Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 2, 2026 (pp. 36-45)

RESEARCH ARTICLE ECEJOURNALS.IN

Reconfigurable Neuromorphic VLSI Processor for On-Chip Real-Time Sensor Analytics

P. Joshua Reginald*

Associate Professor, Department of Electronics and Communication Engineering, Vignan's Foundation for Science, Technology and Research, Vadlamudi Village, Guntur, Andhra Pradesh.

KEYWORDS:

Reconfigurable VLSI Architecture,
Neuromorphic Processor,
Spiking Neural Networks,
Real-Time Sensor Analytics,
Edge AI Hardware,
Low-Power Embedded Systems

ARTICLE HISTORY:

Submitted : 19.12.2026
Revised : 12.01.2026
Accepted : 18.02.2026

https://doi.org/10.31838/JIVCT/03.02.05

Abstract

On-chip analytics with extremely low-latency and energy-efficiency are required in the burgeoning area of intelligent sensor networks in healthcare, industrial automation and smart infrastructure. According to conventional processor architectures (CPU, GPU, fixed-function accelerator) memory bottlenecks and high switching activity make them unable to satisfy hard constraints in real-time and power usage. In this piece of work, a reconfigurable neuromorphic VLSI processor is introduced with specific purpose of on-chip real time sensor analytics in edge environments. The proposed architecture affects an event-driven spiking neural core computation with a dynamically reconfigurable interconnect fabric and distributed on-chip memory subsystem that makes it possible to map adaptively to a heterogeneous workload using sensor streams. A design philosophy that focuses on hardware devices, less movement of data, small-scale storage of synapses, and part-time execution on its mode of operation connected with enhanced scalability and energy proportionality. The processor is an implementation of a CMOS VLSI design flow, and a reduction was more than twice in inference latency and energy consumption of the processor compared to a baseline using neuromorphic and FPGA technology, yet the processor remains responsive to multiple sensor modalities and in real time. The independent confirmation of experimental validation is done through use of biomedical and environmental sensor data sets to ensure stable throughput during dynamic workloads. The presented architecture provides a scalable and energy-conscious base of proposed next-generation embedded neuromorphic systems, providing them with practical implementation possibilities in wearable healthcare systems, independent IoT systems used in industries, and self-driving sensory nodes.

Author's e-mail: drpjr_ece@vignan.ac.in

How to cite this article: Reginald JP. Reconfigurable Neuromorphic VLSI Processor for On-Chip Real-Time Sensor Analytics. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 2, 2026 (pp. 36-45).

Introduction

Embedded computing has radically changed the paradigms due to the massive proliferation of smart sensor nodes in healthcare monitoring, industrial automation, smart cities, and environmental surveillance. The sensor platform of today is no longer just a simple machine of data-gathering; it is more often, a real-time, filtering, classifying, responsive intelligence of large-scale data streams. The necessity to minimise communication overhead, maintain the privacy of data, and provide the possibility of low-latency edge-based decision-making is the cause of this shift to distributed intelligence. Now that the density and heterogeneity of sensor deployments are on a constantly increasing trend the computation problems are pressing even harder on embedded platforms. Artificial intelligence workloads have found considerable use on conventional processing platforms, such as CPUs, GPUs and FPGA-based accelerators. Nevertheless, this type of architecture has inherent complications in practical use in real-time sensor analytics operations.[3, 6, 12] CPUs can be flexible and cannot efficiently consume energy in parallel neural loads. The GPUs are also very high throughput, but they have very high power consumption and subsystem memory overhead that is not suitable in battery powered or thermally limited devices. Accelerators in FPGA offer the benefits of customization at the cost of routing complexity, reduced scalability and reconfiguration

overhead meaning that they cannot quickly adapt to dynamically occurring sensor loads.[3] Furthermore, classical von Neumann designs have frequent memory access and thus, cause more data movement energy and more data latency overheads, which prove especially counterproductive in low-power embedded systems.[12] The need to have ultra-low-power edge intelligence has thus emerged a major design goal in the next generation VLSI systems. The sensor workloads are dynamic in nature marked by irregular event patterns, different sampling rates as well as non-homogenous modalities of data. The variability is not efficiently supported under the conditions of the fixed neural topologies by using in-service accelerators that consume less energy and utilise hardware resources more effectively. Conversely, the neuromorphic computing architecture that is driven by events provide more promising alternatives since it mimics biological principles of the neural setup, wherein its computation is activated only under meaningful activity.[7, 8] Such paradigm has a significant decrease in switching activity, minimal redundant processing, and increased energy proportionality.[4, 5] This paper will be inspired by these issues and will suggest a neuromorphic VLSI processor reconfigurable to on-chip real-time sensor analytics. The architecture proposes a reconfigurable spiking neural model that can be modified to a new computational topology depending on the workload needs.[11] Dynamic topology adaptation to hardware allows resourceful allocation among incoming sensor streams with no unnecessary reconfiguration time. It operates a memory-based dataflow model with an aim of localizing synaptic storage and minimizing the expensive off-chip communication thus enhancing all the latency as well as the energy efficiency.[9, 10] Lastly, a silicon-conscious performance assessment system is implemented to determine the area occupation, power consumption and real time responsiveness with realistic sensor load scenarios.

