

IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays

Sponsor

Standards Technical Committee of the
IEEE Electron Devices Society

Approved June 27, 1991

IEEE Standards Board

Abstract: An introduction to the physics unique to this type of memory and an overview of typical array architectures are presented. The variations on the basic floating gate nonvolatile cell structure that have been used in commercially available devices are described. The various reliability considerations involved in these devices are explored. Retention and endurance failures and the interaction between endurance, retention, and standard semiconductor failure mechanisms in determining the device failure rate are covered. How to specify and perform engineering verification of retention of data stored in the arrays is described. Effects that limit the endurance of the arrays are discussed. The specification and engineering verification of endurance are described. The more common features incorporated into the arrays and methods for testing these complex products efficiently are addressed. The effects that various forms of ionizing radiation may have on floating gate arrays and approaches to test for these effects are covered. The use of floating gate cells in nonmemory applications is briefly considered.

Keywords: Array, floating gate, nonvolatile memory.

The Institute of Electrical and Electronics Engineers, Inc.
345 East 47th Street, New York, NY 10017-2394, USA

Copyright © 1991 by the
Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 1991
Printed in the United States of America

Library of Congress Cataloging in Publication Data
will be found on the following page.

*No part of this publication may be reproduced in any form,
in an electronic retrieval system or otherwise,
without the prior written permission of the publisher.*

IEEE Standards documents are developed within the Technical Committees of the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Board. Members of the committees serve voluntarily and without compensation. They are not necessarily members of the Institute. The standards developed within IEEE represent a consensus of the broad expertise on the subject within the Institute as well as those activities outside of IEEE that have expressed an interest in participating in the development of the standard.

Use of an IEEE Standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE Standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard. Every IEEE Standard is subjected to review at least every five years for revision or reaffirmation. When a document is more than five years old and has not been reaffirmed, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE Standard.

Comments for revision of IEEE Standards are welcome from any interested party, regardless of membership affiliation with IEEE. Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments.

Interpretations: Occasionally questions may arise regarding the meaning of portions of standards as they relate to specific applications. When the need for interpretations is brought to the attention of IEEE, the Institute will initiate action to prepare appropriate responses. Since IEEE Standards represent a consensus of all concerned interests, it is important to ensure that any interpretation has also received the concurrence of a balance of interests. For this reason IEEE and the members of its technical committees are not able to provide an instant response to interpretation requests except in those cases where the matter has previously received formal consideration.

Comments on standards and requests for interpretations should be addressed to:

Secretary, IEEE Standards Board
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
USA

IEEE Standards documents are adopted by the Institute of Electrical and Electronics Engineers without regard to whether their adoption may involve patents on articles, materials, or processes. Such adoption does not assume any liability to any patent owner, nor does it assume any obligation whatever to parties adopting the standards documents.

Library of Congress Cataloging in Publication Data

IEEE Standard definitions and characterization of floating gate semiconductor arrays / sponsor, Standards Technical Committee of the IEEE Electron Devices Society.

p. cm.

"IEEE Std 1005-1991."

"Approved June 27, 1991."

Includes bibliographical references.

ISBN 1-55937-136-6

1. Semiconductor storage devices. 2. Integrated circuits—Very large scale integration—Testing. I. IEEE Electron Devices Society. Standards Technical Committee. II. Title: Floating gate semiconductor arrays.

TK7895.M4I34 1991

621.3973—dc20

91-31636

CIP

Foreword

(This Foreword is not a part of IEEE Std 1005-1991, IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays.)

Over the last few years, floating gate memory arrays have become the overwhelming technology of choice for those applications requiring nonvolatile semiconductor memory whose contents may be altered in the system. However, because this technology raises issues heretofore not commonly addressed by users, there has been uncertainty as to how to best characterize and evaluate products incorporating it for various applications. A complicating factor has been the fact that products using this technology have been available from a number of commercial suppliers who have employed several variants of the basic technology and who have sometimes used inconsistent terminology to describe it.

This standard is intended for users or prospective users of floating gate technology. It provides an introduction to the technology, describes the variants in current use, addresses the salient issues and describes, for new users, tests designed to characterize and evaluate products in light of these issues. It also defines common terms in order to facilitate communication.

