Performance Modeling and Optimization for On-Chip Interconnects in Cross-Bar ReRAM Memory Arrays

Javaneh Mohseni, Chenyun Pan, and Azad Naeemi
Georgia Institute of Technology, 791, Atlantic Drive, NW, Atlanta, GA 30332, USA
Email: imohseni3@gatech.edu, cpan31@gatech.edu, azad@gatech.edu

Abstract—Performance modeling and optimization for on-chip interconnects in resistive RAM (ReRAM) arrays are presented for two memory design technologies of 1T1R and cross-bar arrays. Different memristor characteristics, cell structures and subarray design schemes are investigated to minimize the overall delay and energy-delay product.

Keywords—Interconnect; ReRAM; memory; performance; modeling; delay; power;

I. INTRODUCTION

As the CMOS technology advances to the deep nanoscale era, DRAM scaling faces serious challenges in speed, bandwidth, capacity, and cost [1]. NAND Flash technology has been the leading non-volatile memory technology for many years. However, it is believed that scaling this technology below 25nm has significantly degraded performance and reliability, thus, resulting in significant overhead complexity and computational power-demand from the system controller [2]. To find a solution, many emerging technologies, such as PRAM (Phase-Change RAM), MRAM (Magnetoresistive RAM), FeRAM (Ferroelectric RAM), and ReRAM (Resistive RAM), are being studied. Among these technologies, ReRAM is particularly promising. [2-6].

In addition to utilizing the memory resistor (memristor) technology to build the memory cell, RRAM arrays could be constructed using different arrangements such as the conventional 1T1R cell and the cross-bar cell arrangement. Section I and II present complete investigations of RRAM technology using the conventional 1T1R and cross-bar memory array structures. The benefits and disadvantages of each memory structure are studied and different ways to optimize the memory array are discussed.

I. MODEL APPROACHES AND ASSUMPTIONS

Table I shows the important parameters of the model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip size</td>
<td>10nm×10nm</td>
<td></td>
<td>Technology node</td>
<td>9.5nm</td>
<td></td>
</tr>
<tr>
<td>Cell architecture</td>
<td>1T1R</td>
<td></td>
<td>WL/BL/SL pitch</td>
<td>38/68/68 nm</td>
<td>[8]</td>
</tr>
<tr>
<td>V_{dd}(core/IO)</td>
<td>0.8/1 V</td>
<td>[7]</td>
<td>M1 pitch</td>
<td>58nm</td>
<td>[8]</td>
</tr>
<tr>
<td>Memristor switching time</td>
<td>300ps</td>
<td>[6]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table I. Model parameters.

A. Memory Cell Structure

In the conventional memory architecture, ReRAM cell consists of one access transistor and one memristor. The access transistor enables the connection between the sourceline (SL) connected to the cell memristor and the bitline (BL) which carries data to and from the memory cell. In this structure, the RRAM cell size is determined by the size of the large access transistor required for delivering the large write current to the memory cell. This prevents further increase in cell density and memory capacity.

B. Memory Subarray Architecture

The memory array is divided into a number of banks. The global interconnects transmit address and data from the array input to the bank, forming a hierarchical-tree (h-tree) network [9]. The wordlines (WL) are connected to the gates of the transistors, the SLs are connected to the source of the transistors and also to the cell memristor, and the BLs are connected to the drain of the access transistors and carry the data to and from the memory cell. Elmore delay model is used to calculate the interconnect delay. An optimum number of repeaters are placed along the global interconnects. For the WLs, BLs, and SLs, no repeaters are used due to the cell size limit. To calculate the delay of the wordlines, the capacitance of transistors connected to the wordlines is added to the wire capacitance. Figure 1 shows the ReRAM cell and subarray structures including the different local and global interconnects.

![Image](image1.png)

Fig. 1 ReRAM cell and subarray structures showing different local and global interconnects.

