Functional Read Enabling In-Memory Computations in 1Transistor—1Resistor Memory Arrays

Akhilesh Jaiswal, Robert Andrawis, Amogh Agrawal, and Kaushik Roy, *Fellow, IEEE*

***Abstract*—'In-memory computing' is an emerging paradigm that attempts to embed some aspects of logic computations inside memory arrays leading to higher throughput and lesser energy-consumption. Consequently, various in-memory compute proposals using emerging non-volatile technologies based on *functional read* - wherein multiple word-lines are simultaneously activated and a Boolean function of the constituent activated rows is read, is being extensively investigated. In this brief, we first show that the conventional sensing scheme for such functional reads, operated on 1 Transistor - 1 Resistor (1T-1R) memory arrays, is limited theoretically by low sense-margin. We demonstrate that the sense-margin *does not* improve even if the ON-OFF resistance difference is increased from low values to considerably higher values. Subsequently, we present a new sensing scheme based on skewed sense-amplifiers and staggered world-line activation as a method for enabling functional read operations. We show that in-memory XOR, IMP (implication) and bit-wise comparison can be easily implemented through the proposed scheme.**

***Index Terms*—Functional read, in-memory computing.**

I. Introduction

THE UNPRECEDENTED scaling of the metal oxide semiconductor field effect transistors (MOSFETs) have powered the ever increasing compute efficiency and throughput on silicon chips [1]. Specifically, the drastic improvement in compute efficiency in past few decades can be traced back to the availability of larger memory capacity and the speed-up obtained due to shrinking transistor sizes [2]. However, transistor scaling has slowed down significantly as the dimensions of the transistors approach their physical limits [3]. Further, the existence of the well-known 'memory-bottleneck' inherent in the state-of-the-art von-Neumann computing model poses a major constrain for energy efficient and high throughput computations.

Manuscript received November 5, 2019; revised January 11, 2020; accepted February 13, 2020. Date of publication February 21, 2020; date of current version November 24, 2020. This work was supported in part by the Center for Brain Inspired Computing (C-BRIC), One of Six Centers in JUMP, in part by the Semiconductor Research Corporation Program Sponsored by DARPA, in part by the National Science Foundation, in part by the Intel Corporation, and in part by the Vannevar Bush Faculty Fellowship. This brief was recommended by Associate Editor A. Calimera. *(Akhilesh Jaiswal and Robert Andrawis contributed equally to this work.) (Corresponding author: Akhilesh Jaiswal.)*

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: jaiswal@purdue.edu; randrawi@purdue.edu; agrawa64@purdue.edu; kaushik@purdue.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2020.2975658

Researchers are actively investigating a two-pronged solution to keep-up with the compute requirement for future data-intensive tasks. On one hand, beyond CMOS emerging non-volatile resistive technologies are being widely investigated to achieve high-density, low stand-by power memory systems [4], [5]. While on the other, there is a renewed quest for embedding computations close to the memory units enabling 'in-memory' processing that can to some extent mitigate the 'memory-wall bottleneck' [6]. With respect to in-memory vector Boolean computations, the general approach is to activate two rows of a memory array, simultaneously. In response to the activated rows, bit-wise Boolean operation of the data stored in the two individual rows is performed by minimal modification to the memory array and peripherals.

Intuitively, one would expect that with improvement in ON-OFF resistance ratios of the constituting memory devices, drastic improvements in read sense-margin for functional read schemes can be achieved. We start by pointing out that within the constrain of using the well-known cross-coupled inverter based sense-amplifiers, the conventional sensing scheme is unsuitable for in-memory operations. Furthermore, the sense-margin *does not* improve with improved ON-OFF ratios. As such, whether the ON-OFF ratio is high or low, the conventional sensing scheme would result is high sensing failure rate with respect to in-memory computing. Toward that end, we propose a novel sensing scheme suitable for in-memory computing with the major advantage that the sensing-robustness steadily improves with improving ON-OFF ratios. Furthermore, we only rely on the well-known cross-coupled inverted based sense-amplifiers that are used in almost all large-scale chips fabricated based on emerging devices [7], [8].

The rest of this brief is organized as follows. Section II presents the proposed function read scheme. Section III discusses the result before concluding this brief.

