Design Framework for Reliable, Energy Efficient Cross-Point based Resistive Memory

***Abstract*—With conventional memory technologies approaching their scaling limit, emerging non-volatile memory technologies have attracted considerable attention because of their non-volatility, high access speed, low power consumption, and good scalability. Resistive RAM (ReRAM), with its simple structure, small cell size ($4F^2$), and support for 3D stacking, has been a leading candidate among emerging technologies. A key advantage of ReRAM comes from its non-linear nature, which enable us to build a cross-point RAM array without having a dedicated access transistor in each cell. While cross-point design is effective in improving memory density, it has inherent disadvantages which introduce extra design challenges. In this work, we analyze these challenges. Based on the circuit characteristics of the cross-point array, we propose a mathematical model to perform a comprehensive analysis of issues of reliability, energy consumption and the area overhead. In addition to the cell-level analysis, different programming schemes are also discussed in detail. Based on the study, a detailed design methodology is proposed to enable designers to identify the most energy/area efficient ReRAM organization that meets specific design goals early in the design stage.**

I. Introduction

The scaling of traditional memory technologies, such as DRAM and FLASH, is approaching its physical limit. In the past few years, emerging non-volatile technologies (NVM), such as Phase Change RAM (PCRAM), Magnetoresistive RAM (STT-RAM), and Resistive RAM (ReRAM) have been widely studied as potential candidates for the next generation memory technologies to meet the need of higher density, faster access time, and lower power consumption. Among all of these emerging memory technologies, ReRAM has many unique characteristics, including simple structure, non-linearity and high resistance ratio, making itself one of the most promising technologies. Researchers have shown that the state-of-the-art single-level-cell ReRAM can achieve sub-8ns random access time for both read and write operations with a resistance ratio larger than 100 [**?**]. Also, HP labs and Hynix have already announced plans to commercialize the memristor-based ReRAM and predicted that ReRAM could eventually replace traditional memory technologies [**?**].

Different from other non-volatile memory technologies, ReRAM can be implemented in a cross-point style structure without any access devices. Specifically, in a nano cross-point array, each bistable ReRAM cell is sandwiched by two orthogonal nanowires, without access devices. Thus the area occupied by each cell is literally the area underneath the intersection of wires, which is $4F^2$ per bit. However, the simplicity of access device free, cross-point structure introduces challenges to the peripheral circuit design and memory organization. While there has been prior studies on cross-point ReRAM array [**?**], [**?**], [**?**], [**?**], they do not consider the effect of voltage drivers and programming methods to cells.

In addition, detailed area and energy analysis is also absent. In this work, we address the design challenges of cross-point structure based ReRAM. We build an accurate mathematical model to evaluate memory reliability, energy consumption, and area overhead for different designs and cell parameters. Based on this study, we propose a detailed design methodology which allows for exploring the most energy/area efficient ReRAM design with different design constraints and cell parameters at the very beginning of the design stage. On the other hand, the system designers can also leverage the proposed framework to provide valuable feedback to device researchers who will in turn adjust ReRAM cell design. We believe that this kind of two-way communications will be very helpful to shorten time-to-market of ReRAM memory.

TABLE I
Comparison of Emerging Non-Volatile Memory Technologies

Metric	STT-RAM	PCM	FeRAM	ReRAM
Cell Size(F^2)	$6-20$	$4-8$	15	4
Read Latency(ns)	1-10	20-50	20-80	5-50
Write Latency(ns)	2-20	150	100	5-50
Endurance	10^{15}	10^{8}	10^{12}	10^{8-10}

The rest of this paper is organized as follows. In Section II, the preliminaries of ReRAM technology and cross-point architecture are introduced. Section III discusses the mathematical model we propose for crossbar structure ReRAM and the edge conditions for different write and read schemes. Section IV analyzes different design constraints of write and read operations on the cross-point based ReRAM array. The energy consumption and area overheads are also analyzed in this section. Then in Section V, the design methodology for the ReRAM array is proposed based on our mathematical model and simulation results. Finally, the conclusion is presented in Section VI.

II. Preliminaries

This section provides background of ReRAM technology and cross-point architecture, and discusses their advantages and limitations.

A. Background of ReRAM technology

Table. I compares the properties of state-of-art non-volatile memory technologies. ReRAM and STT-RAM are better than PCM or FeRAM due to their faster access time and high endurance. Although STT-RAM has better read/write delay characteristics, the difference in cell resistance between ON and OFF states is higher in ReRAM. Further, ReRAM has the best memory density. In essence, ReRAM is a leading candidate for next generation storage.

As implied by the name, an ReRAM cell uses its resistance to represent the stored information. A ReRAM cell can be switched between high resistance state (HRS) and low resistance state (LRS) by applying an external voltage across the cell. In general, a ReRAM cell is built on a Metal-Insulator-Metal (MIM) structure. The resistance switching behaviors have been observed in many MIM nanodevices with different metal oxide materials. For example, a particular TiO_2 based MIM structure ReRAM, named ‘memristor’, was developed by HP Labs in 2008 [**?**]. The proposed memristor based ReRAM is considered as the first experimental realization and a theoretical model of the fourth fundamental circuit element, which is predicted by Chua [**?**] about 40 years ago. The memristor based ReRAM has very small cell size with an access time of less than 50ns. Another H_fO_2-based bipolar ReRAM is implemented by ITRI this year with an access time as low as 7.2ns [**?**].

