Reliability of Logic-in-Memory Circuits in Resistive Memory Arrays

Journal:	*Transactions on Electron Devices*
Manuscript ID	Draft
Manuscript Type:	ESSDERC2020
Date Submitted by the Author:	n/a
Complete List of Authors:	Zanotti, Tommaso; Università degli Studi di Modena e Reggio Emilia, Dip. di Ingegneria "Enzo Ferrari" Zambelli, Cristian; University of Ferrara, Engineering Puglisi, Francesco Maria; Universita' di Modena e Reggio Emilia, DIEF Milo, Valerio; Politecnico di Milano, Dipartimento di Elettronica e Informazione Pérez, Eduardo; IHP Microelectronics, Mahadevaiah, Mamathamba; Leibniz-Institut fur innovative Mikroelektronik Ossorio, Oscar; Universidad de Valladolid, Dpto. Electricidad y Electronica Wenger, Christian; IHP Microelectronics GmbH, Materials Research; BTU Cottbus-Senftenberg IKMZ, Pavan, Paolo; Universita di Modena e Reggio Emilia, Dipartimento di Ingegneria "Enzo Ferrari"; Olivo, Piero; University of Ferrara, Engineering Ielmini, Daniele; Politecnico di Milano, Dipartimento di Elettronica e Informazione;
Area of Expertise:	RRAM, Logic-in-Memory, BEOL, SIMPLY, Full adder

Reliability of Logic-in-Memory Circuits in Resistive Memory Arrays

Tommaso Zanotti, *Student Member, IEEE*, Cristian Zambelli, *Member, IEEE*, Francesco Maria Puglisi, *Member, IEEE*, Valerio Milo, *Member, IEEE*, Eduardo Pérez, Mamathamba K. Mahadevaiah, Oscar G. Ossorio, Christian Wenger, Paolo Pavan, *Senior Member, IEEE*, Piero Olivo, and Daniele Ielmini, *Fellow, IEEE*

***Abstract*— Logic-in-Memory (LiM) circuits based on RRAM devices and the material implication logic are promising candidates for the development of low-power computing devices, that could fulfill the growing demand of distributed computing systems. However, these circuits are affected by many reliability challenges that arise from device non-idealities (e.g., variability) and the characteristics of the employed circuit architecture. Thus, an accurate investigation of the variability at the array level is needed to evaluate the reliability and performance of such circuit architectures. In this work, we explore the reliability and performance of SIMPLY (i.e., a recently proposed LiM architecture with improved reliability and performance) on two 4 kbits RRAM arrays based on different resistive switching oxides integrated in the BEOL of the 0.25 μm BiCMOS process. We analyze the trade-off between reliability and energy consumption of SIMPLY architecture by exploiting the results of an extensive array-level variability characterization of the two technologies. Finally, we study the worst-case performance of a full adder implemented with the SIMPLY architecture and benchmark it on the analogous CMOS implementation.**

***Index Terms*— RRAM, BEOL, SIMPLY, Logic-in-Memory, Full adder.**

I. Introduction

Today, there are roughly 17 billion devices at the edge, which causes massive and ever-growing data exchange over communication networks. In this context, edge computing has been identified as a promising solution to relax data transfer and energy consumption limitations, providing advantages for Internet of Things (IoT) applications, smart cities and smart industries, Artificial Intelligence (AI), 5G/6G communications. However, todays ultra-low power hardware solutions are still affected by the von Neumann bottleneck (VNB) [1]–[3]. Specifically, VNB is the time- and energy-demanding process of data transfer between CPU and memory chips, and is the main showstopper for edge computing solutions. As recently suggested in [2], [4], [5], Logic-in-Memory (LiM) circuits that merge together data storage and computation could bypass VNB, thus minimizing the energy and time needed to execute logic functions. Among the most promising solutions, LiM circuits based on resistive random access memory (RRAM) devices and on the material implication logic offer significant advantages by leveraging on the small footprint of RRAMs, on their BEOL integration potential, and on the fact that implication logic is complete, i.e., all possible logic functions can be defined by a sequence of few core operations [2], [4]–[6], namely IMPLY and FALSE operations. However, the reliability of such operations and of the material implication logic circuit tightly depends on the non-idealities of the devices, especially variability [5], [6], and on the characteristics of the employed circuit architecture [5], [6]. Thus, evaluating the potential benefit of introducing the LiM paradigm in edge computing requires an accurate investigation of the variability at the array level. Nevertheless, a clear array-level analysis and demonstration of functionality of RRAM-based LiM solutions is still missing. In this work, we study the performance and feasibility of a recently proposed smart LiM paradigm (named SIMPLY [2], [5]) on two 4 kbits RRAM arrays with different resistive switching oxides integrated in the BEOL of the 0.25 μm BiCMOS process. Previous works [2], [5], restricted the evaluation of the performance of SIMPLY to RRAM technologies taken from the literature for which only a scarce amount of information regarding cycle-to-cycle (C2C) and device-to-device (D2D) variability is available. Here we exploit the extensive array-level variability characterization of the two RRAM technologies to study the performance of the SIMPLY architecture by evaluating the trade-off between reliability and energy consumption. In addition, we estimate the worst-case energy consumption of a 1-bit full adder (FA) implemented in the SIMPLY architecture, and benchmark it against CMOS implementations. Results show the SIMPLY implementation of the 1-bit FA outperforms the CMOS one by more than two orders of magnitude, when the VNB is considered, with significant improvement margins left. The paper is organized as follows: in Section II we introduce the main concepts underlying the material implication logic and the SIMPLY architecture; in Section III we show the results of the array-level variability characterization study; in Section IV we exploit the variability characterization results to discuss the reliability of the SIMPLY architecture; in Section V we benchmark the performance of a 1-bit FA in SIMPLY architecture against the corresponding CMOS implementation. Finally, we draw the conclusions in Section VI.

This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 648635) and in part by the German Research Foundation (DFG) in the frame of research group FOR2093.

T. Zanotti, F.M. Puglisi and P. Pavan are with Dipartimento di Ingegneria "Enzo Ferrari", Universitá di Modena e Reggio Emilia, 41125 Modena, Italy. (e-mail: tommaso.zanotti@unimore.it)

C. Zambelli and P. Olivo are with Dipartimento di Ingegneria, Universitá degli Studi di Ferrara, 44122 Ferrara, Italy.

V. Milo and D. Ielmini are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, 20133 Milano, Italy.

E. Perez, M. K. Mahadevaiah, and Ch. Wenger are with IHP-Leibniz-Institut für innovative Mikroelektronik, 15236 Frankfurt (Oder), Germany.

Ch. Wenger is also with BTU Cottbus-Senftenberg, 01968 Cottbus, Germany.

O. G. Ossorio is with Dpto. Electricidad y Electronica, Universidad de Valladolid, 47011 Valladolid, Spain.

II. SIMPLY Logic-in-Memory Architecture

The revived interest in RRAM technology arises from the possibility of storing and manipulating the information in the same place, by realizing LiM paradigms in which information is not stored as voltage at circuit nodes (like in CMOS logic) but as the resistance value of RRAM devices (with HRS = logic 0 and LRS = logic 1). Specifically, the paradigm based on the material implication logic is among the most effective since it is "complete", thus all the possible logic operations can be implemented as a sequence of two operations, namely the FALSE (i.e., the reset of a single device) and the IMPLY. In the typical arrangement, the IMPLY operation is executed by simultaneously pulsing two devices (P and Q, holding the input bits) with two different voltages (labeled V_{SET} and $V_{COND)}$ in such a way that P holds its state and Q changes state according

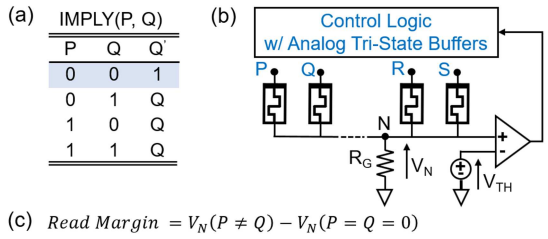

Fig. 1. (a) Truth table of the P IMPLY Q operation. The blue rectangle highlights that the state of Q changes (Q') only when the input combination is P=Q=0. (b) Schematic of SIMPLY architecture on a RRAM array. (c) Read margin (RM) definition considering devices not affected by variability.

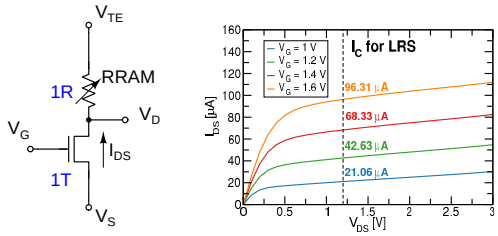

Fig. 2. 1T-1R cell's architecture (left) and I_{DS}-V_{DS} characteristics of the transistor exploited for I_C extraction at different V_G (right).

to the truth table in Fig.1a. However, as thoroughly analyzed in [4], [6], this arrangement suffers from several issues, such as high energy consumption, the degradation of the logic states stored in the devices, and a strong sensitivity to driving voltage variations [6]. Recently, the SIMPLY architecture was introduced to overcome the aforementioned issues [2], [5]. The SIMPLY architecture is sketched in Fig.1b, and can be easily implemented in a 1T-1R array by shunting together the bottom electrodes of a group of RRAM devices. The series transistor is appropriately biased to act as the resistor R_G. The IMPLY operation is performed by: *i)* applying a small V_{read} voltage pulse (200 mV in this work) to both P and Q [2], [5]; *ii)* comparing the voltage at node N (V_N in Fig.1b) against a threshold (V_{TH}) to determine if P=Q=0; *iii)* pulse V_{SET} on Q keeping the driver of P at high impedance only if P=Q=0. In principle, the condition P=Q=0 is easy to detect since V_N is lower in this case than in all other cases, ensuring a sufficient read margin (RM), defined as in Fig.1c. When considering ideal devices, RM is a deterministic quantity dependent on the memory window (i.e., the ratio of HRS to LRS resistance), V_{read}, and R_G. However, the combined effect of D2D and C2C variability, Random Telegraph Noise (RTN), driving voltage variations, and process tolerances results in a relatively wide distribution of RM, potentially impairing the functionality of the circuit. Therefore, the circuit reliability is tightly coupled to the intrinsic variability of the RRAM technology exploited in its integration, and its statistical characterization allows verifying the reliability level that can be achieved by the proposed LiM circuit when implemented in the RRAM technologies under study.

