Energy-efficient Redox-based Non-volatile Memory Devices and Logic Circuits

Rainer Waser,
Vikas Rana, Stephan Menzel, Eike Linn

Jülich-Aachen Research Alliance JARA, Section Fundamentals of Future Information Technology PGI-7, FZ Jülich & IWE2, RWTH Aachen University
ReRAM Group

Regina Dittmann  Eike Linn  Stephan Menzel  Susanne Hoffmann-Eifert  Nabeel Aslam  Ulrich Böttger  Karsten Fleck  Anja Herpers  Annemarie Köhl

Astrid Marchewka  Lutz Nielen  Chanwoo Park  Vikas Rana  Marcel Reiners  Christian Rodenbücher  Bernd Rösgen  Marcel Schie  Kristof Szot

Stefan Tappertzhoven  Ilia Valov  Jan v.d. Hurk  Florian Lentz  Christian Lenser
Acknowledgement for funding
Outline

1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Outline

1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Motivation: The computer challenge

- Computer wins the Chess World Championship
- Computers can generate new jokes which people find really funny
- Computers invent new proofs of mathematical theorems
- Computer wins finale of US Quizz Show Jeopardy!
Motivation: Energy efficiency

Watson:
- 2880 Processors
- ~100 000 kg
- 2 300 000 Watt

Human brain:
- 100 bill. Neurons
- ~ 1,5 kg
- 25 Watt
Alternative devices & architectures?

Devices

Non-volatile switches?
- energy efficient
- better than NAND
- Speed, endurance, scalability

Redox-based Resistive Switching Elements (ReRAM, memristive elements)

Architectures

New storage application
⇒ Storage class memory (SCM)
Beyond von Neumann architecture
⇒ fusion of nv-memory & logic
Neuromorphic computational concepts
⇒ artificial synapses and more
Redox based resistive switching memories (ReRAM)

VCM (Valence change mechanism)
- Bipolar switching
- Based on oxygen vacancy migration

ECM (Electrochemical metallization mechanism)
- Bipolar switching
- Based on Cu / Ag ion migration
Introduction - ReRAM

Low current switching:
- good control of ReRAM device
- high ON-resistance
  → low power operation feasible

The maximum SET current: 1µA

Requirements – binary memories

... to compete with Flash

Endurance: $> 10^7$ cycles  (Flash $10^3 ... 10^7$)

Resistance ratio: $R_{OFF} / R_{ON} > 10$

Scalability: $F < 22$ nm and/or 3-D stacking

Write voltage: approx. 1 ... 5 V  (Flash > 5 V)

Read voltage: 0.1 ... 0.5 V

Write speed: $< 100$ ns  (Flash > 10 $\mu$s)

Retention: $> 10$ yrs

Voltage – time dilemma

Kinetics of switching process requires non-linearity of $> 15$ orders of magnitude
## Future of NAND Flash

### NAND Flash properties:

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>2024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell area</td>
<td>$4F^2$</td>
<td>$4F^2$</td>
</tr>
<tr>
<td>Read time</td>
<td>100 $\mu$s</td>
<td>100 $\mu$s</td>
</tr>
<tr>
<td>Write time</td>
<td>1 ms</td>
<td>1 ms</td>
</tr>
<tr>
<td>Retention</td>
<td>10 years</td>
<td>10 years</td>
</tr>
<tr>
<td>Endurance (cycles)</td>
<td>$10^4$</td>
<td>$5 \times 10^3$</td>
</tr>
<tr>
<td>Write operation voltage</td>
<td>15 V</td>
<td>15 V</td>
</tr>
<tr>
<td>Read operation voltage</td>
<td>1.8 V</td>
<td>1 V</td>
</tr>
<tr>
<td>Feature size 2D/3D</td>
<td>22 nm/-</td>
<td>8 nm/24 nm</td>
</tr>
<tr>
<td>MLC  2D/3D</td>
<td>3/-</td>
<td>4/2</td>
</tr>
<tr>
<td>Layers 3D</td>
<td>1</td>
<td>98</td>
</tr>
</tbody>
</table>