Although there are several variants of ReRAMs, all of them can be classified into two broad categories: unipolar ReRAM and bipolar ReRAM. In an unipolar cell, the resistance switching behaviors do not depend on the polarity of the voltage input across the cell and only relate to magnitude and duration of the voltage input. On the other hand, in a bipolar cell, the voltage polarity for ON-to-OFF

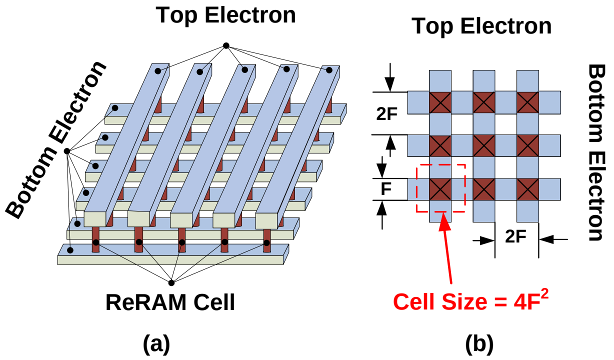

Fig. 1. A schematic view of typical cross-point architecture. (a): Over view of the cross-point architecture. (b): The layout of the cell size of $4F^2$.

switching (RESET operation) is different from OFF-to-ON switching (SET operation). The need for different pulse widths for SET and RESET in unipolar ReRAM means that its write latency will be determined by the longest pulse. Besides, the control of SET, RESET, and read operations without any disturbance is another crucial design challenge, especially in high speed ReRAM design. For these reasons, most high performance ReRAM studies are dominated by bipolar ReRAM [?], [?], [?]. In this study, we perform a detailed analysis of the design challenges of bipolar ReRAM based cross-point array.

B. Cross-Point Architecture

There are two possible memory structures for bipolar ReRAM implementation: traditional MOSFET-accessed structure and the cross-point structure. In the MOS-accessed memory array, a MOSFET is added as an access device for each memory cell. As the size of a MOSFET access device is typically much larger than size of ReRAM cell, the total area of memory array is primarily dominated by MOSFETs rather than ReRAM cells. Hence, ReRAM's area advantage gets lost because of the access device.

On the contrary, the cross-point structure is more area-efficient for the ReRAM based memory array [?]. A schematic view of a typical cross-point memory array is shown in Figure 1(a). In a cross-point array, ReRAM is sandwiched between wordlines and bitlines. Figure 1(b) shows the feasibility of $4F^2$ cells, the theoretical minimum size for a single layer single level memory cell. The memory density can further be improved by using a multi-layer multi-level cross-point ReRAM array [?] [?].

In a cross-point structure, the write operation can either write one bit per access or write several bits attached to a wordline at the same time. Although the second scheme has higher bandwidth, it requires a two-step writing operation to prevent unintentional writing [?], significantly increasing the write latency. For instance, while write to cross-point array, the unselected wordlines and bitlines can either be left floating or half biased. However, while reading a cell, the selected wordline should be biased at read voltage and all other wordlines and bitlines in the array are shunted to ground. The current in each bitline is then sensed and compared to a reference current to determine cell content. However, due to the sneak current existing at the cross-point array, the current in bitline also varies depending upon the values stored in unselected cells. This phenomenon of read disturbance restricts the size of a cross-point array, since sneak current increases with the number of cells attached to wordlines and bitlines. In order to mitigate this problem, a two-step write operation was proposed: In the first step, the background current of the cross-point array will be sensed. Then the total current, comprising of both the background current and current at the selected cell, will be read out.

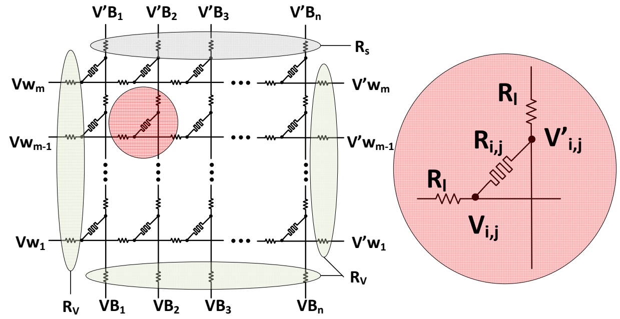

Fig. 2. The basic model of typical cross-point array.

The state of the selected cell can then be determined by computing the difference between the total current and background current. For this scheme to work, the cross-point array should be sized such that the difference between background current and current through the selected cell should be large enough detect. Depending upon the read/write scheme, the size of the array can vary significantly. In this work, we propose a methodology to find the minimum array size that meets specific energy and area constraints based on the worst-case state of the array. This will help designers find an optimal memory organization early in the design flow.

III. Modeling of the Cross-Point Memory

In this section, a detailed mathematical model of the cross-point array is proposed. By using the proposed model with specific parameters and edge conditions, the reliability, energy consumption, and area overheads of different read/write schemes can be easily evaluated.

A. Basic model of Cross-Point Memory

Figure 2 shows the circuit model of an M by N cross-point ReRAM array. The horizontal lines are word lines and vertical lines represent the bit lines. The ReRAM cells are located at each cross-point of word line and bit lines. The resistance of the ReRAM cell at the cross-point of the i^{th} word line and j^{th} bit line is indicated as $R_{i,j}$. We assume the resistance of the interconnect nanowires between two adjacent cross point has the same value of R_{line}. The input resistance of each word line and bit line is R_v and the resistance of sense amplifier is R_s. In order to set up the Kirchhoff's Current Law (KCL) equations, the voltage at each cross point is indicated as $V_{i,j}$ for word line and $V'_{i,j}$ for bit line. A detailed cross point is also shown in right hand figure of Figure 2. Besides, the input voltage for the i^{th} word line is V_{Wi} and for the i^{th} bit line is V_{Bi}. In the case of two side voltage input of word line, the voltage at the other end of the i^{th} word line is denoted as V_{W1}. Finally, the voltage at the sense amplifier is V'_{Bi} during the read operation.