II. Proposed Sensing Scheme for Bit-Wise In-Memory Operations

We start by first describing that the conventional *'simultaneous row activation scheme'* is fundamentally limited by low sense-margin irrespective of the ON-OFF resistance ratio of the constituting resistive memory element.

A. Simultaneous Row Activation

To understand the conventional sensing involving simultaneous row activation, let us assume we have to compute vector

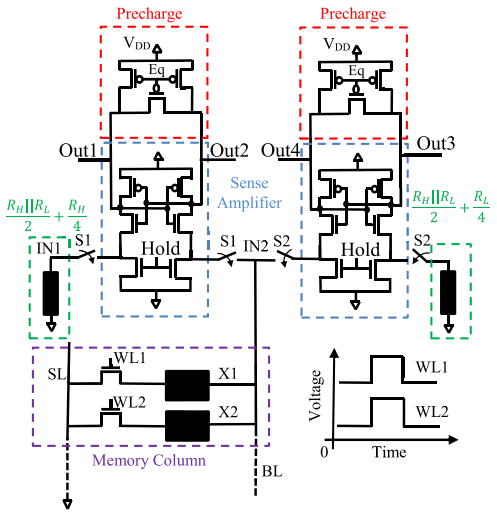


Fig. 1. The circuit schematic of simultaneous memory activation scheme.

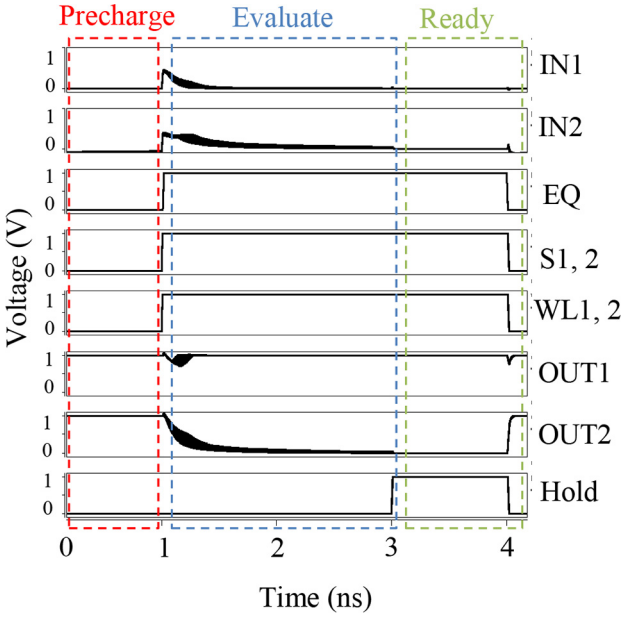


Fig. 2. The time diagram of simultaneous memory activation schema that is simulated by HSPICE Monte Carlo for 1000 samples.

bit-wise Boolean operation corresponding to the bit-stream stored in two arbitrary rows in a memory-array. The simultaneous activation dictates that both the rows storing the data to be computed are activated simultaneously by activating the access transistors through the wordline signals WL1 and WL2 as shown in Fig. 1. With WL1 and WL2 activated, the two memory elements X1 and X2 effectively form a parallel combination (ignoring the presence of the access transistors). This parallel configuration has three possible resistance states - $R_H/2$, $R_L/2$ and $R_H||R_L$ (R_H = High Resistance and R_L = Low Resistance). Assuming R_H represents a digital '1' and R_L a digital '0', if one can sense the presence of $R_H/2$, it would represent an AND gate, since AND (NAND) is '1' only when both the inputs are high. Similarly, if $R_L/2$ is sensed OR (NOR) logic can be accomplished, since OR is zero only when both inputs are zero. The AND (NAND) and OR (NOR) output thus generated can be combined easily to achieve an XOR gate [9].

Thereby, in order to sense $R_H/2$ and $R_L/2$ from among $R_H/2$, $R_L/2$ and $R_H||R_L$ two sense-amplifiers can be used as shown in Fig. 1. Each of the sense-amplifiers has one of its input connected to the bit-line (BL) while the other input is connected to a reference resistance. As shown in Fig. 1, the reference resistance for comparing between $R_H/2$, $R_L/2$ is set to average of $R_H/2$ and $R_L/2$, while the reference resistance for the other sense-amplifier is designed to be the average of $R_L/2$ and $R_H||R_L$. As such, the two sense-amplifiers accomplish two comparisons and ascertain if the parallel combination of the memory elements X1 and X2 is $R_H/2$ or $R_L/2$.