Source: ITRS ERD 2011 / ORTC 2012

Cell properties do not improve much
1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Energy of ReRAM switching

Dissipation of energy in ReRAM write pulse mode

\[ E = \int_{0}^{t} I_{\text{Pulse}} \cdot V_{\text{Pulse}} \cdot d\tau \]

Hermes et al., Fast pulse analysis of TiO$_2$ based ReRAM nano-crossbar devices, NVMTS 2011
Non-linear switching kinetics ("Voltage time dilemma")

Kinetics of resistive switching show extreme non-linearity. Understanding of the origin is lacking.
Non-linearity of switching kinetics

Physico-chemical origins of nonlinearity (rate-determining step)

- **Electron transfer reaction** at the boundaries
  \[ J_{BV} = j_0 \exp\left( -\frac{\Delta W_{BV}}{k_B T} \right) \exp\left( \frac{(1 - \alpha) ze}{k_B T} \eta \right) - \exp\left( \frac{\alpha ze}{k_B T} \eta \right) \]

- **Ion transport (hopping)** within the electrolyte
  \[ J = 2 zecaf \exp\left( -\frac{\Delta W_e}{k_B T} \right) \sinh\left( \frac{aze}{2k_B T} E \right) \]

- **Nucleation probability**
  \[ J_{nuc} = j_0 \exp\left( -\frac{\Delta W_{nuc}}{k_B T} \right) \exp\left( \frac{(N_{crit} + \alpha) ze}{k_B T} \eta \right) \]

- **Phase formation**
  \[ t_{sw} \propto \frac{1}{j_{ion}} \]

→ Field and temperature enhancement possible
Switching kinetics: Field acceleration

- **Electron transfer reaction at the boundaries**
  \[ \alpha \in [0,1]; \, \text{min: } \alpha = 0.5; \, \text{max: } \alpha = 0.1 \]
  \[ m_{BV} = -\frac{\alpha(1-\alpha)ze}{k_BT} \]
  \[ \alpha = 0.5: \, 120 \, \text{mV/dec} \]
  \[ \alpha = 0.1: \, 330 \, \text{mV/dec} \]

- **Ionic transport (hopping) within the SL**
  \[ a \approx 0.5 \, \text{nm}; \, \text{min: } t_{\text{layer}} = 2 \, \text{nm}; \, \text{max: } t_{\text{layer}} = 25 \, \text{nm} \]
  \[ m_{\text{hop}} = -\frac{zea}{2k_BTt_{\text{layer}}} \]
  \[ d = 25 \, \text{nm}: \, 1.37 \, \text{V/dec} \]
  \[ d = 2 \, \text{nm}: \, 237 \, \text{mV/dec} \]

- **Nucleation**
  \[ N_C \geq 1; \, \text{min: } \alpha = 1, \, N_C \text{large}; \, \text{max: } \alpha = 1, \, N_C = 0 \]
  \[ m_{\text{nuc}} = -\frac{(N_C + \alpha)ze}{k_BT} \]
  \[ N_C = 0: \, 28 \, \text{mV/dec} \]
  \[ N_C = 3: \, 7.8 \, \text{mV/dec} \]
Switching kinetics of ECM-type Ag/AgI/Pt cells

Non-linearity experimental observed over 12 orders of magnitude

Simulation of all field acceleration mechanisms

S. Menzel, S. Tappertzhofen et. al., PCCP (2013)
Switching kinetics: Parameter variations

- Variation of $N_{\text{crit}}$
- Variation of nucleation constant $t_0$
- Variation of exchange current density $j_{0,\text{et}}$
- Variation of hopping prefactor $j_{0,\text{hop}}$

$\rightarrow$ Regimes parameter dependent

Switching kinetics: Regimes of RDS

Simulated SET switching kinetics compared to experimental data

Switching kinetics is limited by
- I: Nucleation
- II: Electron transfer reactions
- III: Electron transfer reactions and ion hopping transport

