Nano-Scale Memory Devices:

*Space-Time-Energy Trade-offs*

Ralph Cavin and Victor Zhirnov
Semiconductor Research Corporation
Main Points

- Many candidates for beyond-CMOS nano-electronics have been proposed for memory, but no clear successor has been identified.
  - Methodology for system-level analysis

- How is maximum performance related to device physics?
Three integrated components of a Memory Device:

1) ‘Storage node’
   - physics of memory operation

2) ‘Sensor’ which reads the state
   - e.g. transistor

3) ‘Selector’ which allows a memory cell in an array to be addressed
   - transistor
   - diode

All three components impact scaling limits for all memory devices
Space-Time-Energy Metrics

- Essential parameters of the memory element are:
  - cell size/density,
  - retention time, access time/speed
  - operating voltage/energy.

- None of known memory technologies, perform well across all of these parameters

- At the most basic level, for an arbitrary memory element, there is interdependence between operational voltage, the speed of operation and the retention time.

- More generally, cell dimensions are also part of the trade-off, hence the Space-Time-Energy compromise
Space-Action Principle for Memory

\[ Energy \times time \times Volume = \min \]

\[ E \cdot t \cdot V = \min \]

\[ E \cdot t \cdot L = \min \]

\[ E \cdot t \cdot N_{at} = \min \]

The Least Action principle is a fundamental principle in Physics

\[ E \cdot t = \min \ (\geq h) \]

Plank’s constant
\[ h=6.62 \times 10^{-34} \text{ Js} \]
Three Major Memory State Variables

- **Electron Charge** (‘moving electrons’)
  - e.g. DRAM, Flash

- **Electron Spin** (‘moving spins’)
  - (STT-) MRAM

- **Massive particle(s)** (‘moving atoms’)
  - e.g. ReRAM, PCM, Nanomechanical, etc.
Charge-based Memories

**Requirements:**
1) Efficient charge injection during programming
2) Suppressed back-flow of charge in store/read modes
3) Efficient erase
4) Min. charge/bit: $q = e = 1.6 \times 10^{-19} \, \text{Q}$
Barrier-less Ohmic Transport: The most efficient injection, but...

Write \[ I_{AB} \sim \Delta V \]

… difficult retention

Charge-based memory is a two-barrier system

Example: DRAM
The operation of charge-based memory devices is governed by these basic equations, which put fundamental constraints on device and circuit parameters.

**Boltzmann probability of thermal excitation**

\[
\Pi = \exp\left(-\frac{E_b}{k_BT}\right)
\]

**Heisenberg Relations**

\[
\Delta x \Delta p \geq \hbar
\]

\[
\Delta E \Delta t \geq \hbar
\]

**Current**

\[
I = I_0 \exp\left(-\frac{E_b}{k_BT}\right)
\]

\[
I = I_0 \exp\left(-\frac{2\sqrt{2m}}{\hbar}(a\sqrt{E_b})\right)
\]
What is the minimum barrier height for the charge-based memory?

Thermionic leakage current (ideal case):

\[ J_{th} = J_{th0} \cdot \exp \left( -\frac{E_b}{kT} \right) \]

\[ q = CV \]

\[ 25 \text{fF} \times 0.5 \text{V} = 3.75 \text{fC} \]

<table>
<thead>
<tr>
<th>( E_b, \text{eV} )</th>
<th>Max. retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>4 ms</td>
</tr>
<tr>
<td>0.76</td>
<td>24 ms</td>
</tr>
<tr>
<td>0.8</td>
<td>77 ms</td>
</tr>
<tr>
<td>1.4</td>
<td>1 month</td>
</tr>
<tr>
<td>1.57</td>
<td>12 years</td>
</tr>
</tbody>
</table>

Standard DRAM requirement: 64 ms

High-barrier are needed for Non-volatile memory

Problem: In Si devices \( E_{b,\text{max}} < E_g = 1.1 \text{ eV} \)
Volatile electron-based memory: DRAM

\[ N_{el} = \frac{CV}{e} \]
\[ E = \frac{CV^2}{2} \]
\[ t = \frac{CV}{I} \]

\[ N_{el} \approx 10^5 \]
\[ a = 10 \text{ nm} \]
\[ V_{cap} = 6 \times 10^{-15} \text{ cm}^3 \]
\[ E \approx 10^{-14} \text{ J} \]
\[ t_w \approx 1 \text{ ns} \]
DRAM summary

