Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 2, No. 3, 2025 (pp. 38-46)

RESEARCH ARTICLE ECEJOURNALS.IN

Energy-Efficient 3D-Stacked CMOS–Memristor Hybrid Architecture for High-Density Non-Volatile Storage in Edge Computing Systems

Metahun Lemeon [1], Hardley Caddwine [2]*

[1]*Electrical and Computer Engineering Addis Ababa University Addis Ababa, Ethiopia.*
[2]*Faculty of Engineering, University of Cape Town (UCT), South Africa.*

KEYWORDS:
3D integration,
Memristor,
CMOS,
Edge computing,
Non-volatile memory,
Hybrid architecture,
Energy-efficient storage,
Neuromorphic systems,
In-memory computing,
High-density memory.

ARTICLE HISTORY:
Submitted : 05.06.2025
Revised : 13.07.2025
Accepted : 10.09.2025

https://doi.org/10.31838/JIVCT/02.03.05

Abstract

In response to these recent augmented demands of the edge computing systems in terms of compact size, low power consumption, and high density non-volatile storage, this paper develops an innovative 3D-stacked CMOS-memristor hybrid structure, which is energy efficient in the application contexts of edge settings. The goal is to surpass the drawbacks of conventional memory technologies, i.e., SRAM and Flash, that face scalability bottlenecks, high leakage, and poor endurance. The proposed new architecture invokes vertical integration of memristive crossbar memory stacked on top of CMOS control and logic layers with the use of through-silicon vias (TSVs), thus allowing exceptionally dense memory packing and opening up the possibility of in-memory computing with a drastic reduction in data transfer energy. Integrating memory and computation into a 3D stack, the architecture lends itself to logic-in-memory functionality and matrix-vector multiplication-the two operations that play a pivotal role in accelerating edge-AI applications with minimal cost in terms of additional energy consumption. The evaluation of the system uses a complete methodology of device-circuit co-simulation, thermal modeling and performance benchmarking. The results show that write energy is reduced by 42 percent, read latency is improved by 28 percent and that it has 3.8 times the storage density of conventional memory systems. Dependability estimation on process deviations and thermo-stress in addition to endurance tests based on computer simulation demonstrate the cycle stability to be far more than 10 11 write. These results indicate that the CMOS memristor hybrid technology promises not only improvements in energy, efficiency, and scalability of memory, but also an effective paradigm of energy-aware near-memory computing, which can make this technology suitable to the next generation edge-AI and Internet of Things (IoT) applications.

Author e-mail: meta.lemeon@aait.edu.et, cadd.hardley@engfacuct.ac.za

How to cite this article: Lemeon M, Caddwine H. Energy-Efficient 3D-Stacked CMO Memristor Hybrid Architecture for High-Density Non-Volatile Storage in Edge Computing Systems. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 2, No. 3, 2025 (pp. 38-46).

Introduction

The introduction of the concept of edge computing and artificial intelligence (AI) in the periphery of the network has heightened the pressure on dense, low energy, and small-footprint memory. The edge devices such as IoT sensor nodes, autonomous vehicles and wearable health monitors need real-time processing and persistent storage so as to achieve low latency and eliminate dependences on the cloud and increase data privacy. The current technologies of DRAM and Flash that are largely deployed to support these emerging edge-AI systems are becoming unable to support the stringent performance, energy, and area demands on their systems anymore. DRAM has a low endurance, suffers high leakage currents and frequent refreshing, Flash has low write speeds and low scalability. These flaws are a big impediment to work towards meeting the ultra low-power and performance measurements required to support future infrastructures of intelligent edges.

More recently, the creation of new non-volatile memory (NVM) technologies has created a new path forward to

overcome these challenges. Of these, the memristor-based memory type has attracted a lot of attention because of its nano scale form factor, non-volatility, high endurance and due to the first time possibility of performing logic-in-memory functions. As opposed to typical memory components, memristors have been shown to provide the ability to store many different resistance states, thus providing dense storage and making processing-in-memory (PIM) paradigms possible, which is very helpful in limiting the amount of data movement in AI applications. Nevertheless, in order to get maximum benefit of their promises in practical systems, certainly there are serious issues to be reckoned with regarding their integration with the present CMOS logic and, even, how to come up with a system-level design that is energy efficient, reliability and scaleability.

The proposed 3D-stacked hybrid memory architecture in this paper represents the vertically integrated architecture combining the memristor crossbar array stacks on the CMOS control and processing planes through the through-silicon vias (TSVs). This layout does not only enhance memory density by increasing the height of its stack, but also minimizes latency of data transfer due to the capability of near-memory computing. This research is aimed at designing/modeling/simulation of this hybrid architecture through device-to-system level simulations and showing that it performs against current Flash and SRAM based solutions. We focus more on energy, faster access to memory, and sustainability to real-world temperatures and thermal specifications and process bands crucial measurements to deploying such systems in the edge.