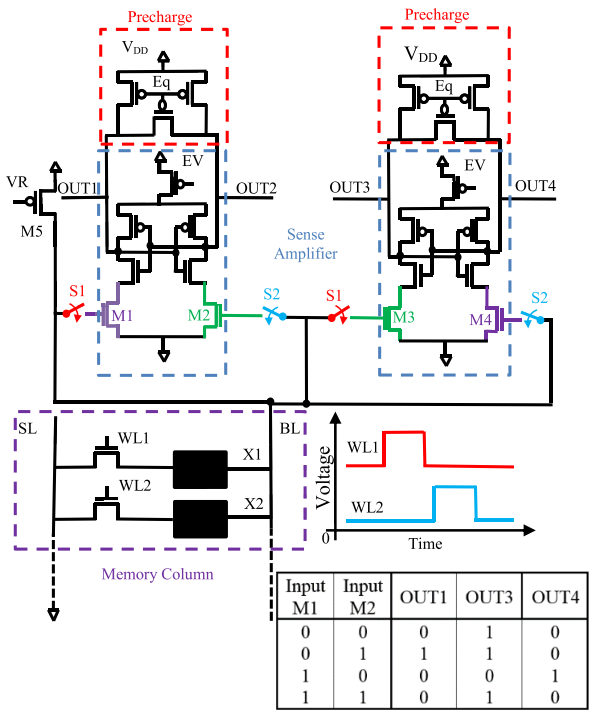
The time sequence of the simultaneous row activation consists of three stages. First, the sense amplifier outputs are precharged to the same voltage level. Second, the memory word-lines WL1 and WL2 are enabled while the precharge circuit has been disconnected. Finally, the memory word-lines WL1 and WL2 are disabled and the Hold signal is enabled to latch the sense amplifier outputs. The time diagram of the simultaneous row activation is illustrated in Fig. 2.

It is important to note that, as a function of the improved ON-OFF ratio between the resistances of the memory elements the sensing robustness for one of the sense-amplifiers in Fig. 1 would steadily improve, while the sensing-margin for the other sense-amplifier stays relatively constant irrespective of improvement in the ON-OFF ratio. Specifically, for the sense amplifier differentiating between $R_H/2$ and $R_H||R_L$, the sense-margin improves due to improvement in the resistance value R_H. In sharp contrast, the sense-margin for the sense-amplifier sensing between the resistance levels $R_L/2$ and $R_H||R_L$ remains almost constant and independent of the increased R_H/R_L ratio. This is because the resistance value $R_H||R_L$ is theoretically bounded by R_L, therefore, irrespective of the improved value of R_H the resistance difference that the sense-amplifier has to sense remains constant. In summary, the conventional scheme wherein both the WLs are simultaneously activated is limited by the poor sense-margin of one of the sense-amplifiers that is independent of the R_H/R_L ratio of the memory element.

B. Proposed Staggered Memory Cell Activation

The proposed staggered memory activation scheme dictates that the WL1 should be activated first for $1ns$ with the switch S1 closed and S2 open and VR is pulled low allowing a voltage-divider to be formed between transistor M5 and the bit-cell. Thereby, the resultant voltage on BL representing the memory bit-cell X1 is stored on the parasitic capacitances at the gate of transistor M1 and M3. Subsequently, WL1 is deactivated and WL2 is activated while switch S2 is closed and S1 is kept open. During this time, the voltage on the BL corresponding to the memory bit-cell X2 is stored on gates of transistors M2 and M4. Thereafter, both S1 and S2 are opened and the sense-amplifiers are ready to be activated.