B. Mathematical Model of the Cross-Point Array

Based on the circuit model shown in Figure 2, the current equations for each cross point can be set following KCL:

$$\Sigma_{I=1}^{k} I_k = 0. \tag{1}$$

All of the cross points have similar structure with no more than three current branches and therefore it is very easy to set up the KCL equations for each cross point. However, we should treat the cross points at the edges of the array specially because there are different conditions for different write/read schemes. For example, the unselected word line for write operation can be either half biased or left floating. Thus, the edge conditions should be adjusted according to each write/read schemes. In particular, all of the cross points in the array can be classified into three major categories: Normal point, Activated point and Floating point.

The normal points locate inside the memory array. In other words, for all of the nodes with $1 < i < m$ and $1 < j < n$, the KCL equations take the form of

$$R_l^{-1}V_{i,j-1}-(2R_l^{-1}+R_{i,j}^{-1})V_{i,j}+R_l^{-1}V_{i,j+1}+R_{i,j}^{-1}V'_{i,j}=0, \tag{2}$$

for the node at word line layer and

$$R_l^{-1}V'_{i-1,j}-(2R_l^{-1}+R_{i,j}^{-1})V'_{i,j}+R_l^{-1}V'_{i+1,j}+R_{i,j}^{-1}V_{i,j}=0, \tag{3}$$

for the node at bit line layer.

The activated point and floating point represent the nodes at the edge of cross-point array with different conditions: an edge point, which is directly connected to the voltage input or to the ground, can be considered as an activated point. Otherwise, it is a floating point. Take the point located at the intersection of i^{th} word line and 1^{st} bit line for example. If the i^{th} word line is activated by a voltage input of V_{Wi}, this cross point is an activated point, and the KCL equation for this point is:

$$-(R_v^{-1}+R_l^{-1}+R_{i,1}^{-1})V_{i,1}+R_l^{-1}V_{i,2}+R_{i,1}^{-1}V'_{i,1}=-R_v^{-1}V_{Wi}. \tag{4}$$

Otherwise, it is floating and has the KCL equation take the form of

$$-(R_l^{-1}+R_{i,1}^{-1})V_{i,1}+R_l^{-1}V_{i,2}+R_{i,1}^{-1}V'_{i,1}=0. \tag{5}$$

For the reasons of clarity, a $2mn \times 1$ vector V is defined to represent all of the variables in the KCL equations:

$$V=[{V_1}^T,{V_2}^T...{V_m}^T,{V_1'}^T,{V_2'}^T...{V_m'}^T]^T, \tag{6}$$

where,

$$V_i=[V_{i,1},V_{i,2}...V_{i,n}]^T, \quad V'_i=[V'_{i,1},V'_{i,2}...V'_{i,n}]^T, \tag{7}$$

for $i = 1, 2...m$. Then all of the KCL equations can be considered as a system of linear equations, which has the form of

$$A \cdot V = C. \tag{8}$$

A is a $2mn \times 2mn$ coefficient matrix, which is determined by Equation(2)-(5). C is a $2mn \times 1$ vector, containing the constant terms of these equations. As shown, the KCL equations for each node have very simple structure and are very similar to each other. Therefore, the linear equation system has a relatively fixed format and simple structure, making it very easy to establish and adjust the coefficients and constants according to different design schemes. The characteristics of the linear system can be summarized as:

1) As shown in Equation (9), the coefficient matrix A can be further partitioned into 4 smaller subblocks :

$$\mathbf{A}=\begin{bmatrix} A1 & A2 \\ A3 & A4 \end{bmatrix}. \tag{9}$$

All of these subblocks have the same size of $m \times n$. Subblock $A2$ and $A3$ are diagonal matrixes and have the value of: $A2_{i,i} = A3_{i,i} = R_{i,i}^{-1}$. Besides, $A2$ and $A3$ do not change their values with different schemas. However, $A1$ and $A4$ are a little bit more complex than $A2$ and $A3$. $A1$ is a tridiagonal matrix and only has nonzero elements at the location in the main diagonal, and the first line below and above the diagonal. Similarly, the A_4 is a special tridiagonal matrix, which has nonzero elements in the main diagonal, the n^{th} line below and above the diagonal, where n is the number of bit line in the cross point model. The value of the elements in $A1$ and $A4$ can be easily derived from Equation (2) and (3). However, as mentioned, the edge condition varies with different program schemes. Therefore, the coefficients related to the edge condition should be update according to the program schemes. Clearly, the four edges shown in Figure 2 correspond to different coefficients in $A1$ and $A4$. Due to the space limitations, we take the nodes at the left edge of the array for example. Coefficients of other edge nodes can be initiated by the same way. The coefficients of nodes at the left edge of the array ($V_{i,1}$) can be set as:

$$A1(k,k)=\begin{cases} -(R_l^{-1}+R_{i,1}^{-1}) & \text{if } floating \\ -(R_v^{-1}+R_l^{-1}+R_{i,1}^{-1}) & \text{if } activated \end{cases} \tag{10}$$

where $k = (n-1)i + 1$ for $i = 1, 2...m$.

2) The constant terms C is a $2mn \times 1$ vector. Equation(2)-(5) show that only KCL equations of the activated points have the constant terms. Therefore, only the following elements in C may have non-zero value: $C((i-1)n+1)$, $C(in)$, $C(mn+i)$ and $C((2m-1)n+i)$ for $i = 1, 2...m$, correspond to the nodes at the four edges respectively. Likewise, we take nodes $V_{i,1}$ for example. The constant correspond to these node can be defined as:

$$C((i-1)n+1)=\begin{cases} 0 & \text{if } floating \\ -R_v^{-1}V_{Wi} & \text{if } activated \end{cases} \tag{11}$$

Therefore, with all of the required parameters, including the resistance of ReRAM cell, resistance of interconnect wires, program voltages and write/read schemes, the voltage at each cross point of the array can be obtained by solving the Equation (8) with simple matrix computations. With the detailed voltage values, $V_{2mn\times1}$, we can analyzed the array at a very fine granularity. Also, these information can be very useful to evaluate the reliability, energy consumption, driven current density, and area overheads of the cross-point array.

