Limitations on Pitch Design due to Thermal Crosstalk in $Pr_{1-x}Ca_xMnO_3$ RRAM Crossbar Arrays

Ujwal Uttarwar*
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
ujwaluttarwar@gmail.com

Kunal Kaushik*
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
kaushikkunal1998@gmail.com

Jayatika Sakhuja
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
18307r013@iitb.ac.in

Vivek Saraswat
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
vsaraswat009@gmail.com

Sandip Lashkare
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
lashkaresandip@gmail.com

Udayan Ganguly
Dept. of Electrical Engineering
IIT Bombay
IITB Centre for Semiconductor Tech. (SEMIX)
Mumbai, India
udayan@ee.iitb.ac.in

***Abstract*—Networks designed with memristor crossbar arrays have been proposed for Artificial Intelligence (AI) and Machine Learning (ML)-based applications for enhanced network efficiency. The area efficiency can be achieved by increasing the density of the crossbar structures (reducing the pitch) while parallel reading multiple memory units can improve the computational performance of networks. When using emerging memories such as Resistive Random-Access Memories (RRAMs) in crossbar architectures, the variability in the memristive weights will impact the network accuracy. One of the sources of weight perturbation is heat generation and dissipation while programming the devices. The phenomenon is famously known as thermal crosstalk. In this work, we have proposed limitations on array pitch design considering thermal crosstalk using an experimentally calibrated TCAD model for $Pr_{1-x}Ca_xMnO_3$ (PCMO) based RRAM arrays. 21% and 48% error for 25nm pitch in computing Multiply and Accumulate (MAC) output is observed for a five-device network for the same and different conductance states respectively.**

***Keywords*—PCMO, RRAM, thermal crosstalk, crossbar arrays, MAC**

I. Introduction

In the era of artificial intelligence and machine learning, integrated memristive crossbar arrays (Fig.1a) have been proposed for applications such as pattern recognition and speech processing to enable area-efficient hardware implementation [1]. These crossbar architectures enable vector-matrix-multiplication (VMM) via multiply-and-accumulate (MAC) operations as shown in Fig.1b, which are fundamental computing tasks [2]. However, the accuracy of the MAC operations is sensitive to the variations in memristive devices [3]. The resistive switching in memristor devices is due to ionic transport and joule-heating phenomena [4]. One of the issue causing device variability is thermal crosstalk, where programming one cell generates heat that distributes to the neighboring cells. This effect is exacerbated by (a) increasing network density (reducing pitch) and (b) parallel read/write operation, which can affect the network accuracy as shown in Fig.1c,d [5].

This study focuses on thermal crosstalk in $Pr_{1-x}Ca_xMnO_3$ (PCMO) based memristor arrays. Here, we demonstrate the impact of the thermal crosstalk on MAC using an RRAM array using TCAD simulations with experimentally calibrated models. First, the individual device is calibrated with the experimental results. The device temperature level and distribution are analyzed over the device area. Second, multiple devices are connected together with a common top electrode (word line) akin to crossbar architecture, and temperature distribution is analyzed for program operation. Finally, MAC operation is performed by current extraction with pitch variation and biasing effects. This is compared with the ideal MAC output without thermal crosstalk and a MAC error of 21% and 48% is observed for 25nm pitch in a five-device network for the same and different conductance states which is high enough to affect network accuracy.

II. Device Details and Experimental Calibration

A. Device Details

The fabricated stack and fabrication process of PMO (x=0) based RRAM is shown in Fig.2a. The platinum (Pt-70nm) metal acts as the noble bottom electrode while Tungsten (W-70nm) is a reactive top electrode. To crystallize PMO(64nm) layer, the sample is annealed at 750°C in an N_2:O_2

This work was supported by the Department of Science and Technology (DST), Prime Minister Research Fellowship (PMRF) and Ministry of Electronics and Information Technology (MeitY). All the authors are with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, 400076, India. (e-mail:udayan@ee.iitb.ac.in)

*Ujwal Uttarwar and Kunal Kaushik contributed equally to this paper.

