Next Generation Nonvolatile Memory
*Its Impact on Computer System*

Dec.04.2013
Sung Hyun Jo and Hagop Nazarian
Next Generation Nonvolatile Memory
*Its Impact on Computer System*

- Challenges of current NVM technology
- Requirements for next generation memory
- Next Generation Memory Developments
- Operation Mechanism of various RRAMs
- Advanced RRAM Technology
- Design & Architectural attributes
- System Benefits
- Comparison with current NVM technology
For several years now, companies have focused on developing a next generation memory technology that will lead to significant improvements in reliability, performance, low power operation and scalability compared to existing non-volatile memories. **Forward Insights believes that RRAM, including Crossbar’s approach, has the potential to succeed NAND flash memory due to its scalability and manufacturability.** – Greg Wong, Forward Insights, August 2013

“The current storage medium, planar NAND, is seeing challenges as it reaches the lower lithographies, pushing against physical and engineering limits. **The next generation non-volatile memory, such as Crossbar’s RRAM, would bypass those limits**, and provide the performance and capacity necessary to become the replacement memory solution.” - Michael Yang, IHS, August 2013
Flash Memory Scaling Challenge

- Information storage in Flash is based on charge density \((C/cm^2)\)
- At 20nm, \(~100\) electrons are stored in the FG \((\Delta V_t = 1V)\)
- Losing a few electrons can cause severe reliability issues
- Scaling = exponentially increasing BER, reduced data retention and cycling
Scaling challenges on BER, and Endurance

BER, ECC, ECC Area Overhead vs. Technology Node

Endurance, BER vs. Technology Node
System Requirements For Next Generation Memory

- Reduce Latency
- Lower Power Consumption
- Improved Reliability and Higher P/E Cycles
- Scalable to several generations
- Embeds in advanced CMOS technology nodes
- Cost effective

- RRAM is the emerging technology with impressive characteristics. It will meet the demands for next generation systems
Introduction to RRAM Technology
**Resistive Random Access Memory (RRAM)**

- Non-charge based emerging nonvolatile memory technology
- Typically two terminal structure
- Information storage based on multiple electrical resistance states
  - Resistance switching by voltage or current signal
- Either bipolar and/or unipolar switching
# Resistance Switching Classification

- RRAM utilizes 1D or 2D effect → ultimate scaling potential

<table>
<thead>
<tr>
<th>Resistance Switching</th>
<th>Valence change</th>
<th>Electro-chemical metallization</th>
<th>Thermo-chemical</th>
<th>Uniform V&lt;sub&gt;o&lt;/sub&gt; exchange</th>
<th>Thermal</th>
<th>Magneto resistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar</td>
<td>Bipolar</td>
<td>Unipolar</td>
<td>Bipolar</td>
<td>Unipolar</td>
<td>Bipolar</td>
<td></td>
</tr>
</tbody>
</table>

- Physical Effect
  - **1D Filament**
  - **2D Interface**
  - **3D Bulk**

- NVM Category
  - **RRAM**
  - **PCRAM**
  - **MRAM**

Images from -
4. Sanchez et al., NCCAVS (2009)
RRAM

- Discrete 1D filament allows low power, high density & reliable RRAM

<table>
<thead>
<tr>
<th>Valence change</th>
<th>Electro-chemical metallization</th>
<th>Thermo-chemical</th>
<th>Uniform $V_o$ exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar</td>
<td>Bipolar</td>
<td>Unipolar</td>
<td>Bipolar</td>
</tr>
</tbody>
</table>

Continuous CF | Discrete CF | Continuous CF | Continuous CF | 2D Interface |

1D Conducting Filament (CF)

Images from -
4. Sanchez et al., NCCAVS (2009)
Valence Change RRAM

- Bipolar switching by *the migration of oxygen* under electric field
- Switching medium – typically transition metal oxide (e.g. TaOₓ, HfOₓ, TiOₓ)
- Electrode – typically inert metal (e.g. Pd, Pt)
- SET – generation of oxygen vacancies and formation of a filament(s)
- RESET – oxidation of the filament(s)

