

In-Memory Computing Based Hybrid RBM-DBM VLSI Architecture for Edge Devices

Ms.V.Aadhana Neya
Dept of Electronics and Communication
Engineering
KIT-Kalaignarkaranidhi Institute of
Technology
Coimbatore, India

M.Sasi kumar
Dept of Electronic Engineering
(VLSI Design and technology)
KIT-Kalaignarkaranidhi Institute of
Technology
Coimbatore, India

K.Dhivaskar
Dept of Electronics Engineering
(VLSI Design and Technology)
KIT-Kalaignarkaranidhi Institute of
Technology
Coimbatore,India

G.Pradeep
Dept of Electronics Engineering
(VLSI Design and Technology)
KIT-Kalaignarkaranidhi Institute of
Technology
Coimbatore,India

Abstract- Edge devices such as wearables, smartphones, and Internet of Things sensors need fast and efficient processing methods in order to process real-time data. In traditional computing systems, data needs to be constantly transported between memory and the processing units, leading to problems such as high energy consumption and increased latency. In-Memory Computing (IMC) has been suggested as a viable answer to these issues. The aim of this project is to implement a hybrid architecture, that incorporates the concepts of VLSI Design, to integrate two popular models: Deep Belief Networks (DBN) and Restricted Boltzmann Machines (RBM). By incorporating RBM-DBN model into memory components, the system enables the efficient extraction of features and advanced learning capabilities. This design therefore has a number of desirable characteristics relating to processing speed, reduced power consumption and suitability for modern edge computing applications.

Keyword- Restricted Boltzmann Machine, Deep Boltzmann Machine, VLSI Architecture, Edge Devices, Neuromorphic Computing, Crossbar Arrays, Energy-Based Models, Hardware Accelerator, Low-power Design.

I. Introduction

Given the rise of edge devices, e.g., wearables, smartphones and IoTs, there is a huge demand to perform the computation quickly and with low power consumption. When training, implementing and inferring deep neural networks which could be highly computational-intensive on these low-power edge devices proves extremely challenging and costly in terms of resources and latency of data shifting in memory. Conventional computing based architecture, especially the von Neumann architecture requires data being constantly shuttled between processor and memory, causing

a bottleneck and waste of power especially when processing large amounts of data with few compute units.

In-memory computing (IMC), on the other hand, is a promising technique that eliminates the data shuttling between memory and processing unit by executing computation within the memory, hence reduces the data movement between components, and accelerates execution with low power. This technique has great applicability for edge AI devices where real-time processing is crucial. It presents an opportunity to design smarter hardware through combining computation and data storage together, reducing the latency while improving power efficiency. [1]

In general, Restricted Boltzmann Machines (RBMs) and deep belief networks (DBNs) are two types of neural network models suitable for the learning of representations or feature extraction. RBMs are one layer learning models whereas DBNs are deep neural networks constructed by stacking multiple layers of RBMs. The hybridization of these two deep learning models can take advantages from both by the accurate and efficient learning capabilities. [2]

Therefore this project mainly proposes a hybrid RBM-DBN system based on IMC technology to further enhance both performance and energy consumption with VLSI implementation. By proposing a hybrid RBM-DBN structure in IMC architecture, we avoid data shuffles for computation in memory and thus achieve high performance in term of latency with low power consumption for edge AI devices application.[3]

II. Related Work

Recent works in the areas of edge computing, VLSI system design, and neuromorphic computing mainly concerned with improving processing efficiency of

computer systems by reducing energy consumption and latency as opposed to merely processing speed. Standard computing paradigms utilize a system of transmitting and receiving information (data) between a memory unit and a processor, which leads to consuming large amounts of energy and processing time. The In-Memory Computing (IMC) has evolved to avoid this difficulty. [4]

It aims to carry out data operations within, or near, its data location, which includes memory, I/O peripheral devices, etc. Several implementations based on IMC using SRAM, RRAM, and memristor based technologies demonstrate potential benefits on reducing data communication and accelerating the execution of machine learning computations and, hence, the proposed implementations are suitable for low-resource edge devices. The Restricted Boltzmann Machine (RBM) has gained considerable interest due to its unsupervised learning capability and the capacity of feature extraction. Various efforts have been made to implement RBM in hardware to accelerate the matrix operations and reduce high computation costs in the existing software approaches. Deep Belief Networks (DBNs) composed of multiple stacked RBM units has proven effective in learning hierarchical feature representations for applications like object classification, face detection, and speech recognition.

