Design of Cross-point Metal-oxide ReRAM Emphasizing Reliability and Cost

Dimin Niu†, Cong Xu†, Naveen Muralimanohar‡, Norman P. Jouppi‡, and Yuan Xie†
†Department of Computer Science and Engineering, Pennsylvania State University
‡Hewlett-Packard Labs

***Abstract*— Metal-Oxide Resistive Random Access Memory (ReRAM) technology is gaining popularity due to its superior write bandwidth, high density, and low operating power. An ReRAM array structure can be built with three different approaches: a traditional design with a dedicated access transistor (1T1R) or an access diode (1D1R) for each cell, or an intrinsic cross-point structure (0T1R), where the metal-oxide is directly sandwiched between the horizontal and vertical wires. Each of these different structures has its advantages and disadvantages, and it is a complicated process to perform a systematic comparison of delay, energy, area, and cost of one over others for a given cell parameters set and technology.**

In this paper, we analyze both advantages and disadvantages for ReRAM arrays built in 1T1R, 1D1R, and 0T1R structures. Based on the analysis, we propose a design flow and provides key insights into architectural tradeoffs. We do this in three stages: first, we use a matrix-based mathematical model to determine the optimal array size, read/write bandwidth, and other key characteristics. This acts as input to the second stage to explore the design space of ReRAM banks and the entire chip. Finally, we estimate the chip-level cost using the area, metal layers, pin count, and cooling requirements. Using the proposed model, we also present a case study in which we compare the energy, performance, and area of a 1D1R cross-point design and a 0T1R design, and show that the 1D1R structure is more promising for a cost-driven memory design.

I. Introduction

Traditional NAND Flash and DRAM technologies are facing scaling limitations beyond the 20nm technology node. The scaling of Flash memory is impeded by several issues, including limited number of electrons stored and the cell-to-cell interference. On the other hand, DRAM cell scaling is challenged by the requirement of large capacitor to store data. Consequently, in the past few years, emerging memory technologies such as Phase-Change Random Access Memory (PCRAM), Spin-Transfer Torque Random Access Memory (STT-RAM), and Resistive Random Access Memory (ReRAM) have been studied intensively as possible alternatives for Flash or DRAM.

Recently, metal-oxide ReRAM has attracted great research attention because of its good scalability, high retention time, excellent cell density, and 3D stackable property. A metal-oxide ReRAM cell can be built with various switching materials, such as Tungsten Oxide (WOx) [1], [2], Hafnium Oxide (HfOx) [3], and Tantalum Oxide (TaOx). ReRAM prototyping chips were built and characterized with difference write/read performance, energy consumption, as well as cost. For example, a 2-layer 32Gbits ReRAM test chip has been demonstrated [4] with 24nm technology. Since this chip is designed as a Flash replacement, density is optimized at the expense of performance, with the read and write latency of the chip are 40us and 230us, respectively. Another TaOx based multi-layered ReRAM macro has been demonstrated in 2012 with 443 Mbits/s write throughput and 8.2ns access latency [5]. This prototype has a comparable performance compared to DRAM, making it a great potential to be a DRAM alternative. An important obstacle that prevent ReRAM from being a good DRAM alternative is the limited write endurance, which is defined as the total number of write cycles to that cell before the cell becomes unreliable.

However, recently published papers have already demonstrated several ReRAMs with very good endurance. For example, a HfOx-based ReRAM prototype was shown to have a good endurance of 10^{10} [3]. In addition, the ReRAM with the best endurance of 10^{12} was demonstrated by Lee *et al.* [6]. Consequently, with the benefits of fast speed, lower power, high capacity, and the relative high endurance, ReRAM technology has a good potential to replace DRAM as an alternative main memory technology. Therefore, in this paper, we focus on the design challenges of ReRAM-based main memory.

Different from DRAM technology, an ReRAM array can be built in three different structures: a traditional 1-transistor 1-ReRAM cell (1T1R) structure, 1-diode 1-ReRAM cell (1D1R) structure, and cross-point structure (0T1R). Since each of them has its own advantages and disadvantages in performance, power consumption, and cost, doing a systematic comparison of delay, energy, area, and cost of one over others for a given technology is a complicated process. For instance, using access transistors increases the array area and therefore results in higher die cost. Although replacing access transistors with diodes can reduce the area overhead of access devices, the diode-accessed ReRAM array requires extra hardware overhead to provide higher operating voltage. In addition, using diodes also increases the fabrication steps. The cross-point structure is much simpler but requires high nonlinearity for ReRAM cells to minimize sneak current and improve noise margin.

In this paper, we address several design considerations of ReRAM arrays with these three structures. We analyze the delay, energy, area, and cost of each structure and based on the analysis, we propose a design flow, which helps design space exploration and provides key insights on architectural tradeoffs. We do this in three stages: first, we use a matrix-based mathematical model to determine the optimal array size, read/write bandwidth, and other key characteristics. This acts as input to the second stage to explore the design space of ReRAM bank and the entire chip. Finally, we estimate the chip-level cost using the area, metal layers, pin count, and cooling requirement. Using the proposed model, we also present a case study in which we compare the energy, performance, and area of a 1D1R cross-point design and a 0T1R design, and show that the 1D1R structure is more promising for a cost-driven memory design.

II. Preliminaries

In this section, we describe the background, array structures, and programming methods of ReRAM technology.