Switching kinetics: Temperature acceleration

Switching time depends exponentially on $1/T$

$$t = t_0(V) \exp\left(\frac{\Delta W}{k_B T}\right)$$

Local temperature increase caused by **Joule heating**.

$$T = T_0 + R_{th} P = T_0 + \frac{\sigma}{8k_{th}} V^2 = T_0 + KV^2$$

*U. Russo et al., T-ED Vol.56 No.2 (2009)*

Switching time

$$t \propto \exp\left(\frac{\Delta W}{k_B (T_0 + KV^2)}\right)$$

Typical values:

- $\sigma = 10^3$ S/m,
- $k_{th} = 1$ W/m K,
- $t_{layer} = 5$ nm

→ Steepness of temperature curve depends on activation energy
Modeling: Switching kinetics of VCM cells

3-D FEM simulation

Conductivity = f(T) - exper. data

Simulation of the thermal, electrical, and ionic transport processes

S. Menzel et al. (Adv. Funct. Mat. 2011)
Modeling: Switching kinetics of VCM cells

3-D FEM simulation

- Joule heating of the conducting filament
- Thermally activated oxygen vacancy drift
- Concentration change affects the electronic conductivity (based on generic lattice disorder model of metal oxides)

Experimental data & simulation

- Pulse width vs. SET voltage experiments
  - Perfect fit to simulation
  - Non-linearity of > 9 orders of magnitude

S. Menzel et al. (Adv. Funct. Mat. 2011)
Outline

1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Scaling towards atomic resolution

VCM cells

ECM cells


-> redox processes can be confined on the atomic scale
Scaling towards atomic resolution

Barrier lowering

Q: How many atoms must be moved?

-> Theory: Displacement of 2 atoms sufficient for ROFF/RON = 470 and barrier > 1.5 eV

Scaling impact on kinetics – ECM

- \( F = 100 \text{ nm}, 50 \text{ nm}, 30 \text{ nm}, 5 \text{ nm} \)
- \( L = F/2 \)
- \( r_{\text{fil}} = 4 \text{ nm}, 2.8 \text{ nm}, 2.2 \text{ nm}, 0.8 \text{ nm} \)

Switching time

\[ \rightarrow \text{The cell performance improves with decreasing the feature size } F \]
Scaling impact on kinetics – VCM

- Feature size $F$ is varied: 100 nm, 50 nm, 30 nm, 20 nm, 10 nm, 5 nm
- Disc thickness = 2·$r_{\text{fil}}$
- $r_{\text{fil}}$ = 4.5 nm, 3.2 nm, 2.5 nm, 2 nm, 1.4 nm, 1 nm

Switching time

Switching energy

→ The cell performance improves with decreasing the feature size $F$
Outline

1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
ReRAM cells in real arrays

Parasitics of empty arrays

ReRAM cells in real arrays – 1R arrays and selectors

Parasitics of empty arrays

Complementary Resistive Switch (CRS)

Four possible 4F² array setups:
A: 1R (non-linear)
B: 1CRS
C: 1S1R
D: 1T1R (vertical)

ReRAM arrays – switching energies of lines

Resulting read time:

\[ t_{\text{read}} = C_{BL} \frac{R_L (R_{\text{cell}} + R_{\text{FET}} + R_{BL})}{R_L + R_{\text{cell}} + R_{\text{FET}} + R_{BL}}. \]

Resulting read energy:

\[ E_{\text{read BL}} = \frac{1}{2} \frac{V_{\text{read}}^2}{R_{\text{on}} + r + R_L} t_{\text{read}} \]
\[ E_{\text{read WL}} = C_{WL} V^2 \]
\[ E_{\text{tot R}} = E_{\text{read BL}} + E_{\text{read WL}} \]

\[ R_{\text{cell}} = R_{\text{ON}} = 100 \text{ k}\Omega \]
\[ R_{BL} = 4.1 \text{ k}\Omega \]
\[ R_{\text{FET}} = 35 \text{ k}\Omega \]
\[ V_{\text{Read}} = 1 \text{ V} \]
\[ V_{\text{Sense}} > 0.1 \text{ V} \]
\[ R_L = 15.4 \text{ k}\Omega \]
\[ C_{WL} = 5 \text{ fF} \]
\[ r = R_{BL} + R_{\text{FET}} \]