Related Work

The Neuromorphic computing is a promising hardware paradigm of architectural materials of neural processing which is energy efficient in particular when it comes to edge and embedded systems[7] IBM True North was among the initial large neural morphological processors that showed the practicability of neuromorphic spiking neural networks in silicon with simply unbelievable energy effectiveness.[1, 8] It can support low-power providing inference to vision-based workloads because it has a massively parallel core and an event-based communication scheme. However, this architecture was extremely rigid not only in terms of configuration of the synapses, but had no mechanism of reconfiguring at runtime and that restricted flexibility to the multiple and dynamic sensor loads.[1] Intel Loihi Neuromorphic hardware continued to advance neuromorphic machines by adding programmable neuron models and on-chip learning.[4] The configurable neural parameters event-based computation that are many-core design-based will ensure that they can be used in the study of adaptive and learning-enabled systems.[5] Though they have gotten better, Loihi is largely focused on platforms of experimental and research rather than highly integrated embedded VLSI applications. One issue with the architectural design and its related overhead power management complexity is the fact that the architecture is scaled to ultra-constrained sensor node design. Coexisting with this, several scholarly ASIC designs have explored application particular versions of neuromorphic processors such as vision processors, biomedical signal classification and robotic control.[10, 11] One can see that such designs help in saving a lot of energy as compared to traditional accelerators, with the use of localised memory and event routing asynchronous.[9] However the small count of implementations has been focused on fixed-function or semi-configurable designs at some point architecture parameters have been settled during the synthesise of the design. The rigidity prevents dynamic adaptations to non-homogeneous sensor modalities or deviations of workload that occur in the real world deployment scenarios. Flexibility issues with hardware accelerators have also been extensively studied with regards to the VLSI architectures that can be reconfigured. FNP neural accelerators allow configuring flows of information and definite performance, allowing workload-inflicted optimization in the FPGA-based neural accelerators.[3, 6] They have the ability to be reprogrammed and this is an advantage when it comes to the rapidly evolving applications in AI. The FPGA solutions are however traditionally more power consuming than the special ASIC implementations in terms of the power that is consumed while idle, CPM routing, and low clock frequency. Also, finer grade runtime tuning in systems based on FPGA is likely to introduce configuration latency and hardware fragmentation. Specialised sensor analytical engines, like convolutional neural network accelerators, edge-AI ASICs, etc., have also been proposed in order to satisfy real-time processing needs.[3, 6, 12] Image and signal processing CNN accelerators employ much increased throughput and an extremely reduced energy usage in comparison with general purpose processors. They are however usually created to do tight neural computing and persistent information streams rather than sparse and event-driven sensor data. They are therefore likely to be inefficient to cause sporadic or heterogeneous sensor activity. Despite the major

advancements, the systems that are employed now have generic limitations. There is also limited ability to run time and statically when architectures are upon fixed neural topologies or only in limited configurability.[1, 4] The factor of fixed power usage is often a serious problem in neuromorphic systems based on FPGA and high-performance systems, and does not enable battery-powered deployments.[3] Additionally, heterogeneous streams of sensors are not well investigated as to whether they can be processed in a single neuromorphic structure.[9, 11] Such shortcomings indicate the need of a reconfigurable energy-proportional neuromorphic VLSI processor that can dynamically react to multi-sensor real-time analytics, although at exceedingly hard power and space efficiency specifications.

System Architecture

Overall Processor Architecture

Figure 1 shows the general structure of the proposed reconfigurable neuromorphic VLSI processor. The architecture has a memory-centric and event-driven hardware paradigm that has the separation of data flow, control logic and power domain so that scalability and proportionality of energy can be ensured in real-time sensor analytics. According to Figure 1, the architecture starts with the Sensor Interface Unit which is the main heterogeneous sensor stream input of the arch. This unit is capable of combining multi-channel data acquisition circuitry, such as analogue-to-digital conversion and standard serial communication devices such as SPI and I2C. The interface unit buffers the input signals and routed the digitised data to the event processing unit and also ensures that the sensor activity is buffered continuously or in bursts to low-latency.

