NAND Flash memory

Samsung Electronics, co., Ltd

Flash design team

2010. 05. 07

Kihwan Choi
Contents

- Introduction
- Flash memory 101
- Basic operations
- Current issues & approach
- In the near future
“Flash memory” in Wikipedia

- Flash memory is a non-volatile computer storage technology that can be electrically erased and reprogrammed.

- Flash memory offers fast read access times (although not as fast as volatile DRAM memory used for main memory in PCs) and better kinetic shock resistance than hard disks.

- Flash memory (both NOR and NAND types) was invented by Dr. Fujio Masuoka while working for Toshiba circa 1980. According to Toshiba, the name "flash" was suggested by Dr. Masuoka's colleague, Mr. Shoji Ariizumi, because the erasure process of the memory contents reminded him of the flash of a camera.

- NAND flash also uses floating-gate transistors, but they are connected in a way that resembles a NAND gate: several transistors are connected in series, and only if all word lines are pulled high (above the transistors' VT) is the bit line pulled low.
“Flash memory” in Wikipedia

- To read, most of the word lines are pulled up above the VT of a programmed bit, while one of them is pulled up to just over the VT of an erased bit. The series group will conduct (and pull the bit line low) if the selected bit has not been programmed. NAND flash uses tunnel injection for writing and tunnel release for erasing.

- One limitation of flash memory is that although it can be read or programmed a byte or a word at a time in a random access fashion, it must be erased a "block" at a time.

- Another limitation is that flash memory has a finite number of erase-write cycles. Most commercially available flash products are guaranteed to withstand around 100,000 write-erase-cycles, before the wear begins to deteriorate the integrity of the storage
“Flash memory” in Wikipedia

- Technology scaling down

Reference: EETimes article on NAND scaling, 3/22/2010
## Major players

### Table 4. NAND Market Share

<table>
<thead>
<tr>
<th></th>
<th>Sales ($M)</th>
<th>Share (%)</th>
<th>GB Equivalent (M)</th>
<th>Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Samsung</td>
<td>5,337</td>
<td>2,007</td>
<td>31.50</td>
</tr>
<tr>
<td></td>
<td>Toshiba</td>
<td>3,570</td>
<td>1,689</td>
<td>26.50</td>
</tr>
<tr>
<td></td>
<td>SanDisk</td>
<td>1,997</td>
<td>1,181</td>
<td>18.50</td>
</tr>
<tr>
<td></td>
<td>Micron</td>
<td>1,203</td>
<td>664</td>
<td>10.40</td>
</tr>
<tr>
<td></td>
<td>Hynix</td>
<td>1,066</td>
<td>347</td>
<td>5.40</td>
</tr>
<tr>
<td></td>
<td>Intel</td>
<td>826</td>
<td>465</td>
<td>7.30</td>
</tr>
<tr>
<td></td>
<td>Numonyx</td>
<td>330</td>
<td>18</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Spansion</td>
<td>37</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Powerchip</td>
<td>3</td>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Renesas</td>
<td>2</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>14,371</td>
<td>6,375</td>
<td>100</td>
</tr>
</tbody>
</table>

Source: Gartner (March 2010)
Now in the market ...

Intel, Micron and IMFT announce world's first 25-nm NAND technology
February 2, 2010, By Sanjeev Ramachandran in Hardware

The NAND flash production scene has received a shot in the arm with Intel Corporation and Micron Technology making it public that the world's first 25-nanometer (nm) NAND technology is now on stream. Significantly enough, the 25nm process is the smallest NAND technology as well as the smallest semiconductor technology in the world.

Hynix Develops 26nm NAND Flash Memory
Tuesday, February 09, 2010

South Korea's Hynix Semiconductor Inc., the world's second-largest memory chipmaker, said Tuesday that it has developed a 26-nanometer based NAND flash memory chip. The company is the world's second flash memory maker to apply the below 30-nanometer technology. Mass production of the new memory will start in in July.