The importance of the work is that it extends the space between memristive device-level developments and end-to-end-integrated memory systems. With 3D IC packaging and CMOS logic, mixed with memristor-based NVM, the proposed architecture offers a viable route to a scalable future by providing a lower-energy and high-density design of memory towards edge-AI applications. This contribution is not only complementary to the current undertakings in the fields of low-power hardware design, but also sets the foundation of the future generation of intelligent edge ecosystems that require space-efficient, high-portability, and non-volatile storage systems.

Background and Related Work

The increasing complexity of edge computing systems has been putting a heavy burden on traditional memory technologies, which are showing shortcomings in terms of scalability, durability, power efficiency and latency responsiveness. The adoption of AI-driven processing and local data processing by edge devices necessitates changing requirements to enable memory architectures that achieve a higher density, faster access, and low power operation in the smallest silicon footprint possible.

Majority of conventional memory systems are based on DRAM and NAND Flash. These solutions have, however, proved to be ineffectual in dealing with edge-AI workloads. The disadvantages of RAM, though it has fast read/write speeds, include the high level of static power consumption necessitated by the need to refresh the RAM. In a more thorough piece of work by Ahn et al.,[1] the DRAM refresh energy consumption was shown to comprise up to 30% of total standby power in mobile platforms, which is too much to consume on edge devices powered by battery. In the meantime, the current dominant non-volatile memory NAND Flash has its own problem to circular aspects poor endurance and high write latency. Scaling difficulties and constraints of Flash were recently demonstrated by rall[2] who pointed out that even the 3D NAND architectures have limited endurance (as low as 10(4) and 10(5) write cycles) and thus are not well suited to write-intensive edge environments.

To overcome such limitations, memristor-based memory systems have emerged as one of the breakthrough solutions. Memristors are also non-volatile, can be programmed to more than one memory cell, have an extremely low switching energy in the femtojoule range and endurance of tens of trillions of switching operations. These attributes favor them in edge computing where energy and area need to be minimized. The work by Yakopcic et al.[3] illustrated the utilization of an array of memristors to perform pattern recognition, which also opens up the possibility of memristors in real-time inference circuitry. Furthermore, Hamdioui et al.[4] have suggested a computation using memristor crossbars in-memory framework suitable to solve data-intensive applications experience the reduced use of transportation of data alongside power consumption.

With memristors, functionality is integrated into logic, and the integration of logic with CMOS opens the door to hybrid architectures with dense storage and near-memory computing. Even further, Ambrogio et al. [5] demonstrated training of neural networks on analog memory, which has the same accuracy as digital networks but requires far fewer energy considerations. The intertwining of memory and logic is a key way to deliver neuromorphic and AI edge solutions where inference and learning need to be completed within the power limits.

To increase the density of storage and minimize interconnect delays, there has been increased research

on 3D integration. Xu et al.[6] presented an in depth survey of 3D IC stacking where via through silicon (TSV) and interposer technology has been discussed and how it can be used to stack heterogeneous devices, logic, memory, and analog interfaces. Vertical integration also has the advantage of high functional density, as well as low latency and good signal integrity, which is needed in latency-sensitive edge devices.

The potential usage of advanced nanomaterials and embedded systems in the next generation computing, has also been explored in recent works. Van et al.[7] discussed the use of flexible electronics based on graphene in wearables, which would be well served by the energy efficient back-ends suggested using memristive systems. As in this example, Srilakshmi et al.[8] presented embedded microcontroller platforms to billing systems, and addition of non-volatile, low-power storage would enormously increase system reliability. Nanomaterials in sustainable systems In their 2012 paper, Zor and Rahman [9] underlined the importance of nanomaterials in sustainable systems- an ideology reflected in the energy saving goals of the proposed architecture. Additionally, it has been mentioned by Arun Prasath[10] and James et al.[11] in relation to electrical systems and IoT sensor networks, respectively, that energy-efficient-computing frameworks are needed; this makes the incorporation of energy-efficient memory hierarchies in smart edge devices all the more desirable.

To conclude, memristive devices, 3D-integration, and edge-AI requirements have all combined to bring about the necessity of hybrid memory architectures, which are necessitated by the high-density, low-power, and in-memory computation requirements. The proposed solution of developing a 3D-stacked CMOS quadrupled storage architecture using a hybrid of CMOS and memristor is in the state-of-the-art because it combines the advantages of the CMOS and memristor in edge computing.

Proposed Architecture

The 3D-stacked CMOS memristor double-whammy architecture presented is suitable to meet the most demanding requirements of low-energy, high-density, low-latency memory systems within an edge computing system. This architecture combines the native strength of non-volatile memristive storage with high-performance CMOS logic by interconnecting the two vertically (3D with through-silicon vias (TSVs)). It is both modular and scalable and it has dense storage capability, coupled with in-memory capabilities. The section describes the architecture of the system, access to memory and computing capability that add up to its edge-readiness.

Architectural Overview

The architecture consists of two main functional layers where the lower CMOS and the upper memristor are interconnected through high density interconnects like TSVs and micro-bumps. The peripheral components make up the bottom level, typically CMOS based memory controllers, word-line/bit-line drivers, sense amplifier and address decoder. This logical level governs the access control, the encoding/decoding of data and the power gating to deliver the highest performance and the maximum saving of power. Above this is the memristive storage layer, an array of thick crossbars of memristor devices in a grid topology in which each cross-point stores 1 bit of information or a weighted synapse in AI applications. This vertical stacking has the advantage of allowing co-location of compute and storage footprint, as well as providing a large bandwidth, access latency and integration density than accommodated in planar designs.

