Flash Memory with Nanoparticle Floating Gates

Sanjay Banerjee
Director, Microelectronics Research Center
Cockrell Chair Professor of Electrical & Computer Engineering
University of Texas at Austin

• Why Nanoparticles in Flash Memory
• Self-assembly schemes
• Device Performance

Acknowledgements: DARPA MARCO, NSF NIRT, Micron
**The Scale of Things – Nanometers and More**

**Things Natural**
- **Ant** ~ 5 mm
- **Fly ash** ~ 10-20 \( \mu m \)
- **Human hair** ~ 60-120 \( \mu m \) wide
- **Red blood cells with white cell** ~ 2-5 \( \mu m \)
- **DNA** ~2-1/2 nm diameter
- **Atoms of silicon spacing** ~tens of nm

**Things Manmade**
- **Head of a pin** 1-2 mm
- **MicroElectroMechanical (MEMS) devices** 10-100 \( \mu m \) wide
- **Pollen grain**
- **Red blood cells**
- **Zone plate x-ray “lens”** Outer ring spacing ~35 nm
- **Self-assembled, Nature-inspired structure** Many 10s of nm
- **Nanotube electrode**
- **Carbon buckyball** ~1 nm diameter

**The Challenge**
Fabricate and combine nanoscale building blocks to make useful devices, e.g., a photosynthetic reaction center with integral semiconductor storage.
Moore’s Law

No exponential is forever!

But can we delay “forever”?
Flash memory applications

Markets
- Consumer Electronics
- Networking
- Wireless
- Industrial Control
- Automotive

Control Gate
Floating gate
Tunnel dielectric
NAND Flash Market Segmentation

- Removable Solid State Flash Cards: 53%
- Digital Settop Box: 0%
- Others: 5%
- MP3 players (Flash based): 22%
- USB Flash Drive: 17%
- Mobile Handset (Embedded): 2%
- Digital Still Camera (Embedded): 1%

2005 NAND Demand TAM Percentage
iSupply: Q3. 2005

NOR Flash had TAM of 7B$, mostly in wireless cell phones.
Very high growth rates!!!
Flash Technology Scaling History

- Flash Invented in mid 1980’s
  - NOR flash evolved from EPROM
  - NAND started as poly-poly erase cell later evolving to present structure
- ~20 years & 10 Generations of ETOX® (Intel NOR) High Volume Production
- 8+ years & 5 Generations of MLC: 2bit / cell

Volume Production Year / Technology Generation

- 1986 / 1.5μm
- 1988 / 1.0μm
- 1991 / 0.8μm
- 1993 / 0.6μm
- 1996 / 0.4μm
- 1998 / 0.25μm
- 2000 / 0.18μm
- 2002 / 0.13μm
- 2006 / 65nm

