EVOLUTION OF
NONVOLATILE SEMICONDUCTOR MEMORY
From Invention to Nanocrystal Memory

S. M. SZE
National Chiao Tung University
National Nano Device Laboratories
Hsinchu, Taiwan
OUTLINE

- INTRODUCTION
- DEVICE PHYSICS
- HISTORICAL DEVELOPMENT OF NVSM
- EMERGING TECHNOLOGIES
- APPLICATIONS AND FUTURE TRENDS
- CONCLUSION
"The principal applications of any sufficiently new and innovative technology always have been — and will continue to be — applications created by that technology." — Herbert Kroemer
EXAMPLES

- The Transistor (1947) was not just a replacement for vacuum tubes, it *created* the modern computer and the new industrial revolution.

- The Heterojunction Laser (1963) has revolutionized the optoelectronics technology, it *created* optical fiber communication, CD, and DVD.

- The Nonvolatile Semiconductor Memory (1967) has revolutionized the information storage technology, it *created* the mobile phone, notebook computer, digital photography, MP3, PDA and GPS.
SEMICONDUCTOR MEMORIES

- VOLATILE MEMORIES — Lose stored information once supply is switched off
  - DRAM (Dynamic Random Access Memory)
  - SRAM (Static Random Access Memory)

- NONVOLATILE MEMORIES — Keep stored information when the power supply is switched off
  - Flash MEMORY
  - EEPROM (Electrically Erasable-Programmable Read Only Memory)
  - EPROM (Erasable-Programmable Read Only Memory)
COMPARISON OF SEMICONDUCTOR MEMORIES

<table>
<thead>
<tr>
<th></th>
<th>DRAM</th>
<th>SRAM</th>
<th>Flash</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T + 1C</td>
<td>6T / 4T + 2L</td>
<td>1T</td>
<td></td>
</tr>
<tr>
<td>Volatile</td>
<td>Volatile</td>
<td>Nonvolatile</td>
<td></td>
</tr>
</tbody>
</table>

- DRAM: Dynamic Random Access Memory
- SRAM: Static Random Access Memory
- Flash: Non-volatile memory
ATTRIBUTES OF AN IDEAL MEMORY

1. Nonvolatility
2. High density (consumes small space/bit)
3. Low power consumption
4. In-system re writability
5. Bit alterability
6. Fast read/write/erase
7. Endurance (high write/erase cycles)
8. Low cost
9. Single-power supply
10. Highly scalable
11. Ruggedness
12. Highly integrable with other system technologies
## COMPARISON OF MEMORY ATTRIBUTES

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Memory</th>
<th>DRAM</th>
<th>SRAM</th>
<th>Flash</th>
<th>EFPROM</th>
<th>EPROM</th>
<th>Hard Disk</th>
<th>Floppy Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nonvolatility</td>
<td>X</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>2. High density</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3. Low power</td>
<td>X</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4. In-system re writability</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>5. Bit alterability</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>6. Fast read/write</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7. High endurance</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8. Low cost</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>9. Single-power supply</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>10. Highly-Scalable</td>
<td>X</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11. Ruggedness</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>12. Highly integrable</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Total Good Attributes</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
COMPARISON OF MEMORY ATTRIBUTES

High density

EPROM

DRAM

FLASH

EEPROM

SRAM

Nonvolatility

Electrical Re-writability
THE FIRST NONVOLATILE SEMICONDUCTOR MEMORY

- The Floating-Gate Concept (1967)

FLOATING-GATE MEMORY DEVICE OPERATION

(a) Programming Mode
(Fowler-Nordheim or Direct Tunneling)

(b) Storage Mode

(c) Erase Mode
(FN or DT)
CHARGE INJECTION MECHANISMS

(a) FOWLER-NORDHEIM TUNNELING

(b) HOT-ELECTRON INJECTION
CURRENT DENSITY EQUATIONS

For Fowler-Nordheim Tunneling

\[ J = C E^2 \exp \left( -\frac{E}{E} \right) \]

\[ E = \frac{V_g}{d_1 + d_2 (\varepsilon_1/\varepsilon_2) - \varepsilon_1 + \varepsilon_2 (d_1/d_2)} \frac{Q}{\varepsilon_1} \]

\[ Q = \int_0^L J dt' = \text{stored charge} \]

For Hot-Electron Injection

\[ J = C_2 E_{in}^2 \exp(-\Phi/\lambda E_{in}) \]

\[ E_{in} = \text{Maximum Channel Lateral Field} \]

\[ \lambda = \text{Inelastic Scattering Length} \]

\[ \Phi = \text{Barrier Height between Si and SiO}_2 \]
GATE CURRENT VERSUS GATE VOLTAGE