A. Background of ReRAM Technology

Different resistance switching behaviors in metal-oxide materials have been observed for more than 50 years. In the past 10 years, with the innovation of simple structures, good performance, and CMOS compatible metal-oxide materials, metal-oxide ReRAM has gained

This work in supported in part by SRC grants, NSF 1218867, 1213052, and by DoE under Award Number DE-SC0005026.

significant research interests, and is considered as a good candidate among different DRAM alternatives. Different from charge-based memory technologies (such as DRAM and SRAM), an ReRAM cell uses its resistance to represent the stored information. An ReRAM cell has a very simple structure: the resistance switching material is sandwiched between two metal electrodes, which is called a metal-insulator-metal (MIM) structure. By applying an external voltage across it, an ReRAM cell can be switched between a high resistance state (HRS) and a low resistance state (LRS). According to the polarity of the programming voltage, ReRAM cells can be classified into unipolar ReRAM and bipolar ReRAM. The resistance switching of a unipolar ReRAM only depends on the magnitude of programming voltage, whereas in bipolar ReRAM, HRS-to-LRS switching (SET operation) and LRS-to-HRS switching (RESET operation) require programming voltages with opposite polarities. Compared to unipolar ReRAM, bipolar is more attractive because of its good cell characteristics, better switching uniformity, and operating margin. Therefore, most recent ReRAM prototypes adopt bipolar ReRAM as their storage cell [3], [5], [7]. In this paper, we focus on bipolar ReRAM due to its aforementioned advantages.

B. ReRAM Array Structures

An ReRAM array can be built in different forms.

- **1T1R:** A traditional 1-transistor 1-ReRAM cell (1T1R) structure requires a CMOS transistor to be integrated with the ReRAM cell as the access device. In a 1T1R structure, the cell is easy to control because of the dedicated access transistor. However, the area overhead of the access device is significant compared to the cell area. Specifically, the array size depends on the area of the access device because even the smallest access transistor ($6F^2$, where F is the minimum feature size) still occupies a larger area than that of the ReRAM cell. For example, the MOSFET-accessed ReRAM built by Sato et al. has a cell size of $15F^2$ [8]. Another prototype of HfOx based ReRAM, which also uses transistors as the access devices, has the cell size of $9.5F^2$ [3]. However, as we will show, 0T1R or 1D1R structures eliminate the area overhead of the access devices. The cell size can achieve as small as $4F^2$, and can be further reduced by 3D stacking [1], [2], [4], [5]. Since the chip cost is directly related to the chip area, the 1T1R structure is not favorable for the market driven by cost per bit. Considering the cost sensitivity of main memory design, we conclude that the 1T1R ReRAM structure is not a suitable choice for ReRAM main memory design.

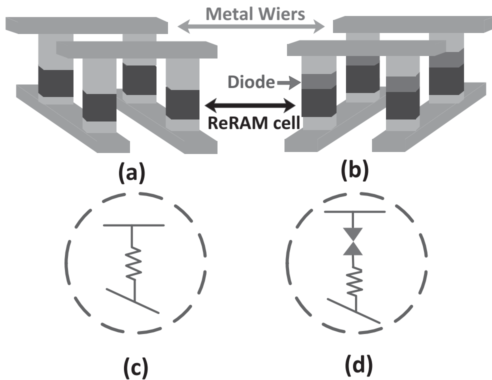

Fig. 1. Schematic view of 0T1R and 1D1R ReRAM structure: array structure of 0T1R(a) and 1D1R (b); circuit diagram of 0T1R (c) and 1D1R (d).

- **0T1R:** The cross-point structure (0T1R) eliminates the requirement of an access device for each cell. Fig. 1(a) and (c) show the schematic view of the 0T1R structure. In this structure, each ReRAM cell is directly sandwiched between two layers of metal wires, which are mutually perpendicular to each other. The function of the access device is realized by the intrinsic characteristics of the ReRAM cell, specifically the nonlinearity. The 0T1R structure occupies much less silicon area compared to 1T1R structure. Nevertheless, the nonlinearity of an ReRAM cell cannot cut off the current through the unselected cells completely, and therefore sneak currents exist in the array, which increases the energy/area overhead of the array.
- **1D1R:** The 1-diode 1-ReRAM cell (1D1R) array is also built in a cross-point structure. As shown in Fig. 1(b) and (d), different from 0T1R, a 1D1R structure contains a bidirectional diode which is built vertically as an access device in series with ReRAM cell. A 1D1R structure can reduce sneak current effectively with similar area as the 0T1R structure. Nonetheless, 1D1R requires a higher programming voltage compared to 1T1R and 0D1R structures, which increases the area overhead of charge pumps.

Considering the cost of a memory chip is directly related to the die area, the 1D1R and 0T1R structures are more cost-efficient than the 1T1R structure. Therefore, for our cost-driven ReRAM design, we focus on 1D1R and 0T1R structures. Fig. 1 also shows that 1D1R and 0T1R share a very similar structure. The only difference is that, in a 1D1R structure, a diode is built in series with the ReRAM cell. For the bipolar cell, a bidirectional diode is required to control the cell with different voltage polarities. This bidirectional diode on ReRAM cell structure is also named "one bipolar selector-one resistor (1S1R)" structure [9]. In this paper, without loss of generality, we use the term 1D1R to represent this structure for both unipolar and bipolar ReRAM. Examples of this bidirectional diode have been demonstrated by Kawahara et al. [5] and Burr et al. [10] with large current, low sneak current, and high voltage margin.

In addition to ReRAM cells, in an ReRAM array, the wordline drivers and bitline multiplexers are connected to each wordline and bitline. Fig. 2 shows a conceptual view of an $M \times N$ ReRAM array. The horizontal lines are wordlines, each of which is connected to a wordline driver located at the edge of the array. Similarly, the vertical bitlines are connected to bitline multiplexers, which are shown in the lower part of the figure. Since each wordline (or bitline) has its dedicated driver (or multiplexer), for an $M \times N$ ReRAM array, there are M wordline drivers and N bitline multiplexers in total. Note that, in the real design, part or all of wordline drivers and bitline multiplexers can be built underneath the cell array, further improving the area efficiency. Fig. 2 also shows that, in addition to the $M \times N$ ReRAM cells located at every cross-point of wordlines and bitlines, the resistance of interconnect line is also modeled as Rw in the figure.