$\alpha = 10 \text{ nm}$  \hspace{1cm} \begin{array}{c} N_{el} \approx 10^5 \end{array} \hspace{1cm} E \approx 10^{-14} \text{ J} \hspace{1cm} V_{cap} = 6 \times 10^{-15} \text{ cm}^3 \hspace{1cm} t \approx 1 \text{ ns} \hspace{1cm} E \times t \times V \approx 10^{-9} \text{ J-ns-nm}^3

DRAM inherent issues:

- Low barrier height - **Volatility**
- Remote sensing – **Large size of Storage node**
Flash: Local Sensing of Memory State
Charge injection problem in high-barrier systems

High-barriers are needed for Non-volatile memory

BUT: Barrier formed by an insulating material (large $E_b$) cannot be suppressed) – charge transport in the presence of barriers: *Non-ohmic charge transport*
Floating gate memory: WRITE and STORE modes

WRITE

STORE

\[ t_w = \frac{q}{I} = \frac{C\Delta V}{I} \]

\[ t_{ret} = \frac{q}{I} = \frac{C\Delta V}{I} \]
We need to create an asymmetry in charge transport through the gate dielectric to maximize the $I_{\text{write}}/I_{\text{ret}}$ ratio.

**Floating gate cell:**

**Write** – triangle barrier

**Retention** – trapezoidal barrier

The asymmetry in charge transport between WRITE and STORE modes is achieved through different shape of barrier (triangle vs. trapezoidal).
Retention Analysis:  
**Minimum Barrier Height and Width:**  

\[
I_{th} = I_{th0} \cdot \exp\left(-\frac{E_b}{kT}\right)
\]

\[
t_s = \frac{Ne}{I_s}
\]

\[
E_{bmin} = 1.8 \text{ eV (400K)}
\]

\[
I_{tun} = A \frac{e^2}{h^2} \sqrt{2mE_b} \frac{V}{a} \exp\left(\frac{2\sqrt{2m}}{h} \cdot a \cdot \sqrt{E_b - \frac{eV}{2}}\right)
\]

\[
a_{min} = 6.9 \text{ nm}
\]

\sim 10 \text{ y}
Floating Gate Cell Retention and WRITE characteristics

**Retention:** direct tunneling

$V_{\text{stored}} < E_b$

The retention time strongly depends on thickness

**Write:** F-N tunneling:

$\Delta V_{F-N} > E_b$

For lower WRITE voltage $E_b$ should be decreased:

$E_{b\text{min}} = 1.8 \text{ eV}$

$V_{\text{min}} > 4 \text{ V}$

---

**Table:**

<table>
<thead>
<tr>
<th>Barrier</th>
<th>$E_b$</th>
<th>$V_{\text{Ret}}$</th>
<th>$a$</th>
<th>$t_{\text{Ret}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si/SiO$_2$</td>
<td>3.1 eV</td>
<td>2 V</td>
<td>4 nm</td>
<td>4.35 min</td>
</tr>
<tr>
<td>Si/SiO$_2$</td>
<td>3.1 eV</td>
<td>2 V</td>
<td>5.4 nm</td>
<td>20 y</td>
</tr>
<tr>
<td>Min. barrier</td>
<td>1.8 eV</td>
<td>0.9 V</td>
<td>6.9 nm</td>
<td>11 y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Barrier</th>
<th>$E_b$</th>
<th>$V_{\text{WR}}$</th>
<th>$a$</th>
<th>$t_{\text{WR}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si/SiO$_2$</td>
<td>3.1 eV</td>
<td>6.8 V</td>
<td>5.4 nm</td>
<td>1h</td>
</tr>
<tr>
<td>Si/SiO$_2$</td>
<td>3.1 eV</td>
<td>12 V</td>
<td>5.4 nm</td>
<td>30 ms</td>
</tr>
<tr>
<td>Min. barrier</td>
<td>1.5 eV</td>
<td>6 V</td>
<td>6.9 nm</td>
<td>40 ms</td>
</tr>
</tbody>
</table>
Voltage-Time Dilemma

- For an arbitrary electron-charge based memory element, there is interdependence between operational voltage, the speed of operation and the retention time.

