Resistive Memories: Multifaceted Impact on Neuromorphic Computing

P. Clémençon[1], T. Hirtzlin[1], F. Rummens[2], T. Dalgaty[2], E. Hardy[1], M. Ezzadeen[1], J. Minguet Lopez[1], O. Billoint[1], L. Grenouillet[1], L. Hutin[1], D. Querlioz[3], E. Vianello[1]
[1]CEA-Leti, Univ. Grenoble Alpes, France, email: elisa.vianello@cea.fr
[2]CEA-List, Univ. Grenoble Alpes, France, [3]Univ. Paris-Saclay, CNRS, France

***Abstract*—Demand for energy-efficient AI has driven significant interest in Compute-In-Memory (CIM) architectures based on memristor devices, which reduce energy consumption and latency by minimizing data movement. However, practical deployment remains limited by device variability. This work explores how such stochastic properties can be harnessed — rather than suppressed — through Bayesian CIM, where memristors naturally encode probability distributions for uncertainty-aware inference. We review recent advances toward enabling on-chip learning at the edge, including hybrid synapses that decouple inference from learning. In addition, we examine algorithmic frameworks such as meta-learning, which retains global training in the cloud while supporting efficient edge adaptation. We also discuss biologically inspired mechanisms to mitigate catastrophic forgetting. Finally, we highlight the role of heterogeneous integration and 3D architectures as key enablers of future scalable, neuromorphic systems.**

***Index Terms*—Resistive Memories, Memristors, Compute-In-Memory, Bayesian CIM, on-chip learning.**

I. Introduction

The energy demands associated with training and deploying artificial intelligence and machine learning (AI/ML) systems are increasing at an unsustainable rate. Simultaneously, the widespread deployment of portable sensory devices — and the accompanying surge in data generation — poses serious challenges in terms of memory density and power efficiency. Conventional AI architectures are not designed to cope with this scale and complexity, rendering the current trajectory increasingly untenable.

These conventional processing systems typically rely on high-density off-chip memory to store large neural network models and require frequent memory access to perform vector-matrix multiplication operations. Data movement — not computation — dominates both energy consumption and latency. For instance, transferring a 32-bit word from off-chip DRAM can consume up to 300 times more energy than a 32-bit floating-point multiplication [1], with access latencies ranging from tens to hundreds of nanoseconds [2].

Neuromorphic computing and memristor technologies, such as resistive RAM (RRAM), magnetic RAM (MRAM), phase-change memory (PCM), and ferroelectric memories have garnered significant interest as potential enablers of energy-efficient AI [3]. In particular, memristor-based compute-in-memory (CIM) architectures present a promising approach. By integrating vector-matrix multiplication operations directly into memory arrays, these architectures reduce the need for energy-intensive data movement. Most implementations employ crossbar arrays, where memristors at each cross-point store synaptic weights as conductance values. Input vectors are applied as voltages across the rows, and the resulting output currents—governed by Ohm's and Kirchhoff's laws—naturally implement the core operations of neural networks: vector-matrix multiplication.

In conventional CPU/GPU architectures, loading and processing a batch of inputs incurs a fixed energy cost, which becomes more efficient on a per-input basis as the batch size increases. However, this amortization is less effective in edge computing scenarios, where batch sizes are typically small due to latency constraints. In contrast, IMC systems rely on a fixed physical array whose energy consumption per inference is independent of the batch size. As a result, the energy required to classify a single input remains nearly constant, providing significant advantages in low-latency, low-power edge applications that operate with small batch sizes.

Despite these advantages, practical deployment faces critical challenges, including device variability, limited scalability, and the absence of robust and efficient on-chip learning mechanisms. This paper explores emerging technologies, device models, and algorithm-architecture co-design strategies aimed at addressing these hurdles to enable scalable, energy-efficient neuromorphic CIM systems.

II. Bayesian Compute-In-Memory

A fundamental limitation of analog compute-in-memory (CIM) based on memristive devices is reduced computational accuracy, primarily due to device variability, programming non-uniformity, and read noise. These challenges stem from the low precision of such devices: when used as multi-level memories, memristors typically support only a limited number of statistically distinguishable conductance levels [4], [5].