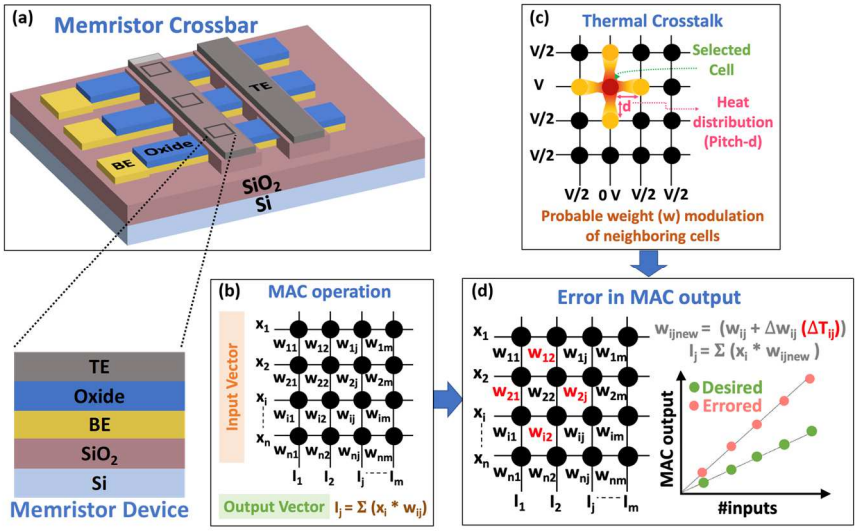


Fig. 1. **Motivation:** (a) Memristor crossbar arrays for accelerated VMM via MAC operations. (b) Typical crossbar structure showing MAC operation in the ideal scenario (absence of thermal crosstalk). (c) Illustration showing tentative heat distribution in neighboring cells on programming/reading a cell in the crossbar. (d) Pitch-dependent thermal crosstalk can produce errors in MAC output. The error tends to increase with the number of inputs.

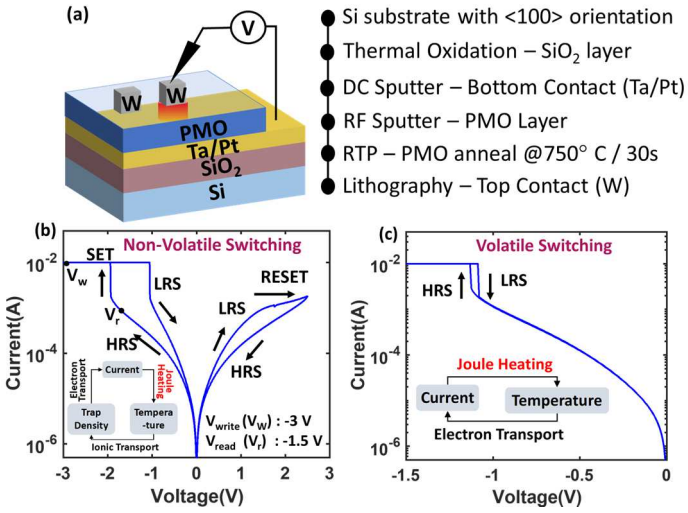
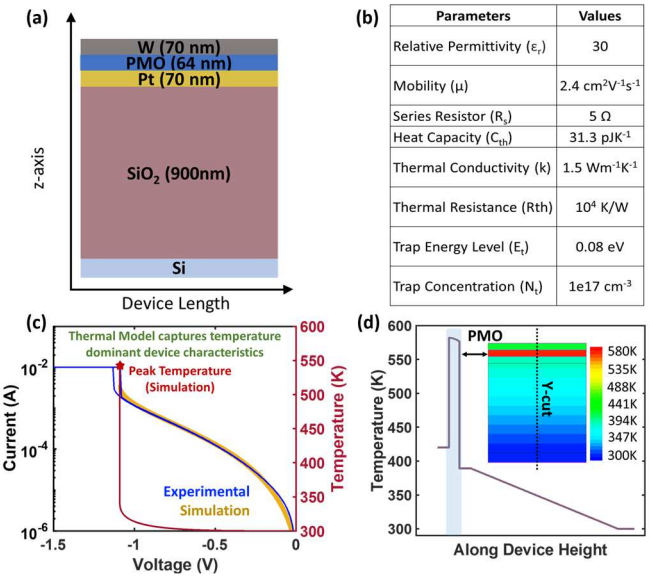


Fig. 2. **Experiment: Joule heating** in PMO RRAM device voltage (a) Device schematic and fabrication process. Heating facilitated switching characteristics (b) Non-Volatile and (c) Volatile

environment [6]. The device's active area of 10x10 μm^2 is defined by photolithography and lift-off. Fig.2b shows the bipolar non-volatile switching behavior of PMO RRAM with RESET process i.e. resistive switching from a low resistance state (LRS) to a high resistance state (HRS) in the positive bias voltage and SET process, i.e. HRS to LRS state change in negative bias voltage [7]. The volatile switching owing to intrinsic self-heating in PMO RRAM in LRS is shown in Fig.2c [8].

In the SET switching, the current starts to rise as the voltage increases which further increases the temperature within the device. Due to the low thermal conductivity of PMO material, heat dissipation is low. The accumulated heat, triggers a current-temperature positive feedback, resulting in a sudden sharp shoot-up in current and temperature [9].

