Securing Flash Technology

Dr. Elena Trichina, Spansion International
Dr. Helena Handschuh, Spansion EMEA
Memory

- All computer-based systems contain memory where information is stored while waiting to be operated on by the Central Processing Unit (CPU) of the computer.

- There are two types of memory: volatile and nonvolatile
  - Operating instructions are in nonvolatile memory
  - RAM used as a read/write working storage
  - Bulk storage in cheaper serial access media (disc or tape)

- Memory may be embedded or discrete

- System design dictates actual memory needs
  - Design parameters: speed, density, cost, von-volatility, read/write, electrical parameters, power, reliability, form-factor, operating conditions, time to market ...
Semiconductor Memories

- Volatile memories, although very fast in writing and reading (SRAM) or very dense (DRAM), lose data content when the power supply is turned off.

- Nonvolatile memories, like EPROM, EEPROM, Flash are able to balance programming and reading performances with the capability to keep the content without power supply.
Flexibility-Cost Analysis of NVM

- Flexibility means the possibility to be programmed and erased many times on the system with minimum granularity (whole chip, page, byte, bit)

- Cost means process complexity/silicon density, or cell size
Flash Technology
Flash Cell Basic Structure

- A flash cell is floating-gate MOS cell, i.e., a transistor with a gate completely surrounded by dielectrics (FG), and electrically governed by a capacitively coupled control gate (CG).

- Being electrically isolated, FG acts as a storing electrode for the cell: charge injected into FG is maintained there.

- The gate dielectric between the transistor channel and the gate is an oxide, 8-10 nm and is called tunnel oxide.

- The dielectric separating FG/CG is formed by oxide-nitride-oxide.
Memory Array

- To achieve a high density of the device with a relative ease of manufacturing, cells have to be arranged in structured manner, normally in an array.
- Each cell can be specified by its row and column address.
- The device translates the linear address presented on I/O lines into a row-column address (decoding).
- Columns of cell all share metal 1 common drain connection called a bitline (BL).
- Rows of cells share common poly-2 gate connection called a wordline (WL).
- Bitlines and wordlines allow a means for placing the appropriate voltages on each cell individually or in groups as required by the operations.
NOR and NAND

- NOR flash name is related to the way cells are arranged in an array (a), through rows and columns in a NOR-like structure: if any memory cell is turned on by the corresponding word-line, the bit-line goes low.

- NOR flash requires 1 transistor for every contact (one drain contact common for two cells); the source electrode is common for all cells.

- Sacrificing random access, denser memory can be achieved by arranging 8/16/32 memory transistors in a series (b). This is typical for NAND flash organization.
Threshold Voltage

- In a neutral state, there is no conductive path between the drain and the source.

- When a positive charge is applied to the gate and the drain, a channel begins to form between the source and the drain.

- When the CG voltage is sufficiently large the channel is formed, and electrons flow from the drain to the source, creating a current, $I_{DS}$.

- The voltage at which the channel forms is called a threshold voltage of a transistor, $V_T$. 

---

12/10/2007
Programmed and Erased Cells: 0/1

- When a charge is stored on a FG, the threshold voltage increases; this increase is proportional to a stored charge.
- We define a logical state “1” from a microscopic point of view as no electron charges stored in the floating gate.
- Vice versa, the logical state “0” is defined by electron charge stored in the floating gate.

![Diagram of control gate and source-drain voltage relationships for programmed and erased states.](image-url)
Read Operation

- The cell has basically two threshold states, a programmed state ($V_{Tp}$) and an erased state ($V_{Te}$).

- The read voltage $V_{read}$ is selected in between $V_{Tp}$ and $V_{Re}$.

- If current flows at $V_{read}$, then the cell is read as erased (“1”).

- Thus the state “1” from macroscopic point of view is defined as a large (tens of microampere) reading current.

- The state “0” is defined as zero current.

- To read a value, we must determine the relative threshold voltage of a cell and compare it to a standard reference.
Sense Amplifiers & Reference Cells

- To read the value of a cell, we must determine the relative threshold voltage of this cell and compare it with a reference.

- For this, we need a means of measuring current and have reference cells for comparison.

- The reference cells dictate the boundaries where the cell is considered programmed or erased by relating the measured $V_T$ of a cell against appropriate ($V_{\text{read}}$, $V_{\text{program-verify}}$, $V_{\text{erase}}$) reference.

- This process is accomplished by sense amplifiers.
Program/Erase Operations

- How charge can be moved to and from a floating gate?