This inherent variability has catalyzed the development of Bayesian compute-in-memory (Bayesian CIM) to implement in hardware Bayesian neural networks. Rather than suppressing stochastic behavior, a Bayesian CIM leverages it to represent and quantify model uncertainty. In this framework, each network parameter — including weights and biases — is modeled as a probability distribution, typically a normal distribution [6]. The intrinsic randomness of memristive devices enables efficient sampling from these distributions. These

Invited paper

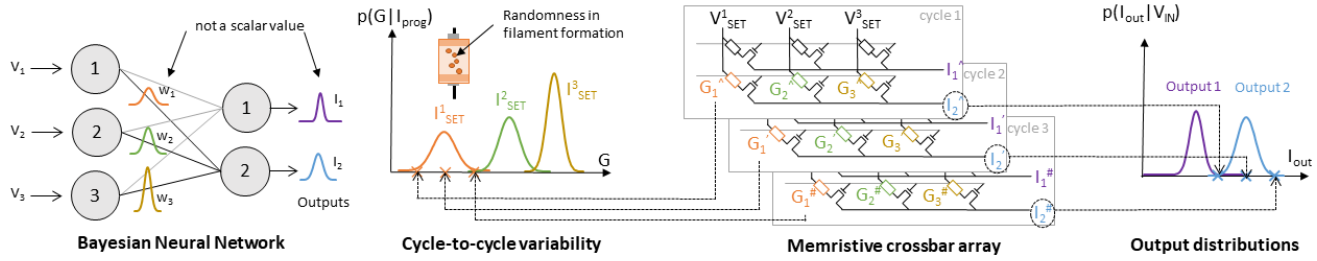

Fig. 1. In a Bayesian neural network, each parameter is represented by a Gaussian distribution rather than a single scalar value. Due to intrinsic variability, a population of nominally identical memristors can inherently encode such distributions. Crossbar arrays simultaneously perform vector-matrix multiplication operations and draw weight samples.

samples are propagated through the network to form ensembles of outputs, allowing predictions to be accompanied by well-calibrated uncertainty estimates — an essential capability for high-stakes applications such as medical diagnosis and predictive maintenance.

The network topology in Bayesian neural networks mirrors that of conventional networks: inputs propagate through successive fully connected, convolutional, or recurrent layers, each followed by nonlinear activations. However, in Bayesian CIM, each weight is sampled from its associated distribution during inference.

Due to cycle-to-cycle variability, a population of nominally identical memristors can inherently encode a probability distribution, making them well-suited for representing Bayesian network parameters. Slight variations among devices naturally form an ensemble that approximates a target probability density (Fig. 1). For example, a small group of devices ($G_1^{'}$, $G_1^{\sharp}$, $G_1^{'}$), programmed under identical conditions, can collectively realize a normal distribution with a defined mean (μ) and standard deviation (σ). The programming conditions determine both μ and σ of the resulting distribution. Crossbar arrays can then perform the required vector-matrix multiplication operations while simultaneously drawing weight samples in the analog domain.

This approach has been experimentally demonstrated with filamentary memristors [7], [8], where the stochastic nature of filament formation leads to a normally distributed range of conductance values after each programming event. Similarly, phase-change memory (PCM) devices exhibit variability due to the stochastic nucleation and growth of crystalline domains during crystallization, resulting in unique microstructures (e.g., grain size, distribution, and orientation) with every programming cycle [8].

Read-to-read noise in filamentary memristors — where repeated reads of the same cell produce independent samples from a narrow distribution centered on the stored value — has also been proposed as a mechanism for Bayesian parameter sampling [9]. An analogous effect occurs in devices based on two-dimensional (2D) materials, where fully exposed channels are especially sensitive to surface traps, leading to distinctive read noise profiles [10].

For practical Bayesian CIM systems, it is crucial to operate over sufficiently large sampling domains where μ and σ can be independently controlled. A common strategy to decorrelate these parameters involves subtracting conductances obtained from two separate sampling operations [8], [10]. An alternative approach is to store μ and σ in separate arrays and use them jointly to generate samples [11].

Flagship experimental demonstrations of Bayesian CIM using emerging nanotechnologies have illustrated the key advantages of Bayesian inference. In particular, the variance of the predictive distribution offers a direct measure of model confidence, enabling out-of-distribution detection [8]. Furthermore, Bayesian methods maintain consistent and reliable performance even when inputs are corrupted by noise [11].

Finally, not all forms of device randomness are well-suited for Bayesian CIM. Predictable and well-characterized sources of variability — such as device-to-device variation and read-to-read noise — can be harnessed effectively. In contrast, uncontrolled phenomena like temporal drift can degrade distribution fidelity and compromise the reliability of probabilistic computation. In this context, accurate physical models that capture and differentiate between various sources of variability are essential for guiding device design and ensuring robust system behavior.