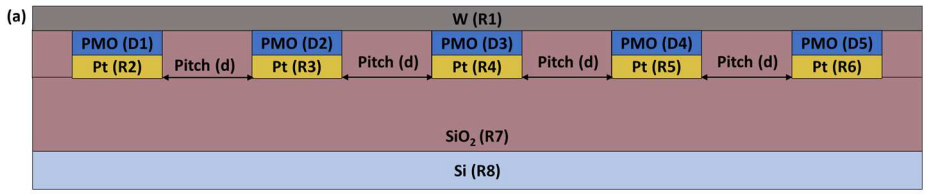
B. Simulation Structure and Experimental Calibration



Parameters	Values
Relative Permittivity (ε_r)	30
Mobility (μ)	2.4 $cm^2V^{-1}s^{-1}$
Series Resistor (R_s)	5 Ω
Heat Capacity (C_{th})	31.3 pJK^{-1}
Thermal Conductivity (k)	1.5 $Wm^{-1}K^{-1}$
Thermal Resistance (Rth)	10^4 K/W
Trap Energy Level (E_t)	0.08 eV
Trap Concentration (N_t)	1e17 cm^{-3}

Fig. 3. **Experimentally calibrated TCAD Thermal Model:** (a) Simulated device structure (b) Simulation Parameters (c) Calibrated LRS-IV (d) Temperature distribution at the point of peak temperature. Maximum temperature in PMO region in LRS state.

The thermal dynamics due to self-heating in PMO device is modeled in Sentaurus device-based TCAD software. The simulated device structure is shown in Fig.3a. The Poisson and thermodynamic models were solved throughout the device, and the Peltier model was used to account for heating effects in the interface region. The model is accurately calibrated with



(b)

Bias Scheme				
Operation	Selected Cell		Unselected Cell	
	W (TE)	Pt (BE)	W (TE)	Pt (BE)
Read	V_{read} (-1.5V)	0	V_{read} (-1.5V)	$V_{read}/2$ (-0.75V)
Write	V_{write} (-3V)	0	V_{write} (-3V)	$V_{write}/2$ (-1.5V)

Fig. 4. (a) **Simulated array structure** with common W Top Electrode (WL) (b) Biasing scheme for read and write operations

the experimental DC-IV characteristics of the PMO device with effective simulation parameters (Fig.3b,c). Fig.3d shows the temperature distribution along the device. Further, the array structure with five PMO devices is simulated with a common top electrode as a word line and separate bottom electrode (bit lines) as shown in Fig.4a. The biasing schemes for read/write operations to select array cells are shown in Fig.4b.

III. Results

Here, different simulations have been performed to analyze the problem of space constraint- pitch dependent density of arrays and time constraint- parallel memory read defining network speed. For the space constraint study, a single device centered in 5 cell array structure was programmed with the write voltage from the biasing scheme in Fig4b, and a temperature profile along the array length for different pitch lengths was observed (Fig5). For the time constraint study, current through multiple devices (having the same as well as different conductance states) was read and compared with the ideal case of no thermal crosstalk (Fig6,7).

A. Sneak Path Heating with Single-Device Programmed

All the devices in the array simulation are in LRS. To analyze the sneak path heating, the center device (D3) in the array structure is programmed by applying V_{write} voltage, while $V_{write}/2$ is applied across the unselected cells. The temperature maps across array length for varying pitch (700nm to 100nm) are shown in Fig.5a. With the decrease in array pitch, the current and peak temperature of the programmed device increases (Fig.5b).

As the devices are in LRS, the programming currents are high, generating more heat and dissipation to neighboring cells via metal lines. The dissipated heat enhances currents in neighboring cells with V/2 voltage drop resulting in more heat generation, hence sneak path heating. The temperature rise in the neighboring cells increases as pitch decreases (Fig.5c). The temperature feedback can result in a state change in the unselected cell. The error in the state of the device will then reflect in array applications like MAC operations.