- This can be done by exploiting different physical effects:
  - The CHE mechanism, where electrons gain enough energy to pass the oxide-silicon energy barrier thanks to the electric field in the transistor channel between source and drain.
  - The Fowler-Nordheim electron tunneling mechanism, i.e., a quantum-mechanical tunnel induced by an electric field. Applying a strong (8-10 MV/cm) electric field across a thin oxide, it is possible to force a large electron tunneling current.
  - The photoelectric effect, where electrons gain enough energy to cross the barrier thanks to the interaction with a photon energy (which corresponds to UV radiation) larger than the barrier itself. This mechanism is the one originally used in EPROM to erase the entire device.

- A common NOR flash is programmed by CHE injection at the drain side and is erased by FN tunneling through the tunnel oxide from FG to the silicon surface.
Programming in the Array (1)

- To program an individual cell, its drain connection (bitline) must be of an intermediate potential and its poly 2 gate (wordline) must be *pulsed* at a high potential

*Programming is self-restricted in nature*
How to prevent other cells in an array from being programmed?

- A high drain bias must be used to cause significant current flow in a channel so that electrons gain sufficient energy.
- All we have to do is to hold the drains of other cells on the wordline at ground potential so that they do not program along with the selected cell. For example, if we program cell (0,2), the bitlines of cells (0,0) and (0,1) must be held at 0V (ground).
Erasing in the Array

- In order for FN tunneling effect to perform, a high negative current must be applied to the gate, and the drain is left floating with the source tied to a positive potential through an erase load.

- At this conditions, the cell X/Y will erase, but what about other cells which share the wordline?

- If they were all programmed, they also begin erase.

- In non-programmed cells, hole-electron pairs also will begin to separate from FG, and the cells become over-erased.

- We can only erase cells on an entire row basis.
- All cells in a row must be programmed before they can be erased.
- Different cells erase at different rate; APDEV read is performed on a cells after every erase pulse so that column leakage can be corrected.
Erase Algorithms and Reliability

- Methods to fix over-erased (2) and under-erased (3) cells

How Overcame Endurance Failures

Stage #1
- Pre-Program
- Verify
- Erase
- Verify

1K Cycle Endurance (No Overerase Cell Correction)

Stage #2
- Pre-Program
- Verify
- Erase
- Verify

100K Cycle Endurance (No Undererased Cell Correction)

Stage #3
- Pre-Program
- Erase
- APDE Verify
  - Yes
    - APDE
  - No
    - Erase Verify
      - Yes
        - Done

1M Cycle Endurance (Undererased Cell Correction)

12/10/2007
Reading the Array

- Each bitline ends up connected to a sense amplifier so that programming, erasing and reading operations can take place.

- To read an individual cell, the read conditions are placed on individual bitlines and wordlines; and the current is measured and compared to read reference cell.

- If the cell Row 0 Column 0 is programmed, its apparent $V_T$ has been raised, thus it has a low $I_{DS}$ current under normal read conditions (this $I_{DS}$ is dictated by the PGMV reference current).
Reading: Effect of Column Leakage

- If the cell at Row 1 Column 0 was over-erased (i.e., there is a positive charge on FG seen as a positive $V_{GS}$ potential), a sense amplifier sees the sum of currents entering bitline from all the cells in that column.

- If this sum is larger than RDV reference, the zero will be interpreted as a false one.

- Thus it is of vital importance to correct any residual column leakage both during and after erase operation.
Typical Array Organization: I/O

- A flash device has 8 or 16 I/O lines through which data can be programmed or read. This data is derived directly from sense amplifiers, so that bitlines carry the data to and from array.

- The individual metal 1 bitlines are grouped into I/O groups; a typical number is 128 bitlines per group → 1024 bitlines/sector.

  Translation from a linear address to X/Y address is called X/Y decoding.

- When voltage is placed on a wordline, all cells in a row see it. If we organized an array so that all 8 data bits of an input byte lay on the same wordline, we can read them at the same time.
Sector Array Organization

Hierarchical organization

- Read delay is proportional to the row’s length ⇒ row length must be limited
- Flexibility of erase operation is another parameter that defines memory organization. Erase is carried out in sectors. Sectors are arranged to form an array made up by vertical and horizontal strips, and can be organized by rows and columns
  - Larger sectors ⇒ faster erase
  - Smaller sectors ⇒ simpler programming and memory management
- The chosen organization affects row and column decoding strategy
- To confine negative effects of gate leakage and stresses, a hierarchical organization is used
Logic Architecture Inside Flash Device