At the time that it prepared this standard, the Floating Gate Array Standards Committee had the following membership:

Richard L. Wiker, *Chairman*

J. E. Brewer
J. Caywood
M. Gill

S. Lai
N. Mielke
W. Obereiner

J. Schoepp
D. Sweetman
H. A. R. Wegener

The balloting committee that approved this document for submission to the IEEE Standards Board had the following membership:

J. E. Brewer
M. Briner
J. Caywood
R. Fedorak
V. Fong

M. Gill
Tr. Hoang
D. Kaputa
N. Mielke
J. Schoepp

D. Sweetman
M. Van Buskirk
H. A. R. Wegener
R. L. Wiker

When the IEEE Standards Board approved this standard on June 27, 1991, it had the following membership:

Marco W. Migliaro, *Chairman*

Donald C. Loughry, *Vice Chairman*

Andrew G. Salem, *Secretary*

Dennis Bodson
Paul L. Borrill
Clyde Camp
James M. Daly
Donald C. Fleckenstein
Jay Forster*
David F. Franklin
Ingrid Fromm

Thomas L. Hannan
Donald N. Heirman
Kenneth D. Hendrix
John W. Horch
Ben C. Johnson
Ivor N. Knight
Joseph L. Koepfinger*
Irving Kolodny
Michael A. Lawler

John E. May, Jr.
Lawrence V. McCall
Donald T. Michael*
Stig L. Nilsson
John L. Rankine
Ronald H. Reimer
Gary S. Robinson
Terrance R. Whittemore

*Member Emeritus

Adam Sicker
IEEE Standards Project Editor

IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays

1. Introduction

1.1 Overview. Section 1 contains an introduction to the physics unique to this type of memory and an overview of typical array architectures. It also defines the limitations of this standard.

Section 2 defines symbols and definitions used in the remaining sections.

Section 3 provides the reader with references that are necessary for implementation of this standard.

Section 4 describes the variations on the basic floating gate nonvolatile cell structure that have been used in commercially available devices.

Section 5 explores the various reliability considerations involved in these devices. It discusses retention and endurance failures and the interaction between endurance, retention, and standard semiconductor failure mechanisms in determining the device failure rate.

Section 6 defines retention of data stored in these arrays. It describes how to specify and perform engineering verification of retention. Some typical production screens and monitors are indicated.

Section 7 defines endurance of floating gate arrays. Effects that limit the endurance are discussed. The specification and engineering verification of endurance are described.

Section 8 deals with testing floating gate arrays. It describes the more common features incorporated into these arrays and how to efficiently test these complex products.

Section 9 treats the effects that various forms of ionizing radiation may have on floating gate arrays and on approaches to test for these effects.

Section 10 sketches briefly the use of floating gate cells in nonmemory applications. Competing electrically-alterable semiconductor nonvolatile memory technologies are also mentioned.

1.2 Basic Tunneling Physics. The fundamental challenge in designing an electronically-reprogrammable nonvolatile memory cell is the necessity of finding a controllable, reproducible effect that has a sufficient nonlinearity such that the cell can be written to a state at one voltage in about 1 ms, and that state can be maintained during read or storage at another voltage for more than 10 years ($\sim 3 \cdot 10^8$ s). One phenomenon that exhibits the necessary nonlinearity is Fowler-Nordheim tunneling, which was first described by Fowler and Nordheim in 1928 and subsequently named after them, see [B14].¹

The basic idea is illustrated in Fig 1.1. The energy difference between the conduction and valence bands in Si is about 1.1 eV. The energy difference between these bands in SiO₂ is approximately 9 eV. When the two materials are joined, the conduction band in SiO₂ is 3.25 eV above that in Si. The differences in valence band energies is even larger (>4 eV). Since the thermal energy of an electron averages only 0.025 eV at room temperature, the chances of an electron in silicon gaining enough thermal energy to surmount the barrier and enter the conduction band in SiO₂ is exceedingly small. This case is illustrated in Fig. 1.1(a).

Fowler and Nordheim (see [B14]) pointed out that in the presence of a high electric field, the energy bands will be distorted as illustrated in Fig 1.1(b). Under these conditions, there is a small but finite probability that an electron in the conduction band in the silicon will quantum mechanically tunnel through the energy barrier and emerge in the conduction band of the SiO₂, as illustrated in Fig 1.1(b).

Fowler-Nordheim emission was observed early in this century for the case of electrons being emitted from metals into vacuum. In 1969, Lenzlinger and Snow observed this phe-

¹The numbers in brackets, when preceded by the letter B, correspond to those in the bibliography in Section 11.