IV. Analysis of Design Constraints - A Case Study

In this section, a typical case of the ReRAM based cross-point array is detailed analyzed by using the proposed mathematical model.

A. Overview

As shown in Figure 2, in order to write or read the cross-point array, the external voltages should be applied at the end of the word line and the bit line. Since several potential read/write schemes can be used to program the memory array, it is quite difficult to point out which scheme is the most proper choice under given design constraints of area/energy/reliability. Therefore, in this section, studies on different operation schemes are conducted. The requirement for array size, energy consumption and area overheads are analyzed in the worst cases scenario. The results of this study can be very useful to guide the design of the cross-point array.

Table II shows the circuit parameter of our baseline 32nm design. The data is derived from the recently published studies on ReRAM [**?**] [**?**]. In the following the reliability, energy consumption, and area overheads for the four write schemes are detailed. Then the sensitivities of these schemes to the data pattern of HRS and LRS ReRAM cells, and non-linearity are studied.

Although the purposes of read and write operations are different, both of them are realized by fully biasing selected cell and floating (or half biasing) the unselected cell. Thus, the set up of the coefficient matrix A and constant vector C are very similar for read and write operations. In addition, the energy consumption and area overheads also have similar trends for them. Therefore, in the next section, we first study the writing operation comprehensively. After that, for the read operation, we mainly focus on the read margin analysis since it is unique to the read operation.

B. Write Operation

To write a ReRAM cell, an external voltage is applied across the cell for a certain duration. Intuitively, there are four possible schemes for the write operation:

1) According the location of the selected cell, activate one word line and one bit line and leave all of the other lines floating (FWFB shemes).
2) Activate the selected word line and bit line. Leaving all the unselected word lines floating and half bias the unselected bit lines (FWHB shemes).
3) In contrast with the scheme 2, activate the selected word line and bit line. Leaving all the other bit lines floating and half bias the other wold lines (HWFB shemes).
4) Activate the selected word line and bit line. Then half bias all other word lines and bit lines (HWHB shemes).

Since the reliability, energy consumption, and area overheads for these schemes are different from each other. We will address these problems separately and then combine all of the constraints to provide a design guideline for write operation.

Reliable Write Operation.

The most important issue for the write operation is the reliability concern. In the ideal condition, the resistances of interconnect wires and the sneak currents at unselected cells are negligible. In this case, all of these write schemes can make sure that the write voltage $V_W(W) - V_B(W)$ is fully applied across the specified cell. However, the realistic circuit is not perfect and the electronic behavior of the cross-point array will deviate from the ideal case with different data pattern stored in the ReRAM cells. A reliable write operation can be defined as: switching the selected cells into required states without disturbing the states of unselected cells. Therefore, there are two potential problems of a write operation: **write failure**, an unsuccessful write on selected cell, and **write disturbance**, an undesirable write on unselected cell. All of the write schemes should meet the reliability requirement all the time. In other words, the designer should make sure there is not any write failure and write disturbance exist even in the worst case. Otherwise, after several unreliable write operation, the data stored at the cross-point array will become unpredictable.

First of all, we use an example to show the inherent problem of FWFB scheme, which may result in severe write disturbance. Figure 3 shows the voltage drop across each ReRAM cell of a 64×64 cross-point array. In this example, in order to write the cell at the cross point of the 32^{th} word line and the 32^{th} bit line, the selected word line and bit line are biased at 2V and 0V, respectively. All of the other word lines and bit lines are biased at 1V. The ReRAM cells at the selected bit line are in the HRS, while all of the other cells are in the LRS. It is clear that the voltage drop across the selected cell ($V_{32,32}$) almost has the same magnitude as the unselected cell at the same bit line, resulting the write disturbance to all of the unselected cells at the selected bit line. Actually, for a $M \times N$ matrix ($M > N$), the worst case voltage drop of the unselected cell can be calculated as:

$$V_{worst} = V_{select} \cdot [1 - \frac{1}{M + (N-1)R_{off}/R_{on}}]. \qquad (12)$$

Considering that the reported On-OFF resistance ratio of ReRAM cell is always > 50 [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], , the worst case voltage drop at the unselected cell is larger than 98% of the voltage at the selected cell, making it is impossible to build a reliable cross-point structure ReRAM with the FWFB scheme. Therefore, in the following discussion, we only compare the results of FWHB, HWFB and HWHB schemes. For each of these three schemes, we can either write the cells at one word line at the same time or only write one bit per access and separate the write operation to several arrays. In the following discussion, we start from one bit per access write operation. And then the results of one word line per access method are discussed.

The write failure mainly results from the voltage drop at the interconnect wires along the word line and bit line. It has been shown that [**?**], for one bit per access write operation, the worst case voltage drop occurs when

$$\begin{cases} R_{M.N} = R_{on} \\ V_{WM} = V_W(W) \\ V_{BN} = V_B(W). \end{cases} \qquad (13)$$

In order to avoid the **write failure** and successfully program the selected ReRAM cell, the driven voltage should be boosted to a higher level, making sure that the voltage across the cell exceed the threshold voltage even at the worst case. Figure 4 shows the lower bound of the driven voltage for different sizes of cross-point array. The minimum word/bit line voltage increases from 2.01 V for a 8×8 array to 4.47 V for a 128×128 cross-point array. Besides, for a memory capability, the cross-point array can be organized with different number of word lines and bit lines. For example, a 4K bits cross-point array can be implemented either by a 64×64 array or by a 32×128 array. In the latter case, the voltage drops along the word line will be much more serious than along the bit line. Figure 5 exams the voltage requirement for different array organizations with different write schemes. The result shows that from the reliability point of view, the cross-point

