TaO$_X$-based ReRAM for Variability-Aware Approximate Computing

Chihiro Matsui, Shouhei Fukuyama, Atsuna Hayakawa, and Ken Takeuchi
Chuo University, Tokyo, Japan
Reliability-Aware Approximate Computing in Storage

Which hierarchy of ReRAM storage has Error Toleration techniques?
How to Relax Reliability for Approximate Computing?

Outline

• Variability-Aware Approximate Computing (V-AC)
• Application-Induced Variability of TaOₓ ReRAM Cell Errors and V-AC Evaluation Platform
• V-AC Strategies of System, Circuit and Device Co-Design (SCDCD)
• Conclusions
Variability-Aware Approximate Computing (V-AC)

System, Circuit, and Device in ReRAM-based storage have Variabilities in nature

By tolerating variability, Performance, Energy, and Cost gain
## Typical Cell Target Strategy of V-AC in ReRAM Storage

### Target Typical ReRAM Cell, NOT Worst Cell

<table>
<thead>
<tr>
<th></th>
<th>Error tolerance</th>
<th>ECC</th>
<th>Set/Reset</th>
<th>LRS Read</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conv. Computing</strong></td>
<td><strong>No error</strong></td>
<td><strong>Worst cell</strong></td>
<td><strong>Slowest cell</strong></td>
<td><strong>Smallest $I_{\text{CELL}}$ cell</strong></td>
</tr>
<tr>
<td><strong>V-AC</strong></td>
<td><strong>1-10 % errors [*]</strong></td>
<td><strong>Typical cell</strong></td>
<td><strong>Typical cell</strong></td>
<td><strong>Typical $I_{\text{CELL}}$ cell</strong></td>
</tr>
</tbody>
</table>

[*] Y. Yamaga et al., *VLSI Tech* 2018.
System, Circuit and Device Co-Design (SCDCD) Platform [*]

![Graph showing application workload and timing]

- Input
- Storage Controller Emulator
- Conv. Exact Computing
- V-AC for ML

Output Results
- IOPS performance
- Energy consumption
- ReRAM Set/Reset cycles

**Application Workload**

<table>
<thead>
<tr>
<th>Time [day]</th>
<th># of requests [x10^5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

- prxy_0 app [4]

**ReRAM Device Emulator**

- TaOx ReRAM
- Sense Amplifier

- 1 sector = 512 Byte
- 1 page = 16 KByte

[*] C. Matsui et al., VLSI Technology 2019.

Set/Reset in TaO$_X$-based ReRAM Cell [*]

- In Set (Reset) operation, LRS (HRS) is formed by moving O$^{2-}$ to TaO$_X$ layer (CF)
- Percolation paths connect (disconnect) between V$_{Os}$ in LRS (HRS)

[*] Z. Wei et al., IEDM 2008.
TaO$_x$ ReRAM Conductive Filament (CF) Model [*]

- **Write-hot data** decrease Vo density in CF by horizontal diffusion
- **Relaxation effect** reconnects percolation paths by interface Vo diffusion to CF
- **Read-hot data** cause weak reset by vertical Vo diffusion
- **Data retention of cold data** causes horizontal Vo diffusion

[*] S. Fukuyama et al., *IRPS* 2019.
Storage System Variability

Non-uniformity of Data Access

<table>
<thead>
<tr>
<th>Sorted address space [GByte]</th>
<th># of requests</th>
<th>Requests of proxy_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10^5</td>
<td>Hot data</td>
</tr>
<tr>
<td>0.2</td>
<td>10^4</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>10^3</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>10^2</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>10^1</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>10^0</td>
<td>Cold data</td>
</tr>
</tbody>
</table>

Set/Reset Cycles Variability

- 95% cycles of Typical Cells
- x10^3 cycles of Worst Cells

- Write-hot data induce large Set/Reset cycle difference in cells without smoothing by Wear-leveling (Wear/L) [*]
- Set/Reset cycles of Typical Cells reduce by 95% while those of Worst Cells increase by x10^3

[*] T. Onagi et al., SSDM 2014.
ReRAM Device-induced BER Variability

- Measured LRS show tail error cells at high Set/Reset cycles [*]
- BER increases with Set/Reset cycles

[*] K. Maeda et al., IRPS 2017.
System-induced BER Variability

Application-induced Variability

Device-induced Variability

System-induced ReRAM Cell BER Variability

For Variability-Aware Approximate Computing (V-AC)
# V-AC Error Tolerance Strategies [*]

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Operation</th>
<th>Conv. computing</th>
<th>Proposed Strategy</th>
<th>Proposed V-AC Technique</th>
<th>Data characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Wear-leveling (Wear/L)</td>
<td>w/ Wear/L</td>
<td>I</td>
<td>w/o Wear/L</td>
<td>Write-hot Read-hot</td>
</tr>
<tr>
<td></td>
<td>ECC</td>
<td>Worst-error target (35-bit correction)</td>
<td>II</td>
<td>Typical-error target (5-bit correction)</td>
<td>✓</td>
</tr>
<tr>
<td>Circuit</td>
<td>Read</td>
<td>NA</td>
<td>III</td>
<td>Adaptive Read</td>
<td>✓</td>
</tr>
<tr>
<td>Device</td>
<td>Set/Reset</td>
<td>Verify NA</td>
<td>IV</td>
<td>w/o Verify</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower $V_{SET}/V_{RESET}$</td>
<td>✓</td>
</tr>
</tbody>
</table>
Strategy I: Wear-Leveling (Wear/L) Elimination

Wear/L Operation [*]

- Overwrite data in Page A
- Wear/L to Page B

Endurance Error Reduction by Wear/L

- Read and write to Low Set/Reset cycle cells

- Wear/L reduces BER of ReRAM storage by smoothing Set/Reset cycles. However, Total Set/Reset cycles increase by extra data copy

[*] T. Onagi et al., SSDM 2014.
Strategy I: Wear-Leveling (Wear/L) Elimination

- Strategy I eliminates Wear/L to remove extra data copy and improves storage performance by 33%
Strategy II: Typical-Error Target ECC

- ECC has trade-off between error correction capability, decoding time and code-rate (cell area)

ECC Structure

- Code word = n, Code-rate = k/n
- User data k, Parity (n-k)

BCH ECC Trade-off

- k = 512 Byte (1 sector)
- Stronger ECC
Strategy II: Typical-Error Target ECC

- ECC code-rate increase and performance improves by 85%
Strategy III & IV: Error Toleration in Circuit & Device

Strategy III (Circuit):
Adaptive Read

- Set/Reset cycles = $10^5$
- $V_{\text{RESET}} = 0.23$ [a.u.]
- 8 Kbits, 210 degC

Retention time:
- LRS
- HRS

- $I_{\text{REF}}$ for cold data
- $I_{\text{REF}}$ for hot data

Read Performance

Strategy IV (Device):
Verify Elimination

- Lower $V_{\text{SET}}/V_{\text{RESET}}$

Write Performance

Selector Scalability

Flash Memory Summit 2019
Santa Clara, CA
Conclusions

- Application-induced Variability-aware Approximate Computing (V-AC) is proposed with System, Circuit and Device Co-Design (SCDCD).
- Performance, Energy, and Cell Area of ReRAM storage improve by \( x7.0 \), 90\%, and 8.5\%.
Conclusions

Thank you for your attention

This presentation is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO)