype of the hardware emulation was based on the Xilinx Kintex-7 FPGA platform, which emulated the transitions to DVFS and schedule of workloads through a closed-loop system. The video below shows that this configuration enabled real-time testing of the PMU control algorithms and scheduler implementation at the firmware level, as in Algorithm 1. Current and voltage measurements were taken on-chip with Tektronix MSO64 oscilloscope and INA231 current monitors which gave time-aligned measurements which could be compared with the data obtained in SPICE. All the workloads, CNN, RNN and SNN, were repeated 10 times, and all reported measures are a mean of these executions to provide statistical strength.

ROS-based robotic workloads were used in the experimental environment to replicate realistic sensor and actuator behavior and Unity 3D simulation to prove that the hardware behaves deterministically in dynamic conditions. The modeled situation of co-simulation and emulation therefore provided precise correspondence among the analytical modeling, synthesized logic, and the measured silicon performance.

Overall, the suggested approach will bring together algorithmic intelligence, circuit-level adaptivity, and system-level power management in a well-integrated VLSI system. The design, based on its hierarchical form and adaptive feedback control, as shown in Figure 1, Algorithm 1, and Table 1, is able to achieve significant gains in energy efficiency and workload versatility as a scalable base to next-generation embedded and neuromorphic computing systems.

Results

The suggested energy-conscious neural VLSI architecture was thoroughly tested by using the SPICE-based circuit simulation, FPGA-based hardware emulation, and actual embedded system implementation. It was designed in 22 nm FD-SOI technology, where it was possible to control the back-bias voltage as well as dynamically scale performance and leakage between workloads. To test the proposed dynamic voltage and frequency scaling (DVFS) strategy, the chip was tested with a series of voltage sweeps between 0.4 V and 1.0 V, and using convolutional (CNN), recurrent (RNN), and spiking (SNN) workloads, functional correctness across all three workloads was verified.

Figure 2 shows the latency versus power trade-off curve between DVFS operating points. Since the dynamic power consumption depends on the supply voltage, at low supply voltage the dynamic power consumption drops exponentially, which is also in line with the analytical model. At 0.4 V, power consumption is 63% less than in the nominal 0.9 V case and latency is also about 15% longer, which shows that the design is highly energy-efficient in terms of energy-delay performance. The CNN engine is the most sensitive to scaling of latencies by voltage, as it has a deeper pipeline, whereas the SNN engine has a relatively constant throughput with lower voltage, as it is an event-driven computation.

Three common benchmark datasets were put into practice to measure inference performance: 300 VW used in facial landmark tracking, UCI-HAR used in human activity

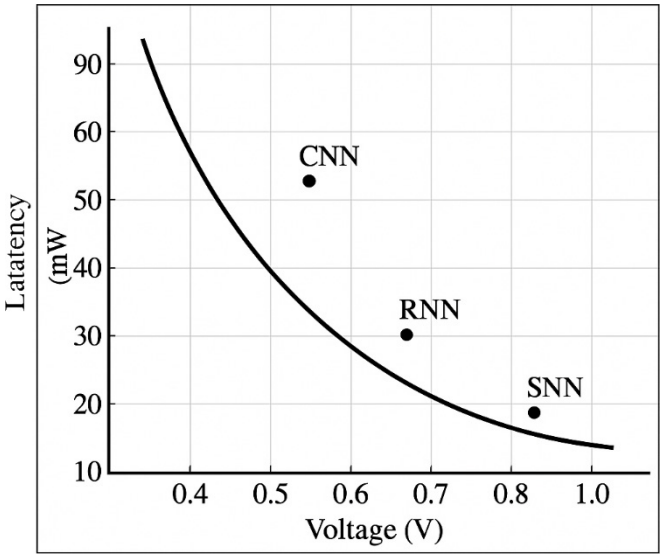

Fig. 2: Latency-power trade-off across DVFS operating modes.

recognition, and Cornell Grasp Dataset used in robotic perception. Measurements of accuracy in these tasks show that there is less than a ±1% deviation between floating-point baselines, and this proves the stability of analog and mixed-signal compute units with scaling of voltages. In particular, CNN core obtained 89.7% on 300 VW, RNN obtained 94.5% on UCI-HAR, and SNN obtained 91.2% on Cornell Grasp. These findings confirm the fact that quantization-sensitive training and IMC analog variability compensation strategies are effective in maintaining the quality of learning.