Note, at this stage the two sense-amplifiers have voltages at each of their input terminals proportional to the data stored in the memory elements X1 and X2, respectively. Four combinations are possible for X1 and X2 $-$ $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$. For the case of $(0, 1)$ or $(1, 0)$ the voltages developed on the



Input M1	Input M2	OUT1	OUT3	OUT4
0	0	0	1	0
0	1	1	1	0
1	0	0	0	1
1	1	0	1	0

Fig. 3. The circuit schematic of staggered memory activation schema.

two inputs of the sense-amplifiers will be different. As such, the sense-amplifiers sense the difference in voltages and results in $OUT1 = 1$ and $OUT3 = 1$ for the case $(0, 1)$ and $(1, 0)$, respectively, as shown in the table in Fig. 3. However, for the case $(0, 0)$ and $(1, 1)$ the voltages on the two sides of the sense-amplifiers would be same since the data in X1 and X2 are the same. This can lead to an 'unpredictable state' for the sense-amplifier output. We circumvent this issue by asymmetrically sizing the sense-amplifiers such that M1 is sized greater than M2 and M4 is sized larger than M3. Because of such asymmetric sizing, for the cases $(0, 0)$ and $(1, 1)$, i.e., when the voltages on both the inputs of the sense-amplifiers are the same, the sense-amplifiers will end up switching according to the *in-built* skew of the constituting transistors. Accordingly, for both the cases $(0, 0)$ and $(1, 1)$ $OUT1 = 0$ and $OUT3 = 1$, respectively, as mentioned in the table in Fig. 3.

The table of Fig. 3 implies that, if we OR OUT1 with the NOT of OUT3, XOR logic can be easily accomplished. Additionally, the logic levels corresponding to OUT3 represents the truth table for IMP (implication logic). Not only XOR is an important functionality for various bit-wise operations, it can also be used for 'in-memory' ADD operations by including additional peripheral circuits. Furthermore, XOR and IMP form a universal basis set, hence, in general, any bit-wise Boolean functionality can be achieved through repeated use of XOR and IMP operations. Another, key aspect of the presented proposal as opposed to previous works is the fact that we can achieve bit-wise Boolean in-memory comparison operation. To understand how this can be achieved, let us observe the truth table of 3. OUT1 is '1' only when the input data is $(0, 1)$ while OUT4 is '1' only for the input data $(1, 0)$. Thereby, OUT1 represent X1 $<$ X2 and OUT4 represents X1 $>$ X2, while if OUT1 and OUT4 are both '0' it indicates X1 $=$ X2. Thus, the proposed sensing scheme allows for in-memory XOR, IMP as well as bit-wise comparison.

It is important to note that the skewing of the sense-amplifier although helps to avoid the 'unpredictable state' for the cases $(0, 0)$ and $(1, 1)$ it slightly degrades the sense-margin for the cases $(1, 0)$ and $(0, 1)$. However, as shown in the next section despite the slight degradation of the sense-margin the proposed scheme performs better than the conventional simultaneous activation and the sense-margin notably improves with improvement in R_H/R_L ratio.

Notably, the most remarkable aspect of the proposed staggered activation is that the sensing-robustness steadily improves with improvement in the R_H/R_L ratio. Intuitively, this is because of the fact that for the staggered activation scheme the sense-amplifiers either compares same voltages and fall to their 'default skewed' state or they compare R_H with R_L the difference of which increases significantly with increase in R_H/R_L ratio. Recall, in contrast the simultaneous activation is constrained by sensing between $R_L/2$ and $R_H||R_L$ the ratio of which remains constant irrespective of the R_H/R_L ratio.

III. Simulation Results and Discussions

In order to ascertain the resilience of the proposed sensing scheme against process variations, a Monte Carlo circuit simulation is conducted considering normally distributed resistance of the storage element within 5% standard deviation (STD). The transistor threshold voltage variations are modeled by a normal distribution with standard deviation $(\sigma\sqrt{\frac{W_{min}L_{min}}{WL}})$, where W, L, W_{min}, and L_{min} are the width, length, minimum width, and minimum length of the transistor, respectively and σ was chosen to be 30mV [10]–[12]. A supply voltage V_{DD} of 1V is used for all simulations in this brief. A total of 10,000 Monte Carlo samples is used for estimating all the results in this brief using the predictive technology model (PTM) for 45 nm node [13], [14].

For a fair comparison, we used similar sizing for all the transistors which is $2w_{min}$ for PMOS and w_{min} for NMOS, where w_{min} is the minimum transistor sizing. This sizing is used for both the simultaneous and staggered sense amplifier except for M1, M2, M3, and M4. The sizing for M1, M2, M3, and M4 are $1.2w_{min}$, $5w_{min}$, $1.2w_{min}$, and $5w_{min}$, respectively. Additionally, the supply voltage is 1V. All switches are transmission gates consisting of NMOS and PMOS transistors. The read error rate is inverse proportional to the time period of the read operation, we strict all the simulation to iso-switching time of 4ns.

