# Sub-Crosspoint RRAM Decoding for Improved Area Efficiency

Ravi Patel and Eby G. Friedman
Department of Electrical and Computer Engineering
University of Rochester
Rochester, New York 14627
Email: (rapatel, friedman)@ece.rochester.edu

***Abstract*—Two sub-crosspoint physical topologies are proposed that places the decode circuitry beneath the metal-oxide RRAM crosspoint array. The first topology integrates only the row decode circuitry, while the second integrates both the row and column decoder. The topology for sub-crosspoint row decoding reduces area by up to 38.6% over the standard peripheral approach, with an improvement in area efficiency of 21.6% for small arrays. Sub-crosspoint row and column decoding reduces the RRAM crosspoint area by 27.1% and improves area efficiency to nearly 100%.**

***Index Terms*—RRAM, resistive memory, memristors, non-volatile**

## I. Introduction

RESISTIVE random access memory (RRAM) is an emerging memory technology poised to replace flash memory as the workhorse for high density solid-state storage. Metal-oxide RRAM is a two terminal resistive device where the steady state resistance is modulated by a voltage or current to store information. Unlike charge based memories, RRAM stores information by modulating the chemical structure of a thin film oxide. As a result, an RRAM cell dissipates no power to retain state and is immune to radiation induced soft errors. Fabrication of these devices can require as little as a single lithographic step [1] and is materials compatible with CMOS [2]. These devices are typically fabricated between metal layers in a fashion similar to interlayer metal vias. Unlike flash memory, which suffers from charge storage issues, the available lithographic feature size (F) is the primary limitation to device density.

For high density applications, physical area is of paramount importance. A standard approach in semiconductor memory is to place the access circuitry, such as the decoders and sense amplifiers, peripherally around the memory cells. The RRAM devices, however, are integrated into the metal layers without using the silicon area beneath the array.

Two topologies are proposed that integrate RRAM within the intermediate metal layers, where the decode circuits are placed beneath the array (which is called here, sub-crosspoint decoding). The peripheral row and column decode circuits are integrated beneath the crosspoint array by introducing crosspoint gaps, and by vertically and horizontally staggering contacts to the rows and columns. A topology where only the row decode circuitry is placed underneath the array exhibits 38.6% reduction in area for a single array with a 21.6% improvement in array efficiency. A second topology, with sub-crosspoint placement of both the row and column decoders, reduces the area of large RRAM crosspoint arrays by 27.1% and improves area efficiency to nearly 100%.

Background on RRAM and crosspoint memories are reviewed in Section II. The physical topology of the RRAM crosspoint array is described in Section III. The proposed topology is evaluated and compared to standard peripheral approaches in Section IV, and some conclusions are offered in Section V.

## II. Background

RRAM memories exhibit different behavior than standard CMOS SRAM and DRAM arrays. RRAM is composed of two terminal devices that enable memory to be configured as a crosspoint array. The electrical behavior and physical structure of RRAM and crosspoint arrays in general are outlined in the subsecuent sections followed by a discussion of related work.

### A. RRAM devices

RRAM devices are typically based on thin film transition metal oxides (*e.g.,* TaO, TiO, HfO, SiO) [3]–[6] with dopants in the form of oxygen vacancies. In the off state, these devices act as traditional metal-insulator-metal tunnel barriers. Under a large applied bias, these oxygen vacancies migrate to form electrical conduction paths through the insulating oxide, changing the resistance of the tunnel barrier. This effect occurs by either changing the effective insulator thickness, or

This research is supported in part by the Binational Science Foundation under Grant No. 2012139, the National Science Foundation under Grant No. CCF-1329374, and by grants from Qualcomm, Cisco Systems, and Samsung.

by providing a resistive short between the two device electrodes.

Migration of these vacancies gives rise to a change in the instantaneous resistance of the device. A positive applied voltage decreases the resistance, while a negative applied voltage increases the resistance. At low voltage bias, this change can be measured without perturbing the resistance state. An advantage of RRAM over other resistive memories is that the resistance can be continuously tuned between a maximum ($R_{off}$) and minimum ($R_{on}$) resistance [7], allowing multiple bits of information to be stored within a single memory cell to improve the effective bit density [8].

### *B. Nonlinear crosspoint array*

RRAM and other memristive devices have been proposed for use in crosspoint arrays. Crosspoints arrays achieve a high cell density by integrating an RRAM device at the intersection of perpendicular metal lines on adjacent metal layers, as illustrated in Fig 1.