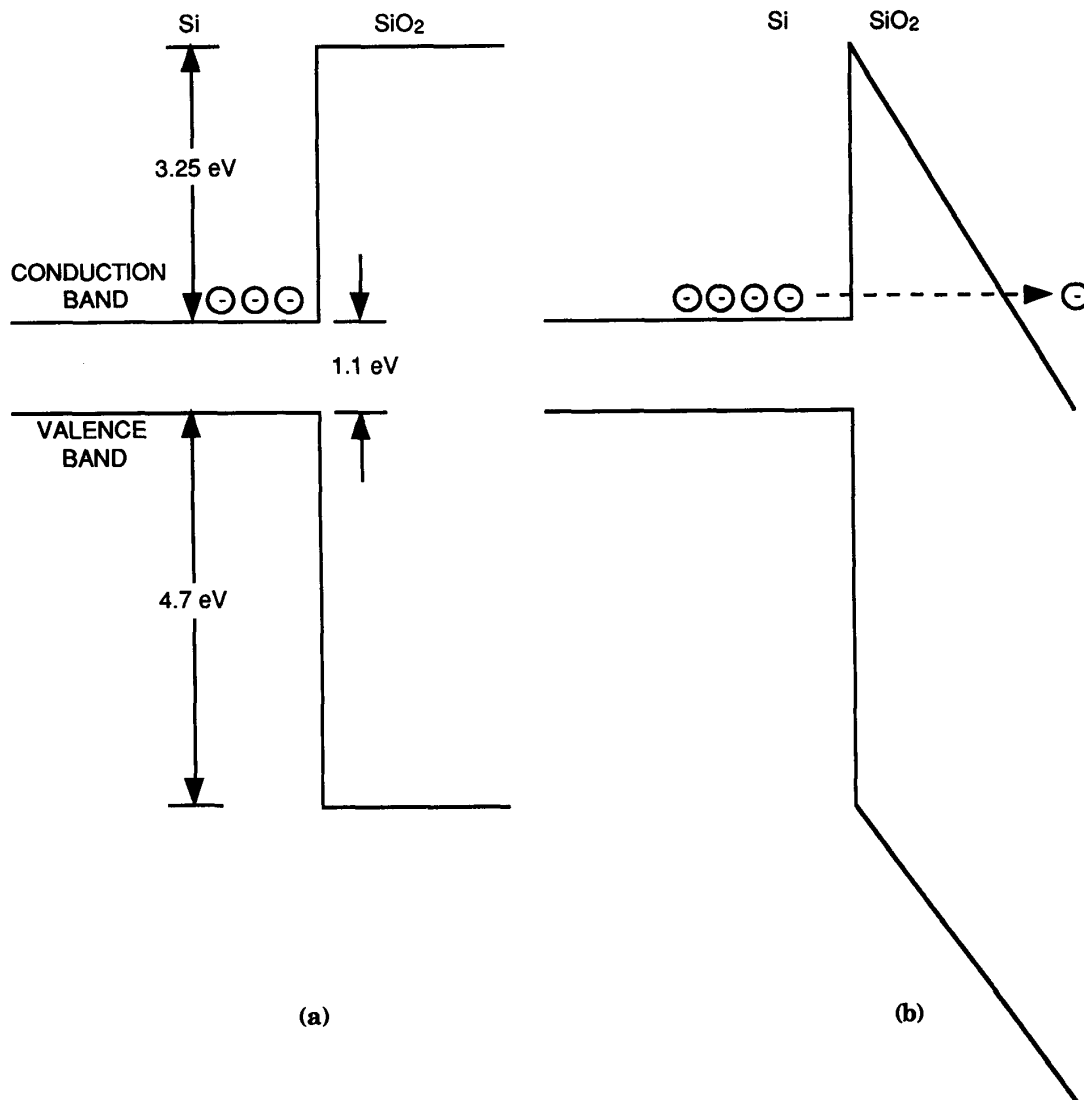


Fig 1.1
Energy Band Diagrams of the Si/SiO₂ System
in the Neutral State (a) and During the Write Operation (b)

nomenon for the Si-SiO₂ system, see [B25]. The Fowler-Nordheim current increases exponentially with applied field and becomes readily observable (i.e., $J \sim 10^{-6}$ A/cm²) for the Si-SiO₂ system when the field at the Si-SiO₂ interface is on the order of 10 mV/cm.

Local fields of this magnitude, at voltages practicable for use on high density integrated

circuits, have been obtained by either applying a voltage across a very thin (~ 100 Å) oxide layer grown on a smooth Si substrate or by applying a voltage across a somewhat thicker oxide (~ 500 – 1000 Å) grown on polycrystalline electrodes. In the latter case, the field enhancement arising from the positive curvature regions of the textured polycrystalline

surface results in tunneling at similar voltages as in the first case, see [B11]. There are several reports of practical memories being fabricated with both techniques, see [B10], [B14], [B17], [B18], and [B21]–[B24].

1.3 Introduction to Floating Gate Memory Arrays. The floating gate devices discussed in this standard program and erase via Fowler-Nordheim tunneling and may require that pulses of the order of 20 V be applied to the cell terminals in order to write to the cell. Moreover, reliability is enhanced if the pulse is shaped to reduce the instantaneous tunnel current spikes. Early memory arrays left to the user the task of supplying these pulses, just as early RAMs lacked onboard sense amplifiers. However, all modern commercial arrays of this type generate the necessary write pulses onboard the chip from a standard 5 V power supply under the control of external logic level signals.

Typical system cycle times are in the order of 50–500 ns. Floating gate memory arrays are made with read access times compatible with system timing requirements. In order to avoid overstressing the tunnel oxide, however, it is desirable to allow the writing to take place over a time on the order of a millisecond. To match the write times with the system cycle times, today's memory arrays control the long internal timings on chip in response to logic signals with timings that are comparable to read access times.