(a) Fowler-Nordheim Tunneling

(b) Hot-Electron Injection

\[ \begin{align*}
V_D &= 10 \text{ V} \\
Z &= 60 \mu\text{m} \\
L_{eff} &= 2.2 \mu\text{m} \\
N_g &= 3.3 \times 10^{18} \text{ cm}^{-3}
\end{align*} \]
PROGRAMMING CURRENT AND STORED CHARGE
THRESHOLD VOLTAGE SHIFT DUE TO CHARGE STORAGE ON THE FLOATING GATE

\[ \Delta V_T = - \frac{Q}{C_{FC}} \]
DEMOnstration of EEPROM operation

(a) Applied gate pulse voltage

(b) Source-drain current
   For $V_G = \pm 50 \text{ V}$, $t_1 = 0.5 \mu\text{s}$
FIRST PAPER ON NVSM (1967)

- The possibility of nonvolatile memory in semiconductors was recognized for the first time, and an experimental electrically erasable-programmable read only memory was demonstrated.

- The paper introduced not only the basic concept of nonvolatility, but also the floating-gate structure which has been the dominant technology for nonvolatile charge storage.

- The paper introduced nonlinear transport processes for programming and erase. These approaches have been adopted for most NVSM devices.
# HISTORY OF FLOATING-GATE NVSM

<table>
<thead>
<tr>
<th>YEAR</th>
<th>DEVICE</th>
<th>INVENTOR(S)/AUTHOR(S)</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967</td>
<td>Floating-Gate Concept</td>
<td>D. Kahng, S. M. Sze</td>
<td>Bell Labs</td>
</tr>
<tr>
<td>1971</td>
<td>EPROM-FAMOS</td>
<td>D. Frohman-Bentchkowsky</td>
<td>Intel</td>
</tr>
<tr>
<td>1976</td>
<td>EEPROM-SAMOS</td>
<td>H. Iizuka et al.</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1978</td>
<td>NOVRAM</td>
<td>E. Harari et al.</td>
<td>Hughes</td>
</tr>
<tr>
<td>1984</td>
<td>Flash Memory</td>
<td>F. Masuoka et al.</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1988</td>
<td>ETOX Flash Memory</td>
<td>V. N. Kynett et al.</td>
<td>Intel</td>
</tr>
<tr>
<td>1995</td>
<td>Multilevel Cell</td>
<td>M. Bauer et al.</td>
<td>Intel</td>
</tr>
</tbody>
</table>
TWO EARLY VERSIONS OF FLOATING GATE MEMORY

In 1971
FAMOS (Floating Gate Avalanche Injection MOS)

In 1976
SAMOS (Stacked Gate Avalanche Injection MOS)
NOVRAM (Non-Volatile RAM) IN 1978
AN ETOX CELL OPERATION (EPROM with Tunnel Oxide) in 1988

(a) Select gate
Floating gate
Source
Substrate
Drain

(b) Select gate
Floating gate
Approximately +6V
Source
Substrate
Drain
Program, hot electron injection

(c) Select gate
Floating gate
+12V
Source
Substrate
Drain
Erase, Fowler-Norvel tunneling
SINGLE-ELECTRON MEMORY CELL
- A Limiting case of the Floating Gate NVSM
Coulomb Blockade: the effect of the electron reduces the potential in the quantum well and blocks the transfer of another electron.
MEMORY DENSITY OF DRAM AND SET

Minimum Feature Size (mm)

- 256T
- 64T
- 16T
- 4T
- 1T
- 256G
- 64G
- 16G
- 4G
- 1G

300K  NOVORAM
77K
SET+FET HYBRID
40K
10K
FET DRAM (SIA Forecast)
MULTILEVEL-CELL
THRESHOLD VOLTAGE DISTRIBUTION
“Moore’s Law” Continues with FG NVM

- Visibility for scaling of floating gate clear to 32 nm
- Forms high “hurdle” for new technologies to leap

90 nm cell is 1/476th of 1.5 μm cell

Source: Intel
8 Gb FLASH MEMORY

(3 V  65 nm  133 mm²  0.02μm²/cell)
CELL SIZE REDUCTION VERSUS YEAR
EMERGING NONVOLATILE MEMORIES

MRAM

OUM

Data Storage Region
Chalcogenide
Phase Change Material
Resistive Electrode
Heater

Polymer

FERAM

RRAM

FE Polymer
EMERGING NONVOLATILE MEMORIES

- **MRAM** (Magneto-resistive Random Access Memory): based on tunneling Magneto-resistance effect.
- **FeRAM** (Ferroelectric RAM): based on remanent polarization in Pervoskite materials.
- **OUM** (Ovonies Unified Memory) also called **PCRAM** (Phase-change RAM): based on reversible phase conversion between the amorphous and the crystalline state of a chalcogenide glass, which is accomplished by heating and cooling of the glass.
- **RRAM** (Resistance RAM): based on change in resistance with applied electric field.
- **Polymer Memory**: based on resistance change of polymer at cross point of metal layers.
- **Millipede Memory**: based on concept similar to punch cards, using thermally assisted, reWritable displacement media such as PMMA.
- **Nanocrystal Memory**: based on distributed nano-floating-gate concept.
NANOCRYSTAL MEMORY

• Nanocrystal memory is a multiple-nano-floating-gate structure. Instead of charge stored in a single floating gate, charge is stored in many floating nanocrystals.

