Energy-Efficient Bayesian Inference Using Near-Memory Computation with Memristors

C. Turck, K.-E. Harabi, T. Hirtzlin, E. Vianello, R. Laurent, Jacques Droulez, Pierre Bessiere, Marc Bocquet, Jean-Michel Portal, Damien Querlioz

Energy-Efficient Bayesian Inference Using Near-Memory Computation with Memristors

C. Turck*, K.-E. Harabi*, T. Hirtzlin†, E. Vianello†, R. Laurent‡,
J. Droulez‡, P. Bessiere§, M. Bocquet¶, J.-M. Portal¶, D. Querlioz*
*Univ. Paris-Saclay, CNRS, C2N, Palaiseau, France. †CEA, LETI, Grenoble, France. ‡Hawai.tech, Grenoble, France.
§Sorbonne Univ., CNRS, ISIR, Paris, France.¶Aix-Marseille Univ., CNRS, IM2NP, Marseille, France.

***Abstract*—Bayesian reasoning is a machine learning approach that provides explainable outputs and excels in small-data situations with high uncertainty. However, it requires intensive memory access and computation and is, therefore, too energy-intensive for extreme edge contexts. Near-memory computation with memristors (or RRAM) can greatly improve the energy efficiency of its computations. Here, we report two fabricated integrated circuits in a hybrid CMOS-memristor process, featuring each sixteen tiny memristor arrays and the associated near-memory logic for Bayesian inference. One circuit performs Bayesian inference using stochastic computing, and the other uses logarithmic computation; these two paradigms fit the area constraints of near-memory computing well. On-chip measurements show the viability of both approaches with respect to memristor imperfections. The two Bayesian machines also operated well at low supply voltages. We also designed scaled-up versions of the machines. Both scaled-up designs can perform a gesture recognition task using orders of magnitude less energy than a microcontroller unit. We also see that if an accuracy lower than 86.9% is sufficient for this sample task, stochastic computing consumes less energy than logarithmic computing; for higher accuracies, logarithmic computation is more energy-efficient. These results highlight the potential of memristor-based near-memory Bayesian computing, providing both accuracy and energy efficiency.**

***Index Terms*—memristor, ASIC, Bayesian inference.**

I. Introduction

Incorporating artificial intelligence (AI) in edge systems can be a game-changer for monitoring human health or the safety of buildings and industrial installations. AI algorithms often consume high energy when operated on conventional hardware, leading to privacy and security concerns as data is often uploaded to the cloud for processing. In-memory or near-memory computing approaches have been proposed to reduce AI's energy consumption by minimizing data movement. These approaches are particularly useful with memristors, an emerging non-volatile memory that can be embedded at the core of CMOS [1]. Until now, most of the research in this field has focused on neural networks. These algorithms excel in many situations; however, they struggle in cases where little data is available and provide non-explicable answers, limiting their use in safety-critical applications.

In our previous work, recently published in Nature Electronics [2], we presented a memristor-based Bayesian machine fabricated in a hybrid 130-nanometer CMOS/memristor process (Fig. 1a). This machine implements Bayesian inference, which does not require large training data and provides explainable results, using principles of stochastic computing. With minimal area requirements, stochastic computing is particularly adapted for near-memory computing; but it suffers from limited precision. In this Late News paper, we present a new fabricated integrated circuit (Fig. 1e), tested recently, where computation is made using another unconventional paradigm – logarithmic computation [3]. On-chip measurements show the viability of both approaches with respect to memristor imperfections. We also demonstrate that our Bayesian machines can operate at low supply voltages and that scaled-up versions can perform a gesture recognition task using orders of magnitude less energy than a microcontroller unit. We compare the energy efficiency of these two Bayesian chips, providing the first explicit comparison of stochastic and logarithmic computing in a near-memory computing IC with nanodevices.

This work was supported by the ERC Grant NANOINFER (715872).