Many other functions intended to make these arrays easier to use, such as the ability to write many bytes per write cycle, the ability to detect the end of the internal write cycle, and the ability to inhibit writing when desired, have been incorporated into some devices. These are discussed in more detail in Section 8 along with methods for testing them.

In response to differing application requirements, various array organizations have been developed. One major difference in organization is whether the array is byte alterable, page (or block) alterable, or chip (or bulk) alterable. (Alterability is the ability to change the designated section from any pattern to any other pattern without changing the data in other sections.) Cells that may be used in byte alterable arrays are described in 4.3 and 4.4. Section 4.5 discusses a cell applicable to chip or block alterable arrays.

1.4 Restrictions. This standard is restricted to consideration of floating gate memory arrays that program and erase via Fowler-Nordheim tunneling. MNOS memory arrays are already covered in another IEEE standard (IEEE Std 641-1987 [B3]). Floating gate memories that program via hot electron injection through the gate oxide are specifically excluded. Hot electron emission, a different programming mechanism, leads to different device physics and reliability concerns. The relative inefficiency of the hot electron programming mechanism may require that a high-voltage pulse be externally supplied to the array. An external high-voltage supply may lead to different applications and different testing issues.

This standard also does not deal with the use of any floating gate memory cell in non-memory applications such as logic or trimming for analog functions.

Such excluded topics are discussed briefly in Section 10.

2. Definitions and Symbols

2.1 Definitions

address. A character or group of characters that identifies a register, a particular part of storage, or some other data source or destination (see JESD 100-82 [B6]).

AOQL. Average outgoing quality level. A statistical measure of outgoing quality.

array. A group of memory cells arranged in a pattern.

bit. (1) An abbreviation of binary digit. (2) A single occurrence of a character in a language employing exactly two distinct kinds of characters. (3) A unit of storage capacity. The capacity, in bits, of a storage device is the logarithm to the base two of number of possible states of the device (see IEEE Std 100-1988 [B1]).

bit-retention. The time interval from the writing of either state into a specified memory location to the earliest appearance of an incorrect state from such memory location as measured on specified output terminal(s).

block. A group of data that is contiguous in nature (see IEEE Std 100-1988 [B1]). As used in this standard, it is understood to imply a subset of an entire memory array. It may also be referred to as a sector.

block clear. Operation that sets a block of data within a memory to a common "1" state without affecting any other memory block.

NOTE: In the field of nonvolatile memories, *clear* conventionally means set to a "1" state. This usage is adhered to in this standard.

block select. An input terminal to which a signal must be applied in order to permit the device to read or write a block of data.

bulk erase. Operation of removing electrons from all of the bits of an array.

byte. A binary bit string operated on as a unit, usually eight bits long, and capable of holding one character in the local character set (see IEEE Std 100-1988 [B1]).

byte clear. Operation that sets all bits in an addressed byte of memory to a common "1" state.

cell. Same as *memory cell*.

charge pump. Circuitry used to create an on-chip voltage greater than the voltage available from the device power supply. This voltage is typically used for the write operation.

chip. A small unpackaged functional element made by subdividing a wafer of semiconductor material. Sometimes referred to as a die (see IEEE Std 100-1988 [B1]). Also used as a modifier to indicate an operation that applies to the entire chip, as in *chip enable* or *chip clear*.

chip clear. Operation that causes the writing of all memory elements to a common "1" state. Terminal to which the clear signal is applied (preferred usage).

chip erase. Same as *bulk erase*. May be referred to as *mass erase*.

data change. The event in which new data is written to a byte or word. This event may be used as a unit of endurance measurement.

data rewrite. The operation in which the array is written, whether measured in terms of data changes or program/erase cycles.

dc static tests. Voltage and current tests to determine the input levels, output levels, and power dissipation characteristics of a device.

differential sensing circuit. A circuit capable of detecting the differential characteristics of a multiple-sense-transistor memory cell.

dose rate. The time rate of deposition of ionizing radiation (see IEEE Std 581-1978 [B2]).

dynamic radiation test. Test of the instantaneous effects of radiation obtained by monitoring electrical and memory properties of interest continuously during and immediately after exposure to a transient radiation pulse (see IEEE Std. 518-1978 [B2]).

EEPROM (E²PROM). Electrically erasable programmable read only memory. A non-volatile, randomly-addressable storage device whose bits can be written into either state electrically. These devices are characterized by much slower write times than read times. The memory is intended for applications specifying a maximum number of write operations (cf., JESD 100-82 [B6]).

E²PROM. Same as *EEPROM*.

EEPROM redundancy. A circuit technique in which extra EEPROM cells are mapped into the memory array address space to replace defective EEPROM cells. The mapping may be controlled through EEPROM or other fuse techniques.

endurance. The measure of the ability of a nonvolatile memory device to meet its data sheet specifications as a function of accumulated nonvolatile data rewrites.