C. Programming ReRAM Array

Reliably programming a cell in a 0T1R or 1D1R ReRAM array is not simple. These structures make it difficult to completely isolate the selected cell from all of the other unselected cells. For example, by activating a wordline associated with the selected cell, all of the unselected cells in this wordline will be biased. To mitigate this impact, a *V-by-2* method is usually adopted. Fig. 2 shows an example of the *V-by-2* programming method. In this method, to apply a write voltage (V_W) across the selected cell, which is located at the cross point of mth wordline and n^{th} bitline in this figure, we have to set voltages on wordline WLm and bitline BLn to V_W and 0,

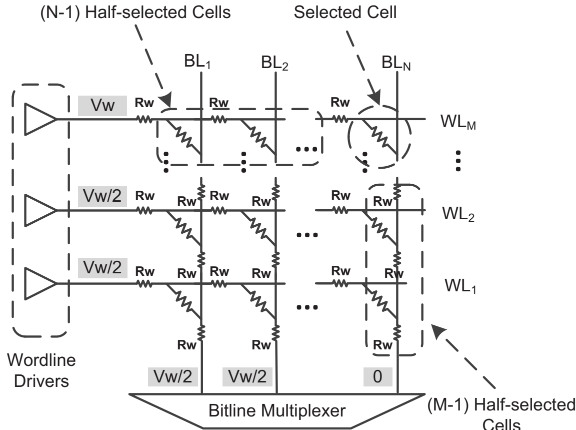

Fig. 2. Programming method of the ReRAM array.

respectively. In addition, in order to reduce the number of disturbed cells, all of the other wordlines and bitlines are biased at $V_W/2$. In this case, cells located at the selected wordline and selected bitline are biased at $V_W/2$, which are called half-selected cells, and all of the other cells have 0 voltage across them. Therefore, for an $M \times N$ array, apart from the selected cell, there are $(M+N-2)$ half-selected cells and $(MN - M - N + 1)$ unselected cells in total.

Similar to the write operations, to read a cell in the array the selected wordline is biased at the read voltage (V_R), with all of the bitlines grounded. Then the current in the selected bitlines are sent to the sense amplifier to determine the value of the stored bit.

III. Methodology

In this section, we describe the proposed cost-optimized design flow of the ReRAM design. Our flow consists of three stages: (1) building the array-level model to determine reliable array configurations under specific constraints; (2) the bank-level design space exploration to figure out the most area-efficient bank organizations; (3) building the chip-level model to evaluate the energy, performance, and cost-per-bit of the entire chip.

A. The First Stage: The Array-Level Model

In the first stage, the array-level optimization is performed. The modeling is based on the circuit shown in Fig. 2. The basic circuit model is built upon Kirchhoff's Current Law (KCL) similar to the work in [11], [12] and its validity can be guaranteed by deductions from the basic circuit theory. The model considers the impact of both the cell characteristics and the interconnection wires on the ReRAM array design and is capable of calculating the detailed DC parameters of each cross point in the array.

First of all, since this is a cost-driven optimization, the area efficiency is a crucial parameter. We define the area efficiency of an ReRAM array as

$$\eta_{Area} = A_{cell}/A_{array}, \tag{1}$$

where A_{cell} is the area occupied by ReRAM cells and A_{array} is the total area of the array, including the area of wordline drivers and bitline multiplexers. Thus, we have

$$A_{array} = A_{cell} + A_{driver} = A_{cell} + (A_{wl} + A_{bl}), \tag{2}$$

where the total driver area A_{driver} includes the area of wordline drivers (A_{wl}) and the area of bitline multiplexers (A_{bl}). Note that, the size of a wordline driver or a bitline multiplexer increases proportionally to the increase of maximum driving current requirement [13].

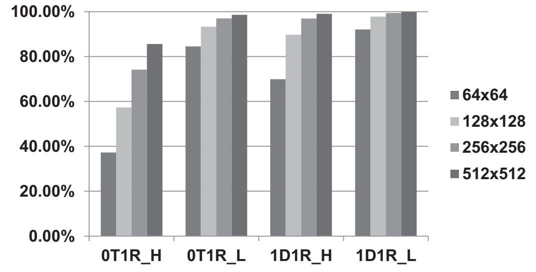

Fig. 3. Area efficiency of cross-point array.

Besides, the write current of ReRAM cell is always much larger than read current. Therefore, the total drivers area A_{driver} depends not only on the array size, but also on the worst case write current requirement of the wordlines and bitlines.

To evaluate the impact of wordline drivers and bitline multiplexers in area efficiency, we start from an ideal cross-point array. In an ideal array, we assume the resistance of interconnect wires and sneak current are negligible.

Therefore, for an $M \times N$ array, the area efficiency can be calculated as

$$\begin{aligned}\eta_{Area} &= \frac{MNA_{c0}}{MNA_{c0} + \alpha MA_{w0} + NA_{b0}} \\ &= \frac{A_{c0}}{A_{c0} + \alpha A_{w0}N^{-1} + A_{b0}M^{-1}},\end{aligned} \tag{3}$$

where A_{c0} is the area of a single cell. A_{w0} is the minimum area of a wordline driver that can provide enough write current for writing a single cell and A_{b0} is the minimum area of bitline multiplexer for a single cell write. α is the maximum number of cells that can be modified during each write operation. Clearly, in the ideal case, increasing either M or N will improve the area efficiency of the array.