Toshiba readies sub-25nm flash memory chip production
By Jose Vilches, TechSpot.com Published: April 5, 2010, 12:14 PM EST

The company produces 32nm and 43nm memory chips, but the plan is to begin production on "sub-25nm" chips that would enable larger storage capacities to be shoved into the same form factors that we use today. Toshiba will begin output of NAND chips with circuitry widths in the upper 20 nanometre range soon, while production of chips with circuitry widths in the lower 20 nanometres is slated to start as early as 2012.

Samsung pioneers 20-nm NAND flash memory technology
April 19, 2010, By Thomas Antony in Storage

Right on the heels of Toshiba's announcement to start on sub-25nm flash memory, Samsung today announced the industry's first production of 20 nanometer class NAND chips for use in SD cards. Samsung's 20nm MLC 32-gigabit NAND chips are sampling now for use in embedded storage and SD memory cards ranging from 4GB to 64GB. This is a significant step forward for Samsung who started its 30nm production just one year ago. The new class of memory chips will allow for higher-density in storage, lower manufacturing costs and 50% higher productivity than 30nm technology.
Flash memory 101
Flash memory cell vs. MOSFET

- Flash cell has a charge storage layer such that Vth of a cell can be changed → memorize information
Flash memory operation

- Write & read binary data to a flash cell
  - data ‘0’ \(\rightarrow\) ‘OFF’ state (program)
  - data ‘1’ \(\rightarrow\) ‘ON’ state (erase)

MOS Transistor

<table>
<thead>
<tr>
<th>No. of cells</th>
<th>Read</th>
<th>Program</th>
<th>Erase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘ON’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘OFF’</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ V_{th} [V] \]

\[ V_{gs} (V_{g} > V_{th1}) \]

\[ V_{gs} (V_{g} < V_{th2}) \]

\[ V_{ds} (>0V) \]

\[ I_{ds1} (>0) \]

\[ I_{ds2} (=0) \]
Flash memory cell structure

- Cell Vth changes depending on the amount of F/G charge
- electrons can be injected(ejected) into(out of) the F/G through Tox with electric field across Tox

Diagram:

- B/L Metal
- Dielectric : ONO
- Charge Storage Node
- Tunnel Ox
- Source
- Drain

- C/G(W/L)
- F/G(F-Poly)
NAND vs. NOR

**Truth table**

**NAND**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**NOR**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Logic gate**

- NAND
  
- NOR

**Circuit**

- NAND
  
- NOR

**Flash**

- NAND
  
- NOR
### Features: NAND vs. NOR

<table>
<thead>
<tr>
<th>Feature</th>
<th>NAND</th>
<th>NOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell Array &amp; Size</strong></td>
<td><img src="image1" alt="NAND Cell Array Diagram" /></td>
<td><img src="image2" alt="NOR Cell Array Diagram" /></td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td><img src="image3" alt="NAND Cross-section" /></td>
<td><img src="image4" alt="NOR Cross-section" /></td>
</tr>
<tr>
<td><strong>Features</strong></td>
<td>Small Cell Size, High Density Low Power &amp; Good Endurance ➔ <strong>Mass Storage</strong></td>
<td>Large Cell Current, Fast Random Access ➔ <strong>Code Storage</strong></td>
</tr>
</tbody>
</table>
Write methods in Flash memory

**Hot Electron Injection**

- $P_{Ed}$: No collision before reaching interface
- $P_{PoC}$: No collision in field reversal region in oxide
- $P_{ob}$: Gains enough energy

Program in NOR Flash
Impact ionization at drain side

**Fowler-Nordheim Tunneling**

- $\Phi_B = 3.1$ eV

Program/Erase in NAND Flash
Erase in NOR Flash

Impact ionization at drain side
NAND Flash: Program & Erase

V<sub>cg</sub> = 18 V

<Diagram>

V<sub>sub</sub> = 0V

Program
F-N Tunneling

Off cell
(Solid-0)

V<sub>cg</sub> = 0V

float

V<sub>sub</sub> = 20V

Erase
F-N Tunneling

On cell
(Solid-1)
Coupling ratio

\[ V_{fg} = V_{cg} \times \alpha_{cg} \]

\[ \alpha_{cg} = \frac{C_{ONO}}{(C_D + C_S + C_B + C_{ONO})} \]