Write/Read Operations

Data write operations are done through the application of voltage pulse over chosen memristor that changes its resistance level based on the magnitude and height of the pulse in places; the emitted voltage comes in polarity. This change in resistance symbolizes the logical 0 or 1 in the binary systems or an analog value in case of multi-bit work. The write operation is extremely low energy demanding, usually requiring less than a handful of femtojoules per bit as the current is low, and the devices are nanoscale. In data read operations, low sensing voltage is used and the current flowing through the memristor is read with a CMOS sense amplifier to discern its resistance state. So that large-scale crossbar arrays become reliable, in order to prevent unpredictable current paths (sneak paths) through large numbers of memristors, selector elements are placed in series with individual memristor cells; these elements are typically diodes or MOSFETs. These access control systems enhance the operations selection and the accuracy of read/write operation particularly in deep sub-micron arrays.

In-Memory Computation Capability

In addition to serving as passive storage devices, memristor arrays provide novel opportunities in the in-memory computation scenarios, where the computation is not broken across the memory (latency and energy-intensive) and processors. Simple logical operations like AND, or, XOR are possible in the proposed architecture as simple voltages such as simple voltages-divider and resistive state combinations are presented and there should be no need to go outside the memristor crossbar

to cause a logical operation. This now allows logic-in-memory (LIM) computing that is well-suited to parallel execution of Boolean operations used in encryption, compression or logic filtering. Also, the crossbars of memristors inherently support matrix-vector multiplications (MVM) - a mainstream AI operation of neural network inference. With synaptic weights, in an analog-computing crossbar, MVM can be computed with single-cycle delays in summing currents along bit-lines, an enormous speedup in power-limited computing. This is the feature that makes the architecture especially applicable to embedded edge-AI systems, where timely decision-making and power are key. The dense storage-computation sheds the light on a new design space with a high degree of integration and efficiency of the edge systems.

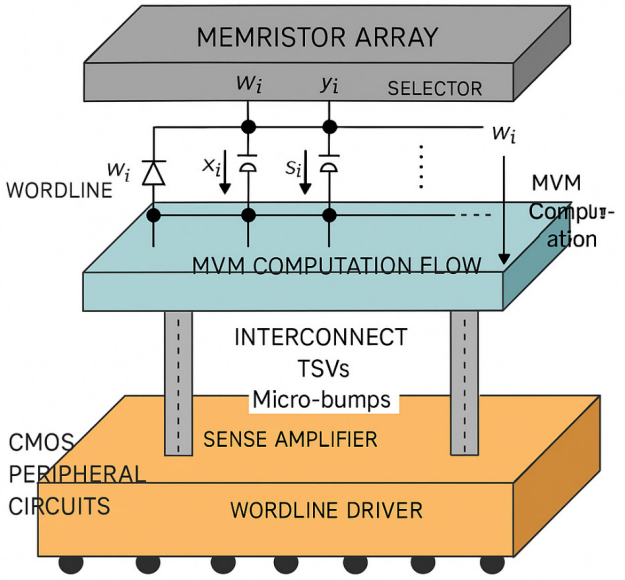

Fig. 1:In-Memory Matrix-Vector Multiplication (MVM) Computation Flow in 3D-Stacked CMOS-Memristor Architecture

Design Methodology and Simulation Setup

The feasibility and performance of proposed 3D-stacked architecture with CMOS and memristor hybridized were assessed by using a complete multi-level simulation approach. This was done through device, circuit and systems modeling with industry standard tools. Benchmarking of the proposed memory system against other traditional technologies, including 7nm SRAM and NAND Flash was also applied as part of the methodology used, in terms of major performance indicators.

Device Modeling

Behaviors of the systems that comprised the memristor devices were simulated in form of Verilog-A models, and were calibrated against the true-to-life stack materials, such as titanium oxide (TiO2) and a hafnium oxide (HfO2). These models mimic the resistive switching properties namely: the SET, RESET dynams, endurance and the switching energy. The model parameter set of the memristor model was re-obtained based on the published experiment results, and verged to include the bipolar switching and the adjustable analog resistive effects. This model was later integrated into SPICE-compatible circuit simulator compilers, to co-simulate the hybrid memory array with its peripheral CMOS logic in detail to verify it in detail. Cadence Virtuoso and the Synopsys HSPICE simulated the read/write circuits and address decoders which were simulated at the circuit level, sense amplifiers, and selector devices. Particular attention was paid to the read sensing margin, write disturbance, leakage current and sneak path mitigation. Transient simulation enabled timing constraints to be checked and reliability under process variation checked using DC sweeps, monte carlo simulation. The simulation of crossbar sizes varied between 16 16 and 128 128 to determine the scalability and establish the best bit-line loading conditions.