*Actual filament(s) growth direction (e.g. BE → TE, TE → BE) depends on several factors such as switching layer material (e.g. oxygen deficient vs. metal deficient, electron affinity) and bias scheme.*
Electrochemical Metallization RRAM

- Bipolar switching by *the migration of metal ions* under electric field
- Electrode – active metal (e.g. Ag, Cu,..)
- Various switching materials such as chalcogenide, amorphous silicon,…
- SET – anodic dissolution of active metal and formation of a filament(s)
- RESET – electrochemical dissolution of the filament(s)

![Diagram showing SET and RESET processes](image-url)

- **SET**
  - active metal (M)
  - inert electrode
  - $\textbf{+}$
  - low R

- **RESET**
  - inert electrode
  - $\textbf{+}$
  - high R

- active metal ion (e.g. \(M^+, M^{2+}\))
- neutral active metal atom

*Actual filament(s) growth direction (e.g. BE → TE, TE → BE) depends on several factors such as metal ion mobility in the switching medium, ion trap density, leakage current density, and bias scheme.*
Thermochemical RRAM

• Unipolar switching (fuse – antifuse) triggered by Joule heating
  - Local dielectric breakdown → heating → local structural modification (local redox reaction)
• Switching medium – some transition metal oxides (e.g. NiO)
• SET – local heating-induced $V_o$ generation or electrode metal diffusion with current compliance
• RESET – thermal dissolution (rupture) of the filament with higher current (larger heating)

*Actual filament(s) formation/rupture process (oxidation vs. metal migration) depends on several factors such as switching material and bias condition.
Signature of 1D Filamentary Switching

- **ON** – area independent
- **OFF** – depends on switching materials and bias conditions

![Diagram showing ON and OFF states for Quantum point contact and Tunneling/Schottky contact](image)

Switching medium examples - HfOx, TaOx

Switching medium examples – a-Si, TiOx

RRAM Scaling

- Area independent conducting filament
- ON/OFF ratio improves as device size decreases
  - Higher sensing margin (faster read speed)
  - Larger array possible
- Sub 10nm scaling potential
Sub-20nm Crossbar RRAM

- Superior performance still maintained in sub-20nm devices
- Large ON/OFF ratio allows ≥ 2bits/cell on the same physical bit

Multi-level cell (MLC) Demo (sub-20nm device)

MLC Cycling (sub-20nm device)
Large endurance $> 10^{10}$ P/E cycles

- Crossbar cell has demonstrated endurance $>10^{10}$ cycles
- ON/OFF ratio of $>100X$ is maintained
Retention > 10yr @85°C

- Large ON/OFF ratio maintained @85°C for 10yrs
- Multiple devices measured under the same conditions, show very similar retention characteristics
Good Thermal Stability

- Cycling parameters show no dependence on temperature dependence
- 100X ON/OFF ratio is maintained across the whole temperature range
Immune to Program and Read Disturb

- No program disturb observed at voltages lower than the programming voltage
- No change in either the program state or the erased state after >10^8 read cycles
- Immunity to read disturb is maintained at 85°C
RRAM Integration – 1 Transistor per 1 RRAM Cell

1T1R
- For high performance (e.g. speed)
RRAM Integration - 1 Transistor per $n$ RRAM Cells

Crossbar (1TnR)

- For high density
Crossbar Architecture – Leakage Current Control

- Reducing leakage current by
  - Non-linear IV (increased R in small bias)
  - Rectifying IV (increased R in reverse bias)
Switching Behavior Modulation

- Both non-linear IV and rectifying switching obtained by switching medium optimization and process control
  - With still high on/off ratio of $10^3 \sim 10^6$

*IV curves obtained from different devices which are designed for different product requirements
Information Storage in Passive Crossbar

Lu et al., Nano Lett. (2012)
Stackable 3D Memory Array

- Simple materials and structure
- Low temperature fabrication process
- Easy integration with standard CMOS logic
- → 3D stackable memory architecture
RRAM for Neuromorphic System
Modern Computer System – Complex & Inefficient

- Computer systems consume several orders of magnitude higher energy than the animal’s brain for complex (multiple inputs) tasks.