EPROM (UV-EPROM). Electrically programmable read only memory. A non-volatile, randomly-addressable semiconductor memory in which data is stored as charge on an electrically insulated gate, programming occurs via hot electron injection, and erasure is implemented by exposure to ultraviolet light.

erase. The operation of removing electrons from the floating gate of the memory cell.

erase margin. The minimum measured difference between the erased states and the sensing level for the array.

FAMOS™. Floating gate avalanche MOS transistor used in EPROMS.

FET. Field-Effect Transistor.

FLOTOX™. Floating gate tunnel oxide.

Fowler-Nordheim tunneling. A quantum mechanical effect in which electrons penetrate through a region of space in which they are energetically forbidden as a result of an externally applied electric field.

functional dynamic tests. Operation of the array at specified dynamic conditions to determine functional performance.

IC. Integrated circuit. A combination of interconnected circuit elements inseparably associated on or within a continuous substrate (IEEE Std 100-1988 [B1]).

IGFET. Insulated-gate field-effect transistor. A field-effect transistor having one or more gate electrodes that are electrically insulated from the channel (see JESD 77-81 [B5]).

inadvertent write protection. Circuitry used to prevent inadvertent write cycles that may occur when the device's control signals enter an indeterminate state during power on/off or noise transitions.

incremental total dose test. Test of permanent changes induced by radiation obtained by a comparison of characteristics before and after exposure to a given integrated total dose level.

integrated circuit. See *IC*.

internal reference voltage. Internally developed bias voltage.

latch-up. A state in which a low-impedance path results from and persists following an input, output, or supply overvoltage that triggers a parasitic four-layer bipolar structure (see JEDEC Standard No. 17 [B4]).

LET. Linear energy transfer. The amount of energy transferred from an ionized particle to the medium through which it passes per normalized unit of distance traveled. Usually expressed in units of MeV·cm²/mg.

LSI. Large-scale integration.

mass erase. Same as *bulk erase*.

memory array. Same as *array*.

memory cell. The combination of one or more single or merged transistors formed to provide a means of accessing, changing, and storing data.

memory capacity. Same as *storage capacity*.

memory location. Addressable portion of the memory that is capable of storing a binary signal.

memory organization. An arrangement of memory cells, either by geometrical arrangement in rows and columns, or by the organization of the data to be stored (see IEEE Std 641-1988 [26]).

MNOS. Metal-nitride-oxide semiconductor transistor (see IEEE Std 581-1978 [B2]).

nonvolatile memory. A memory in which the data content is retained when power is no longer supplied to it (see JESD 100-82 [B6]).

NOVRAM. Nonvolatile random access memory.

NVRAM. Nonvolatile random access memory. (Also known as "Shadow Memory.")

page. A section of the array that may be written to simultaneously. This usually refers to more than one byte.

powerup/down protection. See *inadvertent write protection*.

™ FAMOS is a trademark of Intel Corp.
™ FLOTOX is a trademark of Intel Corp.

program. The operation of injecting electrons onto the floating gate of the memory cell.

program margin. The minimum measured difference between the program states and the sensing level for the array.

program/erase cycle. The event of writing a memory cell from one state (e.g., erased) to the opposite state and back to the original state. This event may be used as a unit of endurance measurement.

radiation-induced data loss characteristics (data upset). Stored data integrity as a function of total dose or dose rate after a given test pattern has been written into the device.

rads in Si. Amount of radiation measured by its ionizing effect in silicon; one rad equals 100 erg of energy deposited in a gram of irradiated solid (see IEEE Std 100-1988 [B1]).

RAM. Random access memory. A memory that permits access to any of its address locations in any desired sequence with similar access time to each location (see JESD 100-82 [B6]).

NOTE: The term RAM, as commonly used, denotes a read/write memory with unlimited data rewrite capability and equal read and write times.

read circuitry. That part of the memory used in transferring the stored data from the memory section to external circuitry.

read disturb. The corruption of data induced by reading the memory.

retention time. The time period from storing data until the data can no longer be read correctly under all the conditions allowed by the device specification (cf., IEEE Std 641-1988 [B3]).

sector. See *block*.

sense amplifier voltage window. Sense amplifier sensitivity between differential state or reference voltage and sensed state.

SEU. Single-event upset. The loss of data caused by passage of a single ionizing particle through an array.

static dose rate test. Test of the permanent changes induced by radiation obtained by a comparison of characteristics before and after exposure at a given dose rate.

storage capacity. The amount of data that can be contained in a storage device (see IEEE Std 100-1988 [B1]).

storage cell. An elementary unit of storage (for example, a binary cell, a decimal cell) (see IEEE Std 100-1988 [B1]).

timing sequence. Sequence of enable, coding, and data pulses to permit writing or reading of information.

total dose. The total amount of ionizing radiation deposited in the active area of a device over a given period of time. The unit of measure for total dose most commonly used in the present context is rads (Si) (see IEEE Std 100-1988 [B1]).