III. Variability Characterization on RRAM Arrays

To statistically assess both the D2D and the C2C variability, we performed electrical characterization measurements on the 4 kbits 1T-1R arrays whose architecture is described in [7]. We remind that the array is based on the 0.25 μm BiCMOS process from IHP and that the select transistor in the 1T-1R cells is an n-MOS with gate width W = 1.14 μm and gate length L = 0.24 μm. The transistor allows modulating the compliance current I_C by tuning V_G during operations, thus enabling a tight control of the cell conductance and enhanced power-control features in LiM circuits. Fig. 2 shows the I_{DS}-V_{DS} characteristics of a transistor in the array exposing the different I_C. The memristive element is connected in series with the drain of the select transistor and is integrated during BEOL on top of the second metal level (M2) featuring a 600 × 600 nm^2 area. To provide an even more solid exploration of the RRAM technology impact on the LiM circuit, we considered two different memristive stacks integrated in separated arrays, namely a TiN/Ti/HfO$_x$/TiN and a TiN/Ti/Hf$_{1-x}$Al$_x$O$_y$/TiN structure. The process characteristics of each stack can be retrieved in [8].

Resistive switching of the memristive element is enabled for all the cells in the array through a Forming operation which consists of the application of the ISPVA algorithm for best yield [9] with duration t_p = 1 μs and a top electrode voltage V_{TE} from 2 V to 5 V in steps of 10 mV. Reset operation is then performed to reach HRS by using a single pulse gate voltage V_G = 2.8 V to minimize transistor resistance and a source voltage V_S = 1.8 V. Same pulse duration is considered. We would like to point that such t_p is chosen to simplify the requirements of the measurement setup, although the functionality of the two RRAM technologies was also proven with t_p = 50 ns [10] (see Fig. 3a). The result of the Reset operation on the 4 kbits array allows extracting the HRS bound dictated by the chosen RRAM technology for the LiM target application. As shown in Fig. 3b, we consider 40 kΩ for both memristive stacks. It is worth to mention that a higher HRS resistance results in a lower power consumption of LiM circuit, so we consider the former value as a worst-case condition that allows speculating on the performance and reliability limits. Concerning the Set operation, we used a V_{TE} = 1.2 V and a single pulse duration t_p = 1 μs associated with four different V_G values from 1 V to 1.6 V in 200 mV steps. This allows a tuning of the LRS on four levels (L_1 to L_4) devising the I_C set by the transistor, while providing a strategy for power consumption reduction policies to be applied on LiM circuits. The LRS read currents (I_{read}) measured with a V_{TE} = 200 mV for L_1 - L_4 correspond approximately to 10 μA, 20 μA, 30 μA, and 40 μA.

Different LRS levels however come with different variability characteristics. Fig. 4 shows a characterization study of the variability for levels L_1 - L_4 of both RRAM technologies. In the figure, the standard deviation σ_R for the C2C and D2D distributions is plotted as a function of the median device resistance indicated as R. Variability data were collected for a subset of 1024 1T-1R devices (i.e., a block of 16 wordlines set with the same LRS level) integrated in the 4 kbits test vehicles and for 1000 consecutive Set/Reset cycles. The cycling routine is performed considering the proper V_G for the Set operation in each subset. As it can be seen, the scatter of the C2C variability dominates over the D2D for all LRS levels following the universal trend for σ_R which is proportional to R^2, as indicated in other studies in literature [11], [12]. However, the D2D variability is the one with the highest σ_R absolute values so that we speculate to be critical for the performance of the LiM circuits like those in this work. Overall, Hf$_{1-x}$Al$_x$O$_y$ material displays a better control of the D2D and C2C variability as demonstrated by the lower scatter of the $< \sigma_R; R >$ points in the plots. LRS level L_1 has however a higher C2C σ_R compared to that of HfO$_x$. While this does not represent a limitation for multi-level storage applications [13], it may pose a limitation in LiM circuits functionality.

IV. Reliability of SIMPLY in RRAM Arrays

The results of the extensive variability characterization in Section III are now exploited to verify the reliability and the energy

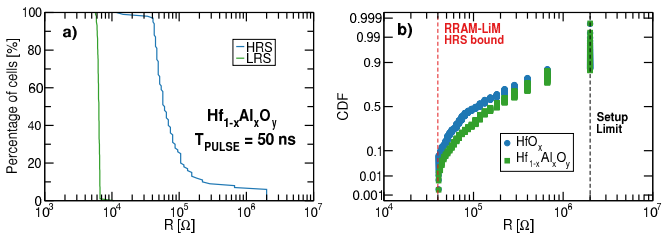

Fig. 3. (a) Demonstration of the functionality of an RRAM array based on $Hf_{1-x}Al_xO_y$ in this work when t_P = 50 ns (distinct HRS and LRS can be achieved). Similar results can be obtained for HfO_x arrays. Adapted from [10]. (b) Cumulative Distribution Function of the HRS extracted from the 4 kbits array evidencing the bound for LiM application at 40 kΩ.

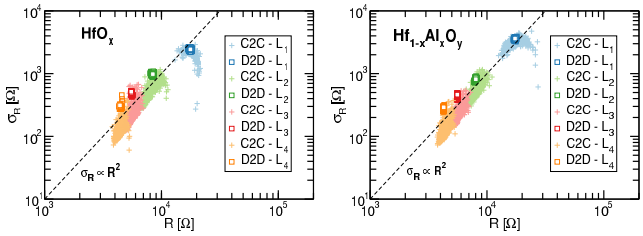

Fig. 4. Standard deviation σ_R of the resistance as a function of the median resistance R for HfO_x (left) and $Hf_{1-x}Al_xO_y$ RRAM technologies (right), respectively.

efficiency of the SIMPLY architecture. In this respect, note that: *i)* the only requirement for a reliable circuit operation of the SIMPLY architecture is that a sufficient RM is available at the input of the comparator in all cases, even in the presence of variability; *ii)* the variability characterization in Section III is performed at the array level, natively including contributions from the device (C2C, D2D and RTN) and from the non-idealities of the peripheral circuits (e.g., possible variations of V_{read}). This makes RM a comprehensive metric of the circuit reliability. Indeed, circuit implementations that result in higher RMs are more reliable and allow using simpler comparator or sense amplifier designs, though more complex designs can be used for smaller RMs. To estimate the RM and compare the performance of the two RRAM technologies, we used the C2C and D2D joint variability data to compute the distributions of V_N when P=Q=0 and when $P \neq Q$ for both the technologies and each LRS level (see Fig.5). The joint probability distribution of LRS for each technology and I_{read} was estimated by combining together 100 random samples for each $< \sigma_R; R >$ pair of Fig.4. To identify the worst-case RM that allows evaluating the performance and reliability limits, we assume HRS fixed at the worst-case of the HRS distribution (R_{HRS} = 40 kΩ for both technologies, as shown in Fig.3) and thus $V_N(P$=Q=0) is constant for each technology and target LRS level and determined by the value of R_G (see Fig.1a). Specifically, the optimal value of R_G that maximizes the RM for each technology and LRS level was chosen using (1):

$$R_G = \sqrt{\left(R_{LRS,MAX}^{-1} + R_{HRS,MAX}^{-1}\right)^{-1} \cdot \frac{R_{HRS,MIN}}{2}} \quad (1)$$

where $R_{LRS,MAX}$ is the μ+3σ value of the joint LRS distribution, and $R_{HRS,MIN}$ and $R_{HRS,MAX}$ are the $\mu \pm 3\sigma$ values of the HRS distribution, respectively ($R_{HRS,MIN}$ = $R_{HRS,MAX}$ = 40 kΩ in this case). As shown in Figs. 5 and 6a, RM grows with I_{read} and the two RRAM technologies show comparable reliability at the same LRS level, with a notable exception at I_{read} = 10 μA where HfO_x devices guarantee a quite larger RM than $Hf_{1-x}Al_xO_y$ devices. This stems from the lower C2C variability of HfO_x devices at low I_{read}, which leads to a smaller tail of L_1 as shown in Fig.4. To compare the reliability metric in Fig.6a to the energy efficiency performance, we report in Fig.6b the worst-case energy consumption of the read step of SIMPLY operation (averaged over the four possible input configurations) at different I_{read}. As expected, the energy per operation increases with I_{read}, which establishes a trade-off between energy efficiency and reliability. No relevant differences between the two technologies are observed.