IBM Blue Gene/P supercomputer
Capable of cat’s cortical level simulation at 83 times slower than the real.
Highly Parallel Computing for Improved Efficiency

- **Sequential processing nature of computers** (inefficient and complex system architecture) ↔ **Highly parallel nature of the neural system** (highly efficient system)
- Key to the high efficiency of bio-systems is the large connectivity between neurons
RRAM Synapse for Neuromorphic System

- CMOS neurons + RRAM synapses in a neuromorphic system
- Crossbar structure for the neural network
Synaptic Function Demonstration by RRAM Synapse

- STDP (Spike Timing Dependent Plasticity) implemented by a hybrid CMOS neuron/memristive device (RRAM) synapse system

- support important synaptic functions
- frame work for neuromorphic systems

Product specifications
Data, Code, and Embedded
## Crossbar Offers Compelling Technical Advantages

<table>
<thead>
<tr>
<th>Applications</th>
<th>Embedded</th>
<th>Code Storage</th>
<th>Data Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>eFLASH</td>
<td>CODE FLASH</td>
<td>Solid State Drive</td>
</tr>
<tr>
<td>256K-4Mbit</td>
<td>Crossbar™</td>
<td>512K-8G</td>
<td>NAND Flash</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>90nm</td>
<td>45nm</td>
<td>Crossbar</td>
</tr>
<tr>
<td>256K-16Mbit</td>
<td>&lt;10nm</td>
<td>&lt;10nm</td>
<td>128Gbit</td>
</tr>
<tr>
<td><strong>Cell Size</strong></td>
<td>18F^2 - 42F^2</td>
<td>6-12F^2</td>
<td>256Gbit</td>
</tr>
<tr>
<td>5.4F^2 - 18F^2</td>
<td>5.4F^2</td>
<td>5.4F^2</td>
<td>20nm</td>
</tr>
<tr>
<td><strong>Program byte</strong></td>
<td>10us</td>
<td>10us</td>
<td>Not Capable</td>
</tr>
<tr>
<td>Program page</td>
<td>2us</td>
<td>300us - 1.4ms</td>
<td>1.2ms</td>
</tr>
<tr>
<td><strong>Erase byte</strong></td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us</td>
</tr>
<tr>
<td>Erase page</td>
<td>2us</td>
<td>2us</td>
<td>16us</td>
</tr>
<tr>
<td>256us 4ms</td>
<td>2us</td>
<td>256us</td>
<td>2us</td>
</tr>
<tr>
<td><strong>Erase block</strong></td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us</td>
</tr>
<tr>
<td><strong>Read Latency</strong></td>
<td>30ns-100ns</td>
<td>100ns</td>
<td>50us</td>
</tr>
<tr>
<td><strong>Endurance Retention</strong></td>
<td>1 million 10Yr@125C</td>
<td>100K 20Yr@55C</td>
<td>&lt;1K</td>
</tr>
<tr>
<td>Density</td>
<td>10Yr@55C</td>
<td>100K</td>
<td>10K</td>
</tr>
<tr>
<td>Technology</td>
<td>10Yr@55C</td>
<td>10K</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Cell Size</td>
<td>10Yr@55C</td>
<td>10Yr@55C</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Program byte</td>
<td>10Yr@55C</td>
<td>10Yr@55C</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Program page</td>
<td>2us</td>
<td>2us/4ms</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase byte</td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase page</td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase block</td>
<td>2us</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Read Latency</td>
<td>30ns-100ns</td>
<td>100ns</td>
<td>50us</td>
</tr>
<tr>
<td>Endurance Retention</td>
<td>1 million 10Yr@125C</td>
<td>100K 20Yr@55C</td>
<td>&lt;1K</td>
</tr>
<tr>
<td>Density</td>
<td>10Yr@55C</td>
<td>100K</td>
<td>10K</td>
</tr>
<tr>
<td>Technology</td>
<td>10Yr@55C</td>
<td>10K</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Cell Size</td>
<td>10Yr@55C</td>
<td>10Yr@55C</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Program byte</td>
<td>10Yr@55C</td>
<td>10Yr@55C</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Program page</td>
<td>2us</td>
<td>2us/4ms</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase byte</td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase page</td>
<td>Not Capable</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Erase block</td>
<td>2us</td>
<td>Not Capable</td>
<td>2us/4ms</td>
</tr>
<tr>
<td>Read Latency</td>
<td>30ns-100ns</td>
<td>100ns</td>
<td>50us</td>
</tr>
<tr>
<td>Endurance Retention</td>
<td>1 million 10Yr@125C</td>
<td>100K 20Yr@55C</td>
<td>&lt;1K</td>
</tr>
<tr>
<td>Density</td>
<td>10Yr@55C</td>
<td>100K</td>
<td>10K</td>
</tr>
<tr>
<td>Technology</td>
<td>10Yr@55C</td>
<td>10K</td>
<td>10Yr@40C</td>
</tr>
<tr>
<td>Cell Size</td>
<td>10Yr@55C</td>
<td>10Yr@55C</td>
<td>10Yr@40C</td>
</tr>
</tbody>
</table>
Design & Architectural Attributes
RRAM Design Suited for Embedded Memory