Source: Intel
### NVM- Long Term Requirements (ITRS 2006)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM ½ Pitch (nm) (contacted)</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>MPU Physical Gate Length (nm)</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Flash technology NOR/NAND – F (nm) [1]</td>
<td>28/25</td>
<td>25/23</td>
<td>22/20</td>
<td>20/18</td>
<td>18/16</td>
<td>16/14</td>
<td>14/13</td>
</tr>
<tr>
<td>Flash NOR cell size – area factor a in multiples of $F^2$ [2, [3], [4], [5]]</td>
<td>10–12</td>
<td>10–13</td>
<td>10–13</td>
<td>11–14</td>
<td>11–14</td>
<td>12–14</td>
<td>12–14</td>
</tr>
<tr>
<td>Flash NAND cell size – area factor a in multiples of $F^2$ SLC/MLC [6]</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
<td>4.0/1.0</td>
</tr>
<tr>
<td>Flash NOR typical cell size ($\mu m^2$) [7], [8]</td>
<td>0.0086</td>
<td>0.0073</td>
<td>0.0057</td>
<td>0.005</td>
<td>0.004</td>
<td>0.0034</td>
<td>0.0026</td>
</tr>
<tr>
<td>Flash NOR $L_g$-stack (physical – $\mu m$) [8], [9]</td>
<td>0.09</td>
<td>0.09</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Flash NOR highest W/E voltage (V) [10], [11]</td>
<td>6–8</td>
<td>6–8</td>
<td>6–8</td>
<td>6–8</td>
<td>6–8</td>
<td>6–8</td>
<td>6–8</td>
</tr>
<tr>
<td>Flash coupling ratio [14]</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
<td>0.6–0.7</td>
</tr>
<tr>
<td>Flash NOR tunnel oxide thickness EOT (nm) [15]</td>
<td>7–8</td>
<td>7–8</td>
<td>7–8</td>
<td>7–8</td>
<td>7–8</td>
<td>7–8</td>
<td>7–8</td>
</tr>
<tr>
<td>Flash NAND tunnel oxide thickness EOT (nm) [16]</td>
<td>6–7</td>
<td>6–7</td>
<td>6–7</td>
<td>6–7</td>
<td>6–7</td>
<td>6–7</td>
<td>6–7</td>
</tr>
<tr>
<td>Flash NOR interpoly dielectric thickness EOT (nm) [17]</td>
<td>8–10</td>
<td>8–10</td>
<td>8–10</td>
<td>8–10</td>
<td>7–9</td>
<td>6–8</td>
<td>6–8</td>
</tr>
<tr>
<td>Flash NAND interpoly dielectric thickness (nm) [18]</td>
<td>9–10</td>
<td>9–10</td>
<td>9–10</td>
<td>9–10</td>
<td>9–10</td>
<td>9–10</td>
<td>9–10</td>
</tr>
<tr>
<td>Flash endurance (erase/write cycles) [19]</td>
<td>1.00E+06</td>
<td>1.00E+06</td>
<td>1.00E+07</td>
<td>1.00E+07</td>
<td>1.00E+07</td>
<td>1.00E+07</td>
<td>1.00E+07</td>
</tr>
<tr>
<td>Flash nonvolatile data retention (years) [20]</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Flash maximum number of bits per cell (MLC) [21]</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Nanoparticle Gate Flash Memory

Conventional Flash Memory
* A defect totally discharges the floating gate
  - Thick tunnel oxide
  - High voltage/ power
  - Low reliability/ speed

Nano-floating Gate Memory
* A defect discharges only one dot
  - High-k tunnel oxide
  - Speed/ power/ density better
  - Reliability improved
  - New phenomena- self-assembly, Coulomb blockade, multi-level cells
SiGe Nanocrystals on High-K Dielectrics

AFM scans (1 micron x 1micron) showing SiGe dots grown at ~ 500ºC for 90 s with 0.75 gas ratio of GeH₄ to Si₂H₆.

Band diagram of HfO$_2$ and SiO$_2$ dielectric at low program voltage

- **SiE$_c$**
- **SiGe E$_c$**
- **SiE$_c$**
- **SiGe E$_c$**

- Direct tunneling (due to small $\phi_b$)

- $J_g$ [A/cm$^2$] vs. $V_G$ [V]
- F-N tunneling
- E.O.T 4nm

- Channel
- Gate
- Nanocrystal Dots
- Interface oxide
- Tunneling oxide (HfO$_2$)
- Tunneling oxide (SiO$_2$)
- Dot

- Channel
- Gate
Program & Erase Transient Characteristics

Kim DW, Kim T, Banerjee SK, ELECTRON DEV 50 (9): 1823-1829 SEP 2003
The Bio-nanometer Perspective

Nanomaterials:
- Nanotubes
- Dendrimers
- Nanopores
- Quantum dots
- Nanoshells
- Nanoparticles

DNA
Protein
Virus
Bacterium
Hair
Fish egg
Human eye

Atom
DNA
Water
Human eye
Hair
10^(-1)
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8

nm
μm
mm

Gate
1.2nm SiO_2
Silicon substrate
Gate/
dielectrics/substrate

Transistor

Interconnect

Integrated circuit

Microprocessor

Slide by C. Li
Fig. 1. Schematic drawings of apoferritin molecule. (a) looking down along four fold axis (b) cross-section of apoferritin (c) looking down along three fold axis.