TABLE I
1-BIT FULL ADDER ENERGY PER OPERATION (INDICATING MIN - MAX RANGE) VS I_{read} WITH t_p = 1 μs

I_{read}	Read	Write	FA
10 μA	(2.4 - 2.6) · 10^{-11} J	(5.1 - 5.6) · 10^{-10} J	(5.3 - 5.8) · 10^{-10} J
20 μA	(3.2 - 3.5) · 10^{-11} J	(1.0 - 1.1) · 10^{-9} J	(1.1 - 1.2) · 10^{-9} J
30 μA	(3.9 - 4.2) · 10^{-11} J	(1.7 - 1.8) · 10^{-9} J	(1.7 - 1.9) · 10^{-9} J
40 μA	(4.4 - 4.8) · 10^{-11} J	(2.3 - 2.5) · 10^{-9} J	(2.4 - 2.6) · 10^{-9} J

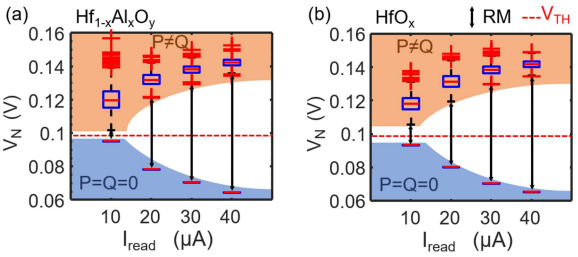

Fig. 5. (a)-(b) Distribution of V_N due to C2C and D2D variability, when P=Q=0 (blue area) and $P \neq Q$ (orange area) for the SIMPLY operation for different I_{read} and RRAM technology. Only the worst-case (i.e., lowest resistance value due to variability) HRS resistance is considered. The read margins (RM black arrows) and the threshold voltages (V_{TH} dashed red) for the comparator are evidenced. Black whiskers indicate the extreme points of the distributions. Red crosses indicate outliers. V_N when P=Q=1 is always much higher than in all other cases (thus is not reported in these box plots).

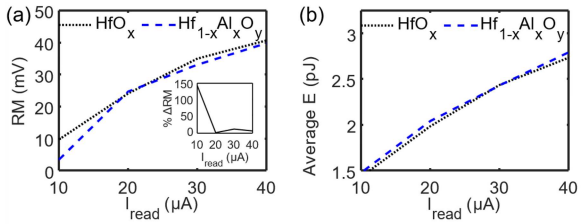

Fig. 6. (a) Worst-case RM (-3σ deviation from the mean of the RM distribution due to variability) for the two different RRAM technologies and different I_{read} values. The inset shows the relative difference between the RM obtained with HfO_x and $Hf_{1-x}Al_xO_y$ devices. (b) Worst-case energy consumption of the read operation during SIMPLY operation for HfO_x and $Hf_{1-x}Al_xO_y$ RRAM technologies at different I_{read} values. It is computed as the average of the worst-case energy consumption (+3σ deviation from the mean of the energy distribution due to variability) of each input configuration.

V. LiM FULL ADDER IMPLEMENTATION

To benchmark the performance of the SIMPLY architecture on more complex operations, we designed a 1-bit ripple carry full adder (FA) with the two RRAM technologies and estimated the worst-case performance. The FA was implemented using the SIMPLY archi-

TABLE II
COMPARISON BETWEEN THE PROPOSED RRAM-BASED FA AND A CMOS-BASED FA WHEN EXECUTING 32 PARALLEL 32-BIT FA OPERATIONS (ON A MEMORY ARRAY INTEGRATING 4096 DEVICES)

	# Devices	Energy	Delay	Energy-Delay Product (EDP)	EDP Normalized to CMOS w/ VNB	Improvement w.r.t. to CMOS w/ VNB
CMOS w/ VNB*	8192 - 28672 FET	$\approx$ 9.4 μJ	$\approx$ 284 μs	$\approx 2.7 \cdot 10^{-9}$ J·s	1	1
CMOS w/o VNB**	8192 - 28672 FET	$\approx 9.7 \cdot 10^{-4}$ - 7.4 pJ	$\approx 5.6 \cdot 10^{-2}$ - 4.8 μs	$\approx 1.7 \cdot 10^{-24}$ - $3.6 \cdot 10^{-17}$ J·s	$6.3 \cdot 10^{-16}$ - $1.3 \cdot 10^{-8}$	$7.5 \cdot 10^{7}$ - $1.6 \cdot 10^{15}$
This work I_{read} **= 10** μA t_P **= 1** μ**s**	**3232 RRAM**	$\approx$ **594 nJ**	$\approx$ **2.9 ms**	$\approx$ **1.7** $\cdot$ 10^{-9} **J·s**	**0.6**	**1.57**
This work I_{read} **= 10** μA t_P **= 50 ns**	**3232 RRAM**	$\approx$ **30 nJ**	$\approx$ **147** μ**s**	$\approx$ **4.4** $\cdot$ 10^{-12} **J·s**	**1.6·10**$^{-3}$	**607**

*,** estimates with (w/) and without (w/o) VNB are performed considering energy and delay overhead for reading one memory page of 4kB data from a NAND flash memory [14]. CMOS FA performances were estimated projecting the time and energies for different 1-bit FA schemes taken from [3], [15], [16] where 0.18 μm, 45 nm, and 10 nm CMOS technology are used.

tecture with 8 RRAM devices performing the sequence of operations reported in [2], which includes 28 steps of core operations (18 IMPLY and 10 FALSE). During a FA cycle, the energy consumption will also depend on the configuration of input bits, since it dictates the number of reset (during FALSE), set (during IMPLY when P=Q=0), and read (during IMPLY in all cases) operations that are executed. Here we consider the worst-case energy consumption during set and reset operations $E = V_P \cdot I_C \cdot t_P$, where V_P and t_P are the applied voltage magnitude and width. We consider the worst-case energy for each input combination also for the read operation. The minimum and maximum worst-case energies per FA cycle are proportional to I_C as reported in Table I. Values on the order of nJ are obtained for a t_P of 1 μs as the one used in Section III ($t_{IMPLY} = 4 \cdot t_P$, $t_{FALSE} = 2 \cdot t_P$, $t_{FA} = 10 \cdot t_{FALSE} + 18 \cdot t_{IMPLY}$).

To show the advantages offered by the proposed LiM scheme in terms of energy efficiency, we compare the performance of the proposed architecture against CMOS FA implementations from [3], [15], [16] with and without considering the VNB [14]). The VNB overhead was computed considering the time and energy required to read a typical flash memory page of 4kB [14]. Although the writing of the results to the memory should also be included, we consider only the overhead of the memory read to take into account the possibility of using smaller page sizes. We estimate the delay and energy required to compute 32 parallel 32-bits FA operations (simple ripple carry) which require slightly less (3132) than the available 4096 devices in the array. To show the potential energy efficiency improvement over CMOS we consider the case I_{read} = 10 μA. As shown in Table II, the largest share of energy consumption and delay for CMOS logic comes from the VNB data exchange overhead [3], [14]–[16]. With the technologies explored in this work, when t_P = 1 μs (i.e., the pulse duration used in the characterization phase in Section III) the energy delay product (EDP) of SIMPLY is only slightly better than its CMOS counterpart when the VNB effect is included. However, the functionality of the RRAM devices considered in this work was proven also with t_P = 50 ns (see Fig.3a), and energy projections with such a t_P show that the proposed LiM scheme outperforms CMOS (when including the VNB overhead) by more than two orders of magnitude (worst-case projection) in energy efficiency with similar computing time, as shown in Table II. In addition, the number of required devices is reduced as compared to CMOS, thus achieving higher integration density.

VI. CONCLUSIONS

In this work, we studied the reliability and performance of SIMPLY architecture on two different 4 kbits RRAM arrays integrated in the BEOL of the 0.25 μm BiCMOS process. We highlighted and evaluated the trade-off between circuit reliability and energy consumption by exploiting the extensive array-level variability characterization of the two technologies. Furthermore, we analyzed the performance of a 1-bit FA implemented on SIMPLY. Even when considering the worst-case, the proposed architecture is $\approx$ 600 times more efficient than the CMOS counterpart including the VNB, while also achieving higher integration density. These results suggest that the proposed solution is a viable technology for the development of ultra-low power computing devices for IoT applications.

REFERENCES

[1] J. Backus, "Can Programming Be Liberated from the Von Neumann Style?: A Functional Style and Its Algebra of Programs," *Commun. ACM*, vol. 21, no. 8, pp. 613–641, 1978, doi: 10.1145/359576.359579.

[2] T. Zanotti, F. M. Puglisi, and P. Pavan, "A Smart Logic-in-Memory Architecture for Low-Power non-von Neumann Computing," *IEEE Journal of the Electron Devices Society*, 2020, (in press), doi: 10.1109/JEDS.2020.2987402.

[3] M. Aguirre-Hernandez and M. Linares-Aranda, "CMOS Full-Adders for Energy-Efficient Arithmetic Applications," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 4, pp. 718–721, 2011, doi: 10.1109/TVLSI.2009.2038166.

[4] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 10, pp. 2054–2066, 2014, doi: 10.1109/TVLSI.2013.2282132.

[5] F. M. Puglisi, T. Zanotti, and P. Pavan, "SIMPLY: Design of a RRAM-Based Smart Logic-in-Memory Architecture using RRAM Compact Model," in *ESSDERC*, 2019, pp. 130–133, doi: 10.1109/ESSDERC.2019.8901731.