\( \alpha_{cg} \) : Coupling Ratio

(From Q=CV and Charge Conservation Law)

For fast programming, high Vfg is required, i.e. either high Vcg or large \( \alpha_{cg} \)
NAND Flash cell structure
Terms in NAND Flash – string, page, block

- **string**: minimum cell array element
- **page**: program unit
- **block**: erase unit

Diagram illustrating the layout of NAND Flash memory, showing strings, pages, and blocks.
NAND Flash chip architecture

- High density & simple architecture (cell efficiency > 65%)

- Block = 128-page (2Mb~4Mb)
  - SLC : 64-page
  - MLC : 128-page

- Control I/O Pad
- WL Decoder & Driver
- Page Driver & Buffer (2KB~4KB)
- Peripheral & Charge Pump
- Data I/O Pad

Samsung 63nm 8Gb MLC NAND
Functional block diagram of NAND Flash

Diagram showing the block diagram of a NAND Flash memory with the following components:

- **X-Buffers Latches & Decoders**
- **Y-Buffers Latches & Decoders**
- **Command Register**
- **Control Logic & High Voltage Generator**
- **Cell Array**
- **Data Register & S/A**
- **Y-Gating**
- **I/O Buffers & Latches**
- **Global Buffers**
- **Output Driver**

Inputs include A12~A30, A0~A11, Command, CE, RE, WE, and outputs are I/O 0 to I/O 7.
Basic operations

NAND Flash
Read operation

- **Bias condition**
  - selected WL: Target voltage ($V_{target}$)
  - unselected WLs: high enough to conduct all cells in a string
Sensing margin

- The ratio of On cell & Off cell current
Read disturbance

- Increasing $V_{\text{read}} \rightarrow$ soft program occurs in the unselected cell of selected string

On cell moves to off cell
Program operation

- **Bias condition**
  - selected WL : Program voltage (Vpgm)
  - unselected WLs : Pass voltage (Vpass)

- Selected WL
- Unselected WLs
- SSL
- GSL

# of cell

on cell becomes off cell

Vcc  0V  Vcc

0V  Vcc  Vcc

- 26/48 -
Program inhibition

Should not be programmed (= program inhibited)

For inhibition only, the higher Vpass, the better Vpgm disturbance
But, higher Vpass causes more Vpass disturbance
Self boosting

Channel potential strongly depends on the cell states in a string
**Vpass window**

- “Optimal Vpass region” considering both Vpgm and Vpass disturbances at the same time

![Diagram showing program cell and inhibit cell voltages with Vpass window curve]
Erase operation

- **Bias condition**
  - all WLs in the selected Block : 0V
  - GSL/SSL : Floating
  - Bulk : Vera

- off cell becomes on cell
Erase disturbance

- "Self boosting" can be used

Old Method

Select Block

SSLn(20V)
WL0(0V)
WL7(0V)
GSLn(20V)
CSL(20V)
GSLn-1(20V)
WL7(20V)
WL0(20V)
SSLn-1(20V)

Unselect Block

Cell Array Substrate 0V 20V

B/L (20V)

New Method

Select Block

SSLn(Floating)
WL0(0V)
WL7(0V)
GSLn(Floating)
CSL(20V)
GSLn-1(Floating)
WL7(Floating)
WL0(Floating)
SSLn-1(Floating)

Unselect Block

Cell Array Substrate 0V 20V

B/L (20V)
Cell Vth distribution

- Cell Vth width requirement
  - 1bit/cell vs. 2bit/cell vs. 3bit/cell

![Diagram showing SLC, MLC, and 3-bit cell Vth distributions with Vth Upper Limit (NAND Flash)]
Cell Vth width control

- Incremental Step Pulse Program (ISPP)
  - programmed state Vth width can be controlled
  - narrower width requires more program loop

Change of cell Vth: slow vs. fast cell

Cell Vth change over the number of loops.

Initial Vth distribution and the change in cell Vth for both fast and slow cells.
Current issues & approaches
Criteria for a NAND Flash device

Cost
- Chip size
- Bit density
- etc...

Performance
- Program time
- Read time
- etc...