If we consider the sneak current at half-selected cells, the area efficiency should be recalculated as

$$\begin{aligned}\eta_{Area} &= \frac{MNA_{c0}}{MNA_{c0} + [\alpha + \frac{(N-\alpha)}{Kr}]MA_{w0} + [1 + \frac{M-1}{Kr}]NA_{b0}} \\ &= \frac{A_{c0}}{A_{c0} + Kr^{-1}(A_{w0} + A_{b0}) + \alpha(1 - Kr^{-1})A_{w0}N^{-1} + (1 - Kr^{-1})A_{b0}M^{-1}}\end{aligned} \tag{4}$$

where Kr is the nonlinearity coefficient of the cell, which is defined as the ratio of current at the selected cell to the current at the half-selected cell. Since the current of half selected cell is always smaller than the current of selected cell, the nonlinearity coefficient Kr is always greater than or equal to 2. Therefore, η_{Area} in Equation (4) is monotonically decreasing with the increase of array size.

To demonstrate our conclusion, Fig. 3 shows the area efficiency of two 0T1R arrays and two 1D1R arrays. For the 0T1R structure, we choose two ReRAM cells with lower nonlinearity and higher nonlinearity. The cell with lower nonlinearity has a higher sneak current and is therefore denoted as 0T1R_H. On the other hand, the cell with high nonlinearity results in very small sneak current and is denoted as 0T1R_L. For the 1D1R structure, the sneak current is determined by the bidirectional diodes. Bidirectional diodes with higher sneak current (1D1R_H) and lower sneak current (1D1R_L) are used. Clearly, the area efficiency increases when the array size increases, which verifies our statement that cost-efficient design favors a large array size. Also, both higher cell nonlinearity and better diode access devices can reduce the peripheral circuit area significantly.

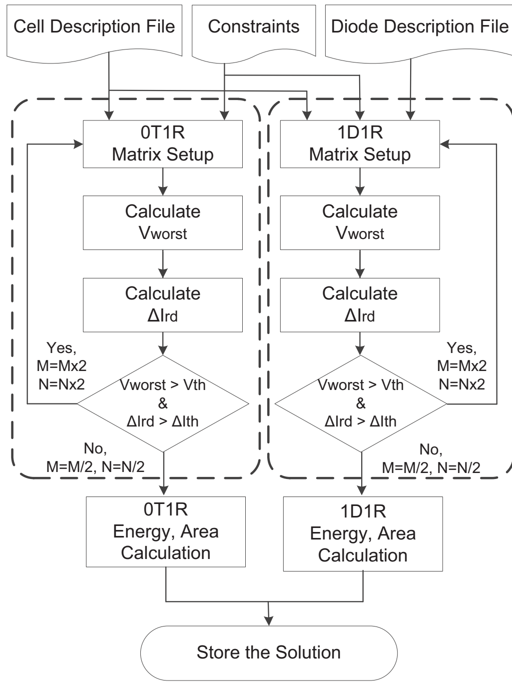

Fig. 4. First stage design flow. (V_{worst}, V_{th}: calculated value and constraint of worst case voltage. $\Delta I_{rd}, \Delta I_{th}$: calculated value and constraint of read noise margin. M, N: number of wordline and bitline.)

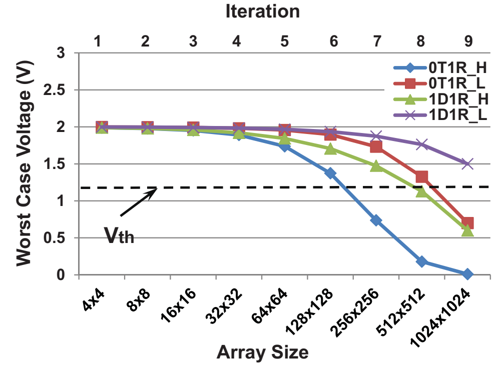

Fig. 5. Array Size Limitation Bounded by Worst Case Voltage.

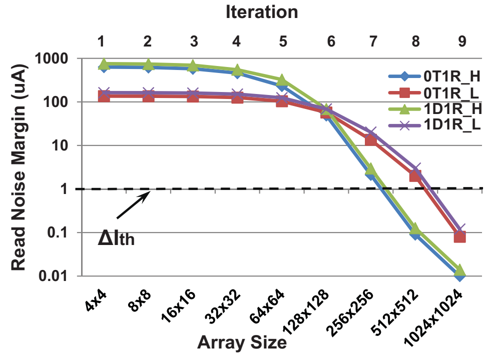

Fig. 6. Array Size Limitation Bounded by Read Noise Margin.

Although the area-optimized (or cost-optimized) design always prefers a large array size, as mentioned in literatures [11], [12], [14], the array size is limited due to the voltage drop along the wordline and bitline. In addition, the large array size has negative impact on the read noise margin, which is defined as the minimum current difference between HRS and LRS during the read operation. Therefore, in this stage, our model is used to decide the maximum array size and to calculate the energy/area of the ReRAM array under the given voltage drop and read noise margin constraints.

The detailed design flow in this stage is shown in Fig 4. At the beginning, cell parameters, diode parameters, and design constraints are provided as the inputs. Then the mathematic models are set up for both 0T1R and 1D1R ReRAM arrays according the input parameters with a relative small array size, for example 4×4. Then the worst case voltage, V_{worst}, and the worst case noise margin, ΔI_{rd}, are calculated. These results are compared to design constraints, write voltage threshold V_{th}, and minimum noise margin ΔI_{th}, which are specified at the beginning of this stage. If $V_{worst} > V_{th}$ and $\Delta I_{rd} > \Delta I_{th}$, the current array size is workable. Then the array size is increased by multiplying the wordline number M and bitline number N by 2 and another iteration is processed. When V_{worst} is smaller than V_{th} or ΔI_{rd} is smaller than ΔI_{th}, the array size is identified as invalid and therefore the array size at the previous iteration is determined as the maximum array size for the ReRAM array.