- Specifically, the nonvolatile electron-based memory, suffers from the “barrier” issue:
  - High barriers needed for long retention do not allow fast charge injection
  - It is difficult (impossible?) to match their speed and voltages to logic
Flash Scaling limits due to ‘Sensor’

Selector

Sensor

Storage Node
Semiconductor Physics sets limits on barrier quality

Depletion width

\[ W \approx \sqrt{2 \varepsilon \varepsilon_0 V_{bi} \frac{N_A + N_D}{N_A N_D}} \]

\[ W_{\text{min}} = \sqrt{2 \varepsilon \varepsilon_0 E_g \frac{N_C + N_V}{N_C N_V}} \approx 10 \text{ nm} \]

\( N_C \) – effective density of states in the conduction band, for Si \( N_C=2.8\times10^{19} \text{ cm}^{-3} \)

\( N_V \) - effective density of states in the valence band, for Si \( N_V=1.4\times10^{19} \text{ cm}^{-3} \)

\( E_g \) – the band gap, for Si \( E_g=1.12 \text{ eV} \)
Flash in the limits of scaling

Electrostatics requires gate oxide scaling to maintain FET charge sensitivity.

- Storage Node
- V = 2 \times 10^{-18} \text{ cm}^3
- Sensor
- a \sim W_{\text{min}}
- Selectors
- \ell \sim 10 \text{ nm}
- ~6\text{nm}
- N_{\text{el}} \sim 10
- ~20\text{nm}
- Flash FET T_{\text{ox}} = \text{const} = 5.4 \text{ nm} – degradation of sensitivity and control
Flash Summary

\( a = 10 \text{ nm} \)

\( N_{el} \approx 10 \)

\( E \approx 10^{-15} \text{ J} \)

\( t \approx 1 \mu s = 1000 \text{ ns} \)

\( V = 2000 \text{ nm}^3 \)

\( E \times t \times V \approx 10^{-9} \text{ J-ns-nm}^3 \)
Conclusion on ultimate charge-based memories

- All charge-based memories suffer from the “barrier” issue:
  - High barriers needed for long retention do not allow fast charge injection
  - It is difficult (impossible?) to match their speed and voltages to logic

  - **Voltage-Time Dilemma**

Non-charge-based NVMs?
Emerging Memory Devices

The Choice of Information Carrier
**Desired: ‘Benchmark’ memory cell**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell size, $l$</td>
<td>$&lt;10$ nm</td>
</tr>
<tr>
<td>Store time, $t_s$</td>
<td>$&gt;10^8$ s</td>
</tr>
<tr>
<td>Write time, $t_w$</td>
<td>$&lt;10^{-7}$ s</td>
</tr>
<tr>
<td>Read time, $t_r$</td>
<td>$&lt;10^{-7}$ s</td>
</tr>
<tr>
<td>Read Voltage, $V_r$</td>
<td>$\sim 1$ V</td>
</tr>
<tr>
<td>Read current, $I_r$</td>
<td>$\sim 10^{-6}$ A</td>
</tr>
<tr>
<td>Read current density, $J_r$</td>
<td>$&gt;10^6$ A/cm$^2$</td>
</tr>
</tbody>
</table>

**Driver: Cell Scaling**

**Driver: Sensing**
Spin torque transfer MRAM
Magnetic storage node (Moving spins): Energy Limit

\[ f_{tr} = f_0 \exp \left( -\frac{E_b}{k_B T} \right) = f_0 \exp \left[ -\frac{KV}{k_B T} \right] \]

\((f_0 \sim 10^9-10^{10} \text{ c}^{-1})\)

\[ t_{store} = \frac{1}{f_0} \exp \left[ -\frac{KV}{k_B T} \right] \]

\(t_{store} > 10 \text{ y}\)

\(E_b = KV > 36k_B T \sim 1.25 \text{ eV}\)