UV-PROM. Same as *EPROM*.

window. The measured difference in voltage or current between the erased and programmed states.

word. A character string or bit string considered as an entity (see IEEE Std 100-1988 [B1]).

word clear. Operation that sets all bits of a word to a common logic "1" state.

word erase. The operation of removing the electrons from all bits in a word.

write. The operation by which a data state is entered into one or more memory cells. This may be a program or an erase operation.

write circuitry. Section of the memory used in the alteration of the stored data during the write operation.

2.2 Letter Symbols for Quantities

C_G	=	gate capacitance per unit area
I_{DS}	=	drain-source current
k	=	Boltzmann's constant
k'	=	gain factor

L	=	length of channel in the direction of current flow
N_{end}	=	endurance
T	=	temperature
t_{ret}	=	retention time
V_{DS}	=	drain-source voltage
V_{GS}	=	gate-source voltage
V_{T}	=	threshold voltage
w	=	width of the channel
μ	=	carrier mobility
ϕ_{f}	=	Fermi potential

2.3 Letter Symbols for Units of Measurement

k	=	kilo (10^3)
ns	=	nanosecond (10^{-9} s)
μs	=	microsecond (10^{-6} s)
ms	=	millisecond (10^{-3} s)
M	=	mega (10^6)
V	=	volt
eV	=	electron volt ($1.602 \cdot 10^{-19}$ J)
$\text{\AA}$	=	angstroms (10^{-10} m)
nm	=	nanometer (10^{-9} m)
rad	=	100 erg/gm

3. References

- [1] JEDEC Standard No. 21, Configurations for Solid State Memory.²
- [2] MIL-STD-883, Test Methods and Procedures for Microelectronics.³

4. Basic Floating Gate Cell Types

4.1 Background. The floating gate FET structure was first proposed by Kahng and Sze in 1967. The first commercial realization of a floating gate memory device was the floating-gate avalanche-injection MOS memory (FAMOS) device developed by D. Frohman-Bentchkowsky in the early 70's. Programming of the FAMOS device was by avalanche breakdown of the drain junction. Hot elec-

trons generated by the process were transported to the floating gate. Hot electrons could overcome the oxide barrier with a relatively low applied voltage. To erase the FAMOS memory, ultraviolet light could be used to excite electrons over the oxide barrier. From the early 70's, FAMOS devices evolved into a generic memory product called electrically programmable read only memory (EPROM), which uses channel hot electron for programming.

The obvious drawback of the EPROM was that it required ultraviolet (UV) light for erasing. This required a quartz window in the package and the removal of the memory component from the circuit board to be erased by a special UV eraser. Furthermore, a high-voltage (12–21 V), high-current supply (> 10 mA) in addition to 5 V was required for programming. Consequently, instead of programming in circuit, programmers were developed for the exclusive purpose of programming EPROMs. Ideally, a nonvolatile memory should be erasable electrically and should not require an extra power supply, so that programming can be done in circuit with existing power supplies. Floating gate EEPROMs satisfy these requirements.

4.2 General Description. The basic floating gate electrically erasable programmable read only memories (EEPROMs) are very similar to EPROMs. Insulated from the channel region of the transistor is a poly silicon field plate usually doped to degenerate N+ by phosphorus. This field plate is completely insulated from all other electrodes and silicon by high-quality insulators (usually silicon dioxide) forming a floating gate. Typically, there are one or more electrodes that are coupled capacitively to the floating gate. Electron conduction of the transistor channel can be controlled by the charge stored on the floating gate. When there is gain in negative charge, the threshold of the transistor is increased in the positive voltage direction. When there is a gain in positive charge on the floating gate, the threshold voltage is increased in the negative direction. By applying a fixed reference voltage to the top control gate, the charge state of the floating gate can be sensed by the channel current, thereby giving the memory states.

The major difference between an EPROM and an EEPROM is that in the EEPROM there

² JEDEC publications are available from JEDEC, 2001 I Street NW, Washington, DC 20006, USA.

³ MIL publications are available from the Director, U.S. Navy Publications and Printing Service, Eastern Division, 700 Robbins Avenue, Philadelphia, PA 19111, (215) 697-2179.

are electrical elements in the structure that allow the transport of electrons to and from the floating gate under high voltages, but that maintain the charge states at normal operating voltages. The transport process is Fowler-Nordheim (FN) tunneling from planar structures or enhanced FN tunneling from non-planar structures (e.g., textured poly). The process requires a very high electric field and, in most cases, the tunneling process is bidirectional. Memory cells are designed in such a way as to couple the necessary high voltages for charge transport across the tunneling element over intervals of milliseconds. The transport of electrons results in a change of potential in the floating gate, which reduces the electric field for tunneling. Therefore, the electron transport process is a self-limiting process. Because the floating gate is conductive, it is possible to have charge transported in a small area away from the channel region. This gives flexibility in the design and layout of the cell, see [B24].