Yamashita, IEDM 2006, p. 447 (Ferritin)

Ishii, Nature Vol. 423, 2003 (Chaperonin)

Fig. 3. (a) Interaction of protein-protein and substrate-protein bestowed by the target-specific affinity peptides. (b) Two-dimensional hexagonally close packed array directly formed on the Si substrate.
Protein assembly of Metal (Co, Au,..) and Semiconductor (PbSe, CdSe, Ge ..) NCs

Schematic structure of Chaperonin 60 (GroEL)
Various semiconductor and metal nanocrystals self-assembled, including Co

Density ~ $10^{12}$ cm$^{-2}$
$\Delta V_{FB}$ comparison of protein-mediated PbSe NC MOS capacitor and control samples

<table>
<thead>
<tr>
<th></th>
<th>PbSe NCs (Annealed)</th>
<th>Protein Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>C</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>D</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Program/Erase time = 1s

![Diagram showing the structure of a semiconductor device with PbSe NCs and various layers including TaN gate, control oxide, tunnel oxide, Si substrate, and protein template.]

![Graph showing $\Delta V_{FB}$ as a function of program voltage for different samples A, B, C, and D.]

Program Voltage (V)

$\Delta V_{FB}$ (V)
Endurance and Retention Characteristics

Devices survived $10^5$ cycles of program/erase operation; no sign of window closure.

Devices have good retention.
Heat Shock Proteins (Trent, NASA)
Coulomb Blockade in SiGe dot on SiO$_2$ and HfO$_2$

SiGe on SiO$_2$ @ 520 ºC

- Limits programming by direct tunneling at low voltages
- Multiple electron storage reduces life-time in nanocrystal

*Putting an electron on nanocrystal raises other energy levels*

\[ \frac{dE}{C} = \frac{e^2}{C} \]

For small capacitor (C \sim 10^{-18} \text{ F}),
\[ dE \text{ can be} > \text{thermal energy, } kT \]

\[ \text{electron} \quad \text{K} \quad \text{K} \]

Large C, high T  Small C, low T

*Low T: Given a Write voltage, fixed no. of electrons in nanocrystal*
*For nanocrystals < 5nm, this effect is significant at room temperature*
Vertical Flash Memory

Schematic side view

Scanning Electron Micrograph
Vertical Flash Retention at 300K

± 9 V, 100 ms tunneling P/E
Nanoparticle Self-Assembly Using PS-\(b\)-PMMA Copolymer

**Process Flow**

Employing a sandwich of organic/inorganic/organic layers (Polyimide/SiO\(_2\)/PS-\(b\)-PMMA) to engineer the aspect ratio of the patterns.

---

**Material Characterization**

SEM micrographs of (a) Copolymer template, (b) transferred polymer patterns into the underlying SiO\(_2\) layer, (c) ultimate array of Ni nanoparticles. (d) Cross-sectional image of the embedded nanoparticles within SiO\(_2\).

Histogram of the copolymer pore size distribution shown in Fig. (a)
Characteristics of MOSCAP Memory Devices

- High frequency C-V Characteristics @ 1MHz
- Memory window for different program voltages
- Transient characteristics of the memory device

Graphs showing the variation of Gate Voltage (V) and Write/Erase time (s) for different memory devices with and without Ni dots.

Variability, DOS?

SONOS trap densities $\sim 10^{19} \text{ cm}^{-3} \text{ eV}^{-1}$
For $\sim 3$ nm layer, $\sim 3 \times 10^{12}$ cm$^{-2}$ traps
spatially and energetically distributed
# Non-MOS Memory Contenders (adapted from Fazio, Intel)