Furthermore, we focus on crossbar-based and neuromorphic computing architectures which leverage specialized hardware to speed up the operations performed on the artificial neural networks. Crossbar arrays can serve as weight storage and perform matrix-vector multiplication through massive parallelism, minimizing data transportation across system components and improving the performance of a range of computer vision and other machine learning tasks. Similar to other approaches, VLSI-based neural accelerators typically resort to parallel computations and appropriate memory architectures to reduce power requirements.[5]

Even though the presented approaches have many merits for designing a low-power and energy efficient system on a VLSI system and an electronic circuit design, however, the complexity hardware, device variation, low precision, limited capacity, memory usage, and scalability are still issues that needs to be addressed. In light of this, this work proposes to combine IMC approach with the design of a hybrid RBM-DBN architecture. The work aims to use memory-resident computing to take the maximum benefits from IMC, in order to improve performance by the ability to carry out computations directly at the storage location of the data and, henceforth reduce the overhead in transporting the data from memory to processing units.

Challenges and hurdles of designing a hybrid RBM-DBN model with a VLSI-orientated IMC structure, accuracy requirement, large memory usage, scalability requirement and hardware complexity requirement need to be considered. Based on the above issues. In summary, this thesis proposes a hybrid RBM-DBN architecture by leveraging the concepts of

In-Memory Computing. This design brings computations to the data locations within memory elements, aiming to minimize data communication latency and reduce power consumption.

The presented architecture aims to enhance execution speed and conserve energy by leveraging the feature learning capabilities of RBM/DBN models combined with the massive parallelism and specialized VLSI hardware aspects of the IMC design targeted for the real-time edgeAI application.

III. Proposed Solution

The design that is proposed herein involves the integration of IMC and a Hybrid system using both RBM and DBN into a single model for edge devices that is optimized for hardware resources. The integration of both processing and storage into one system eliminating repeated transfer of data from one memory to another, that causes serious performance Bottleneck in the conventional designs. It significantly reduces access to memory and thus improving speed and reduces the energy consumption.

Traditional systems requires repeated transferring of data between two distinct units but it can be avoided using an In-Memory computing architecture, so instead of moving data to a unit for computation it moves the operations to data in memory. This reduces enormous amount of data transfer and thus reduces latency and energy consumption for the edge applications where the latency and energy are significant parameters.[6]

The In-Memory computing architecture presented integrates the RBM's within a single memory architecture to perform initial layers of feature extraction. Since RBM can learn hierarchically in layers, these layers are parallel and can perform fast on the input, also the RBMs will feed forward and supply inputs to multiple upper layers that would collectively form a DBN. This hybrid design has a highly efficient hardware implementation that could fit for limited on-edge computations. The whole system is characterized using VLSI and is built using efficient memory usage and reduced hardware complexity that benefits a significant improvement in computing performance.

This hybrid design is implemented using modern memories, including memories like SRAM or novel non-volatile memories. Parallel processing is another mechanism embedded within system to process at faster speeds without significant modification. Optimized dataflow of system would provide better bandwidth.

Also to address low-power edge computing systems efficiently for real time operation system implements data parallel processing and efficient data flows.

Therefore, this work provides efficient and scalable platform to build intelligent edge systems and contribute to development of low-power AI hardware.

IV. Methodology

The system uses a structural approach based on hardware optimization for facilitating deep learning on edge devices. The proposed methodology involves the steps:

Gathering and Preparing Data

Edge sensors (either environmental or biomedical), capture some of the environmental data. Conditioning is performed using filtering/amplifying circuitry and the analog signal is converted to digital by means of an ADC (Analog-to-Digital Converter) that read the processed signal. The data is subsequently processed to ready it for neural processing; steps such as noise filtering and normalization are part of this initial preparation.

Data management and input buffering

The processed data is stored in on-chip memory buffers for further processing. This process ensures efficient and seamless data transfer and synchronization between sensing and computation units, allowing uninterrupted real-time operations and minimizing the delay.

Implementation of In-Memory Computing (IMC)

These computations occur directly in the array of memory cells, which are either SRAM or RRAM crossbars, instead of moving data from memory into the processor. This allows parallel Multiply-Accumulate (MAC) computation by mapping the memorized weights directly to memory elements. In addition, both the performance and energy efficiency of computation have been considerably improved.[7]

Feature Learning Based on RBM

Initial learning process of RBMs. They are employed to extract statistical properties from data via the interaction between visible and hidden units. In an RBM, efficient designs, where the impact of probabilistic nature of computing is reduced and implementation overhead is simplified by hardware friendly approximations, are achieved.