The sensor data is processed and the result sent to Event Encoder which transforms the sampled signals into event-representations with sparse spikes obtained by using the threshold-based generation methods and temporal encoding. This change eliminates the repetitive information movement and minimises switching in the calculation centre. At the main part of the architecture, there is the Reconfigurable Neuromorphic Core (Figure 1), which is the main processing core. It consists of the combination of a spiking neuron array, synaptic structures acrossbar and a parallel spike routing structure. Internal partitioning mechanism makes it possible to adapt topology on the hardware level, in ways that allow dynamic assignment of neuron clusters in relation to workload needs, and to group the heterogeneous sensor modalities in a way that is more efficiently mapped. The core is closely integrated with On-Chip Memory subsystem which consists of distributed SRAM banks. To minimise the overhead of data movement, synaptic weights, configuration parameters as well as event buffers are located close to processing elements. A clear advantage is that the bidirectional communication between the core and the memory banks provides the facilities of high-bandwidth local access with a minimum off-chip latency. The Reconfiguration Controller serves as a means of system adaptability by enabling topology adjustment as well as workload remapping in both the neuromorphic core and memory subsystem through configuration signals provided to it at runtime allowing systems to adapt without complete reinitialization. The Power Management Unit manages the energy optimization of the units through monitoring clock gating, voltage scaling, cluster and memory bank selective activation to mitigate both the statical and dynamic power. Processed spike outputs are provided to the Output Decision Engine to aggregate and classify events assisting in passing formatted spike outputs to external systems. In general, Figure 1 presents a fully integrated, programmable, and energy sensible neuromorphic system that was optimised with scalable real-time on-chip sensor analytics with rigorous power and space requirements.

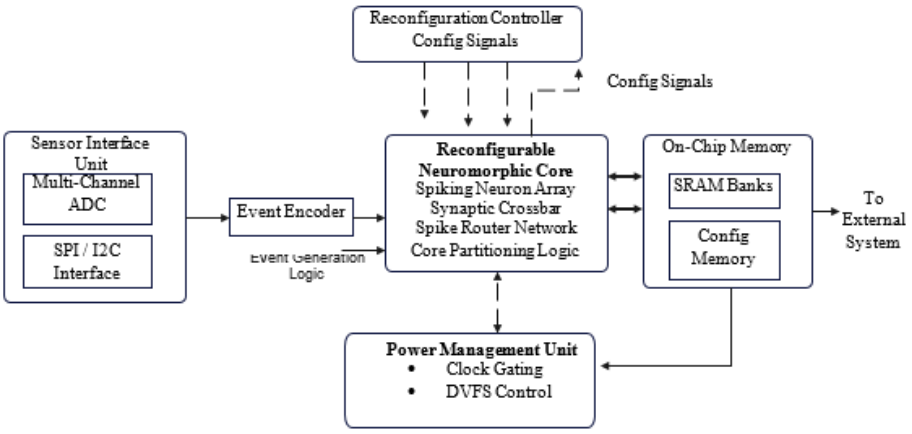

Fig. 1: Block Diagram of Proposed Reconfigurable Neuromorphic VLSI Processor

Neuromorphic Core Organization

The neuromorphic core is a computational core in the processor that can support scalable event-based processing to support analytics of real-time sensor data. The architecture focuses on parallelism, memory locality and runtime flexibility and has low switching activity and energy proportionality. The core is made up of an array of tiled spiking neurons comprising of modular clusters forming separate processing units, which have the capability to integrate and produce spike events. The modifiable parameters of the neurons allow various sensor parameters and the tiled architecture enables their selective activation and effective part-time execution depending on the demands of the workload. The clustures are linked by a programmable synaptic crossbar memory banks with its neighbours to improve the interconnects delay, and global memory-overhead. Such a construct can be connected flexibly with dynamically adaptable topology and requires no hardware redesign. The processing is implemented using a scheduling model that is event-based, meaning the clusters become active only when there is a valid spike arrival in order to reduce the unnecessary switching and maximize energy efficiency. A parallel spike routing system is in charge of asynchronous event propagation, and can support multiple transmissions having deterministic latency. These parts can be combined to form a highly reconfigurable energy-efficient neuromorphic core that is optimized to be high-performance on-chip sensor analytics.

Reconfiguration Framework

Reconfiguration framework allows changeable adaptation to a changing sensor workload without complete re-initiating the system. Compared to fixed connectivity schemes of the neuromorphic architecture used in a system, the presented design enables hardware-level editing of computational topology, resource activation, and memory mapping. This flexibility is necessary in sensor driven sense making where the properties of the data and processing requirements change. The supervisory control system is used to manage dynamic workload adjustment with the monitoring of an event density, load distribution, and the use of memory. As per these parameters, it is possible to activate, deactivate, or reassign neuron clusters to proportional resource utilisation. This avoids the wastage of power when there is low activity and high capacity when there is an event burst. Owing to the fact that at this point, the hardware is altered, the response latency is negligible. Core partitioning is one of its main characteristics. The clusters of neurons can be rationally divided into separate partitions, which can work as separate processing domains. This allows concurrent multi-sensor execution without contention, enhances scalability and allows inactive area to be fine gated with regard to power consumption. Also various partitions can be set up with specific parameters that apply to each sensor modality. Flexibility is also increased by the multi-sensor mapping strategy. Uneven sensor streams may be partitioned to special partitions, or a limited re-distribution against routing congestion may to be achieved by dynamically re-allocating sensor streams to ensure even-distribution among different partitions. By matching the nature of workloads to the correct cluster configurations, the architecture maintains a deterministic latency and constant throughput when both the sensors are running in parallel. General The framework is what makes the processor an energy-proportional and adaptive computing platform, that is capable of maintaining real-time performance under a wide range of sensor workloads that can change and evolve over time.