Reliability
- Data retention
- Program/Read cycle
- etc...

All these are strongly dependent on each other and may become worse as "cell size shrinks down"
Details on programmed Vth

- Factors which affect to programmed Vth width
  - Ideal: ΔVpgm during ISPP program
  - Noise: F-poly coupling, CSL noise, Back pattern dependency
  - Reliability: Endurance, program/read disturbance, charge loss

![Diagram showing program disturbance, read disturbance, charge loss, F-poly coupling, and ideal conditions]
Neighborhood interference

- **F-poly coupling noise**
  - Cell Vth can be raised as neighboring cells are programmed

![Diagram showing neighborhood interference](image)
Cell Vth vs. F-poly coupling

Cell State  # of cells

Vth

Initial  WL  BL  Diag.  Final

WL1  Diag.  WL  Diag.
WL2  BL  V  BL
WL3  Diag.  WL  Diag.

BL1  BL2  BL3
Cell size vs. F-poly coupling

Cell-to-Cell Coupling Trend

- diagonal
- WL-WL
- BL-BL

’09 ISSCC
F-poly coupling reduction – (I)

- **Shadow program**
  - Make final Vth after being coupled

- **Shadow sequence**
  - LSB
  - MSB

- **Conventional sequence**
  - LSB
  - MSB

- Make final Vth after being coupled
F-poly coupling reduction – (II)

- Reprogram
  - Make final Vth after being coupled

Coarse program & coupled

Fine program

# of cells

cell Vth

where P1' < P1, P2' < P2, P3' < P3

cell Vth

Program performance overhead due to “Fine program”
F-poly coupling reduction – (III)

- **All-bit line (ABL) architecture**
  - all cells in a WL are programmed at the same time
  - in SBL, cells in even BL first, cells in odd BL next (BL-coupling exists)

<table>
<thead>
<tr>
<th>SBL (Shielded-bit line)</th>
<th>ABL (All-bit line)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL(e) BL(o)</td>
<td>BL(1) BL(2)</td>
</tr>
<tr>
<td>Physical 8KB cells</td>
<td>One block</td>
</tr>
<tr>
<td>64 cells (one string)</td>
<td>PB</td>
</tr>
</tbody>
</table>

**Features: SBL vs. ABL**

<table>
<thead>
<tr>
<th></th>
<th>SBL</th>
<th>ABL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of PBs (page depth)</td>
<td>4KB</td>
<td>8KB</td>
</tr>
<tr>
<td>pages/block</td>
<td>256 pages</td>
<td>128 pages</td>
</tr>
</tbody>
</table>
In the near future
3D Flash memory

- Bit-Cost Scalable(BiCS) technology

2007 IEDM, Toshiba
3D Flash memory

- TCAT (Terabit Cell Array Transistor) technology

2009 VLSI, Samsung
What is CTF?

- CTF (Charge Trap Flash)
  - SONOS type Flash
  - electron is trapped in nitride trap layer

**SONOS**

- Si (S)
- Oxide (O)
- Nitride (N)
- Oxide (O)

*P.C.Y. Chen, TED, V. 24, pp.584-586, 1977*

**Floating Gate**

- Si
- ON
- F-Gate
- Oxide (O)

*F.Masuoka (Toshiba)*

- Control Gate
- Interpoly Dielectric
- Tunnel Oxide
- Floating gate

Discrete traps

n⁺

p-well

n⁺
Floating gate vs. SONOS

Floating Gate
- Polysilicon Charge Storage Medium (Conductor)
- ONO Interpoly Dielectric
- Floating Gate
- Charge are mobile
- ~100Å Tunnel Oxide
- Leakage path due to SILC, etc.
- Substrate
- All stored charge is lost
- Thicker tunnel oxide
- Device Scaling is difficult

SONOS
- Nitride Charge Storage Medium (Insulator)
- Control Gate
- Immobile Charge
- 50Å~90Å Oxide
- Nitride
- 18Å~70Å Oxide
- Substrate
- Only part of the charge is lost
- Thinner tunnel oxide
- Device Scaling is easy

"No neighborhood coupling"
Thanks for your listening!