Processing Deep Belief Networks (DBN)

A Deep Belief Network is simply a Stack of RBM layers. This helps in hierarchically learning features where each layer extracts features from the output of previous layer. This hierarchical learning generally improves accuracy for pattern recognition applications such as anomaly detection or classification.[8]

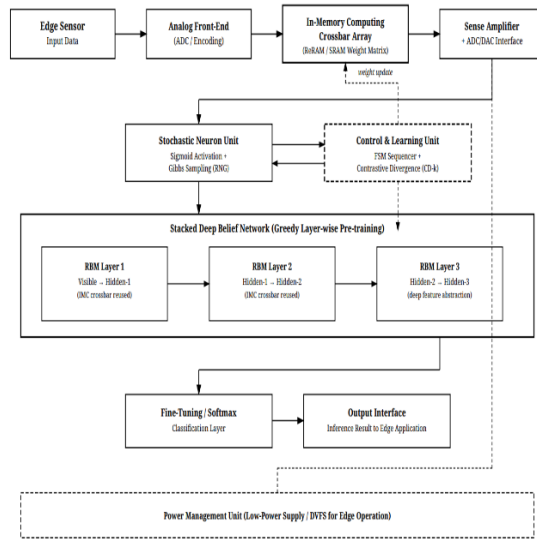


Fig. 1. Block diagram of the proposed In-Memory Computing based Hybrid RBM-DBN VLSI architecture.

Fig. 1. Circuit Diagram of Smart Plant Monitoring System. The components used in the system and their function are listed in the Table I.

TABLE I
Components used in the system and functions

S. No	Component	Function
1.	IMC Core (SRAM/RRAM Crossbar)	☐ Stores weights directly in memory. ☐ Performs MultiplyAccumulate(MAC)inside memory-reduces data movement.
2.	RBM Blocks	☐ Implement visiblehidden layer connection. ☐ Uses probabilistic computation (can be simplified in hadeware)
3.	DBN Block	☐ Stack of RBMs (deep layers) ☐ Performs hierarchical feature extraction
4.	Activation Unit	☐ Lookup tables(LUT) ☐ Piecewise linear circuits
5.	Control unit(implicit)	☐ Clock control ☐ Data flow scheduling

Realization of the Activation Function

To implement nonlinear functions like Sigmoid or ReLU function, lookup tables (LUTs) or piecewise linear approximation circuits are utilized. These options are considered because they reduce the hardware cost and yet maintain an acceptable accuracy level.[9]

Making Decisions and Producing Results

The processed data reached decision unit by using features mined from them performing classification or prediction, depending on the scenario output would be activity recognition, health monitoring or anomalies.

Communication and Output Interface

To transmit the outcomes to outside devices and equipment like display modules or external output ports, low-power communication protocols such as Bluetooth low energy are utilised. The communication ensures real-time data transmission as a result in edge environments.

Strategy for Power Optimization

The low-power system architecture is tailored to the needs of edge devices. This include techniques such as reducing memory footprint, data transfer overhead, parallel processing capability etc. IMC itself is responsible of reducing total energy consumption for that reason as a whole for its architecture based on the IMC concept.

V. Result and Discussion

When compared to conventional architectures of computing hardware the IMC hybrid architecture presented displays enhanced performance attributes. A significant benefit achieved is a reduction in the amount of data transferred from main memory to processor, and accordingly reduced latency and processing time. The architecture also delivers considerable power savings compared to conventional systems from a power usage standpoint.

As the processing is performed within the memory, the frequency of data migration is reduced, thereby making more effective utilisation of the hardware resources.

This also has the advantage of being an ideal fit for low-power edge applications. The utilisation of parallelism in the architectures leads to a throughput increase and a boost in overall performance. The adoption of multiply-accumulate operations within memory allows for simultaneous computations to be carried out and thus enhancing the deep learning execution speed and device response. The model performance evaluation shows an increase in model performance while maintaining an increased level of reliability.

The components used in the system and their function are listed in the Flow chart I.

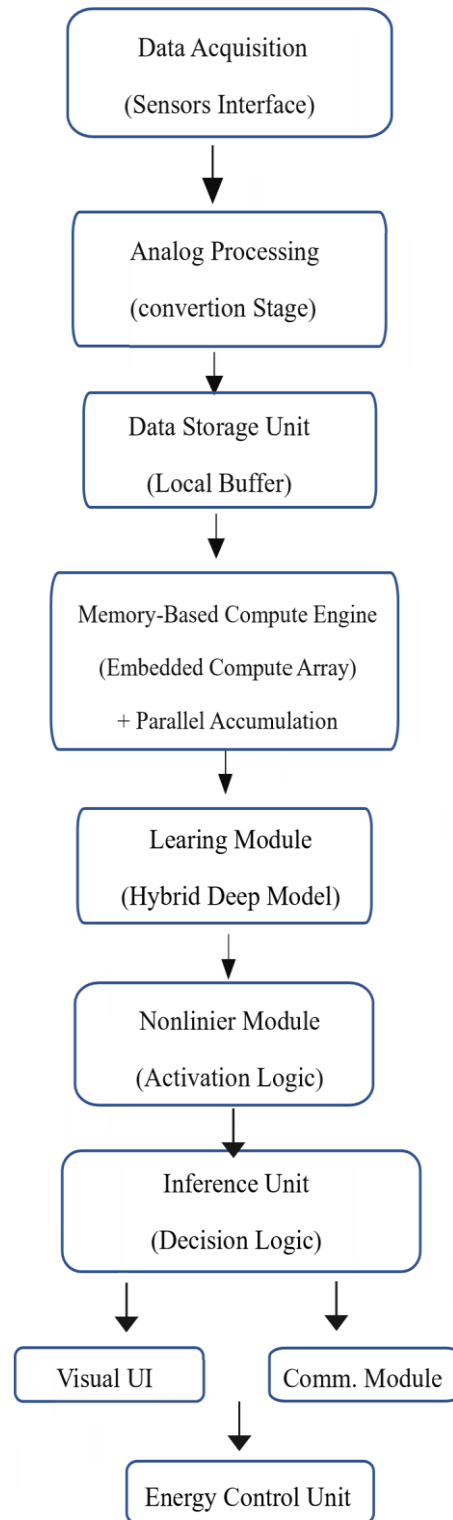


Fig. 2. Flowchart of Smart Plant Monitoring System

In addition the IMC hybrid architecture does not adversely affect the accuracy of the trained models.

Although there exist some advantages as described above, it also poses a variety of practical issues to be solved. System performances are susceptible to factors such as varying memory device properties, noises produced by analog operations, large-scale circuitry designing complexity, etc. Future works should put into design robust, better tune the design in circuit-level optimization and take a lot of error-correcting methods to improve system reliability.[10]

VI.Conclusion

In-Memory computing based Hybrid RBM-DBN VLSI Architecture for the Edge Devices' provide a good solution to improve the drawbacks in existing architectures, especially in the area of edge devices. By carrying computations in or within the memory, the architecture significantly alleviate the data-transfer burden, thus leading to enhance performance and efficiency.

By leveraging the RBM and DBN model in the proposed architecture, feature-learning is done efficiently with retained precision. Lastly, the use of VLSI design techniques contributes to reduce resource-requirement and the architecture obtains balanced performance, low power and high scalability.

VII.Summary

This project proposes an architecture integrated with hardware-in memory computations and hybrid RBM-DBN learning. Our architectural design works to resolve the problems in the conventional architectures by reducing the access latencies and power consumptions.

Due to distributed computations across the memories and the support of parallel computation, our architecture gains great benefits in throughput and fast deep learning operations. The results verify its accuracy and performance in the resource constraint applications IoT and embedded systems.

VIII.Future Scope

Our suggested architecture has several prospects for future enhancement. It involves improving robustness by addressing variability or uncertainty of individual memories to reduce detrimental consequences. Future research may also explore sophisticated mechanisms of error correction and fault tolerance for resilient operation.

Extending our design to accommodate larger deep learning models as well as datasets are also desirable directions.

Beyond this, integration of future memory technologies (e.g. RRAM, memristors) is a promising direction for significantly improve the power-efficiency and performance.

IX. Reference

- [1] Ielmini, Daniele, and Giacomo Pedretti. "Device and circuit architectures for in-memory computing." *Advanced Intelligent Systems* 2.7 (2020): 2000040.
- [2] Lopes, Noel, and Bernardete Ribeiro. "Deep belief networks (DBNs)." *Machine learning for adaptive many-core machines-a practical approach*. Cham: Springer International Publishing, 2015. 155-186.
- [3] Le Roux, Nicolas, and Yoshua Bengio. "Representational power of restricted Boltzmann machines and deep belief networks." *Neural computation* 20.6 (2008): 1631-1649.
- [4] Verma, Naveen, et al. "In-memory computing: Advances and prospects." *IEEE solid-state circuits magazine* 11.3 (2019): 43-55.
- [5] Nagar, Prashray, et al. "Emerging VLSI technologies for high performance AI and ML applications." *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. IEEE, 2024.
- [6] Yu, Wei, et al. "A survey on the edge computing for the Internet of Things." *IEEE access* 6 (2017): 6900-6919.
- [7] Burr, Geoffrey W., et al. "Neuromorphic computing using non-volatile memory." *Advances in Physics: X* 2.1 (2017): 89-124.
- [8] O'Connor, Peter, et al. "Real-time classification and sensor fusion with a spiking deep belief network." *Frontiers in neuroscience* 7 (2013): 178.
- [9] Dai, Weibang, et al. "Lookup Table-based Computing: A Survey from Software Implementations to Hardware Architectures." *ACM Computing Surveys* 58.7 (2026): 1-34.
- [10] Mitra, Subhasish, et al. "Built-in soft error resilience for robust system design." *2007 IEEE International Conference on Integrated Circuit Design and Technology*. IEEE, 2007.