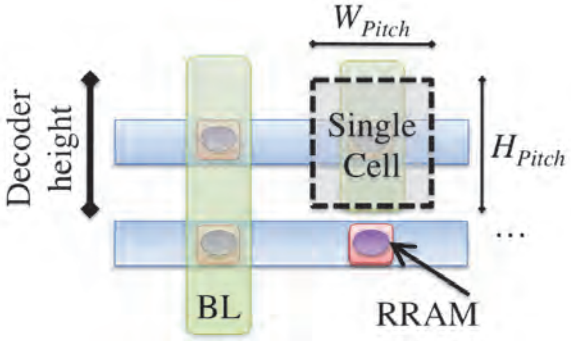

Fig. 1. RRAM crosspoint array

An individual bit is selected by biasing a row and grounding a column within an array. Selecting a single row produces a voltage drop across the unselected rows and columns. This effect produces parasitic sneak currents that propagate through unselected cells, causing a degradation in sense margin and an increase in power consumption [9]. These currents prohibit the use of RRAM-only crosspoints in all but the smallest arrays [10]. Larger arrays utilize a selector device (*e.g.,* a tunneling barrier) in series with the RRAM to ensure that only a small (leakage) current is passed through the unselected rows [11]. Unlike traditional CMOS memories, crosspoint memories also need to be bit addressable. Only a single bit can be written into a crosspoint array during a write operation due to the resistive load of the bit lines in large arrays. This characteristic requires additional area for the peripheral circuitry.

### *C. Related work*

Recently, Liu *et al.* [12] demonstrated a vertically integrated RRAM crosspoint memory that integrates the peripheral access circuitry beneath the RRAM array. The approach places the column and row segmenting circuitry as well as the driver circuitry beneath the array, while placing the column decode circuitry peripheral to the array. The topology proposed here avoids bit line segmentation and integrates both the column and decode circuitry beneath the array and is compatible with the approach described in [12]. Expressions are provided in this paper to size the array according to the physical dimensions of the decoder to ensure that the decode circuitry is beneath the RRAM crosspoint array. Niu *et al.* [13] provide an area and cost model that supports placing the driver circuits beneath the crosspoint array.

## III. Physical design of RRAM crosspoint array

The area efficiency of a memory is the portion of the IC composed of the memory cells as compared to the total area of the memory system including all of the peripheral circuitry. Memories typically exhibit array efficiencies ranging from 30% to 40% for deeply scaled technologies. Only a fraction of the total die area is therefore dedicated to data storage. Higher array efficiencies increase memory capacity without additional die area.

CMOS memory arrays rely on pitch matching of the peripheral circuits, such as the decoders and sense amplifiers, to the width of the corresponding row or column. The height of a row decoder is equivalent to the height of a cell. In a minimum sized RRAM technology, however, the dimensions are significantly smaller (see Fig 1), making pitch matching difficult. The height of a crosspoint cell is 2F to 3F but the minimum height and length of a transistor is generally more than 3F before considering interconnect. The peripheral decode circuitry is therefore placed with staggered interconnect to drive an individual row or column, increasing the area of an array, as illustrated in Fig. 2.

RRAM crosspoint arrays, however, are fabricated within the metal layers and do not utilize the silicon area beneath the memory array. The proposed topologies embed the decoding circuitry beneath the crosspoint array to reduce area, thereby increasing the area efficiency.

The key idea of the proposed topologies is to place the decode circuitry within a grid, beneath the crosspoint array, and to stagger the contacts to ensure that each decode block connects to a single row, as illustrated in Fig. 3. Intuitively, the height of a decoder can be hidden across multiple rows and the width of the decoder can be hidden beneath columns. This structure creates a grid of sub-crosspoint decoders beneath the crosspoint array.