TABLE II
PARAMETERS OF THE BASELINE CROSS-POINT ARRAY

Metric	Description	Values
S_{cell}	Cell Size	$4F^2$
R_l	Interconnection Resistance	1.25Ω
V_{RESET}	Threshold voltage for RESET	$2.0V$
V_{SET}	Threshold voltage for SET	$-2.0V$
V_{READ}	Read Voltage of Cell	$0.5V$
R_{off}	HRS Resistance	$500K\Omega$
R_{on}	LRS Resistance	$10K\Omega$
$V_W(R)$	Word Line Voltage during Read	$0.4V$
$V_W(W)$	Word Line Voltage during Write	$\pm 2V$
$V_W(H)$	Half Selected Word Line Voltage	$1V$
$V_B(R)$	Bit Line Voltage during Read	$0V$
$V_B(W)$	Bit Line Voltage during Write	$0V$
$V_B(H)$	Half Selected Bit Line Voltage	$1V$
M	Number of Word Line	64
N	Number of Bit Line	64

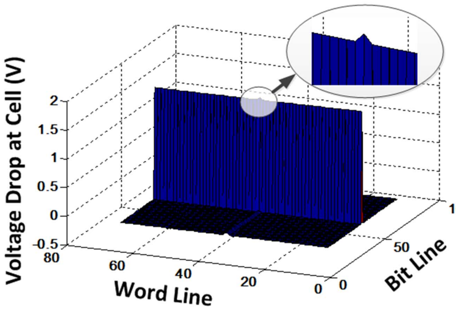

Fig. 3. Write Disturbance for FWFB Schemes. ($V_{W32} = 2V$, $V_{B32} = 0V$. $R_{x,32}$ at HRS, others at LRS.)

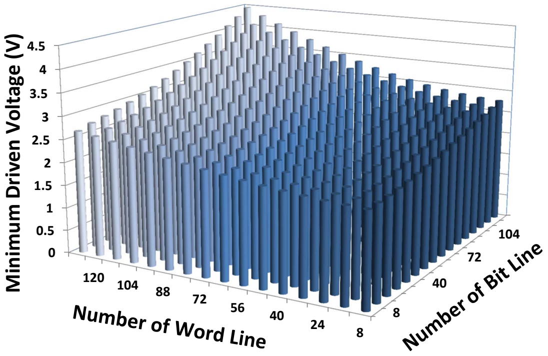

Fig. 4. Write Voltage Requirement (Threshold Voltage = 2V).

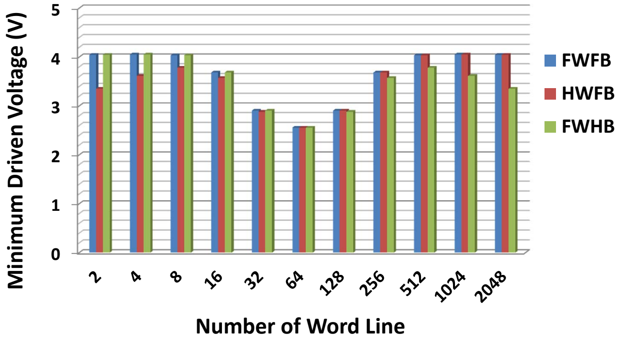

Fig. 5. Write Voltage Requirement with Different Memory Shape. (Array Capacity = 4Kbits, Activated Word Line Voltage = 2V, Activated Bit Line Voltage = 0V.)

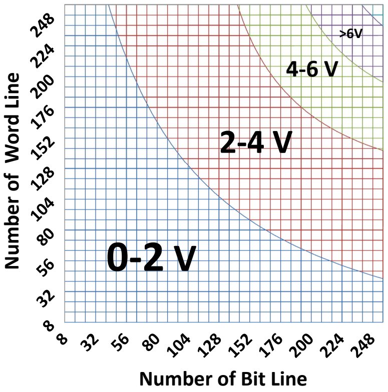

Fig. 6. The Maximum Voltage Applied at Unselected Cells with the Minimum Driven Voltage.

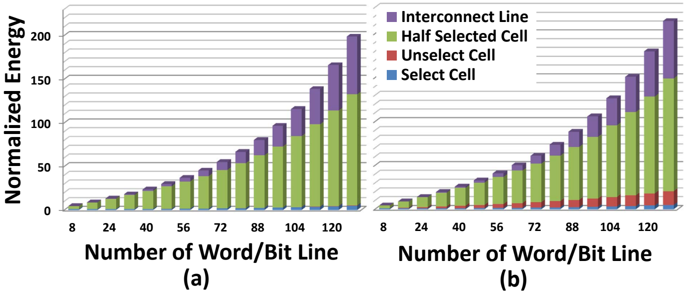

Fig. 7. The Normalized Energy Consumption. (a): HWHB scheme (b): FWHB and HWFB schemes.

array with same numbers of word line and bit line is the best choice. Besides, we also notice that when the array has the same number of word line and bit line, FWFB, HWFB and FWHB schemes have the same minimum driven voltage.

However, boosting the driven voltage also introduces other potential problems for the array. Especially, the increasing of the driven voltage also increases the voltage applied at the unselected cell. Therefore, a **write disturbance** may occur when the voltage applied at the unselected cell exceeds the threshold voltage for SET or RESET operation. Figure 6 shows the maximum voltage applied at unselected cells with the minimum driven voltage, which is determined in Figure 4. Since the threshold voltage of the ReRAM cell is 2V, only the array sizes with worst case voltage less than 2V are allowable. Otherwise, the array is unreliable because it can not avoid write failure and write disturbance at the same time. Therefore, Figure 6 provides the hard constraint of array size, and all of the following energy and area tradeoffs should be bounded by this constraint.

Energy Consumption of Write Operation.