III. Toward On-Chip Learning

Natural organisms do not rely on hardwired circuits for every possible action in every possible environment; instead, they continuously learn new tasks to adapt to changing conditions. Similarly, on-chip learning is essential for adapting to data shifts caused by factors such as sensory or environmental noise, analog hardware degradation, or for performing tasks not addressed during offline training. However, enabling learning with memristive devices presents significant challenges from both hardware and algorithmic perspectives.

From a hardware standpoint, no existing memory technology simultaneously meets the conflicting requirements of inference and training. Inference demands high read endurance and noise robustness to reliably access pre-stored parameters, whereas training requires high write endurance, low programming energy, and fast switching speeds to support frequent, high-precision parameter updates. To address this trade-off, several hybrid memory architectures have been proposed (Fig. 2). These approaches combine memristors—used

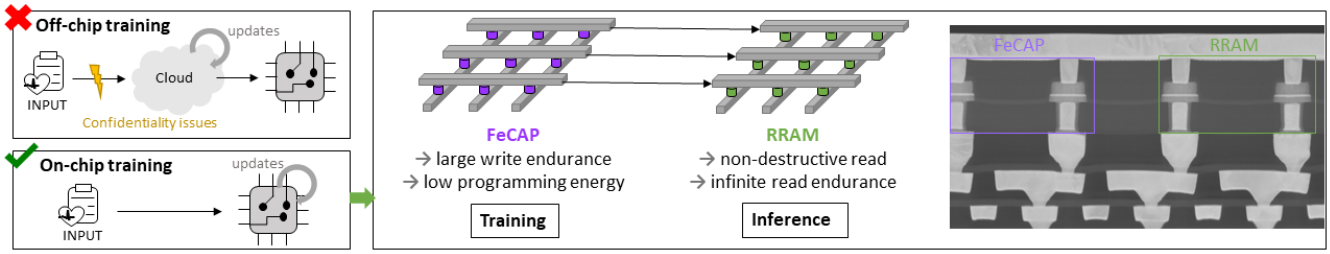

Fig. 2. High-level overview of off-chip and on-chip training procedures. On-chip learning requires a hybrid synaptic architecture combining two device types: one with high write endurance and low programming energy for training, and another with non-destructive readout and virtually infinite read endurance for inference. A unified BEOL-integrated metal–ferroelectric–metal (MFM) stack can function as either a ferroelectric capacitor for training or, after a forming operation, as a memristor for inference.

primarily for inference—with memory technologies better suited for training, such as SRAM [12], ferroelectric memories [13], [14], or gain-cell-based technologies [15].

From an algorithmic perspective, several challenges remain. First, traditional learning algorithms used for off-chip training offload the computational burden to the cloud, reducing on-device complexity and allowing the costly training process to be performed once and reused across multiple users. However, this centralized approach limits adaptability and raises concerns about data confidentiality, as both user data and model parameters must be transmitted and processed externally. A promising alternative is meta-learning — specifically, Model-Agnostic Meta-Learning (MAML) — which enables models to be pre-trained off-chip across a diverse set of tasks (*learning to learn*). This approach allows for rapid adaptation to new tasks on-chip using only a few data samples and energy-efficient weight updates [16]. Notably, MAML has been successfully implemented using both filamentary memristors [17] and phase-change memory (PCM) devices [18], [19].

Another major challenge is catastrophic forgetting, a phenomenon in which artificial neural networks lose previously acquired knowledge when learning new tasks — a limitation that contrasts sharply with the stability of biological learning. The human brain is believed to overcome this issue through synaptic metaplasticity, where synapses dynamically adjust their learning rates based on the importance of prior tasks [20]. This biological principle has inspired hardware-friendly solutions: recent studies have demonstrated that metaplasticity can reduce catastrophic forgetting in binarized and quantized neural networks by using memristors to store the quantized weights, while a separate digital memory holds the corresponding hidden weights as metaplastic variables [21], [22]. These hidden weights represent task importance and guide future updates.

Bayesian learning is grounded in the principles of Bayesian inference, offering multiple ways to incorporate new training data and update model parameters. Given that nanodevice-based hardware is well-suited for Bayesian inference, it also shows strong potential for enabling hardware-native updates of parameter distributions for on-chip learning. One promising approach is to implement Bayesian learning directly in hardware by embedding the stochastic search steps of Markov Chain Monte Carlo (MCMC) methods. Memristor arrays have been used in this context to learn to detect cancer in mammography images with classification accuracy comparable to that of software-based methods [7]. This demonstrates that Bayesian learning is particularly well-suited for efficient, on-chip learning. Moreover, recent work has shown that synaptic uncertainty, stored by exploiting device-level variability, can be harnessed to enable continual learning, allowing systems to adapt to new tasks without catastrophic forgetting [23].