For writing a “0” value to a memory cell, a high voltage is put on the WL, the BL is connected to V_{DD}, and the SL is connected to GND, which results in the cell memristor switching from the low-resistance (LR) state to the high-resistance state (HR). For writing a “1” to a memory cell, the BL and SL voltages are reversed, which causes the cell memristor to switch from HR to LR. The reading process is similar to writing a “0” value, but I_{Read} is much smaller than I_{Write} in order to avoid changing the resistance state of the memristor while reading the cell value.
There are different directions in the field of ReRAM research that include optimization of memristor, cell structure, memory configuration, layout, and interconnects. To find the most crucial research direction, the results for the memory performance are studied in order to find the bottlenecks of the ReRAM performance.

II. MODEL RESULTS AND DISCUSSIONS

Figure 2(a) shows the total memory delay for the read and write operations. The delay of peripheral circuits such as decoders, multiplexers, and amplifiers have been included in the model, but their values are negligible and not visible in the figure. The main contributors to the memory delay are the local and global interconnects, and the cell. As for the interconnects delay, for a small number of banks, since the banks are large and the WLs and the BLs are long, most of the interconnect delay comes from the local interconnects inside the banks; i.e. the WL and BL. For a large number of banks, most of the interconnect delay comes from the global address and data interconnects. Increasing the number of banks reduces the local interconnects delay, but has little impact on the global interconnects delay.

Most of the dynamic power is consumed in the global address interconnects, data interconnects, and the cell as shown in Fig. 2(b). Data interconnects dissipate far more energy than the address interconnects since they transmit decoded data, and as a result require a much larger number of wires. By increasing the number of memory banks, the total memory latency is reduced while the dynamic power consumption is increased, shown in Figures 2(a) and (b). However, the increase in power consumption would not be problematic since, as seen in Fig. 2(c), by increasing the number of banks, the access energy to a memory block is still reduced. This is due to the large reduction in the memory block access time. As the number of memory banks increases, the limitation comes from the added peripheral circuits area since each bank has its own set of peripheral circuits which include decoders, multiplexers, and sense amplifiers. As seen in Fig. 2(d), we have only increased the number of banks to the point that the total peripheral circuits area does not exceed 10% of the memory chip area.

III. CROSS-BAR ReRAM ARRAY

As mentioned before, in the 1T1R memory cell structure, the cell size is determined by the cell access transistor size. The large size of the access transistor capable of delivering the large required current for switching the cell memristor prevents further scaling of the ReRAM cell and increasing the cell density. A solution to this problem is the cross-bar arrangement for memory array. In the cross-bar structure, the access transistors used to enable the connection between BL and cell memristor and the WLs connected to the access transistors are removed. The cell memory cells are placed at the intersections of the BLs and the SLs, and access to a memory cell is realized by applying different voltages to the BLs and the SLs.

The cross-bar memory structure has two important advantages. By reducing the cell size, it increases the cell density and memory capacity. In addition, by removing the WL and the access transistor, it reduces the memory latency and power consumption. The relative improvement in memory delay offered by the cross-bar architecture depends on how dominant the WL delay is compared to 1T1R array. Figure 3(a) shows the delay components for the cross-bar ReRAM array. Figure 3(b) shows the total memory delays for ReRAM arrays using the conventional 1T1R and the cross-bar memory structures. As the number of banks increases, and banks get smaller, the WL delay becomes less dominant, and the difference between the delay of cross-bar and 1T1R arrays becomes smaller.

As for the chip area, using the crossbar ReRAM technology, the cell size is only limited by local interconnects pitch inside the memory bank. This enables the ideal cell size of 4F² for the memory cell where each cell side is only as wide as the local wire pitch (2F). At 9.5nm technology node, assuming die size of 100mm², this increases the memory capacity by 250% from 2.0788 Gbit to 7.2758 Gbit.

In the cross-bar structure, for writing “1” to a cell, the SL connected to the cell is connected to GND and the BL is connected to V<sub>SET</sub>. All the other SLs and BLs are connected to V<sub>SET</sub>/2. This way, the memory cells are divided into three groups that have the voltages of zero, V<sub>SET</sub>/2, and V<sub>SET</sub> across them. The cells with V<sub>SET</sub>/2 across them are a source of concern since they draw parasitic currents and reduce the available voltage across the selected cell memristor. To address this issue, devices with nonlinear I-V characteristics similar to diodes are integrated in series with the memristors that only allow the passage of current when the voltage across the cell is V<sub>SET</sub>, and are called the cell selectors [10].