Fig. 5 and 6 show an example of the search process. In this example, we predefine the write voltage threshold and read noise margin constraints are $V_{th} = 1.2V$ and $\Delta I_{th} = 1uA$. Fig. 5 shows that, due to the worst case voltage requirement, the maximum array size of the 0T1R_H design is only 128×128 while the 1D1R_L design can achieve a size of 1024×1024. However, according to Fig. 6, the size of the 1D1R_L design is limited to 512×512 by the read noise margin requirement. In summary, the voltage drop requirement is more crucial for the 0T1R_H design and the read noise margin requirement is the prime constraint for 1D1R_L design. After obtaining the maximum array size, the worst case drive current, energy consumption, and noise margin are calculated with the corresponding associated data patterns.

B. The Second Stage: Bank Level Design Space Exploration

The performance, energy consumption, and area of the ReRAM chip are highly dependent on the bank level organization. In the second step, we build a bank-level model based on CACTI [15] and NVSIM [16]. Our model is able to estimate the performance, energy consumption, and area of both 0T1R and 1D1R ReRAM banks. Specifically, for our cost-driven optimization, the model is modified to accurately estimate the area of a single ReRAM bank.

To implement the area model of a cross-point ReRAM bank, the following important considerations should be taken into account carefully:

1) Different from DRAM technology, the size of sense amplifier for ReRAM is significantly larger than the cell size. In this case, it is impossible to provide a dedicated sense amplifier for each bitline. We consider two sharing methods to deal with this issue. Firstly, according to the placement of ReRAM array, adjacent arrays share same sense amplifiers located between these arrays. Secondly, by using the bitline multiplexers, only a small portion of cells in the selected row are read out from each ReRAM array. Then the high bandwidth of the bank is realized by activating multiple arrays in the same bank. Therefore, in our model, two parameters, N_{ls} and N_{act}, are used to denote the number of sense amplifiers of each array and the number of array activated at the same time.

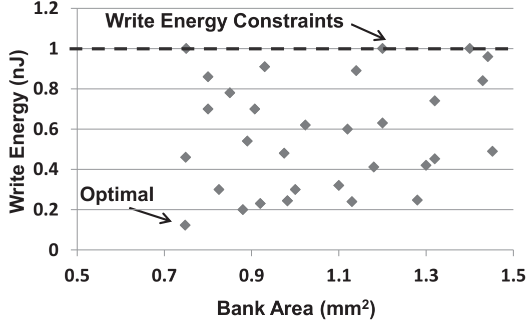

Fig. 7. Design Space Exploration of 128Mbit 1D1R ReRAM bank.

2) As mentioned in Section III-A, the area efficiency of an array is defined as the ratio between the cell area and array area. However, since the ReRAM array can be fabricated on top of CMOS layers, the area efficiency of an array can be improved by building part of the peripheral circuits underneath the ReRAM cell array. Therefore, we define the effective area efficiency as:

$$\eta_{Area_eff} = \frac{A_{cell}}{MAX[A_{driver}, A_{cell}]}. \quad (5)$$

As long as $A_{driver} < A_{cell}$, the area of drivers and multiplexers will be deducted from the total array area. Otherwise, A_{driver} is the actual array area and is used to calculate the total bank area. In our model, the η_{Area_eff} is used to calculate the bank area.

3) For the other part of the bank, we assume the ReRAM bank shares similar structure as DRAM bank, with similar column and row decoders, column and row address buffer, data input/output buffers, and control logics. However, we deduct the area overhead of refresh counter, refresh controller, and other refresh circuits of DRAM bank. The refresh related energy overhead is also deducted in our model.

Using the modified area model, a design space exploration can be conducted. Fig. 7 shows an example of design space exploration of the area and energy trade offs for a 128Mbit ReRAM bank design with ReRAM cell 1D1R_H. During the design space exploration, the ReRAM bank with different design choices, such as inter-array routing method, sense amplifier placement, number of parallelly accessed arrays, interconnect wire type, as well as output buffer design, are evaluated on area, energy consumption, and bandwidth. The simulator only outputs the results that meet the predefined constraints. As shown in this figure, the upper bond of write energy is set to 1nJ. For our cost-driven design flow, the most area-efficient bank organization under certain power budget is chosen.

C. The Third Stage: Chip-Level Cost Estimation

In this stage, the area and cost of the entire chip is calculated to evaluate the cost-per-bit metric.

First of all, the area of a single bank can be obtained from stage 2, and the total ReRAM area (A_{ReRAM}) is

$$A_{ReRAM} = \begin{cases} N_{bank}A_{bank} + A_{mp} + A_{cp} & \text{if } V_W > V_{DD} \\ N_{bank}A_{bank} + A_{mp} & \text{if } V_W \leq V_{DD}, \end{cases} \quad (6)$$

where A_{bank} is the area of one bank obtained from stage 2, N_{bank} is the total number of bank in the chip, A_{mp} is the area of memory peripheral circuits which is similar to DRAM based main memory,

TABLE I
PINCOUNT OF DDR SDRAM.

Standard	DDR	DDR2		DDR3		DDR4	
Type	all	x4 x8	x16	x4 x8	x16	x4 x8	x16
Pin count	72	60	84	78	96	78	96

and A_{cp} is the area of charge pump. Charge pump is only required when the programming voltage is larger than V_{DD}.