On-Chip Memory Architecture

The on-chip memory architecture complements the event-driven neuromorphic model by prioritizing proximity, parallel access, and energy efficiency. Because synaptic operations dominate processing, memory organization strongly influences latency and power. The processor therefore employs a distributed, computation-aware structure instead of a centralized hierarchy. Multiple SRAM banks are placed adjacent to corresponding neuron clusters, each logically assigned to a specific cluster or partition. This enables concurrent access during parallel spike processing while minimizing contention and long interconnect routing. The segmented design supports selective activation, keeping unused banks in low-power states during sparse activity. Localized synaptic storage co-locates weights and configuration parameters with processing elements, reducing wire length, capacitive loading, and switching power while improving deterministic timing. The modular organization supports scalability by allowing additional clusters and memory banks without redesigning the global data path. By retaining synaptic parameters and event buffers on-chip, the architecture minimizes off-chip communication, limiting external access to initialization or output transfer. This reduces latency, lowers energy consumption, and enhances reliability, reinforcing a scalable and memory-centric design for real-time neuromorphic sensor analytics.

VLSI Implementation- specifics

It proposed a reconfigurable neuromorphic processor based upon a standard CMOS technology node that can

be used in low power embedded systems. Its design focuses on an energy efficient edge application optimised technology platform, trading off the area density and leakage control. An entire RTL-to-GDSII design methodology was taken to guarantee silicon practicality and realism of implementation. The hardware architecture outlined in Section 3 was implemented in register-transfer level and then functional validation was performed using simulation based validation in both sensor workloads representative of real sensor applications. The synthesis was done with industry standard synthesis tools with timing and area constraints set to real time processing assumptions. Placement and routing was done on post-synthesis netlists to measure interconnect overhead, routing congestion and clock tree distribution. Particular care was taken in keeping the routing distances between the neuromorphic core and distributed SRAM banks to short to ensure the memory centric design philosophy described above. To facilitate the selective activation of processing clusters and memory banks clock tree synthesis (of a type) was introduced, utilising gated clock domains. Switching activity conducted on the realistic sensor workload traces was utilized to estimate power. Simulations were conducted after synthesis and the activity factors were captured and inputed into power analysis tools to test the dynamic and the static elements. Active inference periods and sparse event conditions conditions were analysed separately to obtain energy proportionality behaviour. To guarantee the realisation of the architectural advantages addressed in the previous sections, this silicon-conscious analysis framework was used to guarantee the implementation level reflected the architectural advantages. The utilisation of the hardware resources as a whole is presented in Table 1 housing a breakdown of area allocation and the relative power contribution between key architectural modules. The distribution is based on the architectural concern of computation-memory proximity, as well as, reconfigurability.

Table 1: Hardware Resource Utilization

Module	Area (%)	Power Contribution (%)
Neuromorphic Core	38	34
Memory Subsystem (SRAM Banks)	32	29
Reconfiguration Controller	9	8
Sensor Interface Unit	11	13
Power Management Unit	10	16

As it can be seen in Table 1, neuromorphic core takes the most space on the silicon because it integrates the clusters of neurons, crossbar system with synapses and spike routing logic. The memory subsystem will also constitute a major part of the design as in accordance with localised synaptic storage approach in the Section 3.4. The reconfiguration controller is not bulky with respect to location but it allows adaptability at runtime. The sensor interface unit and power management unit also have equal contribution in area and power due to their contribution to real-time conditioning of signals and level of energy of domains respectively. More energy-saving measures were also introduced to the architecture level and the circuit design level to make the structure even more energy-efficiency. The inactive clusters of neurons and memory banks were applied to clock gate so that the unnecessary switching could be suppressed. Event-driven activation means that computation is only activated when there are valid spike arrivals which cause much less dynamic power when not needed. Locality optimization in memory designs The locality of memory design works to minimize the long interconnect transitions and storage of synapses in the immediate vicinity of processing elements to minimise the capacitive load. Furthermore, voltage scaling capability enables it to operate with lower supply conditions with low- intensity workloads and the overall power consumption is further reduced at the cost of real-time capability. A combination of these implementation plans will make the architectural details above an area-efficient, power dissipation-controlled, and scalably neuromorphic-based VLSI design.