It is relatively easy to design a single memory cell that has two memory states. It is a very different problem to put the cell in an array and have the cell program and erase in a controlled fashion with no disturbance of other cells. Applying voltage in one way to program a cell in the array would usually mean that another cell would be erased in the array. Special inhibit conditions must be designed into the cell. Several solutions that allow alteration of a single byte are described in 4.3 and 4.4. An alternative that abandons byte alterability in favor of a smaller cell size is described in 4.5.

4.3 Planar Charge Transfer Cell

4.3.1 Tunnel Oxide Defined Lithographically in Drain Extension. The generic name for this memory cell is FLOTOX (floating gate tunnel oxide.), but there are variations of the basic cell that have been reported and used for EEPROM products, see [B16], [B21], and [B31].

4.3.1.1 Charge Transfer. The basic structure of the FLOTOX memory cell is shown in Fig 4.1. The tunnel dielectric (which can be oxide or oxynitride, typically less than 12 nm thick) is grown over an area defined lithographically in the drain region (or an extension of the drain region that may have lighter doping). Under the high-voltage condition, the memory cell is equivalent to a number of ca-

pacitors with the transistor function playing only a relatively minor part in the operation. Charging of the floating gate is achieved by grounding source and drain and applying a high voltage (typically 16-20 V) to the top gate. The cell is designed such that a large fraction (typically >60%) of the applied voltage is coupled across the tunnel dielectric, resulting in transport of electrons from drain to floating gate. Discharge is achieved by grounding the top gate, floating the source, and applying a high voltage to the drain. Again, most of the applied voltage is coupled across the tunnel dielectric (typically >70%), but the field is reversed. Electrons are removed from the floating gate to the drain. The source is floated so that there is no continuous current path. This is an important factor when an internal charge pump is used to generate the high voltage from the 5 V supply.

4.3.1.2 Inhibit Conditions. If a single transistor memory cell is placed in a typical array with drains connected to metal columns and gates connected to common polysilicon word lines, the programming of a cell (with word line grounded) will mean that high voltage is applied to ALL the drains in a common column. Programming can be inhibited in nonselected cells by taking unselected word lines to a high voltage, but unselected cells along the same word line may get charged (defined as electrons on the floating gate). In order to get away from the disturb conditions, in the minimum case, a select transistor is needed to isolate the drain from the column line. In other memory cell designs, there may be more than 1 select transistor. The select transistor will be off for rows that are not selected. The floating gate transistor is only a memory element used for storing the memory state.

4.3.1.3 Reading. During reading, the select transistor serves another purpose. It is possible for the cell to acquire a negative threshold, giving a depletion memory transistor. In an array, this would cause columns to fail. The select transistor will isolate the cells. Reading is accomplished by applying some selected voltage to the gate of the select transistor for memory selection in the row, applying a column voltage on the drain, and sensing the voltage on the floating gate. Electrons on the gate of an *N*-channel transistor will tend to turn it off. A positive gate