[6] T. Zanotti, F. M. Puglisi, and P. Pavan, "Reliability-aware design strategies for stateful logic-in-memory architectures," *IEEE Trans. on Device and Materials Reliability*, vol. 20, no. 2, pp. 278–285, 2020, doi: 10.1109/TDMR.2020.2981205.

[7] A. Grossi, D. Walczyk, C. Zambelli, E. Miranda, P. Olivo, V. Stikanov, A. Feriani, J. Su, G. Schoof, R. Kraemer, B. Tillack, A. Fox, T. Schroeder, C. Wenger, and C. Walczyk, "Impact of Intercell and Intracell Variability on Forming and Switching Parameters in RRAM Arrays," *IEEE Trans. on Electron Devices*, vol. 62, no. 8, pp. 2502–2509, 2015, doi: 10.1109/TED.2015.2442412.

[8] V. Milo, C. Zambelli, P. Olivo, E. Perez, M. K. Mahadevaiah, O. G. Ossorio, C. Wenger, and D. Ielmini, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Materials*, vol. 7, no. 8, p. 081120, 2019, doi: 10.1063/1.5108650.

[9] A. Grossi, C. Zambelli, P. Olivo, E. Miranda, V. Stikanov, C. Walczyk, and C. Wenger, "Electrical characterization and modeling of pulse-based forming techniques in RRAM arrays," *Solid-State Electronics*, vol. 115, pp. 17 – 25, 2016, doi: 10.1016/j.sse.2015.10.003.

[10] E. Perez, O. Gonzlez Ossorio, S. Duenas, H. Castan, H. Garcia, and C. Wenger, "Programming Pulse Width Assessment for Reliable and Low-Energy Endurance Performance in Al:HfO_2-Based RRAM Arrays," *Electronics*, vol. 9, p. 864, 2020, doi: 10.3390/electronics9050864.

[11] A. Fantini, L. Goux, R. Degraeve, D. J. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y. . Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO_2 RRAM," in *IEEE IMW*, 2013, pp. 30–33, doi: 10.1109/IMW.2013.6582090.

[12] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability," *IEEE Trans. on Electron Devices*, vol. 61, no. 8, pp. 2912–2919, 2014, doi: 10.1109/TED.2014.2330200.

[13] E. Perez, A. Grossi, C. Zambelli, P. Olivo, R. Roelofs, and C. Wenger, "Reduction of the Cell-to-Cell Variability in $Hf_{1-x}Al_xO_y$ Based RRAM Arrays by Using Program Algorithms," *IEEE Electron Device Letters*, vol. 38, no. 2, pp. 175–178, 2017, doi: 10.1109/LED.2016.2646758.

[14] S.-Y. Park, D. Jung, J.-U. Kang, J.-S. Kim, and J. Lee, "CFLRU: A Replacement Algorithm for Flash Memory," in *Proc. of the 2006 Int. Conf. on Compilers, Architecture and Synthesis for Embedded Systems*. ACM, 2006, pp. 234–241, doi: 10.1145/1176760.1176789.

[15] A. K. Yadav, B. P. Shrivatava, and A. K. Dadoriya, "Low power high speed 1-bit full adder circuit design at 45nm CMOS technology," in *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*, 2017, pp. 427–432, doi. 10.1109/RISE.2017.8378203.

[16] S. Sharma and G. Soni, "Comparision analysis of FinFET based 1-bit full adder cell implemented using different logic styles at 10, 22 and 32NM," in *2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS)*, 2016, pp. 660–667, doi. 10.1109/ICEETS.2016.7583835.

Reliability of Logic-in-Memory Circuits in Resistive Memory Arrays

Tommaso Zanotti, *Student Member, IEEE*, Cristian Zambelli, *Member, IEEE*, Francesco Maria Puglisi, *Member, IEEE*, Valerio Milo, *Member, IEEE*, Eduardo Pérez, Mamathamba K. Mahadevaiah, Oscar G. Ossorio, Christian Wenger, Paolo Pavan, *Senior Member, IEEE*, Piero Olivo, and Daniele Ielmini, *Fellow, IEEE*

***Abstract*— Logic-in-Memory (LiM) circuits based on RRAM devices and the material implication logic are promising candidates for the development of low-power computing devices, that could fulfill the growing demand of distributed computing systems. However, these circuits are affected by many reliability challenges that arise from device non-idealities (e.g., variability) and the characteristics of the employed circuit architecture. Thus, an accurate investigation of the variability at the array level is needed to evaluate the reliability and performance of such circuit architectures. In this work, we explore the reliability and performance of SIMPLY (i.e., a recently proposed LiM architecture with improved reliability and performance) on two 4 kbits RRAM arrays based on different resistive switching oxides integrated in the BEOL of the 0.25 μm BiCMOS process. We analyze the trade-off between reliability and energy consumption of SIMPLY architecture by exploiting the results of an extensive array-level variability characterization of the two technologies. Finally, we study the worst-case performance of a full adder implemented with the SIMPLY architecture and benchmark it on the analogous CMOS implementation.**

***Index Terms*— RRAM, BEOL, SIMPLY, Logic-in-Memory, Full adder.**

I. Introduction

Today, there are roughly 17 billion devices at the edge, which causes massive and ever-growing data exchange over communication networks. In this context, edge computing has been identified as a promising solution to relax data transfer and energy consumption limitations, providing advantages for Internet of Things (IoT) applications, smart cities and smart industries, Artificial Intelligence (AI), 5G/6G communications. However, todays ultra-low power hardware solutions are still affected by the von Neumann bottleneck (VNB) [1]–[3]. Specifically, VNB is the time- and energy-demanding process of data transfer between CPU and memory chips, and is the main showstopper for edge computing solutions. As recently suggested in [2], [4], [5], Logic-in-Memory (LiM) circuits that merge together data storage and computation could bypass VNB, thus minimizing the energy and time needed to execute logic functions. Among the most promising solutions, LiM circuits based on resistive random access memory (RRAM) devices and on the material implication logic offer significant advantages by leveraging on the small footprint of RRAMs, on their BEOL integration potential, and on the fact that implication logic is complete, i.e., all possible logic functions can be defined by a sequence of few core operations [2], [4]–[6], namely IMPLY and FALSE operations. However, the reliability of such operations and of the material implication logic circuit tightly depends on the non-idealities of the devices, especially variability [5], [6], and on the characteristics of the employed circuit architecture [5], [6]. Thus, evaluating the potential benefit of introducing the LiM paradigm in edge computing requires an accurate investigation of the variability at the array level. **Nevertheless, a clear array-level analysis and demonstration of functionality of RRAM-based LiM solutions is still missing.** In this work, we study the performance and feasibility of a recently proposed smart LiM paradigm (named SIMPLY [2], [5]) on two 4 kbits RRAM arrays with different resistive switching oxides integrated in the BEOL of the 0.25 μm BiCMOS process. **Previous works** [2], [5]**, restricted the evaluation of the performance of SIMPLY to RRAM technologies taken from the literature for which only a scarce amount of information regarding cycle-to-cycle (C2C) and device-to-device (D2D) variability is available. Here we exploit** the extensive array-level variability characterization of the two RRAM technologies ~~is exploited~~ to study the performance of the SIMPLY architecture by evaluating the trade-off between reliability and energy consumption. In addition, we estimate the worst-case energy consumption of a 1-bit full adder (FA) implemented in the SIMPLY architecture, and benchmark it against CMOS implementations. Results show the SIMPLY implementation of the 1-bit FA outperforms the CMOS one by more than two orders of magnitude, when the VNB is considered, with significant improvement margins left. The paper is organized as follows: in Section II we introduce the main concepts underlying the material implication logic and the SIMPLY architecture; in Section III we show the results of the array-level variability characterization study; in Section IV we exploit the variability characterization results to discuss the reliability of the SIMPLY architecture; in Section V we benchmark the performance of a 1-bit FA in SIMPLY architecture against the corresponding CMOS implementation. Finally, we draw the conclusions in Section VI.

II. SIMPLY Logic-in-Memory Architecture

The revived interest in RRAM technology arises from the possibility of storing and manipulating the information in the same place, by realizing LiM paradigms in which information is not stored as voltage at circuit nodes (like in CMOS logic) but as the resistance value of RRAM devices (with HRS = logic 0 and LRS = logic 1). Specifically, the paradigm based on the material implication logic is among the most effective since it is "complete", thus all the possible logic operations can be implemented as a sequence of two operations, namely the FALSE (i.e., the reset of a single device) and the IMPLY. In the typical arrangement, the IMPLY operation is executed by simultaneously pulsing two devices (P and Q, holding the input bits) with two different voltages (labeled V_{SET} and $V_{COND)}$ in

This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 648635) and in part by the German Research Foundation (DFG) in the frame of research group FOR2093.

T. Zanotti, F.M. Puglisi and P. Pavan are with Dipartimento di Ingegneria "Enzo Ferrari", Universitá di Modena e Reggio Emilia, 41125 Modena, Italy. (e-mail: tommaso.zanotti@unimore.it)

C. Zambelli and P. Olivo are with Dipartimento di Ingegneria, Universitá degli Studi di Ferrara, 44122 Ferrara, Italy.

V. Milo and D. Ielmini are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, 20133 Milano, Italy.

E. Perez, M. K. Mahadevaiah, and Ch. Wenger are with IHP-Leibniz-Institut für innovative Mikroelektronik, 15236 Frankfurt (Oder), Germany.