B. Parallel Multi-Device Read for MAC Operation

Fig.6 shows the multi-device read operation for parallel computations. For a given pitch, the heating increase when devices are programmed parallelly as shown in Fig.6a. Here there are two different currents contributing, one is the enhanced current due to thermal crosstalk (for lower pitch) and

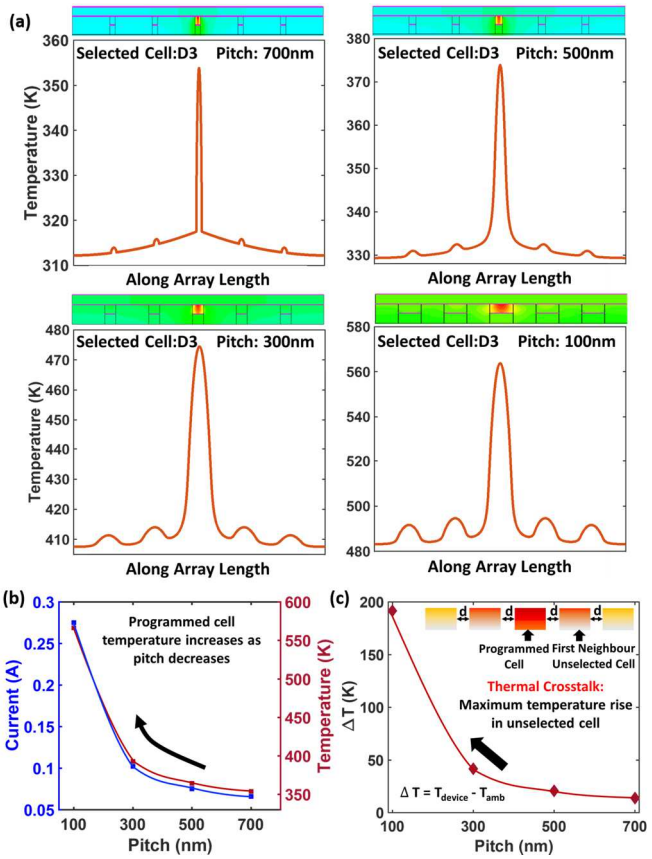


Fig. 5. **Sneak path heating in Array - Pitch Variation (Array Density)** (a) Temperature Maps for varying pitch when the single center device is programmed (b) Extracted peak temperatures of the programmed cell. Peak currents increase with decreasing pitch (c) Temperature rise in neighboring cells due to thermal crosstalk and sneak path currents. Maximum temperature rise for lowest pitch.

the other is the sneak path current due to V/2 voltage drop. The ideal MAC output is obtained from an isolated device structure (PMO with 700nm SiO2 on the sides). For single-device comparison sneak path current addition dominates over thermal crosstalk, thus irrespective of pitch, current levels are equally deviated from ideal. As the number of devices increases, the curve either tries to go away from the ideal (low pitch values) because of thermal crosstalk domination or towards the ideal value (high pitch values) because of a reduction in the number of devices contributing to sneak path current (V/2 drop). Thus the trend shown in Fig.6b. The analysis demonstrates that for a 25nm pitch, the error for

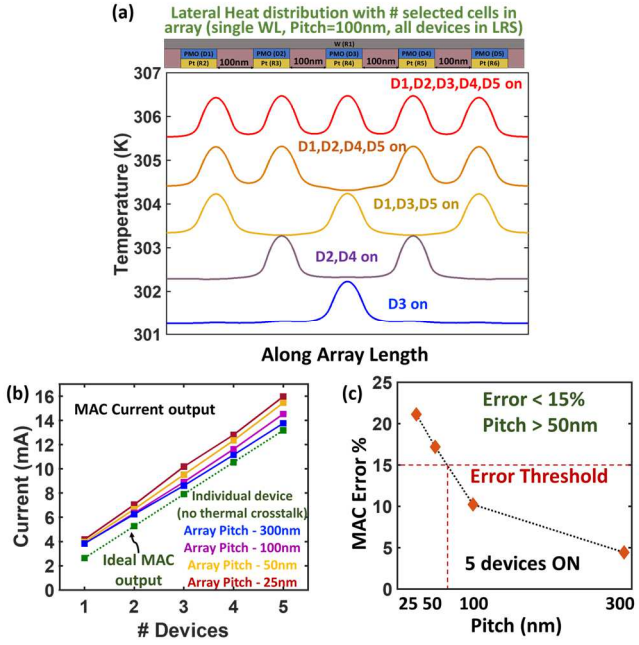


Fig. 6. **Parallel Read for MAC (Same conductance states)** (a) Temperature distribution along array length (b) MAC current output variation with number of devices read in parallel, for different pitch (c) MAC Error