IV. MEMRISTOR CHARACTERISTICS

Even after eliminating the leakage current paths by using cell selectors, the voltage drop along the BL and SL could be problematic since it creates the need for high V<sub>SET</sub> and V<sub>RESET</sub> supply voltages. The severity of this issue depends on the ratio of the sum of the memristor resistance and the local...
interconnects resistance to the memristor resistance. As a result, this is especially challenging during the RESET operation when the memristor is at LR state. Fig. 4(a) shows the required \( V_{\text{SET}} \) and \( V_{\text{RESET}} \) for the case of using memristor with LR and HR of 380\( \Omega \) and 360 k\( \Omega \). As seen in Fig. 4(a), even at the highest bank number of \( 2^{18} \), the required \( V_{\text{RESET}} \) for a memristor with LR of 380 ohm is about 10V, which is too high for an on-chip voltage supply. To address this problem, either the banks should be made smaller with shorter BLs and SLs, or memristors with higher LR values should be utilized. In a 2D memory, the bank size is limited at two levels; i.e. design level and operation level. At the design level, as mentioned before and seen in Fig. 2(d), the bank size reduction is limited by the added peripheral circuits area due to the increasing bank number, assuming a constant memory chip area. At the operation level, the bank size reduction is limited by the memory block size, and reducing the memory block size reduces the system throughput. Figure 4(b) shows the required \( V_{\text{RESET}} \) for different bank numbers and different memristor LR values. As the number of banks increases, and the banks get smaller, the BL and SL resistance becomes smaller, and the ratio of the sum of the memristor resistance and the BL and SL resistance to the memristor resistance approaches one. As a result, for large numbers of banks, the difference of \( V_{\text{RESET}} \) voltage for different LR values becomes smaller. Figure 4(c) shows the minimum limit for memristor LR in order to keep \( V_{\text{RESET}} \) below 3V.

One of the areas that the cross-bar ReRAM could stand out from the previous memory technologies is that in this structure, by adding metal levels, the BL and SL pair levels are built on top of each other with memory cells forming at the interconnects junctions. With this technology, a 3D cross-bar ReRAM array is built, and the bank footprint area could be reduced. As a result, \( V_{\text{RESET}} \) could be reduced without reducing the bank capacity or the memory block size.

![Fig. 4 Required SET and RESET voltages for cross-bar ReRAM array with different bank numbers and (a) memristor LR=380\( \Omega \) and HR=360k\( \Omega \), (b) memristor HR=360k\( \Omega \) and different LR values. (c) Minimum limit for memristor LR to keep the required \( V_{\text{RESET}} \) below 3V.](image)

The advantages of using a memristor with higher LR are lower required \( V_{\text{RESET}} \), lower dynamic power consumption, and lower area of chip peripheral circuits since the memory array could be divided into a smaller number of larger banks. The disadvantages of increasing the memristor LR is an increase in the cell delay. Figures 5(a) and (b) show the impact of memristor LR on memory cell delay and power consumption. Increasing the memristor LR has the largest impact on the cell delay and power consumption during the cell read operation when the cell memristor is in LR state, and the memory cell has the “1” value. The cell write delay and power consumption are not affected considerably by the LR value since for the write operation, the memristor resistance changes from LR to HR or vice versa, and it is the worst case HR value that determines the overall delay because multiple bits of ones and zeros are written simultaneously.

![Fig. 5 (a) Cell delay, and (b) cell dynamic power consumption during the read operation of memory cell with “1” value for different values of memristor LR and memory bank number.](image)

VI. CONCLUSION

A comprehensive interconnect analysis is presented for ReRAM chips, and the limits interconnects impose on the total memory delay, area, and dynamic power consumption are quantified. It is found that global and local interconnects constitute up to 80% of the memory delay. Two ReRAM technologies of 1T1R and cross-bar array are compared. The advantages and challenges of cross-bar ReRAM arrays are investigated, and potential solutions are presented. Impacts of the memristor characteristics on the ReRAM performance are studied. Other challenges of the cross-bar structure such as lower read margin and higher leakage power resulting from the non-ideality of cell selectors will be investigated in future publications.

REFERENCES