- Large area of a standard flash device is occupied by circuitry devoted to implementing various functions as shown in a simplified NOR device block diagram.
High Density Flash Array
Flash Security Features Overview

- **Password Sector Protection**: 64-bit password access
- **Persistent Sector Protection**: In-system virus protection
- **Software Sector Protection**: Software commands to dynamically lock/unlock sectors
- **Standard Security**: 4-6 bus cycle write and erase sequences, lockout voltages, glitch protection, WE# etc.
- **All sectors, including boot sectors**

- **SecSi Sector**: Increasing security
- **WP# pin**: Hardware boot sector locking
- **256 byte OTP**: User programmable or factory programmed random number
- **Boot Sectors**: Hardware-enabled security
- **Software-enabled (VCC) security**: Hardware/Software-enabled (VCC) security
Basic Security Features

- Out-of memory map area used for unique device identification or other security purposes
  - Up to 256 bytes
  - Accessible only by internal FSM
  - One-time programmable by a special (secret) sequence of commands

- Mechanisms protecting from accidental code and data modification
  - No write command is accepted on power-on (noise)
  - The complete control over write operation can be ensured by a special PIN# combination; pins can be protected by packaging
  - No write cycles is accepted when VCC is less than lock-out voltage
  - To protect against glitches, very short pulses (less than 5ns) on pins do not initiate write cycles
Protection Against Malicious Operations

- **HW-based One Time Programmable (OTP) sectors**
  - Can be implemented in HW by setting WE# permanently low after the device has been programmed ➞ prevents entire flash from being re-programmed/erased
  - To add flexibility and prevent only some (e.g., boot code storage) sectors from erase/write a special Write Protect Pin (WP#) is used by manufactures to lock out specified sectors by setting WE# permanently low after programming once

- **Flexible low-level SW-based OTP**
  - A special OTP bit is associated with each sector
    - This bit can only be read by the embedded microcontroller or FSM
    - The OTP protection bits can be only set once by a special (secret) sequence of commands. Once the bit is set, it cannot be unset!
    - When bit is set, the sector can only be read (no access for write/erase operations)
Firmware/SW-based Sector Protection

- To provide better flexibility, the idea of OTP sector protection can be extended in many ways
- Additional re-programmable “soft” protection bit is associated with each sector
  - These bits can be set and re-set only by specially defined (secret) sequence of commands; providing low-level SW-based OTP mechanism
  - After having been set, the chosen combination of soft OTP bits can be “locked” by a special long sequence of commands which cannot be undone, thus “sealing” the access to sectors. This can be done at any stage in a chip life, but only once
- To provide conditional access to otherwise locked sectors, a password-protection mechanism can be implemented
  - A “locked” configuration can be unlocked by presenting a correct (64-128-bit) password. Password is stored outside flash memory address space; password is programmed by a special sequence of commands only once; after having been programmed it is “locked” by the OTP bit. It can be read only by a special command not accessible for applications.
HW & Software Sector Protection

Sectors are protected by Dynamic "OR" Persist. bit

Persist. Lock

PPB Bits

Persist.

Dynamic

U = Unlock

L = Lock

WP# Pin protects two outermost boot sectors

<table>
<thead>
<tr>
<th>Sectors (Size in Kbytes)</th>
<th>N</th>
<th>N+1</th>
<th>N+2</th>
<th>N+3</th>
<th>N+4</th>
<th>N+5</th>
<th>N+6</th>
<th>N+7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>Persist.</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
</tbody>
</table>

U = Unlock
L = Lock
Password Sector Protection

Password

Sectors are protected by Dynamic bit "OR" Persist bit

Password Level Protection
Requires 64-bit valid Password to unlock

U = Unlock
L = Lock

Sectors (Size in Kbytes)
- N
- N+1
- N+2
- N+3
- N+4
- N+5
- N+6
- N+7

Dynamic
- L
- L
- L
- L
- U
- L
- U

Persist
- L
- L
- L
- U
- U
- U
- U

1
Fault Generation (stress)

Apply combinations of strange environmental conditions

- Vcc
- Glitch
- Clock
- Temperature
- UV
- Light
- X-Rays
- ...

and bypass or infer secrets
Decapsulation

Although there are 3-4 metal layers, decapsulation is possible
Low Voltage

Write and erase operations are not possible
High Voltage: Read Operation

Read operation at high voltage results in occasional bit errors

- 20012000FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFxFFFFFFFFDF
- 20014000FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBF
High Voltage: Write Operation

Write operation at high voltage results in correct data being written. However, uncontrolled side-effects are produced (i.e., other cells are being programmed as well).