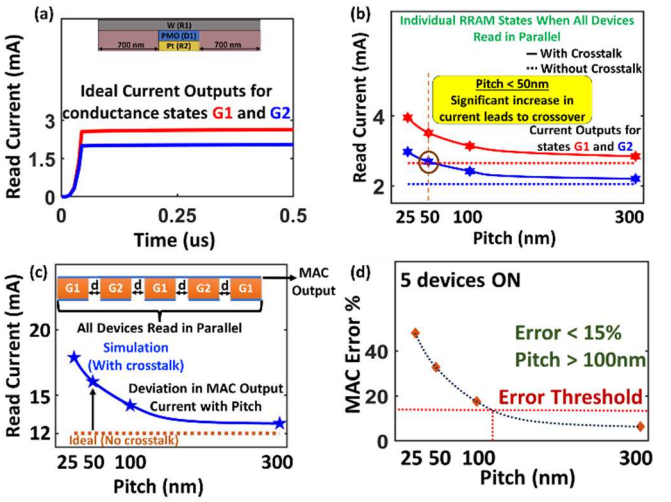


Fig. 7. **Parallel Read for MAC (Different conductance states)** (a) Current output from isolated RRAM with weight stored as G1 and G2 (b) Individual device currents when all devices are read in parallel (c) Effect of thermal crosstalk with pitch variation when all devices are read in parallel (d) MAC Error

programming five devices is maximum (21%). This puts a limit on the array density to minimize computational error. To achieve an error < 15%, the pitch should be more than 50nm (Fig.6c).

Another such simulation was performed, in which devices (as per the structure in Fig.4a) D1, D3, and D5 were at conductance state G1 and D2, and D4 were at conductance state G2. Fig.7a shows the ideal currents corresponding to states G1 and G2, for an isolated device case (inset of Fig.7a). While reading the total MAC current output, individual device currents were tapped for devices D2 and D3 (D1 and D3 had the same current levels with an error of max 2%). It was found out that, currents crossover each other when the pitch is reduced beyond 50nm as shown in Fig.7b. Fig.7c shows the deviation in MAC current output when all five devices, in mentioned states, are read in parallel. The worst-case scenario is observed when all the devices are read in parallel, for different conductance states values. Fig.7d shows limitations on pitch design for the scenario that achieving an error less than 15% requires pitch to be > 100nm.

IV. Conculsion

The study shows that metal lines conduct heat, and V/2 biasing produces sneak path heating in LRS cells for writing as the $V_{write}/2$ sneak current is high. Therefore, programming a device in arrays can cause state variability in neighboring cells. On the other hand, parallel reading that speeds up MAC operation produces increased heating with higher parallelism. This affects the temperature-dependent analog current output. Hence increase in parallel read trades off with MAC error.

Such trade-off requires mitigation strategies. As heating is an issue, pitch selection based on heating considerations is crucial. Hence, the thermal cross-talk and its effective mitigation strategies will determine pitch selection and, thereby array density.

References

[1] Yao, P., Wu, H., Gao, B. et al. Fully hardware-implemented memristor convolutional neural network. Nature 577, 641–646 (2020).

[2] Xia, L., Gu, P., Li, B. et al. Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication. J. Comput. Sci. Technol. 31, 3–19 (2016).

[3] A. P. James and L. O. Chua, "Variability-Aware Memristive Crossbars—A Tutorial," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 6, pp. 2570-2574, June 2022.

[4] N. Panwar, A. Khanna, P. Kumbhare, I. Chakraborty and U. Ganguly, "Self-Heating During submicrosecond Current Transients in Pr0.7Ca0.3MnO3-Based RRAM," in IEEE Transactions on Electron Devices, vol. 64, no. 1, pp. 137-144, Jan. 2017.

[5] Sun, Pengxiao, et al. "Thermal crosstalk in 3-dimensional RRAM crossbar array." Scientific reports 5.1 (2015): 1-9.

[6] S. Lashkare, J. Sakhuja and U. Ganguly, "Voltage Scaling in Area Scalable Selector-Less PrMnO3 RRAM by N2: O2 Partial Pressure Dependent Annealing," 2019 IEEE 9th International Nanoelectronics Conferences (INEC), Kuching, Malaysia, 2019, pp. 1-5.

[7] S. Lashkare, S. Chouhan, T. Chavan, A. Bhat, P. Kumbhare and U. Ganguly, "PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks," in IEEE Electron Device Letters, vol. 39, no. 4, pp. 484-487, April 2018.

[8] J. Sakhuja, S. Lashkare, K. Jana and U. Ganguly, "Investigating Transient Characteristics of Volatile Hysteresis and Self-Heating of PrMnO3-Based RRAM," in IEEE Transactions on Electron Devices, vol. 67, no. 12, pp. 5520-5525, Dec. 2020.

[9] J. Sakhuja, S. Lashkare, V. Saraswat and U. Ganguly, "Thermal Engineering of Volatile Switching in PrMnO3 RRAM: Non-Linearity in DC IV Characteristics and Transient Switching Speed," 2020 Device Research Conference (DRC), Columbus, OH, USA, 2020, pp. 1-2.