Magnetic storage node (Moving spins):

Size Limit

\[ E_b = KV > 36k_B T \approx 1.25 \text{ eV} \]

\[ K \approx 0.1-1 \text{ J/cm}^3 \]

\[ V = L^2 T \approx 2 \times 10^{-19} \text{ cm}^3 \]

*Thin film:* \( T = 2 \text{ nm} \quad L \approx 11 \text{ nm} \)

\[ N_{at} \approx 10^4 \]

\[ N_{spin} \approx 10^5 \]

the anisotropy constant of a material
FET selector is biggest part of STT-MRAM in the limits of scaling

$6 \times 2 \text{ nm} = 12 \text{ nm}$

$V = 1500 \text{ nm}^3$

STT-RAM summary

$V = 1500 \text{ nm}^3$ \hspace{1cm} $t_w \sim 1 \text{ ns}$

$E \approx 10^7 \text{ A/cm}^2 \times 11 \text{ nm}^2 \times 1 \text{ V} \times 1 \text{ ns} \sim 10^{-14} \text{ J}$

$E \times t \times V \sim 10^{-11} \text{ J-ns-nm}^3$
Scaled ReRAM
Moving atoms: ‘Atomic Relay’

Atomic-scale switch, which opens or closes an electrical circuit by the controlled reconfiguration of silver atoms within an atomic-scale junction.

Such ‘atomic relays’ operate at room temperature and the only movable part of the switch are the contacting atoms, which open and close a nm-scale gap.

- Small (~1 nm)
- Fast (~1 ns) - projection
- Low voltage (<1V)

*Nature* 433, 47-50 (6 January 2005)

**Quantized conductance atomic switch**

K. Terabe, T. Hasegawa, T. Nakayama and M. Aono
Two-sided barrier

Both electrodes influence the potential of the electron within the electrode separation. For small gaps, the near electrode electric fields will influence the energy barrier.

Interface-to-interface interaction
Two-side barrier

$\phi_0 = 5.1 \text{ eV}$

$x, \text{ nm}$

$\phi(x), \text{ eV}$

$\alpha = 100 \text{ nm}$

Au

$\phi_0 = 5.1 \text{ eV}$
Two-side barrier

\[ a = 50 \text{ nm} \]

\[ \phi_0 = 5.1 \text{ eV} \]

Au

\[ \phi_0 = 5.1 \text{ eV} \]
Two-side barrier

\( a = 20 \text{ nm} \)

\begin{align*}
\text{Au} & \quad \phi_0 = 5.1 \text{ eV} \\
\text{Au} & \quad \phi_0 = 5.1 \text{ eV}
\end{align*}

\( x, \text{ nm} \)
Two-side barrier

\[ a = 10 \text{ nm} \]

\[ \phi_0 = 5.1 \text{ eV} \]
Two-side barrier

\[ a = 5 \text{ nm} \]

$\text{Au}$

$\phi_0 = 5.1 \text{ eV}$

$\text{Au}$

$\phi_0 = 5.1 \text{ eV}$
Two-side barrier

\( a = 2 \text{ nm} \)

\( \text{Au} \)
\( \phi_0 = 5.1 \text{ eV} \)

\( \text{Au} \)
\( \phi_0 = 5.1 \text{ eV} \)
Two-side barrier

\[ a = 2 \text{ nm} \]

\[ V = 0 \]
Two-side barrier

\( a = 1 \text{ nm} \)

\( V = 0 \)
Two-side barrier

\[ a = 0.5 \text{ nm} \]

\[ V = 0 \]
Two-side barrier

\[ a = 0.5 \, \text{nm} \]

\[ V = 1 \, \text{volt} \]
Two-side barrier

\[ a = 2 \text{ nm} \]

\[ V = 2 \text{ volt} \]
Ultimate ReRAM: 1-atom gap

\[ a \sim n_{Au}^{\frac{1}{3}} = 0.257 \text{ nm} \]

\[ V=0 \]

\[ E_b = 0.64 \text{ eV} \]

\[ d_t = 0.09 \text{ nm} \]

\[ d_t << a \]
Ultimate ReRAM: 1-atom gap

\[ a \sim n_{\text{Au}}^{\frac{1}{3}} = 0.257 \text{nm} \]

\[ \text{ON/OFF} \approx 1.61 \]

\[ V = 0.5 \text{ V} \]

\[ E_b = 0.38 \text{ eV} \]

\[ d_t = 0.075 \text{ nm} \]

\[ d_t << a \]
Ultimate ReRAM: 2-atom gap

\[ a \sim 2n \frac{1}{3} \frac{1}{Au} = 0.514 \text{nm} \]