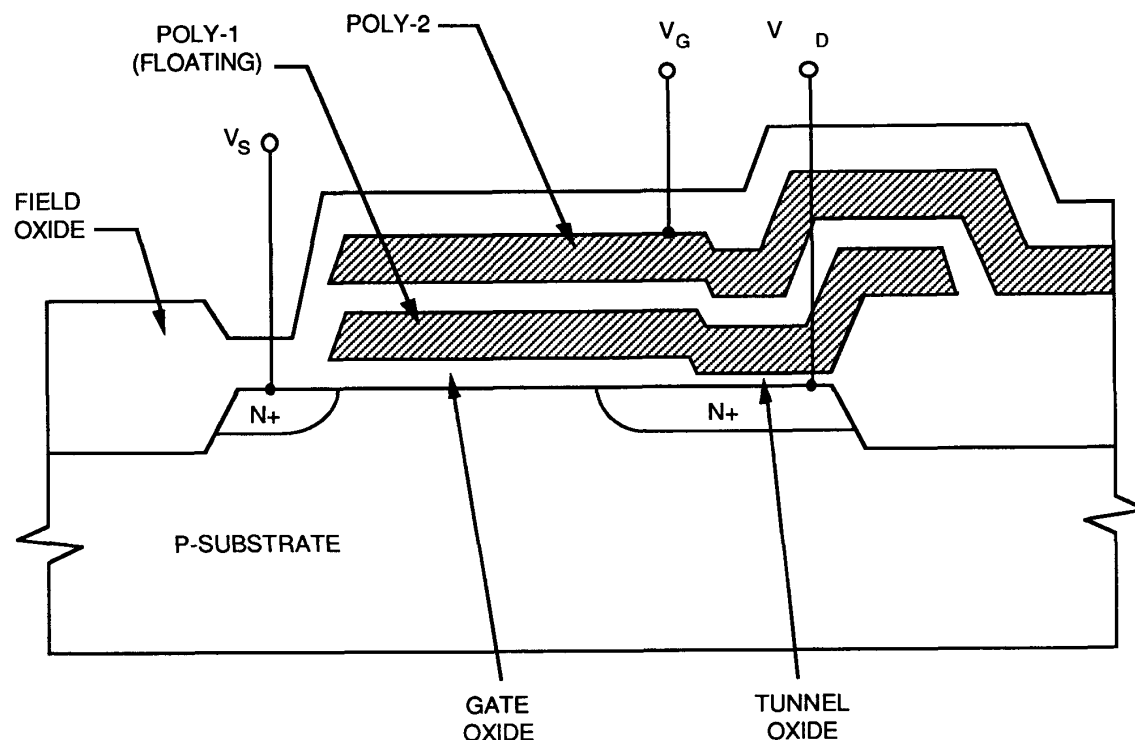


Fig 4.1
Cross-Section View Illustrating the FLOTOX Device Structure

charge will tend to turn it on. Thus, a difference in gate charge of the floating gate memory transistor causes a difference in drain current that can be sensed by the sense amplifier. Because the drain is coupled to the floating gate through the tunnel dielectric, drain coupling is relatively high. A drain voltage increase will decrease the threshold voltage of the memory transistor.

4.3.2 Tunnel Oxide as the Floating Gate Channel Oxide With Drain Overlap. The generic name used for this memory cell is FETMOS (floating gate electron tunneling MOS), see [B22].

4.3.2.1 Charge Transfer. The basic concept of the FETMOS cell is similar to FLOTOX, but instead of a small tunnel dielectric area defined lithographically in the drain, the cell has a tunnel dielectric as the dielectric underneath the floating gate in the channel region instead of a much thicker oxide.

During charging, a high voltage is applied to the top gate. Electrons tunnel from the channel, as well as the S/D region, into the floating gate. During discharging, the top gate is grounded, a fixed low voltage is applied to the source (or floated), and a high voltage is applied to the drain. Tunneling now occurs over the region where the floating gate overlaps the drain region, which is very small. With the high capacitance resulting from the tunnel oxide in the channel, the coupling ratio during charging is usually low; whereas because of the small drain/gate overlap, the programming coupling ratio is very high.

4.3.2.2 Inhibit Conditions. The inhibit conditions are identical to that of FLOTOX in 4.3.1.2.

4.3.2.3 Reading. Reading of the FETMOS cell is identical to reading of the FLOTOX cell. With the very thin tunnel dielectric, the sense transistor usually has high gain and

The cell is designed to have most of the applied voltage coupled across the tunnel/gate dielectric capacitor. This results in the transport of electrons onto the floating gate. For erasure, the top gate is grounded and the source, drain, and well are brought to high voltage. Because the source and drain are brought to high voltage, the channel also will be at high voltage. Most of the applied high voltage will couple across the tunnel/gate dielectric, resulting in the removal of electrons. The erase coupling ratio is the same as for programming because of the symmetry of the cell. This is different from the FLOTOX and the FETMOS cell where there is significant difference between program and erase coupling ratios.

4.3.3.1 Charge Transfer. The basic structure of this cell is shown in Fig 4.2, see [B27]. The tunnel dielectric is defined lithographically as a small area surrounded by thick oxide in the channel region (N channel in the present case). In order for this cell to work, it must reside in a well (P well in present case). Programming is achieved by applying a high voltage to the gate with the channel brought to ground potential by grounding the drain with the source floating.

The cell is designed to have most of the applied voltage coupled across the tunnel/gate dielectric capacitor. This results in the transport of electrons onto the floating gate. For erasure, the top gate is grounded and the source, drain, and well are brought to high voltage. Because the source and drain are brought to high voltage, the channel also will be at high voltage. Most of the applied high voltage will couple across the tunnel/gate dielectric, resulting in the removal of electrons. The erase coupling ratio is the same as for programming because of the symmetry of the cell. This is different from the FLOTOX and the FETMOS cell where there is significant difference between program and erase coupling ratios.

A cross-sectional diagram of a 1T1C1D1 floating gate device. The structure consists of several layers: an N-SUBSTRATE at the base, followed by a P-WELL layer. Within the P-WELL, there are N+ regions. Above the P-WELL is a layer of FIELD OXIDE. A central structure is formed by a TUNNEL OXIDE layer, a POLYSILICON FLOATING GATE, and a GATE OXIDE layer. This central structure is flanked by ALUMINUM CELL-SELECTION GATE regions. The entire structure is covered by an OXIDE/NITRIDE DIELECTRIC layer. Arrows point from the labels to the corresponding layers in the diagram.