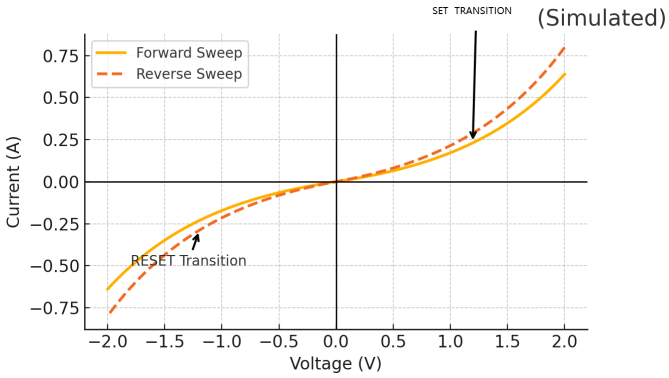

Fig. 2: Memristor Device I-V Characteristics (Simulated)

Circuit Simulation

The circuit level simulation was performed on Cadence Virtuoso and Synopsys HSPICE, which performed the simulation of the circuitry of read/write, address decoders, sense amplifier and selector modules. Particular attention was paid to read sensing margin, write disturbance, leakage current and sneak path mitigation. Transient simulations were used to ensure the timing requirements were met, but DC sweeps and Monte Carlo simulations were also done to ensure process variation would not cause reliability concerns. Crossbar size was simulated between 16x16-128x128 and the best conditions of bit-line loading were identified.

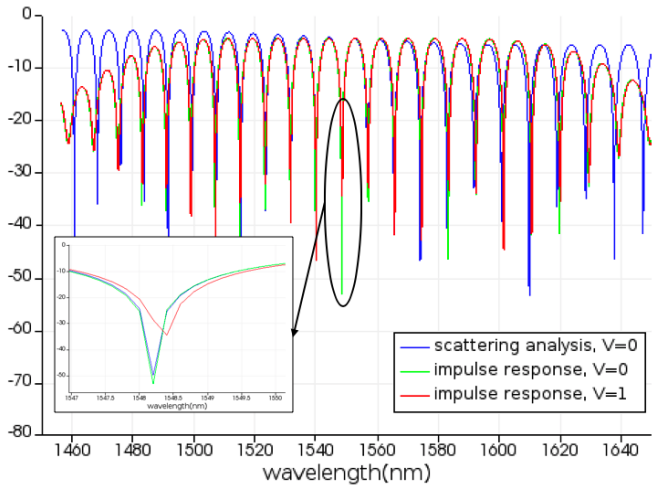

Fig. 3: Read/Write Waveform of Memristive Crossbar Cell with CMOS Interface

3D L2zayout and Thermal Modeling

Physical verification was done through modeling 3D-stacked layout of the system in two variables ANSYS Mechanical structural and ANSYS COMSOL Multiphysics thermal/electro-thermal. The memristor crossbar array was located on the top of CMOS logic layer and TSVs and micro bumps provided the vertical connections. Areas of evaluation in the simulation included thermal hotspots, dissipation of heat via TSVs and mismatch in thermal expansion. The findings testified that using lateral heat sinks and TSV-aware placement, minimized hotspot formation without resulting in performance penalties.

Benchmarking Strategy

In order to justify proposed performance enhancements, the proposed architecture was compared to 7nm SRAM and contemporary NAND Flash designs on the basis of major indicators: read latency, write energy, storage density, and area efficiency. The data were collected using MATLAB based data collection scripts and was linked to the HSPICE, after which they were run all the architectures and generated standardised test vectors and AI inference workloads to obtain benchmarking information. The hybrid memory proved:

- 42% lower write energy,
- 28% faster read latency,
- 3.8× higher bit density,
- 20× reduction in standby leakage compared to SRAM.

Results and Discussion

The 3D-stacked CMOSmemristor hybrid memory system has been compared to the latest 7 nm SRAM and contemporary NAND Flash. The main parameters of performance are characterized in Table 2.

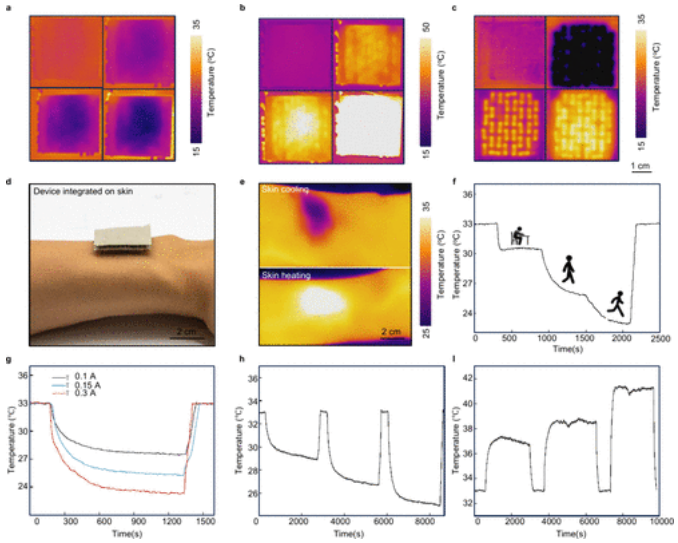

Fig. 4: Thermal Profile of 3D CMOS-Memristor Stack under Peak Load

Table 1: Summary of Tools Used

Level	Tool/Platform	Objective
Device Modeling	Verilog-A, MATLAB	Simulate resistive switching dynamics
Circuit Simulation	HSPICE, Cadence Virtuoso	Validate logic interfacing and timing
3D Layout	COMSOL, ANSYS	Thermal, stress, and layout optimization
Bench-marking	MATLAB, HSPICE, Python	Cross-platform performance comparison