- Suited for high speed embedded memory operation
- Backend process integration. Easier to integrate and less expensive than eFlash
Crossbar array architecture
Suitable for high density memory NOR/NAND

- One transistor selects many RRAMs
- Stackable architecture - Effective cell $4F^2/L$ - $L$ is the number of stacks - $1F^2$ with 4 stacks
- The transistor sizes is not the cell size limiter – No need to down scale the transistors
- Area under the array could be utilized for peripheral circuits – Provides high array efficiency
- Competitive with NOR, NAND, and next generation 3D NAND architectures
- Backend process Integration
RRAM array with linear resistance characteristics or without select device
Linear resistance RRAM in a cross-point array

- Making cross-point 1TnR arrays with linear resistance RRAM cells generates sneak paths (dotted red lines) significantly reducing sensing margin, increasing power, and limiting sector size.
- Biasing is very challenging - Any small potential difference between unselected BL & WL generates will generate very large current consumption.
- Therefore, Linear resistance RRAM utilize 1T1R architecture.
RRAM array with Non-Linear Hysteric IV
Non-linear RRAM in a cross-point array

• RRAM with nonlinear complementary barrier characteristics will:
  • Mitigate the sneak path problem.
  • Yield larger arrays and larger sensing margins, higher programming throughput, and larger array efficiency
Word based Crossbar RRAM Array – Power optimized

Pros:
- Row Alterable for program and erase
- Lower power consumptions – precharge/activate one bank for a byte
- Potentially better immunity to disturb conditions

Cons:
- Slower sensing and pattern sensitive
I/O based Crossbar RRAM Array architecture – Write/Read Speed optimized

Pros:
- Row writable can erase and program simultaneously
- Faster sensing speed and less pattern sensitive

Cons:
- Higher power consumption – Precharges 8 banks for a byte
Each RRAM cell is MLC programmed into different resistance values by limiting the current flowing in each cell during program operation.
Crossbar RRAM and Its Impact on System Performance
# NAND Characteristics, Impact, Remedies, and Trade off

