EEPROM and Flash Memory

Dr. Lynn Fuller

Webpage: http://people.rit.edu/lffeee
Microelectronic Engineering
Rochester Institute of Technology
82 Lomb Memorial Drive
Rochester, NY 14623-5604

Email: Lynn.Fuller@rit.edu
Department Webpage: http://www.rit.edu/kgcoe/microelectronic/
OUTLINE

Introduction
Tunneling Gate Oxide EEPROM Programming
Reading
Test Specification/Results
Fowler-Nordheim Tunneling
Process Variation
Tunnel Oxide Recipe
Flash memory: is a non-volatile electrically erasable programmable memory EEPROM. It is erased and programmed in large blocks making it faster than the early versions where the entire chip had to be erased and then programmed. The flash name is still used.

Early versions stored charge on a floating poly gate between a control gate and the channel by forcing electrons to tunnel through a very thin tunneling oxide. Higher voltages are applied across the gate oxides to cause tunneling and change the charge on the floating gate (write “1” or “0”). While read operations use lower control gate voltages that does not cause any tunneling.

~1970
INTRODUCTION

In certain applications, data must be electrically entered and erased from Read Only Memory (ROM). The procedure can involve the entire ROM sections or one memory cell at a time. From the various technologies available, the FLOTOX EEPROM (FLOating-gate Tunneling Oxide Electrically Erasable Programmable ROM, is shown.

![EEPROM Diagram]

This EEPROM cell has double polysilicon gates, with the top polysilicon as the control gate and the lower polysilicon as the floating gate. A thin tunneling oxide is formed above the drain in the FLOTOX Transistor.
Another form of the FLOTOX Transistor is shown below. The structure is simpler and smaller but is more difficult to manufacture because of the problems associated with diffusion of dopants from the gate poly through the tunnel oxide into the transistor channel region.
FLOATING POLY CHARGE STORAGE

The EEPROM is programmed by transferring electrons between the floating-gate and the substrate, through the tunneling oxide, by means of Fowler-Nordheim tunneling. There are two modes of programming the EEPROM: write and erase. First, in the write mode, the floating-gate is charged negatively by electrons that tunnel from the drain to the floating gate. The charging is done by applying a +15V voltage to the control gate and connecting both the drain and source to ground.

The negative charge stored on the floating gate has the effect of shifting the threshold voltage towards a more positive value. When the floating gate is charged, the normal +5V applied to the control gate during a read operation will not be sufficient for the transistor to conduct channel current. Only when the floating gate is uncharged, will the transistor be able to conduct with +5V on the control gate.
CMOS PROCESS MODIFICATIONS

The fabrication of a FLOTOX EEPROM involves a three modifications to the present (1980’s) CMOS process. An additional n+ drain implant is performed before the polysilicon layers are deposited. A thin 100 Å tunneling oxide is grown above the additional n+ implant region. A second polysilicon layer is deposited for the control gate.
### TUNNEL OXIDE RECIPE – 120Å

<table>
<thead>
<tr>
<th>Step</th>
<th>Gas Flow</th>
<th>Temperature</th>
<th>Time</th>
<th>Boat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Load Tube: N₂ @ 5 lpm</td>
<td>650 °C</td>
<td>?</td>
<td>Out</td>
</tr>
<tr>
<td>1</td>
<td>Push: N₂ @ 10 lpm</td>
<td>650 °C</td>
<td>15 min</td>
<td>In</td>
</tr>
<tr>
<td>2</td>
<td>Stabilization: N₂ @ 15 lpm</td>
<td>650 °C</td>
<td>15 min</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ramp Up: N₂ @ 10 lpm</td>
<td>650 to 950 °C</td>
<td>30 min</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>First Oxide: N₂ @ 10 lpm + O₂ @ 5 lpm</td>
<td>950 °C</td>
<td>15 min</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>First Anneal: N₂ @ 5 lpm</td>
<td>1050 °C</td>
<td>30 min</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Ramp Down: N₂ @ 5 lpm</td>
<td>950 °C</td>
<td>20 min</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2nd Oxide: N₂ @ 10 lpm + O₂ @ 5 lpm</td>
<td>950 °C</td>
<td>10 min</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2nd Anneal: N₂ @ 5 lpm</td>
<td>950 °C</td>
<td>30 min</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Ramp Down: N₂ @ 5 lpm</td>
<td>950 to 650 °C</td>
<td>60 min</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Pull: N₂ @ 5 lpm</td>
<td>650 °C</td>
<td>15 min</td>
<td>Out</td>
</tr>
</tbody>
</table>

Gas Flows Modified 11-20-14

need to verify oxide growth
Entire Test Chip

EEPROM Transistor
EEPROM plus Select
EEPROM Memory Array
Variable Programmable Resistor
Binary-weighted Variable Programmable Resistor
Resistors
Capacitors
EEPROM Transistor

Basic transistor with two layers of poly, tunnel oxide, and 3 connections.
EEPROM Transistor + Select

Basic transistor with two layers poly, tunnel oxide, and 4 connections.
EEPROM MEMORY ARRAY

5V

Control/Read

Control/Read

Control/Read

Control/Read

Word Line 0

Word Line 1

Word Line 2

Word Line 3

B^3  B^2  B^1  B^0

Bit Lines
Memory array of four-by-four EEPROM cells
A variable programmable resistor with equal 1000 ohm resistors in series.
A variable programmable resistor with binary-weighted resistors, 1k, 2k, 4k, 8k .. .. 64 k ohms
Programmed with 3 seconds of +25 volts on gate
Erase with 4 seconds of 25 volts on the Drain
NOR AND NAND FLASH

128G Flash using 10nm technology

512 MB Flash
NOR and NAND flash differ in two important ways: the connections of the individual memory cells are different, the interface provided for reading and writing the memory is different (NOR allows random-access for reading, NAND allows only page access). These two are linked by the design choices made in the development of NAND flash. A goal of NAND flash development was to reduce the chip area required to implement a given capacity of flash memory, and thereby to reduce cost per bit and increase maximum chip capacity so that flash memory could compete with magnetic storage devices like hard disks. NOR and NAND flash get their names from the structure of the interconnections between memory cells. In NOR flash, cells are connected in parallel to the bitlines, allowing cells to be read and programmed individually. The parallel connection of cells resembles the parallel connection of transistors in a CMOS NOR gate.
NOR AND NAND FLASH

In NAND flash, cells are connected in series, resembling a NAND gate. The series connections consume less space than parallel ones, reducing the cost of NAND flash. It does not, by itself, prevent NAND cells from being read and programmed individually. When NOR flash was developed, it was envisioned as a more economical and conveniently rewritable ROM than contemporary EPROM, EAROM, and EEPROM memories. Thus random-access reading circuitry was necessary. However, it was expected that NOR flash ROM would be read much more often than written, so the write circuitry included was fairly slow and could only erase in a block-wise fashion. On the other hand, applications that use flash as a replacement for disk drives do not require word-level write address, which would only add to the complexity and cost unnecessarily. Because of the series connection and removal of wordline contacts, a large grid of NAND flash memory cells will occupy perhaps only 60% of the area of equivalent NOR cells (assuming the same CMOS process resolution, e.g. 130nm, 90 nm, 65 nm).
NAND flash's designers realized that the area of a NAND chip, and thus the cost, could be further reduced by removing the external address and data bus circuitry. Instead, external devices could communicate with NAND flash via sequential-accessed command and data registers, which would internally retrieve and output the necessary data. This design choice made random-access of NAND flash memory impossible, but the goal of NAND flash was to replace hard disks, not to replace ROMs.

Multiple Amounts of Charge
Currently 3 - Levels
ISSCC: SanDisk set to show highest density NAND flash

Peter Clarke
2/22/2012 7:13 AM EST

The chip, set to be discussed in paper 25.8 in the non-volatile memory session of ISSCC, is a 128-Gbit monolithic device that stores 3-bits per memory cell, is the highest density IC ever produced. The chip also has a 3-bit per cell write performance of 18-Mbyte per second and a read throughput of 400-Mbits per second. The chip is a rectangle of silicon of 170 square millimeters area.
3D VERTICAL NAND

~2013

Grey Silicon
Blue Gate Oxide
with traps
Green Poly
Control Gate
3-D VERTICLE NAND FLASH

- NAND String
- Bit line (BL)
- Upper Selection Gate (USG)
- Control Gate (CG)
- Lower Selection Gate (LSG)
- Source Line (SL)
- Vertical SONOS FET
- Vertical FET
In memories, Samsung and a team from Western Digital (now the parent of Sandisk) and Toshiba will reveal separate 512 Gbit NAND flash designs. Both pack three bits per cell and achieve their density using 64 stacked layers. The WD/Toshiba part has a 3.88Gb/mm² areal density and sports a 20% faster read-sensing scheme to improve read throughput over current flash chips. The Samsung device uses a 1.2V I/O voltage supply to increase read bandwidth at lower power. The chips will serve the high end of a solid-state drive market valued at more than $25 billion in 2017.
REFERENCES

HOMEWORK - EEPROM

1.0 None