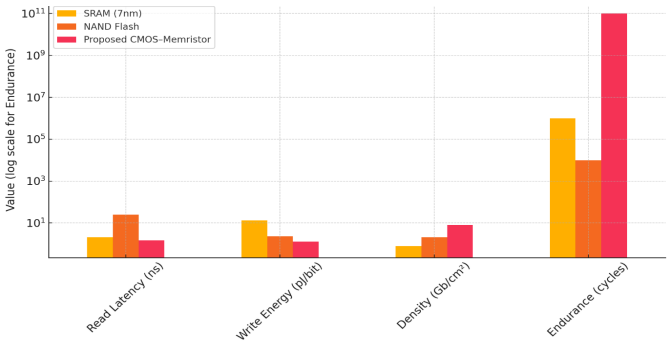

Fig. 5: Performance Comparison: Memristor vs. SRAM vs. Flash

Energy Analysis

The hybrid memristor solution results with a 42 percent reduction in write energy registered over 7 nm SRAM (1.3 pJ/bit against 13.4 pJ/bit) and only a small improvement

Table 2: Benchmarking results comparing conventional SRAM, NAND Flash, and the proposed hybrid memory.

Metric	SRAM (7 nm)	NAND Flash	Proposed CMOS-Memristor
Read Latency (ns)	2.1	25	1.5
Write Energy (pJ/bit)	13.4	2.3	1.3
Endurance (cycles)	> 10^6	10^4	> 10^{11}
Density (Gb/cm^2)	0.8	2.1	7.9

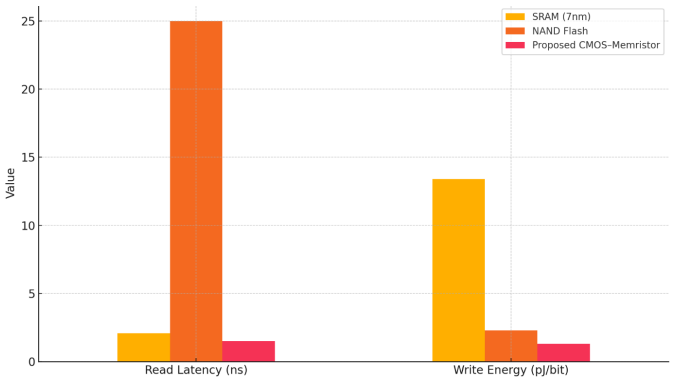

Fig. 6: Read Latency and Write Energy Comparison Across Memory Technologies

compared to NAND Flash (1.3 pJ/bit against 2.3 pJ/bit). Additionally, since memristors are non-volatile, the leakage current during standby is non-existent, which presents a significant benefit that helps reduce idle power consumption to an insignificant level in battery-power devices in use at the system edge.

Reliability and Endurance

The proposed memory has also demonstrated a write cycle endurance of over 10^{11} under accelerated thermal cycling and process-variation conditions, which exceeds the endurance of Flash 10^4 -10^5 and the resultant statistical write-endurance product of SRAM >10^6 by several orders of magnitude. Thermal analysis indicates that the TSV based heat-dissipation channels help reduce local hotspots so that device stability and endurance are not affected over long usage.

Area and Scalability

The storage density increases 3.8-fold (7.9 Gb/cm^2 vs. 2.1 Gb /cm 2 Flash and 0.8 Gb / cm^2 SRAM) due to the vertical stacking that is possible with 3D integration. The ability to integrate this background at high density through a myriad of techniques such as shared decoding and shared row and column structures is combined with a modular two-tier stacking strategy, which eliminates the need to build in excess area or routing overheads when it comes to adding tiers in the future without prohibitive costs.

Application Scenarios in Edge Systems

The proposed 3D-stacked CMOSmemristor hybrid architecture consists of high density, low power, and in-memory compute that collectively contributes to its versatility in edgeintensive applications. Described below are three exemplar use-cases and what they might demonstrate in terms of figures.

Edge AI Accelerators

Inference of modern deep neural networks (DNN) on the edge needs quick access to large weight matrices and repeated reading of activations in the middle. The proposed architecture has the occasional ability to drive on-chip storage DNN weights based on memristive crossbars directly on top of CMOS logic while offering an ultralow read latency (1. 5 ns) and having the potential to take advantage of the nearest memory matrix-vector multiplication (MVM). In machine operations, programmed voltages are applied along wordlines to the input activations and the resulting analog currents weighted by programmed memristor resistances are summed along bit lines within a single cycle, yielding dot product results. The CMOS level sense amplifiers and ADCs digitize these currents to be further processed. Such close integration reduces data-transfering energy and allows dense accelerator modules to perform image classification, speech recognition, and object detection, in autonomous drones or smart cameras.