The area of CMOS compatible charge pump [17] is

$$A_{cp} = k \cdot \frac{N_s^2}{(N_s+1)\cdot V_{DD} - V_{Out}} \frac{I_L}{f}, \quad (7)$$

where k is a constant that only depends on the process technology, N_s is the number of stages of the charge pump, I_L is the load current, f is the charge pump frequency, and V_{out} is the output voltage. Therefore, the area overhead of the charge pump increases with V_{out}.

Considering the possible impact of the bond pads, the entire die area can be calculated as [18], [19]

$$A_{Die} = max(N_{pad}A_{pad} + A_{ReRAM}, (P_{pad}\lceil\sqrt{N_{pad}}\rceil)^2), \quad (8)$$

where N_{pad} is the number of bond pads, A_{pad} is the area of bond pads, P_{pad} is the minimum center to center pitch of bond pads. This equation shows that if the area of the bond pad is larger than the memory area, we have to increase the die area to meet the bond pad area requirement, which is very cost inefficient. Fortunately, in the commercial main memory market, pinout and addressing of the package are usually predefined by industrial standards. For example, the package pinout and addressing organization for each generation of DRAM are defined by the Joint Electron Devices Engineering Council (JEDEC). Table I lists the pin count of the DRAM defined by JEDEC. Clearly, the pin count increases with the DRAM capacity and data interface (x4, x8, or x16). In our model, we assume the ReRAM chip shares the same package design criteria as DRAM. Therefore, according to Equation (8), N_{pad} and the total bond pad area in our model can be calculated at the beginning of the design according to pre-determined memory standard and capacity. Therefore, we define a flag at stage two to denote if die area is determined by the bank area. For example, assuming $V_W > V_{DD}$, flags of all of the bank configurations with area

$$A_{bank} < (P_{pad}\lceil\sqrt{N_{pad}}\rceil)^2 - N_{pad}A_{pad} - A_{mp} - A_{cp} \quad (9)$$

are set to 'false', indicating that all of these configurations have the same and the best area efficiency, because the die size is totally determined by the bond pads instead of the banks.

After the die area A_{Die} is obtained, the total cost of the ReRAM die can be calculated as

$$C_{Die} = C_{Wafer} \times Y_{Wafer}/N_{gd}, \quad (10)$$

where C_{Wafer} is the cost of a wafer, Y_{Wafer} is the wafer yield, and N_{gd} is the number of good dies in the wafer. N_{gd} depends on the die area and can be calculated as [20]

$$N_{gd} = [\frac{\pi d_{Wafer}^2}{4A_{Die}} - \frac{\pi d_{wafer}}{\sqrt{2A_{Die}}}]/[1 + \frac{D_0 A_{Die}}{\alpha}]^{-\alpha}. \quad (11)$$

We assume the cost of mature ReRAM manufacturing process has similar wafer yield and defect density as traditional DRAM technology. However, considering the difference in process procedures, we calculate the wafer cost as

$$C_{Wafer} = C_{Wafer_{Die}} + C_{Wafer_{Die}+} - C_{Wafer_{Die}-}, \quad (12)$$

where $C_{Wafer_{Die}+}$ and $C_{Wafer_{Die}-}$ represent the cost of extra process steps associated with ReRAM fabrication (additional lithography

TABLE II

SURVEY OF BIPOLAR METALCOXIDE RERAM DEVICE CHARACTERISTICS

Reference	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]
Year	2005	2008	2008	2009	2010	2011	2011	2013
Materials	Cu	TaO_x	HfO_x	TiO_x/HfO_x	ZrO_x/HfO_x	$TaOx/Ta_2O_5$	Hf/HfO_x	TaO_x/Ta_2O_5
Technology (nm)	180	180	180	180	N/A	N/A	65	180/110
Write Voltage (V)	<3	2	1.5	2	<1.5	2.5	<3	2.6
Write Current (uA)	45	<170	25	200	50	30-150	50	35

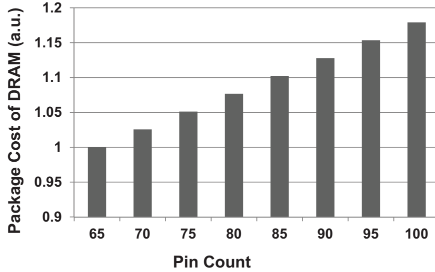

Fig. 8. Relationship between the Cost of DRAM Package and the Pin Count.

TABLE III

RERAM CELL PARAMETERS

Cell	*HCNN*	*HCLN*	*HCHN*	*LCNN*	*LCLN*	*LCHN*
Nonlinearity	none	20	50	none	20	50
$I_{LRS}@V_w(\mu A)$	200	200	200	40	40	40
$I_{Half}@V_w(\mu A)$	100	10	4	20	2	0.8
$R_{LRS}@V_r(K\Omega)$	10	100	250	50	500	1250
$R_{HRS}@V_r(K\Omega)$	500	5000	8000	2500	25000	80000

steps) and the redundant process cost of DRAM compared to ReRAM (cost of trench or stacked capacitors), respectively. We embed the detailed ReRAM stack deposition, isolation, and damascene fabrication into [29] and calculate the total cost of an ReRAM wafer. In addition, we also consider the package cost of the chip in our model. We assume that the ReRAM uses Fine Ball-Grid Array (FBGA) package, the same as the packaging method used in the mainstream DRAM chip. The packaging cost depends on both pin count of the package and the die area. However, according to ITRS [30], the die area of the DRAM chip is kept within a range of $30mm^2$ to $60mm^2$ regardless the technology nodes. In this area range, the package cost is almost independent of the die area and highly depends on the number of pins. Fig. 8 shows the package cost with different pin counts.