Furthermore, gaps are introduced into the array interconnect to improve area efficiency. Despite the slight reduction in cell density, the overall area of an array is reduced. These gaps are strategically introduced into

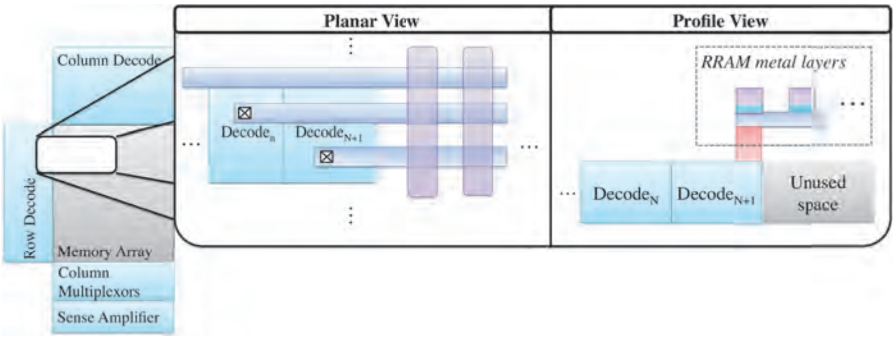

Fig. 2. Planar and profile view of peripheral RRAM crosspoint array.

the array to facilitate access to the columns for column decoding with minimal effect on the physical area.

The decoding circuit used for both topologies is described in Section III-A The proposed topology for sub-crosspoint row decoding is described in Section III-B followed by the topology for sub-crosspoint row and column decoding in Section III-C.

### A. NOR decoder circuit

A NOR-style decoder [14], commonly used in DRAM circuits, provides row and column decoding for both topologies and is shown in Fig. 4. The decoder is modified for resistive memories. The selection transistors, required for address decoding, are shown on the left. The driver circuitry for reads and writes are shown on the right. If all of the inputs are low, indicating a match, the decoding node is pulled high. If either $R_{en}$, $W_{en\_h}$, or $W_{en\_l}$ is enabled, the address is valid and the row or column is driven. The write enable signals ($W_{en\_h}$ and $W_{en\_l}$) are connected to the high and low voltage drivers to enable the bidirectional writes necessary for bipolar RRAM devices.

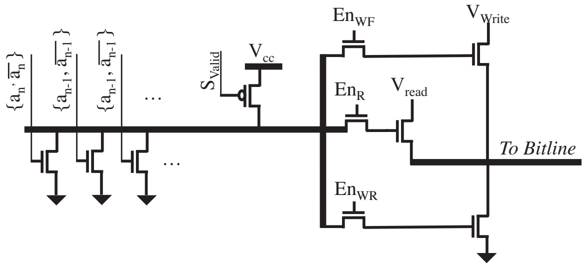
$\{a_n, \overline{a_n}\}$
$\{a_{n-1}, \overline{a_{n-1}}\}$
$\{a_{n-1}, \overline{a_{n-1}}\}$
...
$S_{Valid}$
$V_{cc}$
$En_{WF}$
$V_{Write}$
$En_R$
$V_{read}$
To Bitline
$En_{WR}$

Fig. 4. Decoder circuit

### B. Sub-crosspoint row decoding

A sub-crosspoint row decoder is constrained by the physical size of the minimum sized transistor, size of the output driver, metal routing for the power and ground lines, and the number of columns and rows within the array. The transistor size and metal lines constrain the height of the decoder. The number of columns or rows determines the number of transistors required for decoding, which, in addition to the size of the output driver, determines the width of the decoder.

The height of a decoder is amortized across multiple rows, as illustrated in Fig 3. If the height of a decoder is $H$, $\frac{H}{k}$ decoders are placed side-by-side beneath a crosspoint array, where $k$ is the minimum metal pitch of the technology. For binary decoding, the number of rows is a power of two. The number of rows ($N_{row}$) also indicates the number of decoders placed in parallel, as each decoder connects to a single row. Placing decoders horizontally beneath a crosspoint array allows the global predecoding circuitry to drive a column of decoders, as depicted in Fig. 3.

The number of predecode bits is

$$N_{r\_predec} = \lceil log_2(\frac{H_{sub} + H_{routing}}{H_{c\_pitch}}) \rceil, \qquad (1)$$

where $H_{dec}$ and $H_{routing}$ are, respectively, the height of the decode and routing lines normalized to the feature size of the technology. Hence, $2^{N_{r\_predec}}$ is the number of rows required to "hide" a row decoder. The physical width of a row is

$$W_{row} = 2^{N_{r\_predec}} W_{r\_sub}, \qquad (2)$$

where

$$W_{r\_sub} = W_{drive} + 2W_{tr} log_2(N_{row}) - N_{r\_predec}. \qquad (3)$$

$W_{r\_sub}$ is the width of a single row decoder, $W_{drive}$ is the width of a row driver circuit, $N_{row}$ is the number of rows within an array, and $W_{tr}$ is the width of the selection transistor within the decoder. This expression provides the minimum width of a row. The number of columns to maximize the density of this approach is $\frac{W_{row}}{W_{c\_pitch}}$.