IV. Future: Heterogeneous Systems

Progress in memristor-based neuromorphic architectures requires moving beyond traditional CMOS scaling, driving innovation in materials, device technologies, and integration strategies. Rather than relying on a uniform technology, future systems will embrace heterogeneous devices, each optimized for specific computational roles. This paradigm shift demands a redefinition of process technologies, with a focus on 3D integration — including monolithic 3D integration, where memory layers are sequentially fabricated on a single substrate, and advanced 2.5D and 3D packaging, which combine multiple chips into a unified package [24]. These approaches enable increased on-chip memory density, allowing for the local storage of larger models — a critical requirement for Bayesian models, which are more memory-intensive due to the need to sample and store multiple conductance values per parameter. They also allow the co-integration of diverse memory types to support on-chip learning. In addition, they facilitate the heterogeneous integration of function-specific modules, each fabricated using the most appropriate process node and assembled either side-by-side on a shared interposer or stacked vertically. This architectural strategy mimics the small-world connectivity found in biological brains (see Fig. 3), laying the groundwork for ultra-low-power, bio-inspired computing systems.

For example, the brain of an insect contains specialized regions for parallel signal processing, real-time learning, and multimodal sensory integration. These regions operate using diverse neural and synaptic types and multiple coding strategies — an inherent heterogeneity absent in conventional computing architectures. Emulating this in silicon entails the co-integration of multiple sensory and compute chiplets, each

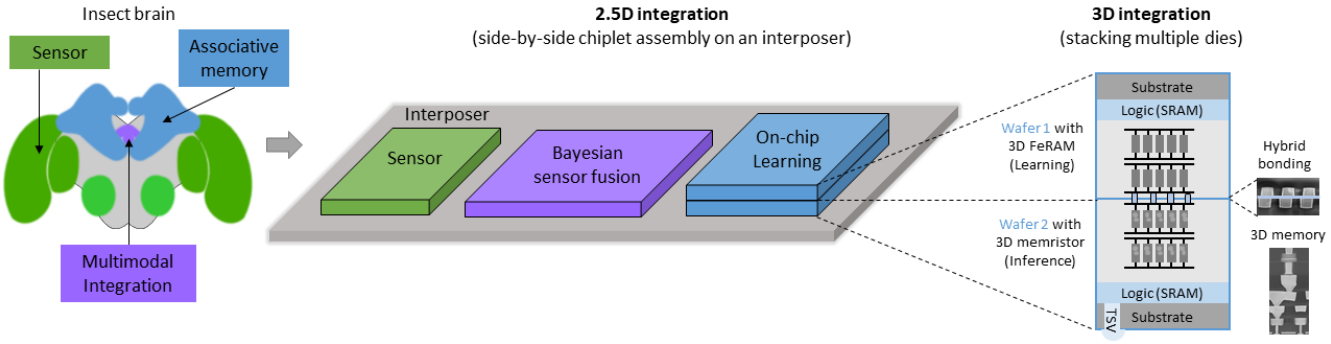

Fig. 3. Future neuromorphic systems, inspired by the modular organization of insect brains, will consist of a heterogeneous collection of chips—each optimized for tasks such as sensing, learning, and sensor fusion—fabricated with different technologies and interconnected through advanced 2.5D and 3D packaging. Increased on-chip memory density will be further enabled by monolithic 3D integration at the single-chip level.

leveraging distinct device technologies, coding schemes, and computational models. In this context, spiking neural networks are well suited for low-latency processing of real-time data from event-based sensors, while Bayesian models provide a principled framework for sensor fusion across heterogeneous and noisy input streams. To support personalization and real-time adaptability, an on-chip learning module becomes an essential component.

V. Conclusion

Harnessing the intrinsic variability of resistive memories through Bayesian Compute In Memory enables uncertainty aware, trustworthy inference at low energy. Coupled with hybrid synapse architectures, new learning algorithms like meta learning, and metaplasticity will enable on-chip edge learning. Looking ahead, heterogeneous 3D integration and modular chiplets-based designs—mirroring biological specialization — provide the scalable, flexible foundation for future neuromorphic edge intelligence.