Ch. Wenger is also with BTU Cottbus-Senftenberg, 01968 Cottbus, Germany.

O. G. Ossorio is with Dpto. Electricidad y Electronica, Universidad de Valladolid, 47011 Valladolid, Spain.

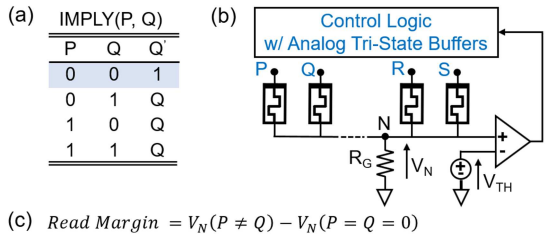

Fig. 1. (a) Truth table of the P IMPLY Q operation. The blue rectangle highlights that the state of Q changes (Q') only when the input combination is P=Q=0. (b) Schematic of SIMPLY architecture on a RRAM array. (c) Read margin (RM) definition considering devices not affected by variability.

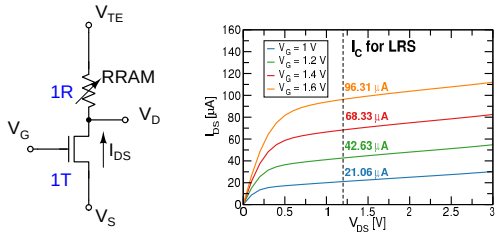

Fig. 2. 1T-1R cell's architecture (left) and I_{DS}-V_{DS} characteristics of the transistor exploited for I_C extraction at different V_G (right).

such a way that P holds its state and Q changes state according to the truth table in Fig.1a. However, as thoroughly analyzed in [4], [6], this arrangement suffers from several issues, such as high energy consumption, the degradation of the logic states stored in the devices, and a strong sensitivity to driving voltage variations [6]. Recently, the SIMPLY architecture was introduced to overcome the aforementioned issues [2], [5]. The SIMPLY architecture is sketched in Fig.1b, and can be easily implemented in a 1T-1R array by shunting together the bottom electrodes of a group of RRAM devices. The series transistor is appropriately biased to act as the resistor R_G. The IMPLY operation is performed by: *i)* applying a small V_{read} voltage pulse (200 mV in this work) to both P and Q [2], [5]; *ii)* comparing the voltage at node N (V_N in Fig.1b) against a threshold (V_{TH}) to determine if P=Q=0; *iii)* pulse V_{SET} on Q keeping the driver of P at high impedance only if P=Q=0. In principle, the condition P=Q=0 is easy to detect since V_N is lower in this case than in all other cases, ensuring a sufficient read margin (RM), defined as in Fig.1c. When considering ideal devices, RM is a deterministic quantity dependent on the memory window (i.e., the ratio of HRS to LRS resistance), V_{read}, and R_G. However, the combined effect of ~~device-to-device (D2D) and cycle-to-cycle (C2C)~~ variability, Random Telegraph Noise (RTN), driving voltage variations, and process tolerances results in a relatively wide distribution of RM, potentially impairing the functionality of the circuit. Therefore, the circuit reliability is tightly coupled to the intrinsic variability of the RRAM technology exploited in its integration, and its statistical characterization allows verifying the reliability level that can be achieved by the proposed LiM circuit when implemented in the RRAM technologies under study.

III. Variability Characterization on RRAM Arrays

To statistically assess both the D2D and the C2C variability, we performed electrical characterization measurements on the 4 kbits 1T-1R arrays whose architecture is described in [7]. We remind that the array is based on the 0.25 μm BiCMOS process from IHP and that the select transistor in the 1T-1R cells is an n-MOS with gate width W = 1.14 μm and gate length L = 0.24 μm. The transistor allows modulating the compliance current I_C by tuning V_G during operations, thus enabling a tight control of the cell conductance and enhanced power-control features in LiM circuits. Fig. 2 shows the I_{DS}-V_{DS} characteristics of a transistor in the array exposing the different I_C. The memristive element is connected in series with the drain of the select transistor and is integrated during BEOL on top of the second metal level (M2) featuring a 600 $\times$ 600 nm^2 area. To provide an even more solid exploration of the RRAM technology impact on the LiM circuit, we considered two different memristive stacks integrated in separated arrays, namely a TiN/Ti/HfO$_x$/TiN and a TiN/Ti/Hf$_{1-x}$Al$_x$O$_y$/TiN structure. The process characteristics of each stack can be retrieved in [8].

Resistive switching of the memristive element is enabled for all the cells in the array through a Forming operation which consists of the application of the ISPVA algorithm for best yield [9] with duration t_p = 1 μs and a top electrode voltage V_{TE} from 2 V to 5 V in steps of 10 mV. Reset operation is then performed to reach HRS by using a single pulse gate voltage V_G = 2.8 V to minimize transistor resistance and a source voltage V_S = 1.8 V. Same pulse duration is considered. We would like to point that such t_p is chosen to simplify the requirements of the measurement setup, although the functionality of the two RRAM technologies was also proven with t_p = 50 ns [10] (see Fig. 3a). The result of the Reset operation on the 4 kbits array allows extracting the HRS bound dictated by the chosen RRAM technology for the LiM target application. As shown in Fig. 3b, we consider 40 kΩ for both memristive stacks. It is worth to mention that a higher HRS resistance results in a lower power consumption of LiM circuit, so we consider the former value as a worst-case condition that allows speculating on the performance and reliability limits. Concerning the Set operation, we used a V_{TE} = 1.2 V and a single pulse duration t_p = 1 μs associated with four different V_G values from 1 V to 1.6 V in 200 mV steps. This allows a tuning of the LRS on four levels (L_1 to L_4) devising the I_C set by the transistor, while providing a strategy for power consumption reduction policies to be applied on LiM circuits. The LRS read currents (I_{read}) measured with a V_{TE} = 200 mV for L_1 - L_4 correspond approximately to 10 μA, 20 μA, 30 μA, and 40 μA.

Different LRS levels however come with different variability characteristics. Fig. 4 shows a characterization study of the variability for levels L_1 - L_4 of both RRAM technologies. In the figure, the standard deviation σ_R for the C2C and D2D distributions is plotted as a function of the median device resistance indicated as R. Variability data were collected for a subset of 1024 1T-1R devices (i.e., a block of 16 wordlines set with the same LRS level) integrated in the 4 kbits test vehicles and for 1000 consecutive Set/Reset cycles. The cycling routine is performed considering the proper V_G for the Set operation in each subset. As it can be seen, the scatter of the C2C variability dominates over the D2D for all LRS levels following the universal trend for σ_R which is proportional to R^2, as indicated in other studies in literature [11], [12]. However, the D2D variability is the one with the highest σ_R absolute values so that we speculate to be critical for the performance of the LiM circuits like those in this work. Overall, Hf$_{1-x}$Al$_x$O$_y$ material displays a better control of the D2D and C2C variability as demonstrated by the lower scatter of the $< \sigma_R; R >$ points in the plots. LRS level L_1 has however a higher C2C σ_R compared to that of HfO$_x$. While this does not represent a limitation for multi-level storage applications [13], it may pose a limitation in LiM circuits functionality.

IV. Reliability of SIMPLY in RRAM Arrays

The results of the extensive variability characterization in Section III are now exploited to verify the reliability and the energy

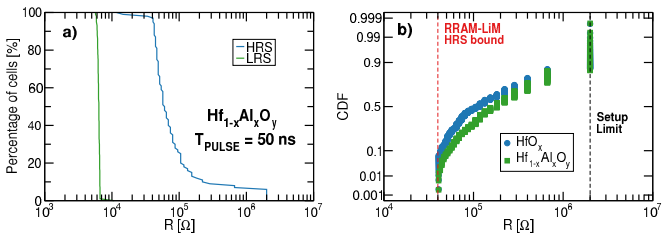

Fig. 3. (a) Demonstration of the functionality of an RRAM array based on $Hf_{1-x}Al_xO_y$ in this work when t_P = 50 ns (distinct HRS and LRS can be achieved). Similar results can be obtained for HfO_x arrays. Adapted from [10]. (b) Cumulative Distribution Function of the HRS extracted from the 4 kbits array evidencing the bound for LiM application at 40 kΩ.