- Dump after written pattern with $V_{cc}=5.8V$
  - :20000000123FFFFFFFEF FF FF FF FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF F FFDA
    :20002000FFFF4567FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF32
  - :20040000FFFFF89ABFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF8A
  - :20060000FFFFFFFCDEFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFE2
    :20080000FFFFFFFFFFFFF9F FFFFFFFFFFFFFFFFFFFFFFFFFF9A
    :200A0000FFFFFFFFFFFF4567FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFB2
  - :200C0000FFFFFFFFFFFFF89ABFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFQOA
    :200E0000FFFFFFFFFFFFCDEFFFFFFFFFFFFFFFFFFFFFFFFFFFF62
    :20010000FFFFFFFFFFFFFFFFFFFFE0129F FFFFFFFFFFFFFFFFFFFFFFFF1 9
    :20012000FFFFFFFFFFFFFFFFFFFFFFFF4567FFFFFFFFFFFFFFFFFFFF31
  - :20014000FFFFFFFFFFFF9DBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF89ABFFFFFFFFFFFFFFFFFFFFE2B
    :20016000FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFCDEFFFFFFFFFFFFFFFFFE1
High Voltage: Erase Operation

Erase operation is successful
However some bits may remain non-erased

- Dump after erase with Vcc 5,8V
  - :20022000FFFFFFFFFFFFFFFFFFFFEEFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFED
  - 20024000FFFFFFFFFFFFFFFFFFFFEEFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFBE
  - :20026000FFFFFFFFFFFFFFFFFFFFEEFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEF9E
  - :20028000FFFFFFFFFFFFFFFFFFFFEEFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEFFEEF5E
High V: Erasing Protected Sector

Sector protection mechanism is robust
Attempt to erase protected sector results in error message
High V - Programming Protected Sector

- Sector protection mechanism is robust: attempt to write into a protected sector results in error message; no write is performed
Erase Protected Sector: High Temperature

Nor flash is resistant to temperature up to 200 °C; afterwards attempts to execute operations result in error message and flash becomes un-operational.
Glitches: Read Operation

- Glitches up to 20 ns have no effect on read operation
- Glitches 20 ns introduce errors
- The higher the voltage the more memory corruption
- Errors may be detected with error detection mechanism
Glitches: Write Operation

- Short glitches have no effect on write operations
- Glitches 100 ns may result in occasional uncontrolled bit error
- Glitches 150 ns result in aborting write operation and error message
- Glitches during write operation on sector protection bits result in error message. However, some unpredictable side-effects arise (e.g., while operation is reported aborted, the protection bit is still set)
Light Attacks

Laser light attacks are more dangerous when light is focused on address decoders, FSM’s, voltage regulators, etc. rather than on memory cells array.
Laser Attack: Read Operation

While it was possible to corrupt data during read operation, EDC mechanism is capable of detecting errors
Laser Attack: Write Operation

- The results of laser attacks during write operation are very mixed

- Depending on the spot where light has been flashed, the consequences can be either:
  - None (operation is performed correctly)
  - Data could be written to another (but similar) address

- To write data in a slightly more controlled way, one needs to know exact device configuration, glue logic, addressing scheme, internal details of implementation of write operation and one must induce a very "strong" error (long duration)

- Some devices in the same series are more resistant to laser attacks during write operation
MirrorBit Flash Memory for Smart Cards

• Latest Flash Technology for Smart Cards
  • 2 bits per cell
  • 90 nm MirrorBit™ Technology
  • next generations: 65 nm and 45 nm

• Submitted to evaluation lab

• Invasive attacks much more difficult on such small technology

• Far less error prone when stressed (FA)

• More resources required for successful invasive attacks
Systems on Flash: High Density Cards

- Next Generation Smart Cards
- High Density cards
  - MegaBytes of Flash memory
  - User data in (OR)NAND Flash (16, 64, 256 MB and more)
  - NO ROM, NO EEPROM
  - Operating system in CodeFlash (512 KB of NOR Flash)
  - Application data and keys in Emulated EEPROM (128, 256 KB NOR Flash)
  - RAM (24, 48, 64 KB)
  - **All this on a Single Die** (75mm$^2$)
  - MMC and USB high speed interfaces
  - On-the-fly encryption of ORNAND Flash
Trademark Attribution

Spansion, the Spansion Logo and combinations thereof are trademarks of Spansion LLC. Other product names used in this presentation are for identification purposes only and may be trademarks of their respective companies.