Acknowledgment

We acknowledge support from the European Research Council (consolidator grant DIVERSE: 101043854) and a France 2030 government grant (BEP: ANR 22 PEEL 0010).

References

[1] M. Horowitz "1.1 Computing's energy problem (and what we can do about it)", In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC).
[2] A. Sebastian *et al.*, "Memory devices and applications for in-memory computing". Nat. Nanotechnol. 15, 529–544 (2020).
[3] M. Lanza *et al.*, "The growing memristor industry" Nature 640, 613-622 (2025).
[4] T.-H. Wen *et al.*, "Fusion of memristor and digital compute-in-memory processing for energy-efficient edge computing". Science 384, 325–332 (2024).
[5] E. Esmanhotto *et al.*, "Experimental demonstration of multilevel resistive random access memory programming for up to two months stable neural networks inference accuracy". Advanced Intelligent Systems, 4(11):2200145 (2022).
[6] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence" Nature 521, 452–459 (2015).
[7] T. Dalgaty *et al.*, "In situ learning using intrinsic memristor variability via markov chain monte carlo sampling" Nat. Electron. 4, 151–161 (2021).
[8] D. Bonnet *et al.*, "Bringing uncertainty quantification to the extreme-edge with memristor-based bayesian neural networks" Nat. Commun. 14, 7530 (2023).
[9] Y. Lin *et al.*, "Uncertainty quantification via a memristor bayesian deep neural network for risk-sensitive reinforcement learning" Nat. Mach. Intell. 5, 714–723 (2023).
[10] A. Sebastian *et al.*, "Two-dimensional materials-based probabilistic synapses and reconfigurable neurons for measuring inference uncertainty using Bayesian neural networks" Nat. Commun. 13, 6139 (2022).
[11] D.-Q. You *et al.*, "14.1 a 22nm 104.5TOPS/w μ-NMC-Δ-IMC heterogeneous STT-MRAM CIM macro for noise-tolerant bayesian neural networks" In 2025 IEEE International Solid-State Circuits Conference (ISSCC).
[12] W.-S. Khwa *et al.*, "A mixed-precision memristor and SRAM compute-in-memory AI processor" Nature 639, 617–623 (2025).
[13] M. Martemucci *et al.*, "Hybrid FeRAM/RRAM Synaptic Circuit Enabling On-Chip Inference and Learning at the Edge" 2023 International Electron Devices Meeting (IEDM).
[14] T. Januel *et al.*, "Dual-Mode 16kb Memory: Transforming a Ferroelectric Capacitor Bitcell into Resistive Filamentary Memory" 2025 IEEE International Memory Workshop (IMW).
[15] S. Liu *et al.*, "Monolithic 3-D Integration of Diverse Memories: Resistive Switching (RRAM) and Gain Cell (GC) Memory Integrated on Si CMOS" IEEE Transactions on Electron Devices, 72, 5 (2025).
[16] C. Finn *et al.*, "Monolithic 3-D Integration of Diverse Memories: Resistive Switching (RRAM) and Gain Cell (GC) Memory Integrated on Si CMOS" In proceedings of the 34th International Conference on Machine Learning, PMLR 70:1126-1135 (2017).
[17] M. Pallo *et al.*, "On Chip Customized Learning on Resistive Memory Technology for Secure Edge AI " In proceedings of the Symposium on VLSI Technology and Circuits (2025).
[18] T. Ortner *et al.*, "Rapid learning with phase-change memory-based in-memory computing through learning-to-learn" Nat. Commun. 16, 1243 (2025).
[19] Z. Yu *et al.*, "Training-to-Learn with Memristive Devices" 2022 International Electron Devices Meeting (IEDM).
[20] S. Fusi *et al.*, "Cascade models of synaptically stored memories" Neuron. 45, 599–611 (2005).
[21] S. D'Agostino *et al.*, "Synaptic metaplasticity with multi-level memristive devices" In Proceedings IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (2023).
[22] A. Laborieux *et al.*, "Synaptic metaplasticity in binarized neural networks" Nat. Commun. 12, 2549 (2021).
[23] D. Bonnet *et al.*, "Bayesian continual learning and forgetting in neural networks" arXiv (2025).
[24] E. Vianello and M. Payvand, "Scaling neuromorphic systems with 3D technologies" Nat. Elec. 7, 419-421 (2024).
[25] P. Sterling P and S. Laughlin, 2015 Principles of Neural Design (MIT Press).