<table>
<thead>
<tr>
<th></th>
<th>MRAM</th>
<th>FeRAM</th>
<th>Phase Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell Size $\lambda^2$</strong></td>
<td>Large ~40 → 20</td>
<td>Large ~25</td>
<td>Small ~6.5</td>
</tr>
<tr>
<td><strong>CMOS Integration</strong></td>
<td>&lt;200C Post Magnetic Tight MJT control</td>
<td>Fe reduces in hydrogen Etching difficult for Fe</td>
<td>Compatible with backend CMOS metal processing</td>
</tr>
<tr>
<td><strong>Read</strong></td>
<td>Non-Destructive, Fast, Low Power</td>
<td>Destructive: Endurance limited read</td>
<td>Non-Destructive, Moderate speed, Scales poorly</td>
</tr>
<tr>
<td><strong>Write</strong></td>
<td>Power constrained, Scales poorly; half select issue</td>
<td>Low power capacitive Theoretically good speed</td>
<td>Power constrained, large drivers, Improves with scaling</td>
</tr>
<tr>
<td><strong>Cycling Endurance</strong></td>
<td>Theoretically Infinite</td>
<td>1e-8 → 1e12; claims made, but limited data</td>
<td>~1e12; claims made, but limited data; Erratic Failures</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>Write current increases with scaling, materials engineering required at each scaling node; superparamagnetic limit?</td>
<td>3D cell required sub 90nm. Material engineering required at each scaling node</td>
<td>No material changes, No physical limits known down to ~5nm</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td>Embedded Working Memory Low Density</td>
<td>Embedded Low Power Low Density</td>
<td>Stand Alone or Embedded High Density Low Cost</td>
</tr>
</tbody>
</table>