### C. Sub-crosspoint row and column decoding

Placing both the row and column decode circuitry beneath the crosspoint array creates an interdependence

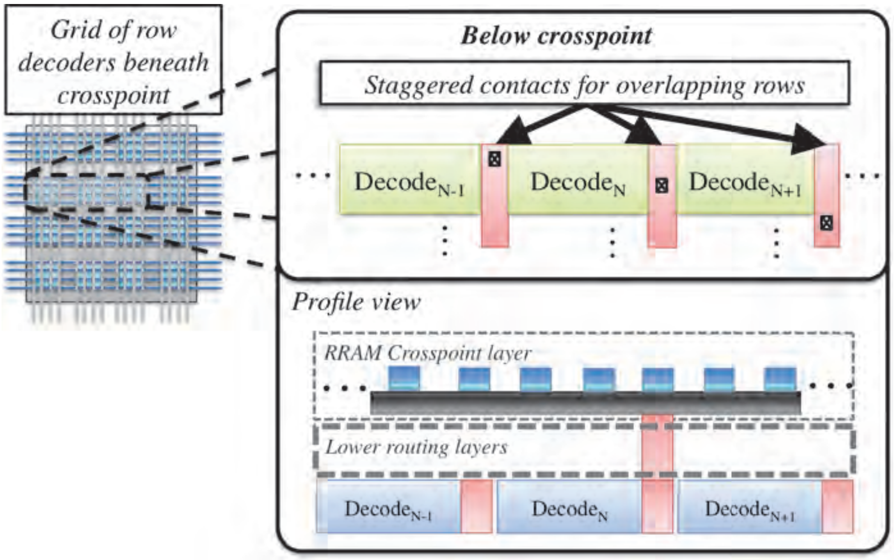

Fig. 3. Planar and profile view of proposed RRAM sub-crosspoint row decoders.

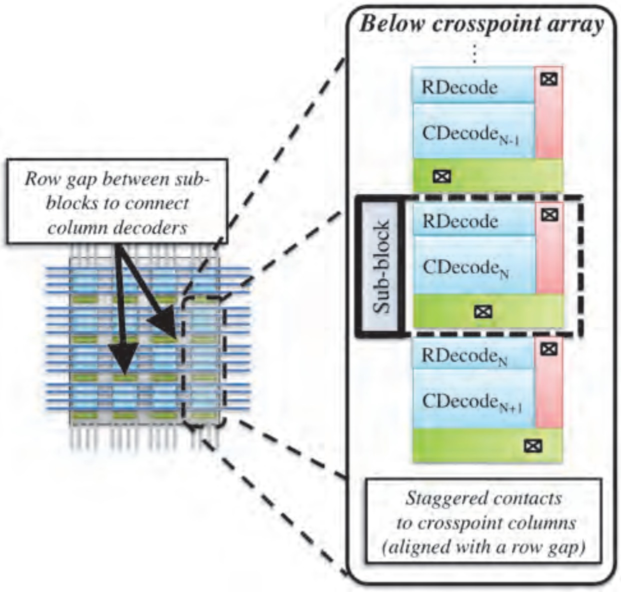

Fig. 5. Physical topology of sub-crosspoint row and column decoders.

between the number of rows and columns. The rows of an RRAM crosspoint array are located directly above the silicon, permitting access from beneath. The row plane of the crosspoint, however, blocks access to the column plane of the crosspoint from beneath. A gap is therefore introduced between the rows to enable the sub-crosspoint decoder to communicate with the column rows. The gap between individual rows provides access to the crosspoint columns using the same metal layer as the row layer.

The physical topology of the sub-crosspoint decoder for both columns and rows is shown in Fig. 5. A decode sub-block consists of a co-located row and column decoder. The sub-blocks are oriented in a grid pattern beneath the crosspoint array. Contacts are staggered horizontally for rows and vertically for columns, as illustrated in Fig. 5. The column decoder is placed below the row decoder to share the power rails with the row decoder, and to ensure that the shape of a row and column decode sub-block is as close as possible to a square.