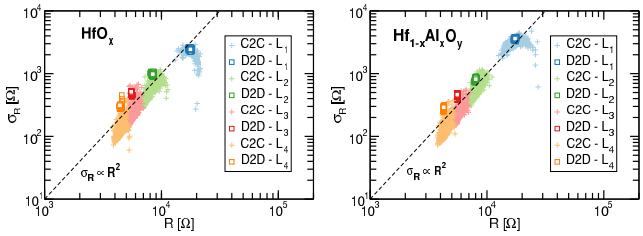

Fig. 4. Standard deviation σ_R of the resistance as a function of the median resistance R for HfO_x (left) and $Hf_{1-x}Al_xO_y$ RRAM technologies (right), respectively.

efficiency of the SIMPLY architecture. In this respect, note that: *i)* the only requirement for a reliable circuit operation of the SIMPLY architecture is that a sufficient RM is available at the input of the comparator in all cases, even in the presence of variability; *ii)* the variability characterization in Section III is performed at the array level, natively including contributions from the device (C2C, D2D and RTN) and from the non-idealities of the peripheral circuits (e.g., possible variations of V_{read}). This makes RM a comprehensive metric of the circuit reliability. Indeed, circuit implementations that result in higher RMs are more reliable and allow using simpler comparator or sense amplifier designs, though more complex designs can be used for smaller RMs. To estimate the RM and compare the performance of the two RRAM technologies, we used the C2C and D2D joint variability data to compute the distributions of V_N when P=Q=0 and when $P \neq Q$ for both the technologies and each LRS level (see Fig.5). The joint probability distribution of LRS for each technology and I_{read} was estimated by combining together 100 random samples for each $< \sigma_R; R >$ pair of Fig.4. To identify the worst-case RM that allows evaluating the performance and reliability limits, we assume HRS fixed at the worst-case of the HRS distribution (R_{HRS} = 40 kΩ for both technologies, as shown in Fig.3) and thus $V_N(P$=Q=0) is constant for each technology and target LRS level and determined by the value of R_G (see Fig.1a). Specifically, the optimal value of R_G that maximizes the RM for each technology and LRS level was chosen using (1):

$$R_G = \sqrt{\left(R_{LRS,MAX}^{-1} + R_{HRS,MAX}^{-1}\right)^{-1} \cdot \frac{R_{HRS,MIN}}{2}} \quad (1)$$

where $R_{LRS,MAX}$ is the μ+3σ value of the joint LRS distribution, and $R_{HRS,MIN}$ and $R_{HRS,MAX}$ are the $\mu \pm 3\sigma$ values of the HRS distribution, respectively ($R_{HRS,MIN}$ = $R_{HRS,MAX}$ = 40 kΩ in this case). As shown in Figs. 5 and 6a, RM grows with I_{read} and the two RRAM technologies show comparable reliability at the same LRS level, with a notable exception at I_{read} = 10 μA where HfO_x devices guarantee a quite larger RM than $Hf_{1-x}Al_xO_y$ devices. This stems from the lower C2C variability of HfO_x devices at low I_{read}, which leads to a smaller tail of L_1 as shown in Fig.4. To compare the reliability metric in Fig.6a to the energy efficiency performance, we report in Fig.6b the worst-case energy consumption of the read step of SIMPLY operation (averaged over the four possible input configurations) at different I_{read}. As expected, the energy per operation increases with I_{read}, which establishes a trade-off between energy efficiency and reliability. No relevant differences between the two technologies are observed.

TABLE I
1-BIT FULL ADDER ENERGY PER OPERATION (INDICATING MIN - MAX RANGE) VS I_{read} WITH t_p = 1 μs

I_{read}	Read	Write	FA
10 μA	(2.4 - 2.6) · 10^{-11} J	(5.1 - 5.6) · 10^{-10} J	(5.3 - 5.8) · 10^{-10} J
20 μA	(3.2 - 3.5) · 10^{-11} J	(1.0 - 1.1) · 10^{-9} J	(1.1 - 1.2) · 10^{-9} J
30 μA	(3.9 - 4.2) · 10^{-11} J	(1.7 - 1.8) · 10^{-9} J	(1.7 - 1.9) · 10^{-9} J
40 μA	(4.4 - 4.8) · 10^{-11} J	(2.3 - 2.5) · 10^{-9} J	(2.4 - 2.6) · 10^{-9} J

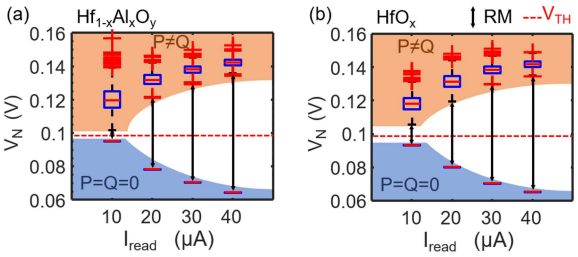

Fig. 5. (a)-(b) Distribution of V_N due to C2C and D2D variability, when P=Q=0 (blue area) and $P \neq Q$ (orange area) for the SIMPLY operation for different I_{read} and RRAM technology. Only the worst-case (i.e., lowest resistance value due to variability) HRS resistance is considered. The read margins (RM black arrows) and the threshold voltages (V_{TH} dashed red) for the comparator are evidenced. Black whiskers indicate the extreme points of the distributions. Red crosses indicate outliers. V_N when P=Q=1 is always much higher than in all other cases (thus is not reported in these box plots).

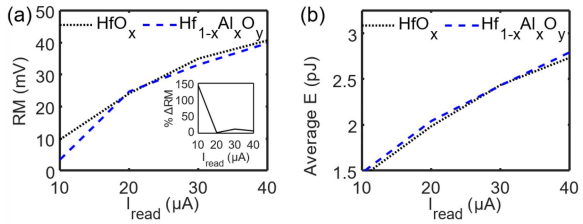

Fig. 6. (a) Worst-case RM (-3σ deviation from the mean of the RM distribution due to variability) for the two different RRAM technologies and different I_{read} values. The inset shows the relative difference between the RM obtained with HfO_x and $Hf_{1-x}Al_xO_y$ devices. (b) Worst-case energy consumption of the read operation during SIMPLY operation for HfO_x and $Hf_{1-x}Al_xO_y$ RRAM technologies at different I_{read} values. It is computed as the average of the worst-case energy consumption (+3σ deviation from the mean of the energy distribution due to variability) of each input configuration.

V. LIM FULL ADDER IMPLEMENTATION

To benchmark the performance of the SIMPLY architecture on more complex operations, we designed a 1-bit ripple carry full adder (FA) with the two RRAM technologies and estimated the worst-case performance. The FA was implemented using the SIMPLY archi-

TABLE II
COMPARISON BETWEEN THE PROPOSED RRAM-BASED FA AND A CMOS-BASED FA WHEN EXECUTING 32 PARALLEL 32-BIT FA OPERATIONS (ON A MEMORY ARRAY INTEGRATING 4096 DEVICES)

	# Devices	Energy	Delay	Energy-Delay Product (EDP)	EDP Normalized to CMOS w/ VNB	Improvement w.r.t. to CMOS w/ VNB
CMOS w/ VNB*	8192 - 28672 FET	$\approx$ 9.4 μJ	$\approx$ 284 μs	$\approx 2.7 \cdot 10^{-9}$ J·s	1	1
CMOS w/o VNB**	8192 - 28672 FET	$\approx 9.7 \cdot 10^{-4}$ - 7.4 pJ	$\approx 5.6 \cdot 10^{-2}$ - 4.8 μs	$\approx 1.7 \cdot 10^{-24}$ - $3.6 \cdot 10^{-17}$ J·s	$6.3 \cdot 10^{-16}$ - $1.3 \cdot 10^{-8}$	$7.5 \cdot 10^{7}$ - $1.6 \cdot 10^{15}$
This work I_{read} **= 10** μA t_P **= 1** μ**s**	**3232 RRAM**	**$\approx$ 594 nJ**	**$\approx$ 2.9 ms**	**$\approx 1.7 \cdot 10^{-9}$ J·s**	**0.6**	**1.57**
This work I_{read} **= 10** μA t_P **= 50 ns**	**3232 RRAM**	**$\approx$ 30 nJ**	**$\approx$ 147 μs**	**$\approx 4.4 \cdot 10^{-12}$ J·s**	**$1.6 \cdot 10^{-3}$**	**607**

*,** estimates with (w/) and without (w/o) VNB are performed considering energy and delay overhead for ~~2 kbits~~ **reading one memory page of 4kB** data **from a NAND flash memory** [14]. **CMOS FA performances were estimated projecting the time and energies for different 1-bit FA schemes taken from** [3], [15], [16] **where 0.18 μm, 45 nm, and 10 nm CMOS technology are used.**

tecture with 8 RRAM devices performing the sequence of operations reported in [2], which includes 28 steps of core operations (18 IMPLY and 10 FALSE). During a FA cycle, the energy consumption will also depend on the configuration of input bits, since it dictates the number of reset (during FALSE), set (during IMPLY when P=Q=0), and read (during IMPLY in all cases) operations that are executed. Here we consider the worst-case energy consumption during set and reset operations $E = V_P \cdot I_C \cdot t_P$, where V_P and t_P are the applied voltage magnitude and width. **We consider the worst-case energy for each input combination also for the read operation.** The minimum and maximum worst-case energies per FA cycle are proportional to I_C as reported in Table I. Values on the order of nJ are obtained for a t_P of 1 μs as the one used in Section III ($t_{IMPLY} = 4 \cdot t_P$, $t_{FALSE} = 2 \cdot t_P$, $t_{FA} = 10 \cdot t_{FALSE} + 18 \cdot t_{IMPLY}$).

To show the advantages offered by the proposed LiM scheme in terms of energy efficiency, we compare the performance of the proposed architecture against CMOS FA implementations **from** [3], [15], [16] ~~(~~with and without considering the VNB ~~energy and time overhead~~ [14]). **The VNB overhead was computed considering the time and energy required to read a typical flash memory page of 4kB** [14]**. Although the writing of the results to the memory should also be included, we consider only the overhead of the memory read to take into account the possibility of using smaller page sizes.** We estimate the delay and energy required to compute 32 parallel 32-bits FA operations (simple ripple carry) which require slightly less **(3132)** than the available 4096 devices in the array. To show the potential energy efficiency improvement over CMOS we consider the case I_{read} = 10 μA. As shown in Table II, the largest share of energy consumption and delay for CMOS logic comes from the VNB data exchange overhead [3], [14]–[16]. With the technologies explored in this work, when t_P = 1 μs (i.e., the pulse duration used in the characterization phase in Section III) the energy delay product (EDP) of SIMPLY ~~is still $\approx$ 5 times higher~~ **is only slightly better** than its CMOS counterpart when the VNB effect is included. However, the functionality of the RRAM devices considered in this work was proven also with t_P = 50 ns (see Fig.3a), and energy projections with such a t_P show that the proposed LiM scheme outperforms CMOS (when including the VNB overhead) by more than two orders of magnitude (worst-case projection) in energy efficiency with similar computing time, as shown in Table II. In addition, the number of required devices is reduced as compared to CMOS, thus achieving higher integration density.