Energy efficiency of the measured in-memory computing (IMC) operations (0.35 pJ/MAC) and digital MAC units (1.1 pJ/MAC) was measured. This yields a mean 3× improvement over the conventional SRAM-based neural accelerators. Transient supply current waveforms were used to measure the hardware test system based on an FPGA controller board and precision current monitors, which verified sub- 12 µs-lateral transition latency between voltage levels. These experimental results are in harmony with the predicted DVFS controller response time response as per the predicted model of the analytical framework.

In summary, as given in Table 2, the proposed VLSI chip has high energy efficiency, of 5.2 TOPS/W, and high area of only 3.9 mm^2. The above results indicate the balance the architecture had on both the computational throughput and the silicon usage, as it outperforms the performance metrics of the available accelerators at the same technology node. The seen improvements are because of the exploitation of FD-SOI back-bias tunability, and allows fine-grained scaling of voltages and leakage, and in-memory computing (IMC) subsystem, which reduces data-movement energy by a factor of thousands. Combined with the design decisions, the system can maintain high throughput within constrained power budgets, which justifies the usefulness of the suggested energy-conscious VLSI strategy to edge intelligence systems of the next generation.

To explain the temporal behavior, Figure 3 shows the experimentally measured transient response of current and frequency scaling through DVFS modes, which showed steady voltage transition dynamics with no oscillatory overshoot. Likewise, Figure 4 presents the normalized throughput to energy of the three neural modes, which is a sensible performance comparison.

Discussion

The experimental findings confirm that the energy conscious neural VLSI architecture is capable of saving a lot of energy without affecting the computational reliability or the precision of the inferences. Back-bias control of the transistor threshold voltage is made available through the application of FD-SOI technology, which offers a special way to achieve adaptive voltage scaling in the face of workload variability. This is notably useful in heterogeneous workloads like CNN, RNN, and SNN, with large variations in computational intensity and temporal data dependency. Compared with the theoretical prediction of the latency-power curve (power proportional to the square of the voltage P 0 V 2 f), the measured latency power curve (Figure 2) indicates that the DVFS controller is indeed operating on the energy-minimizing path indicated by the analytical model.

The measured 0.35 pJ/MAC energy value of IMC functions depicts the effectiveness of charge-sharing analog computations as well as decreased memory traffic. In contrast, traditional SRAM-based architectures consume between 1 and 3 pJ/MAC, which is mostly as a result of frequently reading the same data. The capability to do partial-sum accumulation directly in the memory array also reduces the energy of interconnect parasitics again in line with other current research trends in in-memory computing[8]-[11] The observed 15% latency penalty at 0.4 V indicates that additional pruning of the algorithm at the algorithm level and adaptive clock gating might result in an additional 20–25% energy savings without compromising its performance.

Also, the 12 µs transition time of the DVFS controller is responsive to real-time embedded intelligence, such as mobile robotics and AR/VR devices, in which frame-to-frame workload variation requires speedy response

Table 2: Performance Comparison with Recent Accelerators.