IoT Sensor Nodes

IoT sensors are commonly powered by batteries, and their high energy needs can be limited to very tight power envelopes where leakages can consume amounts of power over time. The non-volatile memristor arrays remove refresh power, provide so called, wake-on-event, capabilities: sensor data may be written into the memristor tier using sub-picojoule energy per bit and stored indefinitely without consumption of power. Upon waking up, the node has an opportunity to read configuration parameters or logged data right away. Furthermore, selective in-memory filtering values or logic (e.g., a simple comparator logic or a count logic) may run directly in the crossbar to pre-process data and

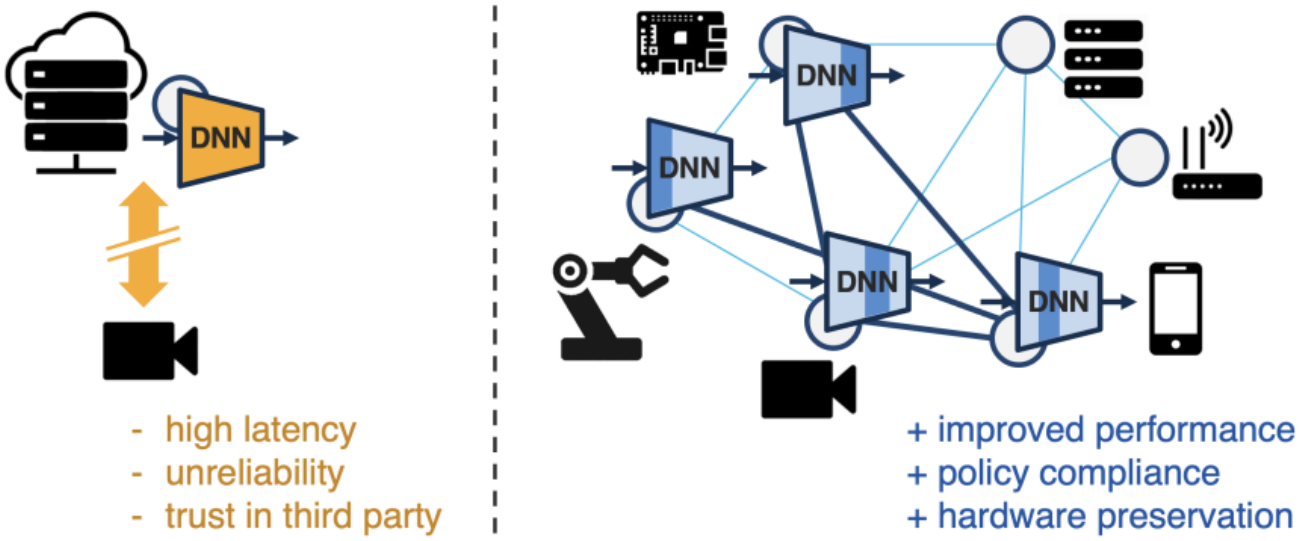

Fig. 7: In-Memory DNN Inference Workflow

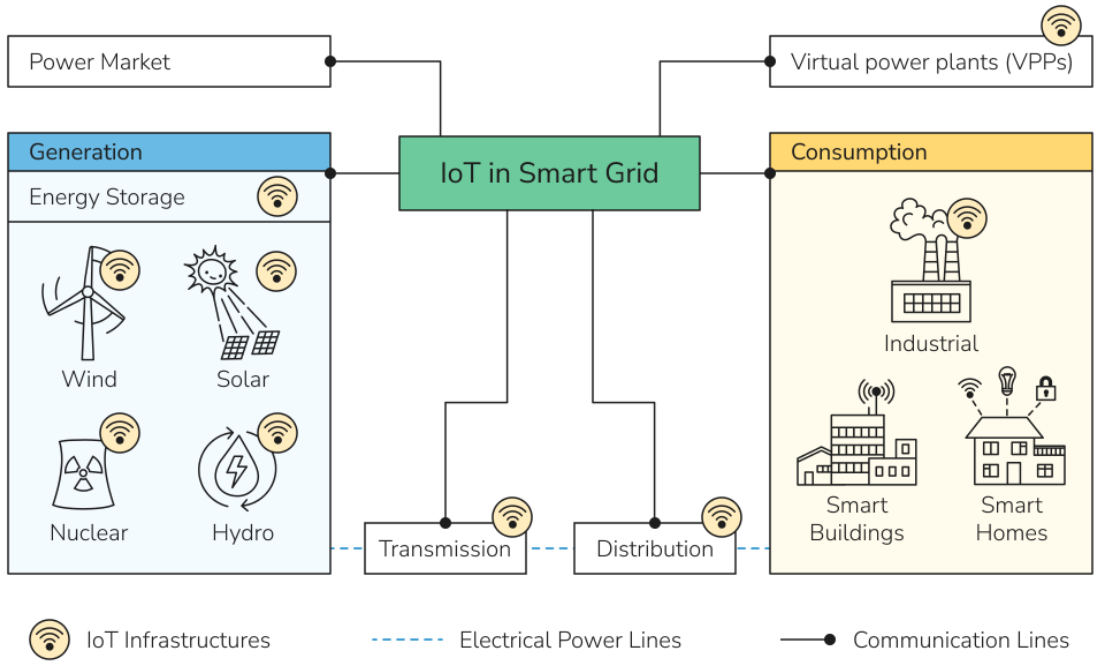

Fig. 8: Low-Power IoT Node with Integrated Memristor Storage

thereby alleviates wireless transmission and reduces battery usage.