The timing diagram of the staggered memory activation is illustrated in Fig. 4a and 4b for the inputs '01' (or '10') and '00' (or '11'), respectively. The proposed staggered read scheme illustrated in Fig. 3 starts by connecting the two sense amplifier terminals to the same BL through two different transmission gates that are enabled in staggered order. First, while WL1, WL2 and VR are '1', '0' and '0', respectively, the BL is charged by the transistor M5 and the developed voltage at (BL) is stored in the gate capacitance of M1. Second, while WL1, WL2 and VR are '0', '1' and '0', respectively, the BL

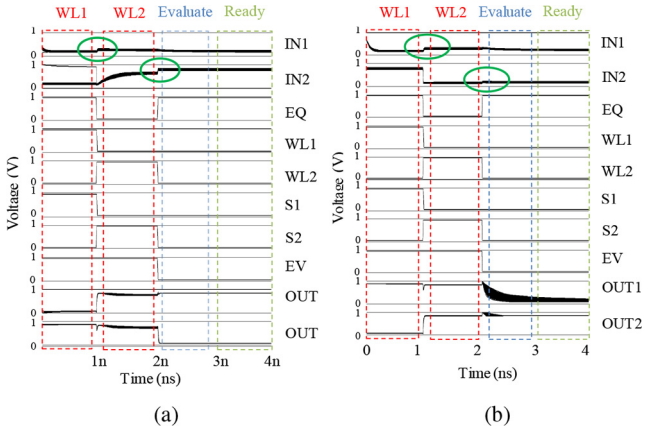


Fig. 4. The time diagram of staggered memory activation schema for '01' simulated by HSPICE Monte Carlo for 1000 samples. The charge sharing and charge injection happen due to switching off the transmission gates are marked by the green oval. (a) The input case '01'. (b) The input case '00'.

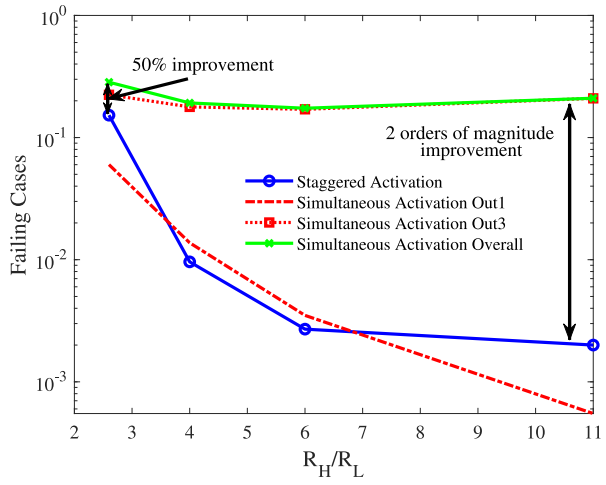


Fig. 5. The number of failing cases for staggered and simultaneous memory activation schemes as a function of different ON-OFF resistance ratios 25°C.

is charged by the transistor M5 and the developed voltage at (BL) is stored in the gate capacitance of M2. Intuitively, the stored charges in M1 and M2 gate capacitance are proportional to the resistance of the first and second memory cell, respectively and hence, to the logical value stored in the cell. Finally, the charging circuit is disabled to start the evaluation phase to develop the final values of OUT1 through OUT4 which is the result of the comparison of the voltage developed by the two memory cells at BL. The values of OUT1 and OUT2 are retained by the latch that consists of the two cross-coupled NOT gates. It is worth noticing from the timing waveforms of Fig. 4(b), OUT1 falls to zero due to in-built skew of the sense-amplifier. On the other hand the output of Fig. 4 is decided based on the difference in the voltage levels at the gate of transistors M1 through M4. Note, although not shown, OUT3 and OUT4 would produce similar timing behavior except that the sense-amplifiers are skewed in opposite direction.