The energy consumption of a write operation for a cross-point array can be calculated as:

$$E_{write} = E_{select} + E_{unselect} + E_{halfselect} + E_{line}, \qquad (14)$$

where the E_{select} is the energy consumed to change the state of the selected cell, the $E_{unselect}$ and $E_{halfselect}$ are the undesired energy wasted at the half selected and unselected cells. The energy consumed by the interconnect lines are represented by E_{line}. Figure 7 shows the decomposed energy consumption for the cross-point array. Note that, the E_{line} and $E_{halfselect}$ take a large amount of the total energy consumption. Besides, the energy wasted during the write operation takes a great part of the total energy for large array size. For example, the undesired energy consumption for writing a 128×128 array is more than 1000 times larger than the 8×8 array. We also notice that, since the impact of sneak paths for floating schemes (FWHB and HWFB) is more serious, the energy consumed at unselected cells for floating schemes are larger than the half-biased scheme. Due to this reason, the total energy consumptions for FWHB and HWFB schemes are at least 10% larger than that of HWHB scheme.

Area cost of Write Operation.

The write operation for a $M \times N$ array requires totaly $M + V$ voltage drivers. Therefore, the average number of ReRAM cells per voltage driver can be calculated as $mn/(m+n)$. Given the array capacity of C_{array}, it is easy to find that the optimal array organization can be achieved when $M = N = \sqrt{C_{array}}$ and the maximum number of cells per voltage driver is: $\sqrt{C_{array}}/2$. However, the area overhead of a voltage driver is also related to its current drive capability. Figure 8(a) shows the maximum write current with different ReRAM array size. According to the current requirement, the area of the voltage driver can be directly calculated, which will be shown in the following discussion.

Discussion on Multi-Bits Write Operation.

So far, we only discuss the one bit per access write operation. In this section, the difference between one bit per access and one word line per access write operations are discussed. Firstly, writing a word line at a time will worsen the voltage drop along the word line. Therefore, as shown in Figure 9, the reliable size of the cross-point array will be further reduced. The maximum array size reduces from 116×116 to 100×100 for HWFB and HWFB schemes.

In order to fairly compared the energy consumption, we compared the energy-per-bit instead of the total energy. For example, in order to write a word line with size of 128, the energy-per-bit can be calcu-

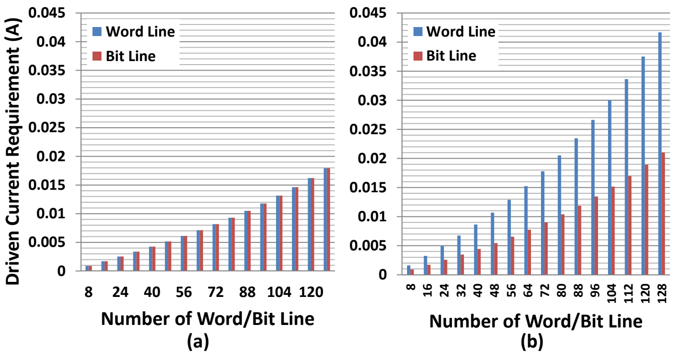

Fig. 8. The Driven Current Requirements for Word Lines and Bit Lines. (a) One Bit Writing; (B) Multi Bit Writing.

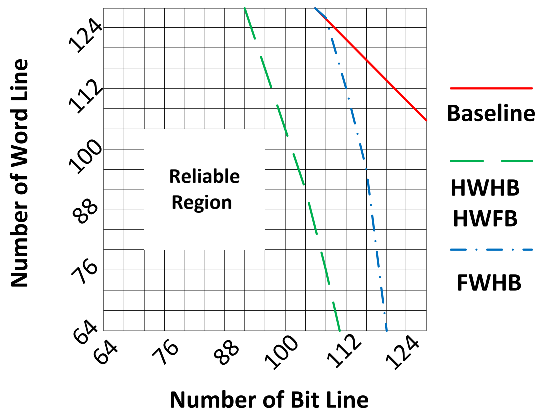

Fig. 9. The Array Size Requirement for the Cross-Point Array with Different Write Schemes. (Baseline: one bit per access. HWHB, HWFB and FWHB: one word line per access.

lated as: $E_{ave} = E_{total}/128/2$. Figure 10 shows the energy-per-bit of the multi-bit write operation. The energy shown in this figure is normalized to the same unit as Figure 7 for easier comparison. The results show that for large size of the cross point array, the multi-bit write operation is much more energy efficient. This is because the energy wasted at the unselected and half-selected cells are shared by multi bits and the average energy for one each bit is therefore reduced. However, although the multi bit write operation has the advantage of energy consumption, the maximum current requirement for each word line also increases. As shown in Figure 8(b), the maximum driven current for bit line is almost the same as one bit writing, the drive capability for wold line is almost doubled for multi-bit writing. Since the area of the voltage driver increases proportional with its driven capability, the area overhead for multi-bit writing is about 50% larger than one bit writing.

Non-linearity of the ReRAM Cell.

One of the most distinct feature of ReRAM is its non-linearity. Take the memristor based ReRAM for example, the non-linearity is observed at LRS and can be described as: the resistance of the memristor cell is not constant but varies with the applied voltage. The non-linearity coefficient is defined as: $K_r(p, V) = p \times R(V/p)/R(V)$, where $R(V/p)$ and $R(V)$ are equivalent resistance of the memristor biased at V/p and V [**?**]. Normally, the $K_r(p, V)$ value for memristor based ReRAM is larger than 20, meaning that the resistance of half-biased cell is at least 10 times larger than full-biased cell. Clearly, the ReRAM cell with larger non-linearity coefficient is more suitable to be employed as the memory cell since the current in the sneak path will be significantly reduced. Besides, the increased resistance at half-selected and unselected cell can also mitigate the voltage drop along the activated word line and bit line. Figure 11 shows the influence of different non-linearity coefficients on the array size requirements for one bit HWHB writing scheme. In this figure, the maximum array size increases from 112 × 112 to 340 × 340 when the non-linearity coefficient K_r increases to 10. Similarly, the non-linearity can also increase the maximum array size for other write schemes.