Completely hiding a sub-block requires $2^{N_{r\_predec}}$ rows. If the same methodology is applied to a column decoder, $2^{N_{c\_predec}}$ columns are required. The number of rows and columns of an array is $2^{N_{r\_predec}+N_{c\_predec}}$, ensuring that the array has an equal number of rows and columns.

Expression (1) is used to determine the number of row predecode bits based on the height of a sub-block. The required number of column predecode bits is

$$N_{c\_predec} = \lceil log_2(\frac{W_{r\_sub}}{W_{c\_pitch}})\rceil, \tag{4}$$

where

$$W_{r\_sub} = W_{drive} + 2W_{tr}N_{r\_predec}. \tag{5}$$

The height of an array is $2^{N_{c\_predec}}(2^{N_{r\_predec}}H_{c\_pitch}+1)$, and the width is $2^{N_{r\_predec}}2^{N_{c\_predec}}W_{c\_pitch}$.

Assuming the same pitch for both the crosspoint rows and columns, the total area of a memory array with sub-crosspoint decoder circuitry is

$$A = 2^{N_{c\_predec}+N_{r\_predec}}H^2_{c\_pitch}(2^{N_{r\_predec}}+1). \tag{6}$$

Note that (1) and (4) contain integer ceiling functions,

TABLE I
PARAMETERS

| | |
|---|---|
| Feature size (F) | 22 nm |
| Metal pitch | 3F |
| Routing metal layers | 1 to 2 |
| Crosspoint metal layers | 3 to 4 |
| $R_{on}$ | 34.9 kΩ |
| RRAM write voltage | 3 V |
| Tunnel barrier thickness | 1.15 nm |
| Tunnel barrier bandgap | 0.6 eV |

ensuring that the decode circuitry occupies less planar area than the crosspoint array. This constraint as well as the $2^n$ growth of the columns and rows with predecode bits produces unused space beneath the sub-block. This space can be utilized to increase the write drivers and reduce the resistive load. Note that these expressions produce a unique array size that is ultimately dependent on the height of the sub-block. The array is therefore no longer a function of the number of rows, as in row-only sub-crosspoint decoding.

## IV. EVALUATION

Sub-crosspoint decoding is evaluated and compared to standard peripheral approaches in the following section. Note that this evaluation only considers the array efficiency of individual memory arrays and does not consider the global logic, decoders, and routing. The cell layout is based on 45 nm FreePDK design rules and is scaled to a feature size of 22 nm. It is assumed that an additional intermediate metal layer is available in 22 nm technology (see Table I). The layout of the decoding circuitry is constrained to the first two metal layers (see Figs. 6 and 7).

The RRAM parameters are based on [15] and scaled to 33 nm (3F is the minimum metal pitch for the local and intermediate metal layers and defines the dimentions of the RRAM element). A Simmons tunnel barrier model [16] is used to simulate the selector device. *In lieu* of high voltage transistor models, the write driver is sized according to a 0.9 volt, 22 nm PTM model and scaled to provide double the current required to apply 3 volts to an RRAM device in the on state. No more than 10% of the resistive load is due to the bit lines to ensure that at least 3 volts are dropped across the RRAM device during a write. The area of the peripheral sense amplifiers and column multiplexors is modeled using the methodology provided in CACTI [17].

A comparison of the peripheral approach with the sub-crosspoint row decode topology for a rectangular array is listed in Table II. A rectangular array integrates as many columns as possible. For smaller array sizes, the proposed topology reduces the overall area by 38.6% and improves area efficiency by more than 20%. For large arrays with 2,048 rows, the area advantage decreases to 6.4% as the area of the array dominates the structure.

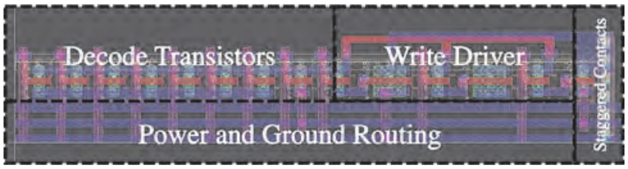

Fig. 6. Layout of row decoder in 45 nm CMOS.