ReRAM arrays – switching energies of lines

Inherent minimum write energy per cell:
Gap: 1 nm x 1 nm x 1 nm: 64 atoms
\[ E_{\text{filam}} = 64 \ E_A \approx 14 \ \text{aJ} \]

**But:** Array write energy is defined by ON resistance:
\[ E_{\text{write}} = \frac{V_{\text{write}}^2}{R_{\text{ON}}} \cdot t_{\text{write}} = 40 \ \text{fJ} \]
(\( V_{\text{write}} = 2 \ \text{V}, \ R_{\text{ON}} = 100 \ \text{k}\Omega, \ t_{\text{write}} = 1 \ \text{ns} \))

Read energy for comparison:
\[ E_{\text{totR}} (5 \ \text{nm}, \ N=128) \approx 5 \ \text{fJ} \]

→ Array/circuitry contribution dominates !.
ReRAM arrays – impact of selectors

Select device can have considerable impact on the array energy consumption:

- **Series resistance** (e.g. of the transistor $R_{\text{FET}}$) increase $t_{\text{Read}}$

- OFF/ON ratio defines additional **sneak current** induced power losses

- The 'turn-on' necessitates higher $V_{\text{read}}$

- **Slow devices** increase $t_{\text{Read}}$

- **Variability** may require large safety margins
1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Example – associative memories

Associative memories – content addressable memory (CAM)
• enables **parallel search** of a memory
• e.g. routing applications, pattern matching, neuromorphic applications

State-of-the-art:
10-T SRAM implementation

Ultra-dense 2S2R and 2CRS approaches are feasible too

Example – associative memories

Associative CRS-based Capacitive Network
• no ML precharge
• uses non-destructive readout
• Low impact of $R_{ON}$ and $R_{OFF}$ variation

Applications area:
- Pattern recognition
- Fast routing

Energy per synaptic function

1 J
Software Simulation

100 µJ
Simplified Software Simulation
Blue Gene/p

10 nJ
Conventional Hardware

100 pJ
ReRAM-based Hardware

10 fJ
Synapse

System Power dissipation

Simulation of the brain: 1 GW..1 TW

Watson: 2,5 MW (selected functionality)

Brainscale
- SRAM Cells (4Dbits) with DAC
- capacitive storage (synapses)
- Floating Gate Cells (Neurons)
- 200k Neurons, 50M Synapses

ReRAM

Artificial brain?

Humans brain: 25 W
- \(10^{11}\) Neurons
- \(10^{15}\) Synapses

K. Meier, *IBM MRC Workshop on Materials, 2012*
Outline

1. Motivation and Introduction

2. Ultra-nonlinear Kinetics of the Switching Process
   * Impact of kinetics on energy efficiency
   * Results for ECM systems
   * Results for VCM systems

3. Scaling of ReRAM Concepts
   * Ultimate physical limits of scaling
   * Impact of scaling on switching energy

4. Array Considerations
   * Selectors
   * Energy of charging lines

5. Towards Neuromorphic Computing

6. Conclusions
Challenges

- **Strong competition for ReRAM expected**
  ... from 3D-NAND

- **Trade-off between energy efficiency and retention**
  ... Physics must be understood in order to resolve this issue

- **Variability**
  ... and its impact on energy efficiency

Prospects

- **Technologically compatible to CMOS interface**

- **Ultimately high scaling potential**
  ... with further improved energy efficiency

- **Functions beyond pure memory**
  ... from FPGA type logic to neural functions for cognitive computing
Further reading …..

Third, completely revised edition

April 2012

Student-friendly price (Euro 85,-)
Thank You!