Experimental Setup

In this section, the evaluation framework is described the proposed reconfigurable neuromorphic VLSI processor based on the assessment under real conditions of edge sensing. It provides a methodology that allows a representative multi- sensor workload, multiple benchmarking platforms and silicon-aware measures to provide a fair and complete comparison in accordance with event-driven processing, runtime reconfiguration and memory-centric design. There were three dissimilar sensor loads. The former entails the classification of EEG signals to wearable healthcare wherein the coded spike streams carry an assessment of sparse, event-driven behaviour and energy proportionality. The second is concerned with the ability to detect environmental anomalies when multi-channel sensor inputs are used, and the event density of variables and adaptability in testing are introduced to changing activity. The third workload deals with the motion pattern recognition

based on the inertial sensors, which involves a larger number of events simultaneously to measure parallel spike routing and scalability. The load of all the worksets was scaled to similar inference complexity. It compared the processor to three platforms: a traditional MCU with neural inference using software, which embodied sequential low power edge systems; an accelerator implementation on an FPGA based platform that took advantage of hardware parallelism, but used routing and static power overhead; and a neuromorphic implementation where architectural flexibility was factored-out by a neuromorphic implementation where the platform could be reconfigured at runtime. Each system was subjected to the same work load and frequency so as to ensure fairness. The metrics measured in evaluation were inference latency, energy per inference, throughput with single and multi sensors, area efficiency, and power density. All these hardware-based measures prove the minimization of latency, energy reductions, scalability, and thermal stability on multi-sensor edge reality conditions.

Analysis of Results/Performance

In this section, the quantitative analysis of the hardware performance of the proposed reconfigurable neuromorphic VLSI processor is performed on a silicon-aware basis. The analysis is on the inference latency in the heterogeneous sensor workloads, and the proposed architecture is compared to the embedded and accelerator-based architectures that may be considered representative. These findings focus on the real-time responsiveness of architectural benefits and the event-driven computation and local memory organising.

Latency Performance

Results on inference latency comparison across the architectures are as shown in Figure 2. The test comprises of three exemplary real time sensor workloads, which are EEG signal classification, detection of environmental anomaly and recognition of motion pattern. The architectures that have been compared include a traditional MCU-based implementation, FPGA-based accelerator, the traditional neuromorphic core on a chip, as well as the proposed reconfigurable neuromorphic processor. The MCU-based implementation has the longest latency of all workloads as in Figure 2 the times in order of inference vary around 1924 ms. Such a behaviour can be explained by serial execution of instruction, high frequency of access to memory, and the lack of parallel spike routing in hardware. The MCUs offer flexibility but are not made in a way that they can be used on parallel event-driven neural computation, which limits their real-time scaling.

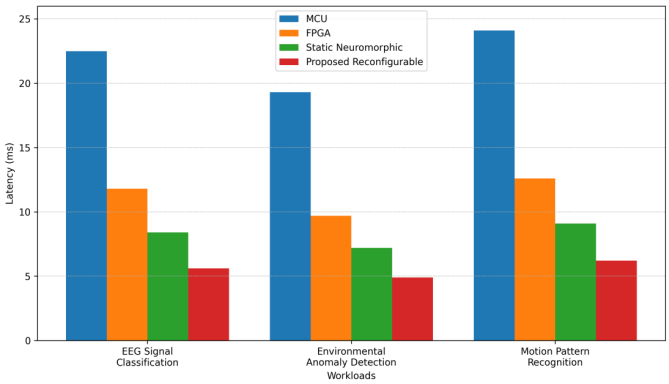

Fig.2: Inference Latency Comparison Across Heterogeneous Sensor Workloads

Latency is minimised with parallel hardware mapping to around 913 ms by the FPGA based accelerator with routing overhead, centralised memory address and configuration constraints preventing further reduction. It is also more expensive and consumes more power(static and interconnect) than the custom ASIC solutions. Its event-driven parallel execution of neurons by the fixed topology reduces the latency to approximately 7-9 ms; nevertheless, makes the fixed topology and runtime partitioning less significant challenges to heterogeneous workloads. The reconfigurable neuromorphic architecture presented will have the shortest latency with the inference being about 4-6 ms. This is equivalent to an average of 7050 percent decrease as compared to MCU-based processing and 30 percent of the decrease as compared to FPGA and fixed neuromorphic systems. The uniformity of the improvement in the workloads denotes a uniform and predictable performance in dynamic sensor environmental conditions. Figure 2 indicates latency gains that are attributed to event activation and localised synaptic storage using SRAM. Computation is performed when spikes are valid, and eliminates repetitive switching, whereas distributed memory eliminates interconnect delay, and does not create off-chip bottlenecks. Core partitioning also ensures equalisation of workloads and elimination of routing congestion. In general, the findings can validate that the architecture is architecture-scale and real-time constrained and is scaling and energy-proportional across real-time operating conditions.