IV. EXPERIMENTAL RESULTS

In this section, we demonstrate the design space exploration of ReRAM memory array at a 45nm technology node with 6 different configurations. In addition to the energy, performance, bandwidth, and area overhead, the cost-per-bit results are also presented.

A. Experiment Setup

Table II summaries ReRAM cell parameters from recently published papers. According to this table, in our experiments, we assume that the write voltage and the read voltage for the ReRAM cell are 2V and 0.7V, respectively. Also, since all of the write current is in the range of 25uA to 200uA, without loss of generality, we choose ReRAM cells with high write current ($200\mu A$) and low write current ($40\mu A$) in our experiment to demonstrate different optimization results. As shown in Table III, We adopt 6 ReRAM cells with different nonlinearity, resistance, and write/read current. Cell *HCNN* (high write current, no nonlinearity), *HCLN* (high write current, low nonlinearity), and *HCHN* (high write current, high nonlinearity) have large write current of $200\mu A$ which corresponds to the LRS resistance of $10K\Omega$. Cell *LCNN*, *LCLN*, and *LCHN* have smaller write current and higher LRS resistance of $50K\Omega$. In addition, as implied by their names, cells also have different nonlinearities, which directly affect the sneak current at unselected cells. For the access diode in 1D1R structure, we adopt a similar design as shown in Burr's work [10] with a sneak current of $0.1\mu A$ and assume that the bidirectional diode can provide enough current for the write operation ($> 200\mu A$). As mentioned in Section III, the chip size of the DRAM technology is kept constant in the range of $30mm^2$ to $60mm^2$ regardless of the technology nodes. Therefore, we assume the ReRAM chip size is also within this range.

B. Experiment Results

Table IV shows the simulation results of the 6 different cells with the parameters specified in Table III. The chip area maintains in the range of $30mm^2$ to $60mm^2$, while different chips may have different capacities. Our simulation results show that, cells with large write current (cells *HCNN*, *HCLN*, and *HCHN*) should be built using a 1D1R structure. The reason is that the voltage drop along the wordline and bitline for these cells are very severe and the array sizes are therefore bounded by the worst case voltage drop. In our simulation, if the 0T1R structure is used for these cells, the maximum array size cannot exceed 128×128, 256×256, and 512×512, respectively. We find that the area efficiency for the 0T1R structure with a small array is worse than a 1D1R structure with a relatively larger array size. For cell *LCNN*, although the write current is smaller, the sneak current at all of the half-selected cells is significant. Therefore, the total wordline/bitline current is increased. Similar to high current cells, the array size of a 0T1R structure for cell *LCNN* is limited by the voltage drop and cannot exceed 256×256. Consequently, a 1D1R structure is necessary for cell *LCNN*. However, the current on both selected cells and half-selected cells for cells *LCLN* and *LCHN* is much smaller than the other cells. In this case, a large 0T1R array is workable. On the other hand, the reduced current means that these cells have larger resistance on both LRS and HRS. Thus, the read noise margin requirement becomes the dominant factor. As shown in [10], the access diode, which is connected in series with the ReRAM cell, shows very high resistance during the read operation. Therefore, the read current for 1D1R structure is reduced, resulting in a degradation of read noise margin. For these reasons, cells *LCLN* and *LCHN* adopt 0T1R structure.

The area and cost per bit results are also shown in Table IV. Area efficiencies for *HCHN* and *LCHN* are much better than other cells. The reason is that, for cells *HCNN*, *HCLN*, *LCNN*, and *LCLN*, the large wordline drivers and bitline multiplexers cannot be totally hidden under the cell array. For *HCHN* and *LCHN*, the large array size reduces wordline drivers and bitline multipliers area overhead. For example, a 1024×1024 array requires 1024 wordline drivers and 1024 bitline multipliers. However, with the same capacity, four

TABLE IV
RERAM CHIP DESIGN OPTIMIZATIONS WITH DIFFERENT CELLS

Cell	Structure	Capacity (Gbits)	Die Area (mm^2)	Array Size	Cost per bit (Microcent)	Write Latency (ns)	Read Latency (ns)	Write Energy (nJ)	Read Energy (pJ)
HCNN	1D1R	2	57.10	512×512	0.171	51.16	28.60	18.32	0.63
HCLN	1D1R	2	54.34	512×512	0.158	51.05	28.55	13.17	0.35
HCHN	1D1R	4	53.67	1024×1024	0.083	53.98	29.25	22.28	0.54
LCNN	1D1R	2	56.21	512×512	0.157	51.08	28.78	6.52	0.87
LCLN	0T1R	2	58.49	512×512	0.168	50.88	29.31	3.91	1.08
LCHN	0T1R	4	54.55	1024×1024	0.081	54.20	31.08	5.53	1.59

512×512 arrays need 2048 wordline drivers and 2048 bitline multipliers in total. In addition, as mentioned in Section III, the small current of cells *HCHN* and *LCHN* are also helpful to further reduce the area of wordline drivers and bitline multiplexers. Therefore, our simulation results show that for *HCHN* and *LCHN*, all of wordline drivers and bitline multipliers can be built underneath the cell array. Furthermore, since the area efficiency is related to the cost per bit, as shown in the table, chips *HCHN* and *LCHN* are also superior in terms of cost.