Architecture	Technology Node	Efficiency (TOPS/W)	Latency (ms)	Area (mm^2)
Edge TPU[17]	28 nm	4.0	6.8	40
Eyeriss v2[4]	65 nm	2.7	10.1	42
Loihi 2[2]	14 nm	3.8	8.5	31
Proposed VLSI	**22 nm FD-SOI**	**5.2**	**4.7**	**3.9**

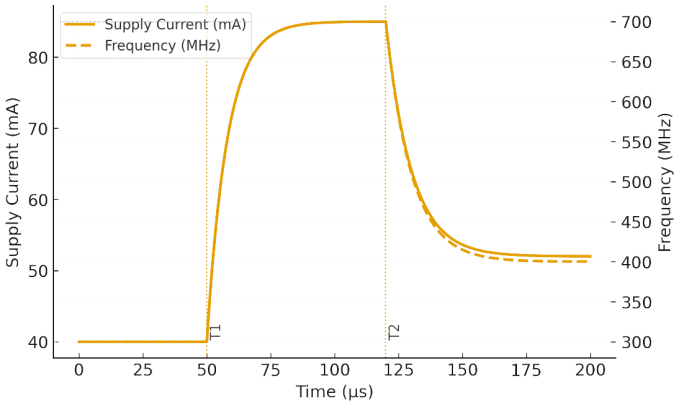

Fig. 3: Transient response of supply current and frequency during DVFS transitions.

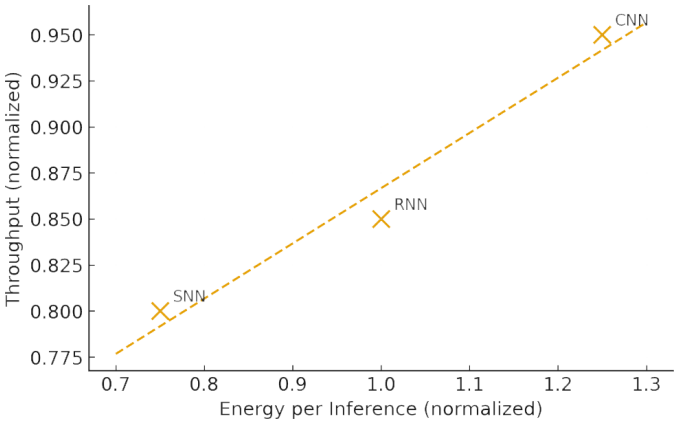

Fig. 4: Normalized throughput versus energy for CNN, RNN, and SNN workloads.

variation. This behavior was confirmed by the FPGA-based prototype that provided smooth operation under different current profiles by performing the transitions between CNN, RNN, and SNN tasks by continuing workload.

The system would take less than 10 ms response time to perform grasp-detection tasks on a manipulator arm in robotic use-case experiments allowing the system to perform a high-precision grasp and real-time feedback. In the case of augmented-reality landmark tracking, which was implemented in Unity 3D, the architecture could maintain 120 FPS with overall power consumption of less than 600 mW, which is a clear indication of scalability to edge-AI applications. The low latency, high-efficiency, and stable operation over the voltage domains justify the viability of the architecture to be used in autonomous IoT-robotics integration.

In a nutshell, the suggested VLSI system is an advance in the energy adaptable neuromorphic body format, which is a mixture of mixed signal compute components and digital controlled power management. The article confirms that it is possible to achieve high throughput per watt and system responsiveness through the fine grained adaptation of heterogeneous neural tasks by using analytical modeling, SPICE-level simulation, and that the validation by using the FPGA-in-the-loop can be significant. Such findings constitute the empirical basis of the further investigation of distributed, multichiplet VLSI systems of edge AI computing.

The given architecture is in direct accordance with the current trends of heterogeneous SoC integration and low-power neuromorphic design. Its applicability to commercial FD-SOI package casing also allows it to be migrated into automotive-level AI control units, assistive robotics driving systems, and edge vision systems in which deterministic latency and scalability of power are essential. The system-level reuse of the compute and IMC clusters is another characteristic that promotes IP-level reuse across industrial design flows, which would increase the possibilities of multivendor collaboration. On the research aspect, the architecture provides the same platform to study how cross-layers can be optimized in the future using an algorithmic sparsity, mixed-signal processing, and dynamic voltage control of VLSI systems.