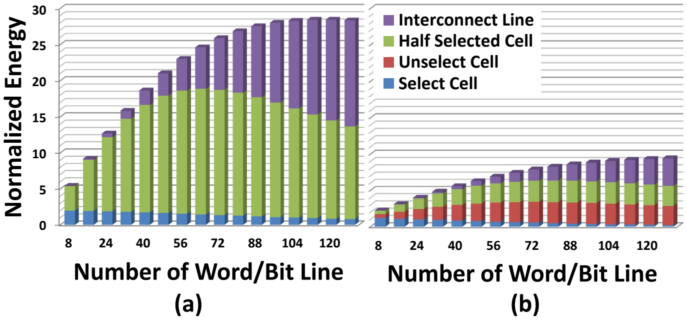

Fig. 10. The Normalized Energy Consumption per Bit for Multi-Bits Write Operation. (a): HWHB and FWHB schemes; (b): HWFB scheme.

On the other hand, the non-linearity can also reduce the energy consumption and area overheads of the cross-point array. Take the 128 × 128 array for example. As shown in Figure 12, the energy consumption for the write operation decreases dramatically with the increase of non-linearity coefficient K_r. As K_r increases from 1 to 40, the write energy is reduced by 98.3%. The driven current requirement is shown in Figure 13(a), and the corresponding area overheads of the voltage drivers are compared to the array size at Figure 13(b). The baseline design is unacceptable because the area of voltage drivers is about 11.6 times larger than the area of cross-point array. In this case, the area efficiency of ReRAM's $4F^2$ cell size will be offset by the extreme huge area overhead of the voltage drivers. However, with the increase of non-linearity, the area of voltage drivers becomes comparable to the array area. Therefore, we can conclude that, the ReRAM cells with small non-linearity coefficient are not suitable for the cross-point structure based memory array. Besides, we also study the area overhead of multi bit write. Figure 14 shows the normalized areas of the voltage drivers for one bit and multi bit write operations. As mentioned, multi bit write operation requires larger driven current. Therefore, the area of voltage driver for multi bit write operation is much larger than that of one bit write operation. Besides, normalized areas of the one bit and multi bit write operations have opposite trends as the array size increases. Normalized area

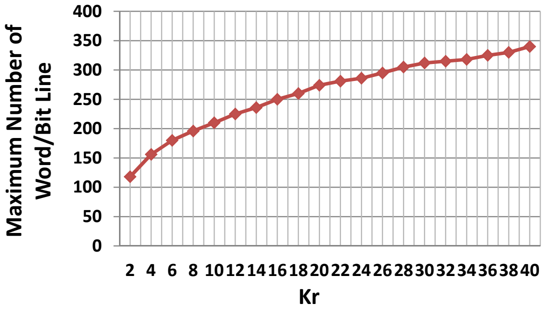

Fig. 11. The Maximum Array Size with Different Non-linearity Coefficient.

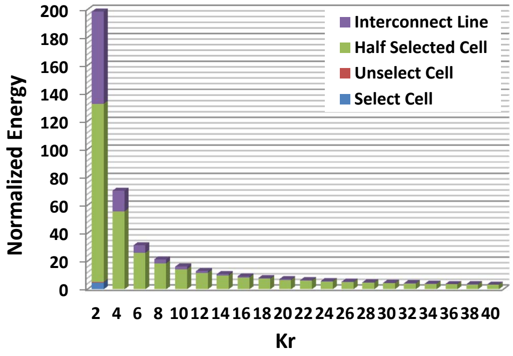

Fig. 12. The Normalized Energy Consumption with Non-linear ReRAM Cells.

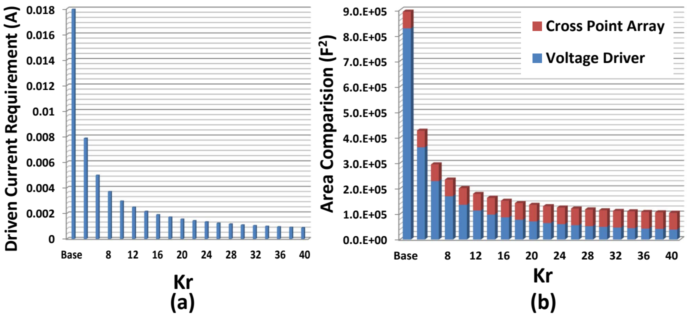

Fig. 13. The Driven Current Requirements and Area Overheads with Different Non-linearity Coefficients

for one bit write operation increases with the array size. On the contrary, normalized area for multi bit write decreases as the array size increase.

C. Read Operation

In this section, the current sensing scheme is used to determine the current in each bit line. In order to read cell $R_{i,j}$, the i^{th} word line is biased at V_{READ} and all of the other word lines and bit lines are grounded. Then the state of the selected cell is read out by measuring the voltage across R_s. The energy consumption for read operation can be analyzed by the same way as that of the write operation. Considering the read voltage is much smaller than write voltage, the read energy is expected at least one order smaller than write operation. Besides, since the read voltage/current is much lower than the write, we believe that the voltage drivers can always provide enough current for the read operation if they meet the current requirement for write operation. Therefore, we can conclude that the area overhead of voltage drivers is determined by the write current. However, the reliability of read operation is different from the write operation. The read reliability is determined by the voltage swing for reading HRS and LRS cells. Figure 15 (a) shows the voltage swing with different array sizes and K_r values. Large array sizes and large non-linearity are harmful to the voltage swing: on the one hand, larger array has more sneak paths,making the output voltage very sensitive to the data pattern of unselected cells; on the other hand, the non-linearity increases the resistance of LRS and therefore reduces the ON-OFF resistance ratio. In order to improve the reliability of the read operation, a two-step sensing scheme can be applied, which senses the current of unselected cell first, then the overall current is sensed, and after that the current difference is converted to the output voltage. The voltage swing of this two-step sensing scheme is shown in Figure 15 (b). By using the two steps sensing schemes, the voltage swing is doubled for the array with same size and non-linearity coefficient.