**Color Code**
- Poor
- OK
- Good

**Diagram:**
- Data Storage Region
- Chalcogenide
- Poly Crystalline
- Amorphous
- Heater
- Resistive Electrode
- Phase Change Material
- Applied Electric Field
- NiFe
- Cu
- Memory storage
- Read
<table>
<thead>
<tr>
<th>Memory Type</th>
<th>SRAM</th>
<th>DRAM</th>
<th>MRAM Industry</th>
<th>NOR</th>
<th>NAND</th>
<th>FeRAM</th>
<th>PCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Elements</td>
<td>6T</td>
<td>1T1C</td>
<td>IR</td>
<td>1T</td>
<td>1T</td>
<td>1T1C</td>
<td>1T1R</td>
</tr>
<tr>
<td>Feature Size F, nm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>90/65</td>
<td>65</td>
<td>180</td>
<td>65/57</td>
<td>65/57</td>
<td>130</td>
<td>90</td>
</tr>
<tr>
<td>2018</td>
<td>18</td>
<td>18</td>
<td>22</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>18</td>
</tr>
<tr>
<td>Cell Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>140 F²</td>
<td>7.5 F²</td>
<td>25 F²</td>
<td>9-11 F²</td>
<td>4 F²</td>
<td>34 F²</td>
<td>7.2 F²</td>
</tr>
<tr>
<td>2018</td>
<td>140 F²</td>
<td>5 F²</td>
<td>16 F²</td>
<td>4/1 F²</td>
<td>16 F²</td>
<td>4.7 F²</td>
<td></td>
</tr>
<tr>
<td>Read Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>0.4 ns</td>
<td>&lt;15 ns</td>
<td>&lt;25 ns</td>
<td>14 ns</td>
<td>70 ns</td>
<td>80 ns</td>
<td>60 ns</td>
</tr>
<tr>
<td>2018</td>
<td>70 ps</td>
<td>&lt;15 ns</td>
<td>&lt;0.5 ns</td>
<td>2.5 ns</td>
<td>12 ns</td>
<td>&lt;20 ns</td>
<td>&lt;60 ns</td>
</tr>
<tr>
<td>W/E Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>0.4 ns</td>
<td>&lt;15 ns</td>
<td>&lt;25 ns</td>
<td>1 μs/10 ms</td>
<td>1ms/0.1 ms</td>
<td>15 ns</td>
<td>50/120 ns</td>
</tr>
<tr>
<td>2018</td>
<td>&lt;0.1 ns</td>
<td>&lt;15 ns</td>
<td>&lt;0.5 ns</td>
<td>1 μs/10 ms</td>
<td>1ms/0.1 ms</td>
<td>1 ns</td>
<td>Not known</td>
</tr>
<tr>
<td>Retention Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Voltatile</td>
<td>64 ms</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
</tr>
<tr>
<td>2018</td>
<td>Voltatile</td>
<td>64 ms</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
<td>&gt; 10 y</td>
</tr>
<tr>
<td>Write Cycles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>&gt;3E16</td>
<td>&gt;3E16</td>
<td>&gt;1E16</td>
<td>&gt;1E5</td>
<td>&gt;1E5</td>
<td>&gt;1E13</td>
<td>&gt;1E12</td>
</tr>
<tr>
<td>2018</td>
<td>&gt;3E16</td>
<td>&gt;3E16</td>
<td>&gt;1E16</td>
<td>&gt;1E5</td>
<td>&gt;1E5</td>
<td>&gt;1E16</td>
<td>&gt;1E15</td>
</tr>
<tr>
<td>Write operating voltage (V)</td>
<td>1.2</td>
<td>2.5</td>
<td>1.8</td>
<td>7.9</td>
<td>15-17</td>
<td>0.9 - 3.3</td>
<td>3</td>
</tr>
<tr>
<td>2018</td>
<td>0.7</td>
<td>1.5</td>
<td>&lt;1.8</td>
<td>6-8</td>
<td>15-17</td>
<td>0.7 - 1</td>
<td>&lt;3</td>
</tr>
<tr>
<td>Read operating voltage (V)</td>
<td>1.2</td>
<td>2.5</td>
<td>1.8</td>
<td>2.5</td>
<td>2.5</td>
<td>0.9 - 3.3</td>
<td>3</td>
</tr>
<tr>
<td>2018</td>
<td>0.8</td>
<td>1.5</td>
<td>&lt;1.8</td>
<td>1.2</td>
<td>1.2</td>
<td>0.7 - 1</td>
<td>&lt;3</td>
</tr>
<tr>
<td>Write energy (J/bit)</td>
<td>7E-16</td>
<td>1E-16</td>
<td>1E-10</td>
<td>8E-15</td>
<td>8E-15</td>
<td>2E-14</td>
<td>1E-10</td>
</tr>
<tr>
<td>2018</td>
<td>2E-17</td>
<td>4E-17</td>
<td>2E-11</td>
<td>3E-15</td>
<td>3E-15</td>
<td>4E-15</td>
<td>Not known</td>
</tr>
<tr>
<td>3D Potential</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Cell Stacking</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not known</td>
</tr>
<tr>
<td>MLC Potential</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not known</td>
</tr>
<tr>
<td>Die Efficiency %</td>
<td>70</td>
<td>56-63</td>
<td>Not known</td>
<td>35-46</td>
<td>54-60</td>
<td>Not known</td>
<td>Not known</td>
</tr>
</tbody>
</table>
- End-of-roadmap (2020, $L=6$ nm) NAND $F= 14$ nm, $Tox= 6$ nm, Cell area $\sim 1000$ nm$^2$; (3X for NOR)
- NC spacing $\sim$ tunnel oxide thickness $\sim 4$ nm
- Spacing & NC size/DOS affects charge capture cross-section & Read
- Optimal spacing depends on dielectric & NC band diagram, retention ($10^8$ s), Erase/Write times ($10^{-6}$ s) $\Rightarrow E_{barr} = 1$eV
- For NC densities of $\sim 10^{12}$ cm$^{-2}$, room for $\sim$10 NC
- Variability is a challenge for NC & trap-based cells
- Self-assembly and non-planar structures?
There is plenty of room at the bottom! Feynman

The fundamental parameters of the human brain\(^{132}\) are estimated to be:

- Number of neurons—2E10
- A single neuron can make 100 to 10,000 synaptic connections
- Mass—1.3 kg\(^{133}\)
- Volume—600 cm\(^3\)
- Power consumption—15–30 Watts
- Information stored—1E12 (short term) bits
- Information processed—1E16 bits/second

- Will these devices make it? The answer is a very definite…..

Maybe!!