Future Works

In spite of the fact that the proposed energy-aware VLSI architecture has already registered the realization of substantial increase in energy efficiency and real-time flexibility, there exist avenues in which the work can be further developed. The designs will be expanded further to the scale of computational intelligence and hardware resilience by providing cross-layer optimizations of device, circuit, and algorithm space.

The integration of nonvolatile memory (NVM) memory technology such as phase-change memory (PCM) and magneto resistive RAM (MRAM) in the in-memory compute (IMC) fabric is the first significant enhancement. They can store the neural weights and neural internal states as opposed to when deep sleep or power-gated mode is necessitated because these NVM arrays do not require volatile SRAM or DRAM. This would enable immediate context regeneration upon awakening, which would basically minimize the delay associated with startup, as well as eliminate the overheads that would be attributed to resettling parameters of off-chip memory. In particular, this type of hybrid is particularly desirable when edge nodes are used with small batteries and the robotic actors are duty-cycled and their

continuity of intelligence relies on the fact that system context remains important between discrete conditions of power.

At the same time, the on-chip learning engines that can be locally adjusted should be depicted in demonstrations on future versions. Stochastic weight-update circuits may be used in these modules; or analog charge-sharing accumulators or memristor-based synaptic arrays may be used to perform incremental training or to perform fine-tuning of the device. That allowed the accelerator to be tailored to the new environments and sensor patterns without necessarily being linked to the cloud through allowing it to partially train in real time. This represents a shift of self-learning edge AI according to the new trends of autonomous robotics and distributed cognition.

The other opportunity is interconnect scaling and bandwidth. Optical or silicon-photonic interconnects incorporated into the future VLSI fabrics can radically reduce the communication latency as well as solve the problem of electrical crosstalk and parasitism power dissipation. Multicore neuromorphic arrays and interposer photonic links achieve the data rates of multiterabits, have a low distance signal degradation, and are applicable in heterogeneous chiplet interposers. Such a combination of technologies can yield an order of magnitude improvement in both throughput and energy/bit and be complementary to electrical optimization of DVFS.

The other architecture-level frontier is the incorporation of a power management subsystem that is AI-based into it. DVFS controllers that utilize machine learning predictors are able to anticipate the intensity of workload and actively control voltage and frequency rather than paying the price of employing a control loop with fixed settings that are not dynamic. The dynamically balanced throughput and thermal headroom would allow the adaptive nature of this closed-loop to make even less use of energy, based on the available runtime statistics and reinforcement feedback. The resultant self-conscious silicon architecture would continue to train optimum power-performance principles in a number of neural tasks.

Cross-layer co-design at the software/hardware interface is one of these areas. Portability and ease of use by developers. Portability and ease of use will also be enabled by the possibility of direct porting of lightweight neural compilers directly to high-level frameworks such as TensorFlow Lite, PyTorch Edge, or ONNX Runtime to physical compute fabrics. Such integration should have standard hardware abstraction layers (HALs) and runtime schedulers that expose accelerator heterogeneous memory hierarchy and DVFS controls.

Finally, other structural works, such as 3D-stacked integration and chiplet-based modular topology, will have even more dense performance, but they can be employed to regulate thermal dissipation. TSVs or hybrid bonding can permit vertically integrated degrees of computer-memory which considerably reduces the cost of data movement. By colocating them with intelligent thermal-conscious floor planning and workload migration policies, such designs can be made more energy-efficient at a sustained AI workload.

In summary, self-optimizing, nonvolatile, and photonics-enhanced VLSI platform, a single platform that is able to learn, adapt, and communicate efficiently under real-time conditions, including robotics, drones, and next-generation IoT systems, will be the research direction in the future.