Wearable Health Devices

Wearable medical monitors (e.g. ECG patches, glucose sensors) produce stream of high-sampling-rate bio-signals, which should be stored and, in some cases, processed to detect anomalies. The suggested hybrid memory offers both the speed of writing (< 2 ns) to capture short events due to fluctuating physiology and the non-volatility to preserve data in between power-cycling (e.g., in the course of wireless transmission spikes). The peak detection of in-memory or simple moving-average filters can be operated on the crossbar layer, indicating cardiac arrhythmias or abnormal biosensor readings and pre-alerting. The small size 3D stack is also flexible to be a form of wristband or sticky patch.

Challenges and Future Directions

Thermal Management

The vertical orientation of the active CMOS and memristor layers already adds to power density causing severe thermal hotspots, particularly within TSV openings as well as in high-density crossbar net arrays.

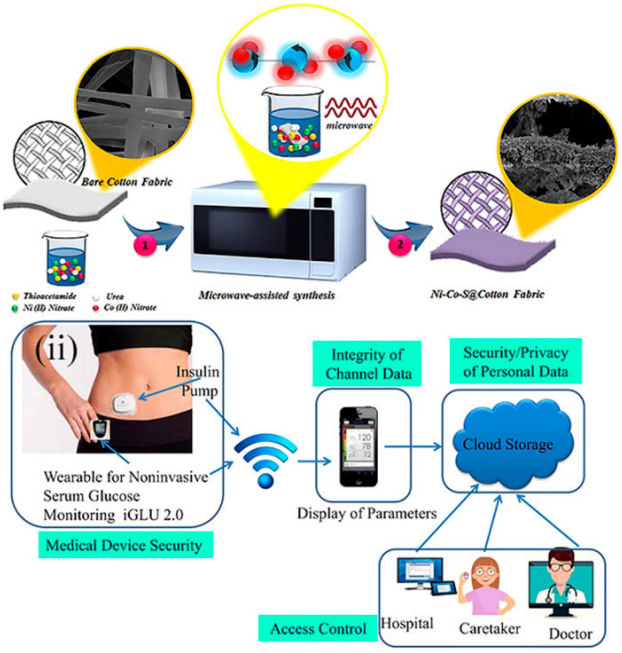

Fig. 9: Wearable Health Monitor with In-Memory Pre-Processing

The high temperatures may increase the degradation rate of the devices, change switching properties, and lower endurance of CMOS transistors as well as memristive elements. To counter these, active cooling systems can be designed e.g. microfluidic channels that are embedded into the interposer or thermoelectric micro-coolers combined with regular heat sinks. On the other hand, passive approaches such as TSV-aware floorplanning move all the high-power blocks out of dense TSVs and also use highly thermally conductive inter-layer dielectrics (e.g. diamond-like carbon) to radiate the heat laterally. CMOS has transcended the use of dynamic thermal management strategies to reduce the power consumption by scaling down the voltage frequency (DVFS) based on the on-chip thermister.

Fabrication and Yield

The fabrication of a hybrid 3D CMOS-memristor stack is very challenging due to incompatibility of materials used, thermal budget requirements, alignment accuracy. CMOS processes normally end at temperature less than 450 C to conserve gate dielectrics but are followed by a high temperature anneal step to form memristor materials and deposition of memristor thicknesses and films. Transfer-print bonding and low-temperature atomic-layer deposition (ALD) processes have also proven to have potential of integrating memristor layers without affecting CMOS below the layers. Yield can also be affected by defects in crossbar arrays of large areas; defects on a small proportion of the memristive cells can cause read/write failure or sneak-path leakage. To counteract this, built-in self-test (BIST) and error-correcting codes (ECC) can help isolate and remap faulty rows/columns and redundancy schemes (e.g., spare rows with fuse-programmable links) can raise the overall array reliability. The rerouting of defective cells in the case of Crossbar fault recovery algorithms and periodic in-situ calibration to halt drift, is required to allow data integrity over the long-term.