VI. CONCLUSIONS

In this work, we studied the reliability and performance of SIMPLY architecture on two different 4 kbits RRAM arrays integrated in the BEOL of the 0.25 μm BiCMOS process. We highlighted and evaluated the trade-off between circuit reliability and energy consumption by exploiting the extensive array-level variability characterization of the two technologies. Furthermore, we analyzed the performance of a 1-bit FA implemented on SIMPLY. Even when considering the worst-case, the proposed architecture is $\approx$ 600 times more efficient than the CMOS counterpart including the VNB, while also achieving higher integration density. These results suggest that the proposed solution is a viable technology for the development of ultra-low power computing devices for IoT applications.

REFERENCES

[1] J. Backus, "Can Programming Be Liberated from the Von Neumann Style?: A Functional Style and Its Algebra of Programs," Commun. ACM, vol. 21, no. 8, pp. 613–641, 1978, doi: 10.1145/359576.359579.

[2] T. Zanotti, F. M. Puglisi, and P. Pavan, "A Smart Logic-in-Memory Architecture for Low-Power non-von Neumann Computing," IEEE Journal of the Electron Devices Society, 2020, (in press), doi: 10.1109/JEDS.2020.2987402.

[3] M. Aguirre-Hernandez and M. Linares-Aranda, "CMOS Full-Adders for Energy-Efficient Arithmetic Applications," IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 19, no. 4, pp. 718–721, 2011, doi: 10.1109/TVLSI.2009.2038166.

[4] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies," IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 22, no. 10, pp. 2054–2066, 2014, doi: 10.1109/TVLSI.2013.2282132.

[5] F. M. Puglisi, T. Zanotti, and P. Pavan, "SIMPLY: Design of a RRAM-Based Smart Logic-in-Memory Architecture using RRAM Compact Model," in ESSDERC, 2019, pp. 130–133, doi: 10.1109/ESSDERC.2019.8901731.

[6] T. Zanotti, F. M. Puglisi, and P. Pavan, "Reliability-aware design strategies for stateful logic-in-memory architectures," IEEE Trans. on Device and Materials Reliability, vol. 20, no. 2, pp. 278–285, 2020, doi: 10.1109/TDMR.2020.2981205.

[7] A. Grossi, D. Walczyk, C. Zambelli, E. Miranda, P. Olivo, V. Stikanov, A. Feriani, J. Su, G. Schoof, R. Kraemer, B. Tillack, A. Fox, T. Schroeder, C. Wenger, and C. Walczyk, "Impact of Intercell and Intracell Variability on Forming and Switching Parameters in RRAM Arrays," IEEE Trans. on Electron Devices, vol. 62, no. 8, pp. 2502–2509, 2015, doi: 10.1109/TED.2015.2442412.

[8] V. Milo, C. Zambelli, P. Olivo, E. Perez, M. K. Mahadevaiah, O. G. Ossorio, C. Wenger, and D. Ielmini, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," APL Materials, vol. 7, no. 8, p. 081120, 2019, doi: 10.1063/1.5108650.

[9] A. Grossi, C. Zambelli, P. Olivo, E. Miranda, V. Stikanov, C. Walczyk, and C. Wenger, "Electrical characterization and modeling of pulse-based forming techniques in RRAM arrays," Solid-State Electronics, vol. 115, pp. 17 – 25, 2016, doi: 10.1016/j.sse.2015.10.003.

[10] E. Perez, O. Gonzlez Ossorio, S. Duenas, H. Castan, H. Garcia, and C. Wenger, "Programming Pulse Width Assessment for Reliable and Low-Energy Endurance Performance in Al:HfO_2-Based RRAM Arrays," Electronics, vol. 9, p. 864, 2020, doi: 10.3390/electronics9050864.

[11] A. Fantini, L. Goux, R. Degraeve, D. J. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y. . Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO_2 RRAM," in IEEE IMW, 2013, pp. 30–33, doi: 10.1109/IMW.2013.6582090.

[12] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability," IEEE Trans. on Electron Devices, vol. 61, no. 8, pp. 2912–2919, 2014, doi: 10.1109/TED.2014.2330200.

[13] E. Perez, A. Grossi, C. Zambelli, P. Olivo, R. Roelofs, and C. Wenger, "Reduction of the Cell-to-Cell Variability in $Hf_{1-x}Al_xO_y$ Based RRAM Arrays by Using Program Algorithms," IEEE Electron Device Letters, vol. 38, no. 2, pp. 175–178, 2017, doi: 10.1109/LED.2016.2646758.

[14] S.-Y. Park, D. Jung, J.-U. Kang, J.-S. Kim, and J. Lee, "CFLRU: A Replacement Algorithm for Flash Memory," in Proc. of the 2006 Int. Conf. on Compilers, Architecture and Synthesis for Embedded Systems, ACM, 2006, pp. 234–241, doi: 10.1145/1176760.1176789.

[15] A. K. Yadav, B. P. Shrivatava, and A. K. Dadoriya, "Low power high speed 1-bit full adder circuit design at 45nm CMOS technology," in 2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE), 2017, pp. 427–432, doi. 10.1109/RISE.2017.8378203.

[16] S. Sharma and G. Soni, "Comparision analysis of FinFET based 1-bit full adder cell implemented using different logic styles at 10, 22 and 32NM," in 2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS), 2016, pp. 660–667, doi. 10.1109/ICEETS.2016.7583835.

We thank the reviewers for their careful comments and suggestions that helped us improving the quality of the manuscript. Please find in the following a detailed point-by-point response (in red) to the reviewers' comments. Edited parts in the manuscript are reported in bold and highlighted in yellow.

- Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author
This work presented a study on reliability of RRAM based logic-in-memory circuits, where full adders are benchmarked for energy and delay. Overall the article is missing several important details (refer to the comments below), and the implementation and results are relatively weak.

1. While this manuscript is based on the same authors' prior works [2][5], there is no publication statement that describes the differences and additions of this manuscript.

We thank the reviewer for his/her comment, that gives us the chance to better clarify the novelty of this work. In the previous works mentioned by the reviewer [2][5], we restricted the analysis to the evaluation of the performance of SIMPLY when using RRAM technologies taken from the literature. In this respect, only a scarce amount of information regarding cycle-to-cycle and device-to-device variability was considered, as variability data available in the literature were very limited. However, variability is a critical aspect to be considered to fully evaluate the viability of SIMPLY and the benefits it brings. In addition, a clear array-level analysis and demonstration of functionality of RRAM-based Logic-in-Memory solutions (not just SIMPLY) is still missing. In this study we consider devices fabricated in a 4kbit array and perform an extensive array-level variability characterization on two different RRAM flavors and in different programming conditions. By exploiting the extended and self-consistent variability dataset, we prove the actual viability and the strong reliability of SIMPLY in a real 4kbit array, including the worst-case and array level non-idealities. We revised the manuscript to better stress the novelty of this contribution in the introduction.

2. Is IMPLY an acronym? It should be spelled out in the first occurrence.

We thank the reviewer for this observation. Actually, IMPLY is not an acronym. It rather is the term commonly used in the field to refer to the material implication operation and also to the circuit arrangement that is typical of RRAM-based Logic-In-Memory circuit based on the material implication logic.

3. Section III is in a single paragraph, which is very long. It should be broken down into several paragraphs based on the content.

We thank the reviewer for his/her suggestion, that gives us the possibility of improving readability. We revised the manuscript and divided Section III in three paragraphs (discussion of the resistive stack materials and integration features, forming/set/reset operation voltages and timings for multi-level operation, variability discussion), as per the suggestion of the reviewer.

4. Regarding Table II, what exactly “CMOS w/ VNB” and “CMOS w/o VNB” means should be more clearly described. Are those data directly from [3] or [14], or are these numbers obtained the authors themselves? If the latter, which CMOS technology was used, and which exact circuit are these values based on?

We thank the reviewer for his/her comment. When discussing the energy and delay of the CMOS w/ VNB solution, what we mean is the actual energy and delay introduced by the combination of the CMOS circuitry that has to perform the operation (in this example a 32 parallel ripple-carry 32-bit full-adders) and of the data transfer from the memory to the CPU (I.e., VNB). The CMOS circuitry per se (I.e., CMOS w/o VNB) is fast and energy efficient, and the VNB (the need to transfer the data from the memory to the CPU) definitely acts as an efficiency bottleneck. In Tab. II, we estimated the time and energy overhead brought by the VNB by taking the information from [14]. Also, we considered a 180nm CMOS technology [3]. Based on the reviewer’s comment, we improved the manuscript by revising the discussion of Table II and by evaluating more CMOS technologies (at different technological nodes, down to 10 nm) to provide a larger overview of how the otherwise excellent CMOS performance is drastically limited by the VNB. The CMOS w/o VNB entries in Table II are computed scaling the performance of 1-bit CMOS full-adders as reported in the literature to 32 parallel 32-bits ripple-carry adders (x32x32 for the energies, and x32 for the delays).