<table>
<thead>
<tr>
<th>NAND Characteristics</th>
<th>Impact to Storage System</th>
<th>Improved by</th>
<th>Trade off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Retention &amp; high BER</td>
<td>Reduces lifetime</td>
<td>ECC (BCH, LDPC)</td>
<td>Controller Overhead &amp; Cost Power consumption</td>
</tr>
<tr>
<td>Low P/E Cycles</td>
<td>Reduces lifetime</td>
<td>Wear Leveling</td>
<td>Performance &amp; Controller Overhead &amp; Cost</td>
</tr>
<tr>
<td>No ReWrite feature</td>
<td>Write amplification</td>
<td>Garbage Collection</td>
<td>Performance &amp; Controller Overhead &amp; Cost</td>
</tr>
<tr>
<td>No page alterable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Page erase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow page read</td>
<td>Random Read Performance &amp; Latency</td>
<td>None</td>
<td>Performance</td>
</tr>
</tbody>
</table>
Present NAND FLASH Technology Trends

**BER, ECC, ECC Area Overhead vs. Technology Node**

**Endurance, BER vs. Technology Node**
NAND’s Re-Write limitation, and Data Revision process

- NAND cannot revise or alter data on a page level – Erase is performed through the bulk substrate which is common to the entire NAND block of the memory cells
- The entire block of memory need to be erased before revising a data. This process will take long time and will accelerate device reliability degradation due to excessive Program Erase (P/E) cycles
- To circumvent excessive P/E cycles revised data is programmed in an erased location
- Logical to Physical mapping (L2P) is also generated and stored in a DRAM location to direct the controller to the address of the revised data
- The controller has to update and maintain this every time data revision is performed
Garbage Collection & Write Amplification

<table>
<thead>
<tr>
<th>Block m</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
<tr>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
<tr>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
<tr>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
</tbody>
</table>

| Block n | Free | Free | Free | Free | Free | Free | Free | Free |

- Block m, and Block 1 are initially erased (all free)
- 8 pages (A-H) are written to Block m

<table>
<thead>
<tr>
<th>Block m</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>B'</td>
<td>C'</td>
<td>D'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E'</td>
<td>F'</td>
<td>G'</td>
<td>H'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Block n | Free | Free | Free | Free | Free | Free | Free | Free |

- 8 additional pages (I-P) are programmed
- Pages (A-H) are revised to (A'-H') and written to the free erased pages

| Block m | Free | Free | Free | Free | Free | Free | Free | Free |

| Block m | Free | Free | Free | Free | Free | Free | Free | Free |

| Block n | Free | Free | Free | Free | Free | Free | Free | Free |

- Valid pages of Block m are moved to controller buffer, and reprogrammed back to Block n
- Block m is erased
- As a result 8 pages were freed or reclaimed

- In this example, 16 pages of data had to be moved from block 0 to block 1 to free up 8 pages that were occupied by stale data. Write Amplification for such a storage device is equal to $24 \text{ (total pages in a block)} / 8 \text{ (freed up pages)} = 3$. 
SSD System NAND-Based

NAND Shortcomings: L2P Mapping, Garbage Collection, Wear Leveling, Bad Block Management, ECC Complexity
SSD System  RRAM-Based

RRAM-Based SSD substantially reduces NAND shortcomings, thus significantly reducing controller complexity.
SSD System Write Performance with NAND & RRAM

<table>
<thead>
<tr>
<th>NAND Spec.</th>
<th>MLC</th>
<th>SLC</th>
<th>RRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND bus freq DDR (MHz)</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Bus width (bits)</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Page Size (KB)</td>
<td>16</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Shift Time + Overhead (us)</td>
<td>100</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>Program Time (ms)</td>
<td>1.5</td>
<td>0.3</td>
<td>0.032</td>
</tr>
<tr>
<td>Read Latency (us)</td>
<td>50</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Write Amplification</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Effective Write xfer rate (MB/s)</td>
<td>32</td>
<td>53</td>
<td>160</td>
</tr>
</tbody>
</table>

- Maximum utilization of the channel
- 5X performance improvement
In Summary

- RRAM provides future Systems
  - Superior performance and lower system power consumption
  - Better reliability
  - Larger densities with 3D integration
  - Embedded memory in advanced CMOS nodes
  - Ease of manufacturability with standard CMOS compatible material
  - Scalability sub <10nm nodes
  - New system architectures