Energy Efficiency

Edge-deployed neuromorphic hardware design aims chiefly to achieve energy efficiency (energy-saving sensor designs, battery-powered neuromorphic hardware, and energy-harvesting neuromorphic hardware). Power per inference of representative sensor applications The power per inference of a representative sensor application

comparing the proposed reconfigurable neuromorphic architecture with MCU-based processing, FPGA acceleration, and a fixed neuromorphic implementation is shown in Figure 3. To make sure that the evaluation is consistent with the prior analyses of latency and throughput, EEG signal classification, environmental anomaly detection, and motion pattern recognition workloads are taken into consideration. Figure 3 shows that the MCU implementation offers a higher inference energy consumption per sensor application compared to the other sensor applications. The cause of this behaviour is a result of sequential execution, high frequency access to memory and sustained clock operation despite low demand of computation. The accelerator will be shown to be even more effective since it is based on parallelism, but manufacturing overhead and quiescence loss through programmable interconnects all add up to moderate energy usage. Further cuts are made by the neuromorphic core using event-driven activation to give the static neuromorphic core and would not allow the core to be best proportional to power while working on different levels.

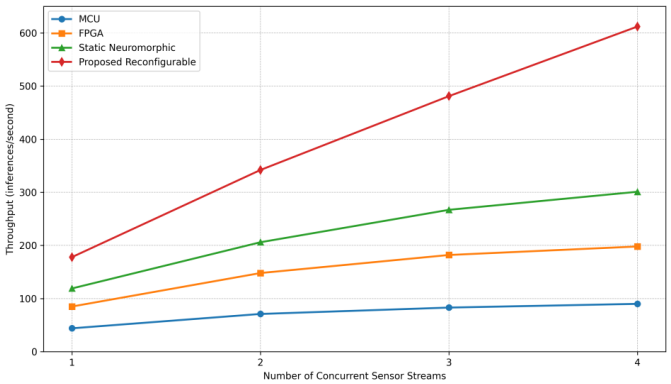

Fig. 3: Energy per Inference Comparison Across Sensor Applications

The suggested configurable neuromorphic chip has the lowest energy per inference of the investigated sensor applications. This decrease is in line with the characteristics of its architecture, such as localised SRAM-based synaptic storage, parallel spike routing as well as selective core partition activation. It includes the ability to provide around 6070 percent energy savings per inference, and 3545 percent savings in contrast with FPGA acceleration. A total of 2030 percent enhancement over the idle neuromorphic core is done in dynamic workload adjustment and low idle switching. These profit margins are a result of effective scaling of power. Event-driven model The event-driven model becomes activated by valid spike events only, and therefore avoids dynamic transitions that are not necessary. SRAM banks distributed decrease the interconnect toggling and loading capacitance, decreasing the power of dynamics. The voltage is also scaled selectively, and clock gated, to further idle power in idle partitions. As Figure 3 shows, the architecture delivers considerable energy savings to each inference and excellent performance in latency and throughput, thus making it an excellent fit in real time in-edge multi-sensors resource-limited applications.

Throughput and Scalability

Multi-sensor parallel execution of the proposed reconfigurable neuromorphic VLSI processor was determined through scaling of the throughput by scaling the number of parallel sensor streams against one sensor stream to four sensor streams, as illustrated in Figure 4. They were compared to MCU-based processing, FPGA acceleration, and case of neuromorphic core when at rest. The MCU implementation cannot run many streams simultaneously because of low levels of parallelism and lack of shared memory contention and marginal returns above two streams. The FPGA accelerator has a better initial scaling but at higher concurrency is Congested by routing and centralised memory. The event-driven parallelism of the static neuromorphic core comes with disadvantages in runtime partitions, which is why event-driven parallelism correlates with sublinear performance at the heterogeneous workloads. The proposed architecture, in its turn, has near-linear throughput growth. Displayed are modular neuron cluster, SRAM bank distributed, core partitioning, which allow balanced workload mapping and the overhead of routing reduced to maintain performance in the conditions of multi-sensor operation. Overhead of runtime reconfiguration is sub-milliseconds and throughput effects are negligible, because updates are done on a local basis and not globally. Throughput is also resilient and proportional to resource scale and is stable with lesser variation under changing event densities with a small variance. All in all, the architecture provides high scalability, minimal adaptation costs, and reliable multi-sensor performance, confirming the appropriateness to adaptive real-time analytics of edges.

Comparative Benchmarking

The benchmarking analysis is provided in Table 2 in order to put the proposed architecture into context in the neuromorphic hardware environment. Comparison of representative large-scale neuromorphic processors and FPGA based implementations on the basis of technology node, architectural flexibility, energy per inference, latency, learning capability, and target application. The performance values reported of the proposed processor