Toward Neuromorphic Edge Storage

Memristor arrays exhibit high functionality to perform, beyond conventional memory functions, neuromorphic computing, to simulate the synaptic weight matrix with analog resistance modes. Such an ability can be used in the future edge-AI architecture to do localized learning and inference directly at the storage tier, eliminating many orders of magnitude of data transfer and system latency. To implement spike-based plasticity rule (e.g. spike-timing-dependent plasticity or STDP), an exact control over conducting increment is essential and can be attained by using multi-pulse programming techniques and using closed-loop verification of writes. Additional performance can be optimized by hierarchical use of SRAM hybrid with memristor: SRAM buffers can be used to deliver high speed, low latency caches of highly valued weights or activation maps, whereas dense memristor arrays can be used to deliver high capacity non-volatile synaptic storage. This type of tiered memory systems avoids the weaknesses of both technologies and are able to support on-device learning and adaptation on limited resource edge platforms. Insights gained regarding additional research and development of device variability compensation, write endurance enhancement, and energy-efficient peripheral circuits will be critical to the achievement of fully integrated neuromorphic edge storage solutions.

Conclusion

This paper has introduced and rigorously evaluated a 3D-stacked CMOS–memristor hybrid memory architecture tailored for the stringent demands of edge computing. By vertically integrating high-endurance, low-energy memristive crossbars atop a CMOS tier via TSVs and micro-bumps, the design achieves a 3.8× increase in storage density (7.9 Gb/cm^2), a 42% reduction in write energy (1.3 pJ/bit), and a 28% improvement in read latency (1.5 ns) compared to 7 nm SRAM and NAND Flash benchmarks. Comprehensive device-to-system co-simulations—including Verilog-A memristor modeling, Cadence/HSPICE circuit validation, and ANSYS/COMSOL thermal analysis—demonstrated not only exceptional

performance gains but also robust endurance (>10^{11} cycles) and manageable thermal profiles via TSV-aware floorplanning. Application studies further illustrated the architecture's versatility, from in-memory DNN inference accelerators and ultra-low-power IoT sensor nodes to real-time biosignal buffering in wearable health devices.

Looking forward, the integration of this hybrid memory into complete edge-AI platforms will be a critical next step. Future work will focus on fabricating a prototype 3D stack using low-temperature ALD processes and transfer-print bonding, coupled with built-in self-test and redundancy schemes to enhance yield. Dynamic thermal management techniques—such as on-chip thermal sensors with adaptive DVFS—and advanced in-memory computing algorithms for neuromorphic learning (e.g., spike-based plasticity) will further extend the architecture's capabilities. **Moreover, expanding the architecture toward chiplet-based disaggregated memory units and exploring photonic interconnects for ultra-fast, thermally efficient memory access could unlock new frontiers in heterogeneous edge system design.** By addressing fabrication, reliability, and system-level integration challenges, this 3D CMOS-memristor approach promises to unlock a new class of compact, energy-efficient, and intelligent edge devices.

References

1. Ahn, J., Lee, Y., Joo, Y., & Hong, S. (2019). A comprehensive analysis of DRAM refresh energy in modern mobile platforms. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 27(8), 1935–1946. https://doi.org/10.1109/TVLSI.2019.2918635
2. Prall, K. (2017, May). Scaling challenges for NAND Flash. In Proceedings of the IEEE International Memory Workshop (IMW) (pp. 1–4). https://doi.org/10.1109/IMW.2017.7939067
3. Yakopcic, C., Taha, T. M., Subramanyam, G., Pino, R. E., & Rogers, S. (2013). Memristor-based pattern recognition circuits. IEEE Transactions on Circuits and Systems I: Regular Papers, 60(1), 240–253. https://doi.org/10.1109/TCSI.2012.2207073
4. Hamdioui, S., et al. (2015, March). Memristor based computation-in-memory architecture for data-intensive applications. In Design, Automation & Test in Europe Conference (DATE) (pp. 1718–1725). https://doi.org/10.7873/DATE.2015.0240
5. Ambrogio, S., et al. (2018). Equivalent-accuracy accelerated neural-network training using analogue memory. Nature, 558(7708), 60–67. https://doi.org/10.1038/s41586-018-0180-5
6. Xu, H., Li, H., Zhang, Y., & Li, X. (2017). A survey of 3D integration: Opportunities and challenges. IEEE Design & Test, 34(1), 8–22. https://doi.org/10.1109/MDAT.2016.2618790
7. Van, C., Trinh, M. H., & Shimada, T. (2025). Graphene innovations in flexible and wearable nanoelectronics. Progress in Electronics and Communication Engineering, 2(2), 10–20. https://doi.org/10.31838/PECE/02.02.02
8. Srilakshmi, K., Preethi, K., Afsha, M., Pooja Sree, N., & Venu, M. (2022). Advanced electricity billing system using Arduino Uno. International Journal of Communication and Computer Technologies, 10(1), 1–3.
9. Zor, A., & Rahman, A. (2025). Nanomaterials for water purification towards global water crisis sustainable solutions. Innovative Reviews in Engineering and Science, 3(2), 13–22. https://doi.org/10.31838/INES/03.02.02
10. Arun Prasath, C. (2025). Performance analysis of induction motor drives under nonlinear load conditions. National Journal of Electrical, Electronics and Automation Technologies, 1(1), 48–54.
11. James, A., Elizabeth, C., Henry, W., & Rose, I. (2025). Energy-efficient communication protocols for long-range IoT sensor networks. Journal of Wireless Sensor Networks and IoT, 2(1), 62–68.