• Nanocrystal memory can extend the floating-gate scaling by allowing a further decrease in the tunnel oxide thickness.

• Nanocrystal memory offers enhanced robustness and fault-tolerance of distributed charge storage.
Advantage of discrete storage nodes: robustness to stress induced leakage current → scaling of the tunnel oxide
TEM OF NANOCRYSTAL MEMORY

EFTEM tuned to the SiO₂ plasmon energy

n⁺ poly-Si

SiO₂

n-Si

20 nm

Si dot

CVD SiO₂

3 nm tunnel SiO₂

Si substrate

3 nm

HR-TEM
NANOCRYSTAL FABRICATION BY POLYMER SELF ASSEMBLY

Fig. 1. (a-d) Schematic diagrams of the process flow used to create an array of Si nanocrystals beginning with a porous self-assembled polymer template.

Fig. 2. (a-c) 200–200 nm top-down SEM images of an porous polymer film on SiO$_2$ (b) porous dielectric after pattern transfer into the SiO$_2$, and (c) Si nanocrystal array. (d) Distribution of polymer pore, dielectric pore, and Si nanocrystal diameters.

"Low voltage, scalable nanocrystal FLASH memory fabricated by templated self assembly", K.W. Guarini et al., 2003 IEDM Technical Digest, Session 22.2.
FLASH APPLICATIONS
MEMORIES FOR MOBILITY

- Nonvolatile = low power: power down to save power
- Low cost: one memory for program and data storage
- ≠magnetic hard disk: need solid state, low power, rugged
- Small form factor and light weight
MOBILE LIFESTYLE : DATA ANYTIME, ANYWHERE
APPLICATIONS OF NVSM

- **Communications**
  Cellular Phone, Cordless, Bluetooth, Pagers, Modems, Internet Appliance, Line Card, PBX, Set Top Box, LAN Modules, Network System, Network Adapters, Data Communication (5.6 Billion Units from 1999 to 2004).

- **Computer and Peripheral**
  PC, Flash-Drive Notebook, USB Drive, Personal Digital Assistant, Hard-Disk Drive, Portable DVD, Graphics Card, FAX, Printer (3.5 Billion Units).

- **Consumer**
  Digital Camera, Digital Voice Recorder, MP3 Music Player, Digital Camcorder, Telephone Answering Device, Games, Dictionary, Organizer, Toy, Electronic Book, Directories (2.5 Billion Units).

- **Transportation**
  Automotive systems, Global Positioning System (Auto, Marine, Aviation) (570 Million Units).

- **Industrial**
  Meter, Sensing Device, Embedded System, POS, Solid State Drive (630 Million Units).
CELLULAR PHONE BLOCK DIAGRAM
GLOBAL TELEPHONE SUBSCRIPTIONS
END-USE APPLICATIONS OF NVSM (1999-2010)
SEMICONDUCTOR MEMORY MARKET

The diagram illustrates the sales volume (in billions) of different types of semiconductor memory over a period from 1990 to 2010.

- DRAM
- Flash
- SRAM

The sales volume shows significant fluctuations, with peaks and troughs indicating changes in market demand and technological advancements over the years.
TECHNOLOGY DRIVER FOR ELECTRONIC INDUSTRY

(Application Created by the New Technology)
CONCLUSION

- The nonvolatile semiconductor memory (NVSM) was invented in 1967 based on the floating-gate concept for charge storage and nonlinear transport processes for programming and erase.

- Because of its attributes of high density, low-power consumption, nonvolatility, and electrical rewritability, NVSM has surpassed SRAM in the global memory market in 1999, and is poised to eclipse DRAM in the near future.

- NVSM has revolutionized electronic technology and enabled tremendous advances in portable electronic systems such as the mobile phone, notebook computer, digital photography, PDA, GPS, Smart IC card, etc.