II. Memristor-Based Bayesian Machines

Bayesian inference aims to evaluate the probability of an event Y based on a collection of observations $O_1, ..., O_n$ using Bayes' law [4] [5]. If all observations are conditionally independent, Bayes' law is simplified to multiplications between prior $p(Y)$) and likelihood factors $p(O_i|Y)$:

$$p(Y|O_1, .., O_n) \propto p(O_1|Y)... \times p(O_n|Y) \times p(Y). \quad (1)$$

The memristor-based Bayesian machines architecture for both stochastic and logarithmic designs (see Fig. 1b and Fig. 1f) is obtained by implementing equation 1 topologically, using near memory computing paradigm. Each likelihood factor is stored in an independent memory array using eight-bit integer representation, and computations are performed physically near these memory arrays. The computation result is then passed to the following memory array. The observations $O_1, ..., O_n$ act as addresses for the memory arrays, telling the likelihood value to be read. The concept of distributed near-memory computation allows the circuit to function with minimal energy consumption due to minimal data movement.

In the stochastic design, we use a pseudo-random number generator (LFSR) and a custom comparator to generate a stochastic bitstream corresponding to probability stored in likelihood memories; therefore, area-expensive multiplications are replaced by AND logic gates (see Fig. 1c). In the logarithmic design, the probabilities are stored in the memory arrays in the logarithmic domain; therefore, multiplications are replaced by integer additions (see Fig. 1g). We fabricated a fully-functional prototype circuit of a stochastic Bayesian machine (see Fig. 1a),

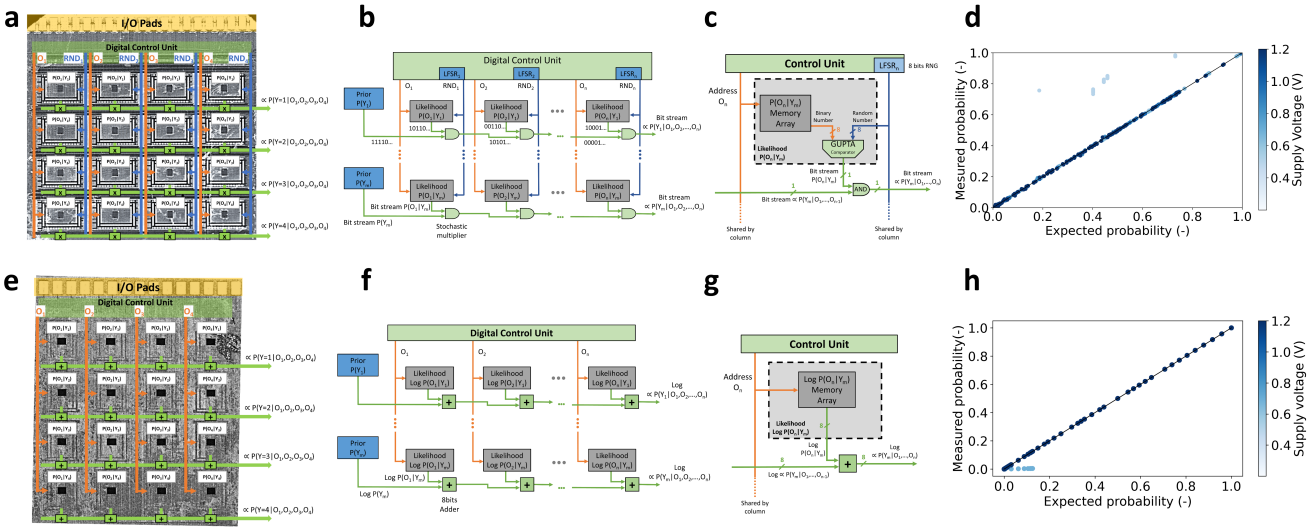

Fig. 1. **Fabricated and tested Memristor-based Bayesian Machines:** Optical microscopy photograph of the **(a)** stochastic and **(e)** logarithmic Bayesian Machine die. General architecture of the **(b)** stochastic and **(f)** logarithmic Bayesian machine. Schematic of Likelihood architecture used in the **(c)** stochastic and **(g)** logarithmic Bayesian Machine. Inference measurements (measured output as a function of expected result) on the fabricated **(d)** stochastic and **(h)** logarithmic Bayesian machine. In the logarithmic case, all points for supply voltages ranging from 0.7 to 1.2 V are superimposed.