The comparison of number of failing cases of the staggered and simultaneous memory cell activation is illustrated in Fig. 5. The results are presented for different R_H/R_L ratio ratios. Few key observations from Fig. 5 are as follows: 1) For

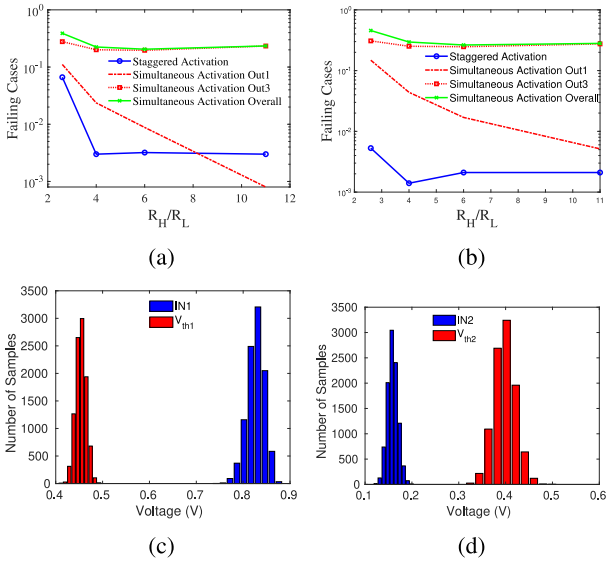


Fig. 6. The number of failing cases for staggered and simultaneous memory activation schemes (a) at 50°C. (b) The same plot as (a) at 10% lower supply voltage. (c) The distribution of the threshold voltage variations and the voltage input of the transistor M1. (d) The distribution of the threshold voltage variations and the voltage input of the transistor M2.

all cases the overall error for staggered activation is better than the conventional simultaneous activation. 2) The overall error for simultaneous activation remains relatively constant across all R_H/R_L ratio values. This is because although OUT1 (i.e., sensing between $R_H/2$ and $R_H||R_L$) improves due to increasing R_H, OUT3 (i.e., sensing between $R_H||R_L$ and $R_L/2$) remains almost constant. 3) In contrast, the number of failing case decreases steadily with the increase in R_H/R_L ratio for the proposed staggered sensing scheme. Of particular interest is the fact that by enabling staggered activations of the word-lines the presented proposal can be used in conjunction with well-known cross-coupled inverter based sense-amplifiers to accomplish in-memory XOR, IMP, bit-wise comparison and also more complex operations like ADD as shown in [15]. For comparison the power consumption for both the simultaneous and staggered are 26.74 μW and 14.56 μW, respectively. The energy of the simultaneous and staggered activation is 106.9 fJ and 58.24 fJ, respectively. The area of the staggered sense circuit is 44 μm^2. The area of simultaneous activation sense circuit is 22 μm^2. To quantify the temperature variations effect we repeated the same simulation at 50°. The simulation results shows almost the same performance for both architectures, as illustrated in Fig. 6. The performance of the staggered activation is still dominating the simultaneous activation at 10% lower supply voltage, as illustrated in Fig. 6. The distribution of the threshold voltage and sense amplifier inputs for the staggered memory activation are illustrated in Fig. 6 for on-off ratio of 10.

The voltage stability is determined by two factors the value of gate capacitance and the leakage current. Based on simulations, the time constant of the gate was found to be 16μs according to SPICE simulations. This means that the time needed for 1% change in gate voltage by leakage current is 100ns. The entire read operation requires 4ns which is 25x less than the gate retention time. Additionally, such sensing

schemes wherein a voltage is stored as a charge on gate capacitance during sensing has been explored in various fabricated chips in prior literature (although not in the context of in-memory computing). One such work is [16] where charges were stored on the gate capacitance of the sense amplifier one by one to enable robust sensing for VCMA based MRAM cells. Finally, we would like to mention that the proposed staggered functional read scheme is valid, in general, for any class of 1T-1R arrays. We evaluated the staggered activation using the filament-based resistive RAM variations from [17]. The device in [17] has large variations and a high on-off ratio of 600. The error rate reaches zero for such a high on-off ratio.