V. CONCLUSION

ReRAM technology has great potential to become a replacement for DRAM and/or Flash. The success of a memory technology primarily depends on its cost per bit. However, since ReRAM design has a large number of design choices, such as array structure, array size, bank organization, and interconnect wire, the most cost efficient design is nontrivial at an early design stage. In this paper, we propose a cost-driven design flow for the metal-oxide ReRAM design. We begin with a circuit level array model in which we identify the optimal array size and architecture that meet the reliability requirements. We then consider various array and bank organizations to further reduce cost. We study two common implementations of ReRAM: 1D1R and 0T1R structures. Our simulation results show that for the cell with small nonlinearity or large current, the 1D1R structure is more cost efficient, in spite of the extra area overhead of the charge pump required by the higher operating voltage of the 1D1R structure. However, for ReRAM with low operating current, the introduction of diode will negatively impact the read noise margin, hence a 0T1R structure is preferred in this case.

REFERENCES

[1] C. Ho *et al.*, “9nm half-pitch functional resistive memory cell with $<$1ua programming current using thermally oxidized sub-stoichiometric woxfilm,” in *IEDM*, 2010, pp. 19.1.1–19.1.4.
[2] W. Chien *et al.*, “A forming-free woxresistive memory using a novel self-aligned field enhancement feature with excellent reliability and scalability,” in *IEDM*, 2010, pp. 19.2.1–19.2.4.
[3] S.-S. Sheu *et al.*, “A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability,” in *ISSCC*, 2011, pp. 200–202.
[4] T.-Y. Liu *et al.*, “A 130.7mm^2 2-layer 32Gb ReRAM memory device in 24nm technology,” in *ISSCC*, 2013, pp. 210–211.
[5] A. Kawahara *et al.*, “An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput,” in *ISSCC*, 2012, pp. 432 –434.
[6] M. Lee *et al.*, “A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5-x/TaO2-x bilayer structures,” *Nature Materials*, vol. 10, pp. 625–630, 2011.
[7] M. Kim *et al.*, “Low power operating bipolar TMO ReRAM for sub 10 nm era,” in *IEDM*, 2010, pp. 19.3.1 –19.3.4.
[8] Y. Sato *et al.*, “Sub-100uA reset current of Nickel Oxide resistive memory through control of filamentary conductance by current limit of MOSFET,” *IEEE Trans. on Electron Devices*, vol. 55, no. 5, pp. 1185–1191, 2008.
[9] J.-J. Huang *et al.*, “One selector-one resistor (1S1R) crossbar array for high-density flexible memory applications,” in *IEDM*, 2011, pp. 31.7.1–31.7.4.
[10] G. Burr *et al.*, “Large-scale (512kbit) integration of multilayer-ready access-devices based on mixed-ionic-electronic-conduction (MIEC) at 100% yield,” in *VLSIT*, 2012, pp. 41 –42.
[11] O. Kavehei *et al.*, “An analytical approach for memristive nanoarchitectures,” *IEEE Trans. on Nanotechnology*, vol. 11, no. 2, pp. 374 –385, Mar. 2012.
[12] D. Niu *et al.*, “Design trade-offs for high density cross-point resistive memory,” in *ISLPED*, 2012, pp. 209–214.
[13] X. Dong *et al.*, “Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement,” in *DAC*, 2008, pp. 554–559.
[14] J. Liang and H.-S. Wong, “Cross-point memory array without cell selectors -device characteristics and data storage pattern dependencies,” *IEEE Trans. on Electron Devices*, vol. 57, no. 10, pp. 2531 –2538, Oct. 2010.
[15] S. J. E. Wilton and N. Jouppi, “Cacti: an enhanced cache access and cycle time model,” *IEEE Journal of Solid-State Circuits*, vol. 31, no. 5, pp. 677–688, 1996.
[16] X. Dong *et al.*, “Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory,” *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994 –1007, Jul. 2012.
[17] G. Palumbo and D. Pappalardo, “Charge pump circuits: An overview on design strategies and topologies,” *IEEE Circuits and Systems Magazine*, vol. 10, no. 1, pp. 31 –45, quarter 2010.
[18] P. Sandborn, M. Abadir, and C. Murphy, “The tradeoff between peripheral and area array bonding of components in multichip modules,” *IEEE Trans. on Components, Packaging, and Manufacturing Technology, Part A*, vol. 17, no. 2, pp. 249 –256, jun 1994.
[19] J. Zhao, X. Dong, and Y. Xie, “Cost-aware three-dimensional (3d) many-core multiprocessor design,” in *DAC*, 2010, pp. 126 –131.
[20] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*. Prentice-Hall, 2003.
[21] A. Chen *et al.*, “Non-volatile resistive switching for advanced memory applications,” in *IEDM*, 2005, pp. 746–749.
[22] Z. Wei *et al.*, “Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism,” in *IEDM*, 2008, pp. 1–4.
[23] H. Lee *et al.*, “Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust HfO2 based RRAM,” in *IEDM*, 2008, pp. 1–4.
[24] Y. S. Chen *et al.*, “Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity,” in *IEDM*, 2009, pp. 1–4.
[25] J. Lee *et al.*, “Diode-less nano-scale zrox/hfox rram device with excellent switching uniformity and reliability for high-density cross-point memory applications,” in *IEDM*, 2010, pp. 19.5.1–19.5.4.
[26] Y.-B. Kim *et al.*, “Bi-layered RRAM with unlimited endurance and extremely uniform switching,” in *VLSIT*, 2011, pp. 52–53.
[27] B. Govoreanu *et al.*, “10x10nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation,” in *IEDM*, 2011, pp. 31.6.1–31.6.4.
[28] A. Kawahara *et al.*, “Filament scaling forming technique and level-verify-write scheme with endurance over 10^7 cycles in ReRAM,” in *ISSCC*, 2013, pp. 220–221.
[29] IC Knowledge LLC., “IC cost model revision 1105.” [Online]. Available: http://www.icknowledge.com
[30] ITRS, “International technology roadmap for semiconductors,” 2010.