TABLE II
COMPARISON OF RECTANGULAR ARRAYS

| | Peripheral array | | | Sub-crosspoint row decoder array | | |
|---|---|---|---|---|---|---|
| Number of rows | Number of column | Area ($\mu m^2$) | Array efficiency | Area ($\mu m^2$) | Area efficiency | Area reduction |
| 128 | 167 | 298.3 | 35.7% | 183.3 | 57.3% | 38.6% |
| 256 | 197 | 590.8 | 42.9% | 430.4 | 58.0% | 27.2% |
| 512 | 228 | 1,206.3 | 48.5% | 992.2 | 58.2% | 17.8% |
| 1,024 | 258 | 2,518.7 | 52.5% | 2,244.0 | 58.3% | 10.9% |
| 2,048 | 288 | 5,352.2 | 55.0% | 5,009.5 | 58.4% | 6.4% |

The area of a traditional square array with an equal number of rows and columns is listed in Table III. Similar to a rectangular array, the physical area for smaller arrays exhibits an improvement of 36.0% and 16.8%, respectively, for area and area efficiency. While the reduction in area follows a similar trend in rectangular arrays, the proposed approach maintains an area efficiency advantage unlike with rectangular arrays. This approach also demonstrates that sub-crosspoint row decoding utilizes only a small portion of the area under a crosspoint array for larger array sizes. At 2,048 columns and rows, 76% of the area under a crosspoint array is unused, permitting additional peripheral logic to be placed under the array to improve the area efficiency of larger size arrays.

The column decode circuitry can be integrated beneath the crosspoint array in the manner described in Section III. The sub-block decoder is shown in Fig. 7. As listed in Table IV, the array efficiency of this approach is nearly 100%. Moreover, a 27% reduction in area is produced.

### A. Implications of sub-crosspoint decoder on array size

Sub-crosspoint row decoding is best applied to smaller RRAM arrays, where the array size and peripheral cir-

TABLE III
COMPARISON OF SQUARE ARRAYS

| | Peripheral array | | Sub-crosspoint row decoder array | | | |
|---|---|---|---|---|---|---|
| Number of rows and columns | Area ($\mu m^2$) | Area efficiency | Area ($\mu m^2$) | Area efficiency | Sub-crosspoint unused space | Area reduction |
| 128 | 249.2 | 31.8% | 159.5 | 48.6% | - | 36.0% |
| 256 | 705.3 | 44.9% | 493.9 | 62.2% | 77.1 | 30.0% |
| 512 | 2,109.9 | 59.1% | 1,622.9 | 74.5% | 660.9 | 23.1% |
| 1,024 | 6,760.0 | 72.3% | 5,657.8 | 84.0% | 3,477.4 | 16.3% |
| 2,048 | 2,3168.3 | 82.6% | 20,707.2 | 90.5% | 15,833.5 | 10.6% |

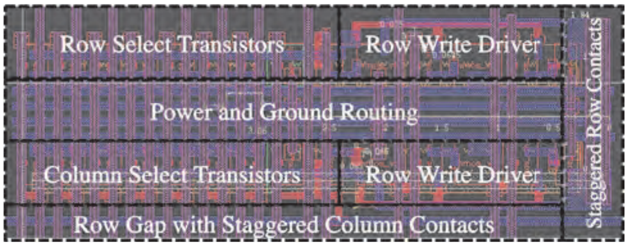

Fig. 7. Layout of decoder sub-block in 45 nm

TABLE IV
SUB-CROSSPOINT ROW AND COLUMN DECODER, SQUARE ARRAY

| Number of rows and columns | Area $\mu$m$^2$ | Area efficiency | Area reduction |
|---|---|---|---|
| 2,048 | 19,412.4 | 99.9996% | 27.1% |

cuitry are comparable. For square arrays greater than 512 x 512, the unused space beneath the array is at least 31.3% of the array area and grows as high as 76% in 2,048 x 2,048 arrays. For an array size of 256 x 256, the unused area is 11% with a 17.3% improvement in area efficiency.

The sub-crosspoint row and column decoding approach presented here presents an inflection point at an array size of 2,048 x 2,048 that achieves near 100% area efficiency. A smaller number of rows and columns exhibits lower area efficiency. While sub-crosspoint topologies produce smaller arrays than the standard peripheral approach, the reduction in area efficiency degrades the storage capacity of an individual memory. Arrays larger than 2,048 x 2,048 are dominated by the area of the crosspoint array and result in a negligible improvement in array efficiency. While additional sub-blocks are required to decode larger arrays, the area of the memory cells is larger than the area of the sub-blocks. Thus, additional unused area is available beneath the array, although with increased resistive and capacitive impedances within the crosspoint array.