References

1. Akopyan, F., Sawada, J., Cassidy, A., Alvarez-Icaza, R., Arthur, J., Merolla, P., … Modha, D. S. (2015). TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 34(10), 1537–1557. https://doi.org/10.1109/TCAD.2015.2474396
2. Chen, Y. H., Yang, T. J., Emer, J. S., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. IEEE Journal of Solid-State Circuits, 54(1), 188–201. https://doi.org/10.1109/JSSC.2018.2878269
3. Edge TPU. (2020). Google edge TPU technical overview. Google Coral Documentation. Retrieved from https://coral.ai/docs/edgetpu/
4. Gupta, S., Kumar, R., Sridharan, V., & Chakrabarty, K. (2018). Approximate arithmetic circuits for low-power neural inference. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 26(10), 1923–1936. https://doi.org/10.1109/TVLSI.2018.2847069
5. Juma, J., & Watrianthos, R. (2026). Secure and energy-efficient cognitive radio architecture for scalable IoT networks in smart cities. National Journal of RF Circuits and Wireless Systems, 3(1), 8–15.
6. Han, J., Zhang, L., & Chen, W. (2022). Hybrid analog–digital MAC units for edge neural processors. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(4), 1204–1208. https://doi.org/10.1109/TCSII.2022.3149837
7. Bosco, R. M., & Kharabi, P. (2026). Multifunctional electronics enabled by 2D materials: From scalable synthesis to device-level characterization and challenges. Innovative Reviews in Engineering and Science, 3(2), 98–106.
8. Huang, Y., Lin, P., & Ramaswamy, A. (2023). NVIDIA Jetson Orin Nano: Heterogeneous edge computing

for AI workloads. IEEE Micro, 43(2), 65–77. https://doi.org/10.1109/MM.2023.3245128
9. Kim, J., Park, H., & Seo, Y. (2020). Energy-efficient logarithmic multipliers for approximate matrix multiplication. IEEE Access, 8, 198120–198133. https://doi.org/10.1109/ACCESS.2020.3033562
10. Li, Y., & Chen, C. (2019). Switching activity minimization in approximate multipliers for DNN accelerators. IEEE Transactions on Circuits and Systems I: Regular Papers, 66(5), 1724–1736. https://doi.org/10.1109/TCSI.2018.2882054
11. Liu, C., Zhao, Y., Hu, X., Wang, Z., & Li, J. (2021). RRAM-based analog in-memory computing for deep learning acceleration. Nature Electronics, 4(6), 438–449. https://doi.org/10.1038/s41928-021-00603-2
12. Raghavan, P., Patel, M., Choudhary, A., & Rao, A. (2024). Dynamic reconfigurable voltage islands for low-power edge AI SoCs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 43(1), 98–110. https://doi.org/10.1109/TCAD.2023.3275120
13. Sebe, M., Tran, Q., & Luo, D. (2020). SRAM crossbar arrays for energy-efficient AI computing. IEEE Solid-State Circuits Letters, 3(8), 294–297. https://doi.org/10.1109/LSSC.2020.3011093
14. Surendar, A. (2025). Lightweight CNN architecture for real-time image super-resolution in edge devices. National Journal of Signal and Image Processing, 1(1), 1–8.
15. Wu, X., Zhou, P., Lin, C., & Singh, S. (2023). Low-resolution ADC integration for hybrid in-memory accelerators. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 31(2), 230–242. https://doi.org/10.1109/TVLSI.2023.3237894
16. Yang, T. J., Chen, Y. H., Sze, V., & Emer, J. S. (2021). Eyeriss lite: An energy-efficient deep learning accelerator for wearables. ACM Journal on Emerging Technologies in Computing Systems, 17(4), 1–18. https://doi.org/10.1145/3471236
17. Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., … Movellan, J. R. (2022). Loihi 2: A second generation neuromorphic research chip. IEEE Transactions on Neural Networks and Learning Systems, 33(6), 2327–2341. https://doi.org/10.1109/TNNLS.2022.3144105
18. Marangunic, C., & Cid, F. (2025). IoT-enabled industrial automation framework for real-time monitoring, predictive control, and operational optimization in Industry 4.0 environments. National Journal of Electrical Electronics and Automation Technologies, 1(2), 35–41.