The AND, OR, NAND, and NOR can be supported with slight modifications. In table in Fig. 3, if the input to M2 and M4 is negated then OUT4 is an AND, OUT3 is a NAND, OUT1 is a NOR, and OUT2 is an OR function. The negation of the memory voltage level could be achieved by the circuit in [11]. The delay of the negation circuit (NOT gate) is neglected compared to the memory access time. For all these functions, the memory cell access is done once. Additionally, IMP and XOR form a complete basis set hence any Boolean function can be achieved by repeated use of IMP and XOR functionality. If the input of the second memory cell is replaced by a reference cell that contains a known value then the proposed scheme becomes similar to a conventional read based sensing scheme. Furthermore, as long as the ON-OFF ratio is high the in-built skew of the sense-amplifier would not lead to any read failure concerns since the input voltage to the sense-amplifiers will be large enough to over-ride the in-built skew.

IV. Conclusion

In this brief, we have proposed a new functional sensing scheme for in-memory Boolean operations based on 1T-1R memory arrays. We rely on the well-known cross-coupled inverter based sense amplifiers as the basic sensing circuit. We first show that the conventional scheme with the aforementioned sense-amplifiers are of limited applicability for in-memory Boolean operations due to low-sense-margin. Further, the sense-margin does not improve with improvement in ON-OFF ratios. As such, to exploit high ON-OFF ratios of emerging memory technologies, we propose use of staggered activation of the word-lines and skewing the sense-amplifiers for robust read operations. We show in-memory XOR, IMP and bit-wise comparison can be easily accomplished by the use of presented staggered activation scheme. A key highlight of the presented proposal is the fact that as opposed to the conventional scheme, the sensing-robustness of the proposed scheme steadily increases with increase in the ON-OFF ratio of the constituting memory elements. Keeping in view the extensive research investigation for improved ON-OFF resistances we believe the presented proposal is well-suited for future non-volatile in-memory bitwise Boolean operations.

References

[1] M. M. Waldrop, "The chips are down for moore's law," *Nat. News*, vol. 530, no. 7589, p. 144, 2016.

[2] S. Borkar and A. A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, 2011.

[3] J. M. Shalf and R. Leland, "Computing beyond moore's law," *Computer*, vol. 48, no. 12, pp. 14–23, Dec. 2015.

[4] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electron.*, vol. 125, pp. 25–38, Nov. 2016.

[5] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nat. Nanotechnol.*, vol. 10, no. 3, pp. 191–194, 2015.

[6] P. G. Emma, "Understanding some simple processor-performance limits," *IBM J. Res. Develop.*, vol. 41, no. 3, pp. 215–232, May 1997.

[7] H. Noguchi, K. Ikegami, N. Shimomura, T. Tetsufumi, J. Ito, and S. Fujita, "Highly reliable and low-power nonvolatile cache memory with advanced perpendicular STT-MRAM for high-performance CPU," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2014, pp. 1–2.

[8] H. Noguchi *et al.*, "7.5 A 3.3 ns-access-time 71.2 μw/mhz 1mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2015, pp. 1–3.

[9] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory boolean computations in CMOS static random access memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 12, pp. 4219–4232, Dec. 2018.

[10] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 12, pp. 1859–1880, Dec. 2005. [Online]. Available: http://ieeexplore.ieee.org/document/1542241/

[11] Q.-K. Trinh, S. Ruocco, and M. Alioto, "Dynamic reference voltage sensing scheme for read margin improvement in STT-MRAMs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 4, pp. 1269–1278, Apr. 2017.

[12] A. Jaiswal, R. Andrawis, and K. Roy, "Area-efficient nonvolatile flip-flop based on spin hall effect," *IEEE Magn. Lett.*, vol. 9, no. 9, pp. 1–4, Apr. 2018.

[13] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proc. 7th Int. Symp. Qual. Electron. Design (ISQED)*, 2006, pp. 585–590. [Online]. Available: http://ieeexplore.ieee.org/document/1613201/

[14] J. P. Duarte *et al.*, "BSIM-CMG : Standard FinFET compact model for advanced circuit design," in *Proc. 41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2015, pp. 196–201.

[15] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic ram," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 3, pp. 470–483, Mar. 2018.

[16] H. Noguchi *et al.*, "Novel voltage controlled MRAM (VCM) with fast read/write circuits for ultra large last level cache," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2016, p. 27.5.

[17] A. Grossi *et al.*, "Fundamental variability limits of filament-based RRAM," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2016, pp. 4.7.1–4.7.4.