ON/OFF \sim 476

V = 0.5 V

\[ E_b = 2.63 \text{ eV} \]
\[ d_t = 0.37 \text{ nm} \]
\[ d_t < a \]
Ultimate Atomic Relay: 4-atom gap

\[ a \sim 3n_{\text{Au}}^{\frac{1}{3}} \sim 1\text{nm} \]

Energy barrier for diffusion \( E_b \)

\[ t_{tr} = t_0 \exp \left( \frac{E_b}{k_B T} \right) = l_0 \sqrt{\frac{m}{k_B T}} \exp \left[ -\frac{E_b}{k_B T} \right] \sim 2s \]
Ultimate Atomic Relay: 4-atom gap

\[ a \sim 4n_{Au}^{\frac{1}{3}} \sim 1nm \]

\[ T_k = 3.2 \times 10^5 \exp \left( -\frac{E_b}{k_B T} \right) \]

\[ t_{tr} = t_0 \exp \left( \frac{E_b}{k_B T} \right)^3 = l_0 \sqrt{\frac{m}{k_B T}} \exp \left[ -\frac{E_b}{k_B T} \right] > 10^y \]

\( E_b (\text{energy barrier for diffusion}) \)

\( 0.5-1 \text{ eV} \)

\( (E_b=0.5 \text{ eV, } n>5) \)
Ultimate ReRAM: A summary

\begin{align*}
\alpha &= 1 \text{nm} \\
V &= 1 \text{nm}^3 \\
N_{\text{at}} &\sim 100 \quad (64) \\
E &\sim N_{\text{at}} \times 1 \text{eV} \sim 10^{-17} \text{J} \\
t &\sim 1 \text{ ns} \\
E \times t \times V &\sim 10^{-17} \text{ J-ns-nm}^3
\end{align*}
## Summary

<table>
<thead>
<tr>
<th></th>
<th>(N_{\text{carriers}})</th>
<th>(V, \text{nm}^3)</th>
<th>(E_w, J)</th>
<th>(t_w, \text{ns})</th>
<th>(\text{Space-Action, J-ns-nm}^3)</th>
<th>(\text{Biggest component})</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>10^5</td>
<td>10^5</td>
<td>10^{-14}</td>
<td>1 ns</td>
<td>(~10^{-9})</td>
<td>Storage Node</td>
</tr>
<tr>
<td>Flash</td>
<td>10</td>
<td>10^3</td>
<td>10^{-15}</td>
<td>10^3 ns</td>
<td>(~10^{-9})</td>
<td>Sensor</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>10^5</td>
<td>10^3</td>
<td>10^{-14}</td>
<td>1 ns</td>
<td>(~10^{-11})</td>
<td>Selector</td>
</tr>
<tr>
<td>ReRAM</td>
<td>100</td>
<td>1</td>
<td>10^{-17}</td>
<td>1 ns</td>
<td>(~10^{-17})</td>
<td>Selector</td>
</tr>
</tbody>
</table>

Constraints by sensor not considered

Main constraints due to sensor
Back-Up slides
FET or Diode selector is biggest part of ReRAM in the limits of scaling
Scaling Limits of Diodes

\[ W \approx \sqrt{\frac{2\varepsilon\varepsilon_0 V_{bi}}{N_D}} \sim 10\, \text{nm} \]

\[ N_d \leq N_C \]

\[ N_d \geq N_C \]

pn-diode → Esaki tunnel diode

Schottky diode → Ohmic contact

\[ N_C \] – effective density of states in the conduction band, for Si \( N_C = 2.8 \times 10^{19} \, \text{cm}^{-3} \)
Space-Action: Flash