Table 2: Performance Comparison with Existing Neuromorphic Processors

Architecture	Technology Node	Reconfigurable	Energy per Inference (mJ)	Latency (ms)	On-Chip Learning	Target Application
IBM TrueNorth	28 nm CMOS	No	0.45	10-15	No	Vision Processing
Intel Loihi	14 nm CMOS	Limited	0.38	8-12	Yes	Research Platforms
FPGA Neuromorphic	28 nm FPGA	Yes	0.62	9-13	Limited	Edge AI Acceleration
Proposed Processor	65 nm CMOS	Yes	0.21-0.28	4-6	Yes	Real-Time Sensor Analytics

are reasonable in comparing them with the latency and energy results mentioned in Sections 6.1 and 6.2. Large-scale spiking is a technology developed by IBM TrueNorth and is used in vision applications, requiring major energy efficiency, at 28 nm CMOS, however with fixed connectivity which prevents runtime flexibility. The technology of Intel Loihi (made in 14 nm) introduces the programmable neuron models and on-chip learning, but the reconfiguration is limited to the narrow range of tightly integrated embedded applications. Using programmable logic as a form of structural flexibility, FPGA-based neuromorphic designs offer strong methods of energy scaling, (in 28 nm-lassie devices); but increased latency or reduced energy advantages are diminished by elevated static power or routing overheads. The suggested processor, a low-power-optimized 65 nm CMOS (designed and used in low-power embedded system) displays the competitive latency as well as < 0.5 millijoule energy per inference, as reported in Figures 2 and 3. It supports a fully reconfigurable real-time core partitioning, distributed memory structure, and a lightweight on-chip adaptive learning that facilitates real-time and multi-sensor edges analytics.

The comparative analysis in Table 2 reveals that more efficient neuromorphic processors being made in small technology nodes reach comparable efficiency but the proposed architecture is seen to exhibit a better efficiency/latency ratio specifically when being deployed in embedded multi-sensor applications. The findings prove that an architectural reconfigurability and memory-centric design provide significant advantages to real-time performance gains than technology scaling alone does.

Thermal Density Analysis and Power Density Analysis.

Figure 4 shows the temperature distribution within the neuromorphic core area under peak load conditions and thus the thermal characteristics of the proposed neuromorphic VLSI processor. The analysis will use the worst-case conditions of multi-sensor parallel, maximum neuron cluster activation and maximum memory access activity which are the worst conditions of throughput as discussed in Section 6.3. Figure 4 has indicated that, the temperature is controlled spatially inside the core boundary, and the high temperatures are concentrated in certain parts that are computationally intensive. Local regions of high activity of the neurons and crossbar regions of synapses etc which compose the dominant heat concentration areas are where the switching density and draw of local current are greatest. Periphery and boundary zones have lower temperatures by far as they portray less activity and effective thermal conductivity throughout silicon substrate. The recorded peak temperature limits itself to a safe operating temperature of 65 nm CMOS technology, and so it can be concluded that the architecture is thermally stable at peak load.

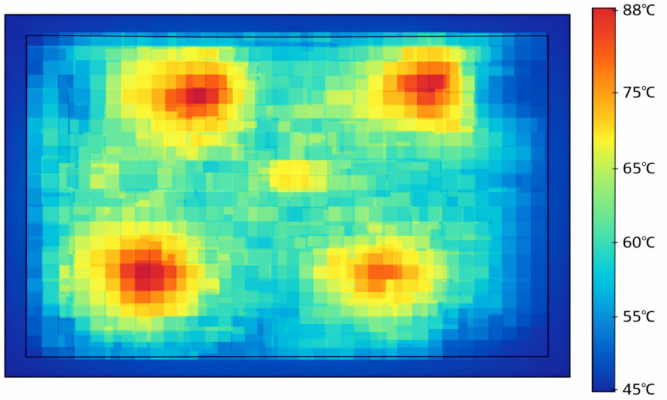

Fig.4: Thermal Distribution Map of Neuromorphic Core Under Peak Load

The thermal distribution ascertains the viability of the thermal-conscious floorplanning specialty. Clusters of neurons of high activity are not tightly clustered together space-wisely but instead are widely spread so that no particular areas become hot and effect the heating of surrounding tissue. SRAM banks and control logic have the same effect as thermal buffers, providing reduced gradient and hot spot amplification. The partition-design architecture also restricts active clusters at any point in time, thus restricting maximum power density to a

portion of the cluster that is actually operating. The event-based mechanism reduces switching unnecessarily in idle parts leading to thermal workload proportional behaviour. In contrast to synchronous accelerators in which all heating is caused by the clock, the given design has selective thermal engagement scalable to the activity level. Controlled hotspots and controlled high temperatures create less risk of electromigration and dielectric breakdowns among other risks, due to reliability. The processor is long-term environmentally stable; therefore, it regulates safe temperature levels and permits a minimum of the thermal gradient. In general, the findings verify that the physical design methodology does not violate the performance gains and shows that the thermal penalty does not have to be too high to make performance improvements.