5. CMOS w/o VNB outperforms the proposed work by a large factor. Many ASIC designs try to co-locate memory and logic, to eliminate the von Neumann bottleneck. To that end, compared to those practical ASIC designs, what advantage does this work have? It seems EDP is much worse.

We thank the reviewer for his/her comment, that gives us the opportunity to better clarify the advantages of the proposed Logic-in-Memory solution. It is true that ASIC designs implementing near memory computing can result in better EDP performances, although lacking reconfigurability. Also, such designs are often based on SRAM memories, which require large chip area, dissipate static power, and need to be rewritten after each power down event due to their volatility. There are also 3D integrated designs for Logic-In-Memory applications relying on the Through Silicon Via approach with DRAM like the one provided in “A 3D-Stacked Logic-in-Memory Accelerator for Application-Specific Data Intensive Computing”, 2013 IEEE International 3D Systems Integration Conference (3DIC), but its integration cost is extremely high and the manufacturing process in sensible to yield loss issues and variability. SIMPLY do not require specific designs but adds computing capabilities to existing non-volatile memory arrays. SIMPLY allows the deployment data intensive computations to the memory array. Another advantage is that SIMPLY can be reconfigured by programming the control logic. Also, to ensure that our estimates of the SIMPLY FA implementation are below the lower bound of the performance, we consider a worst-case in which the HRS resistance is fixed to the lower bound of the distribution, and the SET and RESET energies are overestimated. Thus, real circuit implementations are expected to consume less energy. Also, due to our test setup the analysis was limited to pulse widths of 50ns but switching times <1ns were demonstrated on other RRAM technologies. Shorter pulses would further improve the EDP. In addition, increasing the array size would result in an increased throughput and lower EDP.

6. What exactly is designed?
Did the authors simulate 32 32-bit full adders, using the RRAM characteristics obtained by

the authors own previous works?
Or is any new device or array newly implemented for this submitted work?

We thank the reviewer for his/her comment, that gives us the chance to better clarify the novelty of this work. In this study we fully characterized the array-level variability (both cycle-to-cycle and device-to-device) of a 4kbit array of RRAM devices to verify the viability of the latter to support the SIMPLY Logic-In-Memory paradigm and to evaluate the energy efficiency and overall performance. We did not actually simulate the 32 ripple-carry 32-bit full adders, but rather we simulated the performance of a single full-adder by accounting for the worst-case scenario identified in the array-level variability analysis, and projected the results to the case of 32 parallel 32-bit ripple-carry full adders. We considered 32 parallel 32-bit full adders as they require using approximately 4k RRAM devices, i.e. the size of the fabricated array, and used the estimates of the VNB overhead from [14] when comparing the performance of the proposed solution vs. CMOS. We revised the manuscript to include these considerations in Section V and in the Table II caption.

7. When the authors say “32-bit full adder”, does that mean this is a 32-bit ripple-carry adder? How is the carry communicated among different bit positions in the RRAM based design?

The reviewer is correct; we mean a 32-bit ripple-carry full adder. The output carry bit is stored at the end of each 1-bit full addition as the resistance of a device in the RRAM array. The following 1-bit full addition uses this device as the carry input bit.

8. For a full adder operation, how many read and write operations are required?
Can the authors breakdown the 1-bit full adder energy in Table I, between read energy and write energy?

We thank the reviewer for his/her observation. Table I was updated in the revised manuscript to highlight the read and write energy per operation.

For the RRAM design energy, is the energy of the driver for high write voltage or level shifter (t convert between low voltage and high voltage) included?

The energy values reported in the manuscript take into account the energy of the comparator. The energy of the control logic and of the analog buffers was not considered because is much lower than that of the RRAM device SET and RESET operations (especially when considering the worst-case), therefore bringing a negligible contribution. In fact, we considered the results reported in “Overhead Requirements for Stateful Memristor Logic,” IEEE T CIRCUITS-I, vol. 66, no. 1, pp. 263–273, 2019 (the table with the results is replicated below), where the IMPLY architecture including the control logic was simulated using a general-purpose memristor compact model calibrated on a $\approx 2 \cdot 10^3$ times higher current compliance. By rescaling their results to our current compliance and clock frequency we see that our worst-case estimates result in a higher EDP, thus the overhead of the control logic

is negligible. Finally, the write voltages are relatively low (few volts) and so no high write voltages are needed.

TABLE VII
ENERGY-DELAY PRODUCT FOR MEMRISTORS AND CMOS

Function	Memristor EDP (J·s)	CMOS EDP (J·s)
NAND	$2.58 * 10^{-13}$	$1.87 * 10^{-24}$
AND	$3.95 * 10^{-13}$	$3.59 * 10^{-24}$
NOR	$1.22 * 10^{-13}$	$2.53 * 10^{-24}$
OR	$3.80 * 10^{-13}$	$4.46 * 10^{-24}$
XOR	$7.25 * 10^{-12}$	$9.31 * 10^{-24}$
XNOR	$6.04 * 10^{-12}$	$9.15 * 10^{-24}$
Full Adder	$1.93 * 10^{-11}$	$3.32 * 10^{-23}$

Which tp value is Table I assuming?

The value of tp used in Table I is 1µs (the same used in the variability analysis of the array). This information has been added to the caption of Table I in the revised manuscript.

- Reviewer: 2

Comments to the Author
This paper investigated the impact of RRAM variabilities on the functionality of the previously proposed logic-in-memory architecture. The paper is in good shape and the impact of D2D and C2C variation are well analyzed with two RRAM devices with different stackings. The reviewer suggests ACCEPT with several minor comments.

1) As the resistance drift is a typical nonideality of RRAM and it's hard to deal with as a temporal error, is it possible for authors to discuss the impact of retention on the read margin?

We thank the reviewer for his/her constructive comment. The measurement setup exploited in this work for array measurements already considers the worst-case resistance drift. The set/reset operations with M-ISPVA and ISPVA are performed sequentially cell-by-cell and immediately after their end the array read takes place sequentially with the same cell's order. This ensures that the time interval between the set/reset operation and the read is the same experienced for all the cells. That idle time interval has been proven to cause a negligible retention impact on the LRS and HRS levels as demonstrated in "Toward Reliable Multi-Level Operation in RRAM Arrays: Improving Post-Algorithm Stability and Assessing Endurance/Data Retention", IEEE JEDS, vol. 7, pp. 740-747, 2019

2) Regarding the performance comparison between RRAM-based LiM with CMOS-based logic, I'm assuming the results of CMOS-based designs are based on the same technology (0.25um). If so, this may not be a fair comparison as CMOS technology can be well scaled while the scalability of RRAM process is relatively limited. Could the authors add more discussions from this perspective?

We thank the reviewer for his/her constructive comment, which gives us the opportunity of improving the comparison with CMOS solutions. We would like to point out that, although we are considering a 0.25 um technology for our RRAM array, RRAM devices scaling potential goes down to the sub-10 nm nodes. Particularly, filamentary devices are not expected to behave much differently at ultra-scaled nodes, due to the resistive switching being localized and confined at the filament location. If anything, less impact from parasitics is expected which will improve the circuit-level performance. As far as the comparison with CMOS solutions is concerned, we considered a 0.18um technology for CMOS, but in the revised version of the manuscript we extended the analysis by including the data for more scaled CMOS nodes (down to 10 nm). The results reported in the table below, show that the EDP for CMOS implementations considering 10nm, 32nm, 45nm, 0.18um technology nodes ranges from $1.70 \cdot 10^{-24}$ to $3.55 \cdot 10^{-17}$J·s without considering the VNB overhead.

CMOS Full-Adders implementation	**EDP (J·s) 32x 32-bit parallel Full-Adders (min - max)**
0.18um [Aguirre 2011]	$1.5 \cdot 10^{-19}$- $1.7 \cdot 10^{-18}$
45nm [Yadav2017_1]	$1.70 \cdot 10^{-24}$ – $7.09 \cdot 10^{-24}$
32nm [Yadav2017_2]	$2.22 \cdot 10^{-24}$
10nm [Sharma2016]	$3.55 \cdot 10^{-17}$

[Aguirre2011] M. Aguirre-Hernandez and M. Linares-Aranda, “CMOS Full-Adders for Energy-Efficient Arithmetic Applications,” IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 19, no. 4, pp. 718–721, 2011, doi: 10.1109/TVLSI.2009.2038166.

[Yadav2017_1] A. K. Yadav, B. P. Shrivatava and A. K. Dadoriya, "Low power high speed 1-bit full adder circuit design at 45nm CMOS technology," 2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE), Bhopal, 2017, pp. 427-432, doi: 10.1109/RISE.2017.8378203.

[Yadav2017_1] A. Yadav, B. P. Shrivastava, and A. K. Dadoria, “Low power high speed 1-bit full adder circuit design in DSM technology,” in 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Aug. 2017, pp. 1–6, doi: 10.1109/ICOMICON.2017.8279098.

[Sharma2016] S. Sharma and G. Soni, “Comparision analysis of FinFET based 1-bit full adder cell implemented using different logic styles at 10, 22 and 32NM,” in 2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS), Apr. 2016, pp. 660–667, doi: 10.1109/ICEETS.2016.7583835.