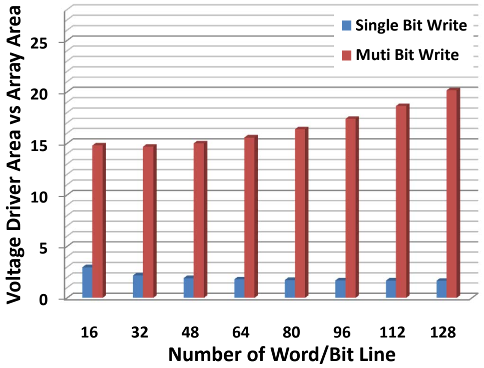

Fig. 14. The Normalized Area Overhead of Voltage Drivers ($K_r = 20$, the areas are normalized to the area of cross-point array).

V. Design Methodology

Based on the analysis of Section IV, a new design flow are proposed to explore the design space of the cross-point ReRAM array, which is shown in Figure 16. Generally, the flow can be summarized as two stages: initialization stage and the computation stage. At the initialization stage, the physical parameters, including the resistances of the ReRAM cell, interconnect wires and pull up resistors, the threshold voltage of the ReRAM cell, as well as non-linearity coefficients, are firstly input. All of this parameters are determined by materials and the process technology. Thus they and can not be changed easily during the design of memory array. After that, the design constraints should be decided according to the area/energy budgets and different applications. Then, based on these data, the original version of coefficients matrix A_{basic} and the vector of constant terms C_{basic} are set up. Since the value of A_{basic} and C_{basic} do not consider the edge conditions of the write/read schemes and therefore do not change during the design space exploration, their value are saved at this step. Then the programming schemes are chosen. The designer can either explores all of the possible programming schemes or chooses proper schemes according to the prior experience observed in our study (For example, we have already shown that the one bit write operation is more suitable for an area-constrainted design than multi bit write). Then the final step of the initialization stage is to adjust the coefficients in A_{basic} and C_{basic}

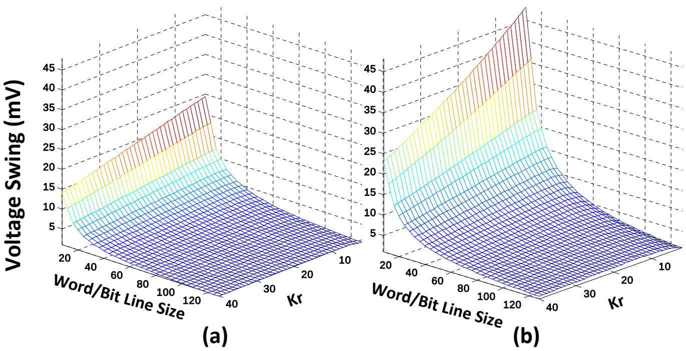

Fig. 15. Relationships among the Voltage Swing, Array Size and Non-linearity. (a) Normal Sensing Scheme; (b) Two-step Sensing Scheme

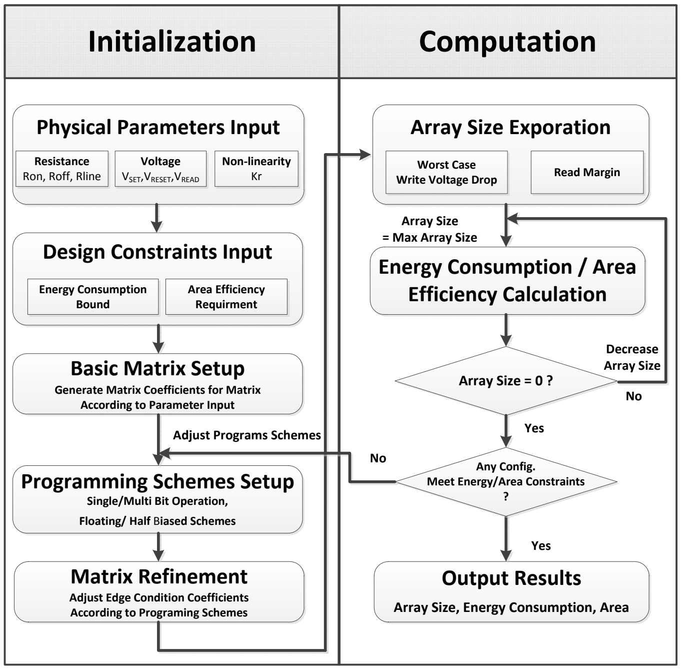

Fig. 16. The Proposed Design Flow of Design Space Exploration for ReRAM based Cross-point Array.

base on edge conditions. At the beginning of the computing stage, the reliable array size is obtained by examining the worst case voltage drop and the read margin requirement. Then an iteration is performed to calculate the energy consumption and area overheads for each array size. The result are collected after the computation. If there is any array organization that meets the design constraints provided at the initialization stage, the allowable array sizes with their energy consumption and area overheads are summarized as the design space for the given constraints. Otherwise, the programming schemes are not workable. Designers should adjust the programming schemes for a new round of evaluation.

VI. Conclusion

The ReRAM is a promising candidate of the next-generation non-volatile memory technology. The area efficient cross-point structure is the most attractive memory organization for the ReRAM based memory design. However, intrinsic problems of the cross-point structure, such as the existence of sneak current and the voltage drop along the nanowire introduce extra challenges to the design of reliable ReRAM based memory array. In this paper, a mathematical model for the cross-point array is proposed. We show that the propped model has a vary simple structure and is flexible to evaluate different write/read schemes. By using this model, the design constraints, including the array size, energy consumption, and area overheads, are analyzed in details. Based on the results of our study, a detailed design methodology is proposed, which can help designers explore the most energy/area efficient ReRAM design with different design constraints and parameters at the very early design stage.