## V. CONCLUSIONS

Two physical topologies for sub-crosspoint decoding of an RRAM based memory are demonstrated. The two approaches are sub-crosspoint row decoding, and sub-crosspoint row and column decoding. Expressions are provided for both topologies to size a crosspoint array as well as the column and row decode circuitry. Sub-crosspoint row decoding reduces area by up to 38.6% over the standard peripheral approach, with an improvement in area efficiency of 21.6% for small 128 x 128 arrays. For large 2,048 x 2,048 square arrays, area is reduced by 10.6% with a corresponding improvement in area efficiency of 8.0%. Sub-crosspoint row and column decoding reduces the RRAM crosspoint area by 27.1% and improves area efficiency to nearly 100%.

## ACKNOWLEDGMENT

The authors thank Shahar Kvatinsky, Avinoam Kolodny, and Uri Weiser of Technion—Israel Institute of Technology for helpful discussions and feedback.

## REFERENCES

[1] Q. Xia. *et al.*, "Self-Aligned Memristor Cross-Point Arrays Fabricated with One Nanoimprint Lithography Step," *Nano Letters*, Vol. 10, No. 8, pp. 2909–2914, June 2010.

[2] Q. Xia *et al.*, "Memristor-CMOS Hybrid Integrated Circuits for Reconfigurable Logic," *Nano Letters*, Vol. 9, No. 10, pp. 3640–3645, September 2009.

[3] Z Wei *et al.*, "Highly Reliable $TaO_x$ ReRAM and Direct Evidence of Redox Reaction Mechanism," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–4, December 2008.

[4] T. Prodromakis, K. Michelakisy, and C. Toumazou, "Fabrication and Electrical Characteristics of Memristors with $TiO_2/TiO_2$ Active Layers," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1520–1522, May 2010.

[5] H. Y. Lee *et al.*, "Low Power and High Speed Bipolar Switching with a Thin Reactive Ti Buffer Layer in Robust $HfO_2$ Based RRAM," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–4, December 2008.

[6] Y. F. Chang *et al.*, "Study of $SiO_x$-based Complementary Resistive Switching Memristor," *Proceedings of the IEEE Device Research Conference*, pp. 49–50, June 2012.

[7] F. Miao *et al.*, "Continuous Electrical Tuning of the Chemical Composition of $TaO_x$-Based Memristors," *ACS Nano*, Vol. 6, No. 3, pp. 2312–2318, February 2012.

[8] R. Patel and E. G. Friedman, "Arithmetic Encoding for Memristive Multi-Bit Storage," *Proceedings of the IEEE/IFIP International Conference on VLSI and System-on-Chip*, pp. 99–104, October 2012.

[9] M. A. Zidan *et al.*, "Memristor-Based Memory: The Sneak Paths Problem and Solutions," *Microelectronics Journal*, Vol. 44, No. 2, pp. 176–183, February 2013.

[10] S. Shin, K. Kim, and S. Kang, "Data-Dependent Statistical Memory Model for Passive Array of Memristive Devices," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 57, No. 12, pp. 986 –990, December 2010.

[11] Y. Deng *et al.*, "RRAM Crossbar Array With Cell Selection Device: A Device and Circuit Interaction Study," *IEEE Transactions on Electron Devices*, Vol. 60, No. 2, pp. 719–726, February 2013.

[12] T. Liu *et al.*, "A 130.7$-mm^2$ 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology," *IEEE Journal of Solid-State Circuits*, Vol. 49, No. 1, pp. 140–153, January 2014.

[13] D. Niu *et al.*, "Design of Cross-Point Metal-Oxide ReRAM Emphasizing Reliability and Cost," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 17–23, November 2013.

[14] B. Jacob, S. Ng, and D. Wang, *Memory Systems: Cache, DRAM, Disk*, Morgan Kaufmann, 2010.

[15] J. P. Strachan *et al.*, "State Dynamics and Modeling of Tantalum Oxide Memristors," *IEEE Transactions on Electron Devices*, Vol. 60, No. 7, pp. 2194–2202, July 2013.

[16] J. G. Simmons, "Generalized Formula for the Electric Tunnel Effect between Similar Electrodes Separated by a Thin Insulating Film," *Journal of Applied Physics*, Vol. 34, No. 6, June 1963.

[17] Hewlett-Packard Western Research Laboratory, Palo Alto, *CACTI 3.0: An Integrated Cache, Timing, Power, and Area Model*, 2001.