- Many emerging NVSM technologies have been investigated to extend the floating-gate scaling to sub-10 nm regime. A major candidate is the nanocrystal memory.
ACKNOWLEDGEMENTS

R. Bez et al  ST Microelectronics
Stefan Lai  Intel
S. Lombardo et al  CRN-IMM
F. Masauoka et al  Toshiba
Sang Lyul Min  Seoul University
A. Niebel  Web-feet Research
UNCONVENTIONAL APPLICATIONS OF NVSM

- Medical nanorobots
  
  Destrobot
  DNA Repair Machine
  Digital DNA

- Sensor networks in general and battlefield applications in particular
FeRAM

- **Operation**
  - Selected PZT crystalline materials have bi-stable center atom
  - Data is stored by applying an voltage to polarize the internal dipoles “Up” or “Down”
  - Reading by sensing the displacement current
  - Fast read/write < 100 nSec
  - Destructive read, limited number of read cycles due to fatigue
  - Very low power consumption
MRAM

- **Operation**
  - Cell is 1 MJT + 1 Transistor
  - Electric current switches the magnetic polarity of sense layer
  - Change in magnetic polarity sensed as resistance change in tunnel junction
  - Non destructive read
  - Very fast read/write performance, < 10 nSec reported
  - Unlimited number of R/W endurance
  - Relatively high write current

- **FET cell (1TJ-1T)**
  - Larger cell area (>8 $F^2$)
  - Faster random access time
  - SRAM like ($\leq$10ns)
Ovonics Unified Memory

Operation
- Chalcogenide material alloys used in re-writable CDs and DVDs
- Electrical energy (heat) converts the material between crystalline (conductive) and amorphous (resistive) phases
- Cell reads by measuring resistance
- Non-destructive read
- $\sim 10^{12}$ write/erase cycles
- Medium write power
- Relatively easy integration with CMOS
Complex Metal Oxide RRAM

- **Operation**
  - PCMO material, Complex metal oxide studied for high temp superconductivity
  - Change in resistance with applied electric field
  - Low resistance with forward bias, high resistance when reverse electric field applied
  - Low field read by measuring resistance
  - Relatively low power write

*Fig.8 Positive pulse writes the resistor to high resistance state. Negative pulse resets the resistor to low resistance state. Programming efficiency depends strongly on pulse width*
Programmable Metallization Cell

- **Operation**
  - Silver "dissolved" in chalcogenide
  - Change in resistance with applied electric field driving silver to form low resistance path
  - Reversible with reverse field
  - Low field read with no disturb
  - Very fast write and relatively low power
Resistance Polymer Memory

- **Operation**
  - Polymer material with special formulation
  - Change in resistance due to ionic transport with applied electric field
  - Low resistance when ionic conductance paths formed, high resistance when process is reversed when paths broken
  - Low field read with no read disturb
  - Low cost with polymer
Carbon Nanotube Switches

- **Operation**
  - Carbon nanotube suspended in crossbar architecture
  - Electrostatic field attracts the nanotube and held together by van der Waals force
  - Reverse bias repulse the wires
  - Reading by low vs high resistance
  - Array Architecture TBD
Molecular Memory

- **Operation**
  - Molecules in cross point array
  - Change in electronic states in Redox process with applied voltage
  - Memory state sensed by charge displacement
  - More than 1 state can be stored by design of molecule (up to 8 demonstrated)
Molecular Tunnel Memory

- **Operation**
  - Molecules in cross point array
  - Change in resistance with applied electric field
  - Memory state sensed by measured resistance change
Ferroelectric Polymer Memory

- **Operation**
  - Polymeric Ferroelectric RAM (PFRAM)
  - Polymer chains with a dipole moment
  - Data stored by changing the polarization of the polymer between metal lines
  - Zero transistors per bit of storage
  - Polymer layers can be stacked
Millipede Memory

**Operation**

- Basic concept similar to punch cards
- Thermally assisted, rewriterable displacement media, PMMA
- Large array of independently Z-axis controlled tips
- Low moving mass
- Direct point to point motion, no rotational latency
- Low power, no motors, actuators
- High data transfer rates, concurrent data transfer to/from multiple tips
- Packaging similar to hermetic devices
3D One Time Program Memory

- **Operation**
  - Multi-layer one time programmable diodes at minimum lithography dimensions
  - Simplified architecture with minimum number of metal lines
  - Zero transistor per cell
  - Standard CMOS transistors under and outside array, process cost amortized over multiple memory cells
  - Significant lower cost compared to single layer
Spin Polarized E Beam Magnetic Memory

- **Operation**
  - Basic concept similar to Magnetic Hard Drive
    - Mechanical arm -> spin polarized E Beam
    - Spinning Magnetic Media -> steered E Beam
  - Much lower first read latency, first access limited by E Beam steering; fast read with E Beam
  - Lower power, no motor, no arm actuator
  - No moving part, rugged, vacuum space required
  - Not as small or compact as all solid state memories