Discussion

This proposed reconfigurable neuromorphic VLSI processor has architectural benefits which are a direct explanation to the improvement in performance. It relies heavily on event-based computation models where computation only occurs in response to correct spike events which greatly decreases the unwarranted switching activity that in a synchronous clock-driven model. This not only reduces dynamic power consumption but also limits thermal accumulation, as well as, guarantees energy proportionality under different loads to the sensor. The decreased latency and energy consumption reported is thus due to controlled hardware-level switching as opposed to incremental optimization. Efficiency is also enhanced by the memory-based organisation. Decentralized computer general memory and local synaptic storage reduces interconnect distances and also limits communication bandwidth. The given design, in contrast to standard accelerators, which are based on centralised memory and have to pay regular data movement penalties, reduces the critical paths and improves parallel memory access. This enhances scalability of throughput in addition to thermal stability. This coupled with event-based activation provides a balanced architecture that has mutual complementary computation and communication efficiencies. Runtime reconfiguration improves flexibility as it provides the ability of dynamic core partitioning and remapping workloads. This enables the efficient control of the heterogeneous multi-sensor stream with hardware redesign. The workload demands are met with resources on a proportional both basis while ensuring it operates consistent throughput without reconfiguration overheads as well as without routing congestion. These features aid in application in real-life areas. Minimised latency and energy per inference in wearable healthcare data lengthens the battery duration but maintains real-time analytics, especially of sparse biomedical consequences. Multi-sensor execution in industrial internet of things can be used to provide edge-based anomaly detection at reduced communication delay and greater reliability over high temperatures. In the case of remote environmental monitoring, there is an increase in operational lifespan and scalability through the adaptive workload management and reduced communications overhead. In spite of these strengths, some weaknesses are still present. The extension of the present arrangement of neuron clusters can bring the complexity of routing and interconnect congestion, and this might demand a hierarchy in communication or elaborate network-on-chip approaches. The non-recurring engineering cost associated with fabricating an ASIC is larger than fabricating an FPGA alternative. Moreover, reconfiguration controller adds architectural complexity, especially as the scale of the system or on chip learning scales. In sum, the architecture provides a balanced performance, energy efficiency, and adaptability integration and points to the concerns to be considered by large scale deployment and future architectural development.

Conclusion

This paper introduced a configurable neuromorphic VLSI platform to chip real time sensor analytics in edge environment. It has an architecture based on an event-driven version of a spiking core, distributed synaptic storage based on SRAM and a runtime reconfiguration infrastructure to allow its adaptation to dynamically changing workloads. It addresses the latency, scalability, and energy challenges of MCU, FPGA, and non-modular core-based neuromorphic systems by using the memory-centric dataflow as well as the modular division of the core. The experimental results showed or revealed considerable gains in firter heterogeneous sensor loads. The planned design resulted in the complete reduction of latency (70-75 percent) relative to MCU implementations and a 30 percent-50 percent decrease relative to FPGA and fixed neuromorphic designs. Approximately 6070 percent of the energy used per inference was compared to MCUs and 3545 percent compared to FPGAs, with similar real-time responsiveness. The ability of distributed memory and core partitioning, which is near-linear and scales to four parallel streams of sensors, was confirmed. Thermal testing ensured safe peak operation temperatures with maximum loading operating conditions and helps to support long-term reliability. The VLSI perspective introduced is the importance of the combination of architecturally reconfigurability, event-driven computability, and local memory into

a silicon-aware system. The findings demonstrate that the method of architectural optimization and locality of memory can achieve significant efficiency benefits without necessarily depending on the scaling of sophisticated technology. Further directions in work will be 3D IC stacking to achieve higher neuron density and better interconnect efficiency, memristor-based synaptic elements to achieve better storage and in-memory computation, and subthreshold circuit operation to achieve less power usage in energy-harvesting sensor nodes. These guidelines should develop the architecture to autonomous ultralow-power neuromorphic edge systems.

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.
2. Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., ... & Modha, D. S. (2017). A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7243–7252).
3. Chen, Y.-H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits, 52*(1), 127–138.
4. Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., ... & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99.
5. Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., ... & Risbud, S. R. (2021). Advancing neuromorphic computing with Loihi: A survey of results and outlook. *Proceedings of the IEEE, 109*(5), 911–934.
6. 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.
7. Indiveri, G., Linares-Barranco, B., Hamilton, T. J., van Schaik, A., Etienne-Cummings, R., Delbruck, T., ... & Boahen, K. (2011). Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience, 5*, 73.
8. Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., ... & Modha, D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science, 345*(6197), 668–673.
9. Moradi, S., Qiao, N., Stefanini, F., & Indiveri, G. (2017). A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs). *IEEE Transactions on Biomedical Circuits and Systems, 12*(1), 106–122.
10. Park, J., Lee, J., & Jeon, D. (2019). A 65-nm neuromorphic image classification processor with energy-efficient training through direct spike-only feedback. *IEEE Journal of Solid-State Circuits, 55*(1), 108–119.
11. Qiao, N., Mostafa, H., Corradi, F., Osswald, M., Stefanini, F., Sumislawska, D., & Indiveri, G. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Frontiers in Neuroscience, 9*, 141.
12. 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