incorporating 2,048 memristors and 30,080 transistors, and a logarithmic Bayesian machine (see Fig. 1e), with 2,048 memristors and 35,400 transistors, using a special low-power 130-nanometer CMOS process, where hafnium-oxide memristors are fabricated in place of vias between metal layers 4 and 5.

III. Measurements Results

Both designs were probe tested, using an STM32 microcontroller unit to send and receive the inputs and outputs signals. Figs. 1d,h show the measured output probability as a function of the expected one, for supply voltages ranging between 0.65 and 1.2 V. Measurements follow the ideal x=y curve for supply voltages up to 0.7 V, and start to show errors below. This result shows the potential of our IC for low-voltage operation. The memristors do not suffer from read disturb issues: after over 5 million whole-array consecutive readings with a 1.2 V supply voltage, no changes in the memory values were found.

IV. Energy Comparison Results

We also designed a scaled-up version of the Bayesian machine, with 6x4 arrays of 4,096 bits of memory, to perform a gesture recognition task [2]. Tab. I lists the accuracy on gesture recognition and energy consumption for different situations, obtained using the Cadence Voltus power integrity solution framework. Conventional stochastic computing refers to doing the number of cycles and then deciding based on the maximum number of ones; in the power-conscious mode, the computation is stopped when the first one is out. We see that this method achieves superior energy efficiency for an equivalent accuracy. Moreover, Tab. I shows that logarithmic computing performs better in energy consumption than both stochastic computing approaches, for accuracies higher than 86.9%. In the last column, we added the reading energy and can see that it is dominant for the computation in most cases. Stochastic computing energy could be reduced by using another type of random number generator: in our design, the consumption related to random number generation is 60% of the total consumption.

V. Conclusion

We have reported two fabricated integrated circuits in a hybrid CMOS-memristor process, enabling Bayesian inference with stochastic and logarithmic computation. Our results show that memristor-based near-memory Bayesian computing is a viable solution for energy-efficient machine learning systems. Stochastic computing is more energy efficient for lower-accuracy inference (up to 86.9% for our gesture recognition task) and has an inherent robustness to single-event upsets [2]. Logarithmic computing has lower latency (one cycle).

Architecture	Inf. cycles	Accuracy (%)	Energy (nJ) Inf.	Energy (nJ) Inf. & Read
Stoch Conv.	255	90.0	2.17	2.47
Stoch Conv.	50	86.7	0.43	0.73
Stoch Conv.	25	82.9	0.21	0.51
Stoch PC	255	86.9	0.10	0.40
Stoch PC	50	84.4	0.06	0.36
Stoch PC	20	80.2	0.04	0.34
Logarithmic	1	90.6	0.20	0.50

TABLE I
Comparison of the two Bayesian machines on the gesture recognition task. Conv: conventional stochastic computing. PC: power-conscious stochastic computing. Inf: inference

References

[1] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, no. 6, p. 333, 2018.

[2] K.-E. Harabi, T. Hirtzlin, C. Turck, E. Vianello, R. Laurent, J. Droulez, P. Bessière, J.-M. Portal, M. Bocquet, and D. Querlioz, "A memristor-based bayesian machine," *Nature Electronics, accepted. Available at arXiv:2112.10547.*, 2022.

[3] L. Sousa, "Nonconventional computer arithmetic circuits, systems and applications," *IEEE Circ. Syst. Magazine*, vol. 21, no. 1, pp. 6–40, 2021.

[4] E. T. Jaynes, *Probability theory*. Cambridge Univ. press, 2003.

[5] P. Bessière *et al.*, *Bayesian programming*. CRC press, 2013.