Chapter 1

G. Groeseneken,
H. E. Maes,
J. Van Houdt,
and J. S. Witters

Basics of Nonvolatile Semiconductor Memory Devices

1.0. INTRODUCTION

Since the very first days of the mid-1960s, when the potential of metal–oxide semiconductor (MOS) technology to realize semiconductor memories with superior density and performance than would ever be achievable with the then commonly used magnetic core memories became known, chip makers have thought of solutions to overcome the main drawback of the MOS memory concept, that is, its intrinsic volatility. The first sound solutions to this problem, with applicability beyond the mere read-only memory (ROM) function, were the floating gate concept [1.1] and the metal–nitride–oxide–semiconductor (MNOS) memory device [1.2], both of which were proposed in 1967. A 1 Kbit UV-erasable programmable read-only memory (PROM) (EPROM) part, based on the floating gate concept, became readily available in 1971, shortly after 1 Kbit random access memories (RAM) came on the market.

The ultimate solution—a genuine nonvolatile RAM that retains data without external power, can be read from or programmed like a static or dynamic RAM, and still achieve high-speed, high-density, and low-power consumption at an acceptable cost—remains unfeasible to this day. Yet tremendous progress has been made over the years in realizing the "alternative best" idea of a reliable, high-density, user-friendly reprogrammable ROM memory. During the last decade, these reprogrammable memories have constituted an almost steady 10% of the total semiconductor memory market. This can be seen from Fig. 1.1, which shows the increase in the

Petitioner Micron Ex-1024, 0001

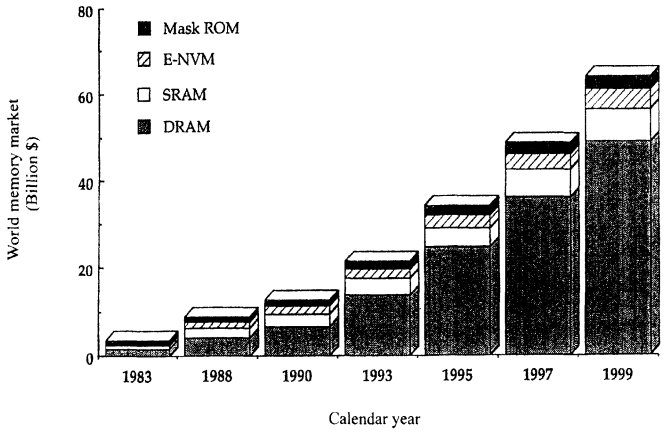


Figure 1.1 The increase of the world memory market during the last decade, the forecast for 1997 and 1999, and the share in this market of the different major classes of memories.

world memory market during the last decade, the forecast for 1997 and 1999, and the share of this market for the different major classes of memories. Until 1992, this 10% share came almost entirely from the least sophisticated and least functional version of this class of memories—that is, EPROMs, which do not allow in-system reprogrammability and are used mainly for standard program storage.

Reprogrammable nonvolatile memories can be subdivided into the following classes:

1. UV-erasable EPROM and one-time programmable (OTP) devices.
2. EEPROM memories, which can be further subdivided into full-feature electrically erasable programmable read-only memory (FF-EEPROMs) and Flash EEPROMs.
3. Nonvolatile RAM (NOVRAM), which combines the nonvolatility of EEPROM with the ease of use and fast programming characteristics of static RAM.
4. Ferroelectric RAM (FRAM).

Adding in-system reprogrammability to PROM memories (leading to FF-EEPROMs and to Flash EEPROMs), however, yields increased system flexibility and opens a broad new range of applications such as intelligent controllers; self-adaptive, reconfiguring, and remotely adjustable systems; programmable/adaptable logic; artificial intelligence; and numerous others [1.3]. The term *Flash* refers to the fact that the contents of the whole memory array, or of a memory block (sector), is erased in one step.

Petitioner Micron Ex-1024, 0002

In 1983, 16 Kbit EEPROMs based on both the MNOS [1.4] and the floating gate concept [1.5] were introduced, many analysts projected that EEPROMs would grow into a high-volume market and gradually even replace EPROM as the standard program storage medium in microprocessor-controlled systems. Figure 1.2 shows the actual and projected growth of the EPROM, FF-EEPROM, and Flash EEPROM markets over a 16-year period. In 1984, it was forecast that the EEPROM market would really start to take off around 1985, with projected global sales on the order of $2.5 billion by 1988. It is clear from Fig. 1.2 that this predicted significant increase in the EEPROM market was delayed by more than six years. Moreover, the increase has not been as strong as it was then anticipated. The growth of the EPROM market has, however, slowed down and recently reversed. In 1994, and certainly 1995, the EEPROM market surpassed that of EPROM. Figure 1.2 also shows the emerging domination of Flash EEPROMs for the programmable ROMs for the next generations. They are at present the fastest growing MOS memory segment, and it is expected that they will eventually constitute the third largest segment behind DRAM and SRAM.

The EEPROM market did not grow as previously predicted because of their high cost per bit as compared to EPROM, the lack of large-scale applications for full-featured EEPROMs, and the poorly understood reliability of these components. The reliability issues of EEPROMs and Flash memories have, however, recently been thoroughly investigated and are now much better known and documented. In addition, recent lower pricing and increased performance of Flash memories have stirred new interest in these parts. New large-scale applications are emerging

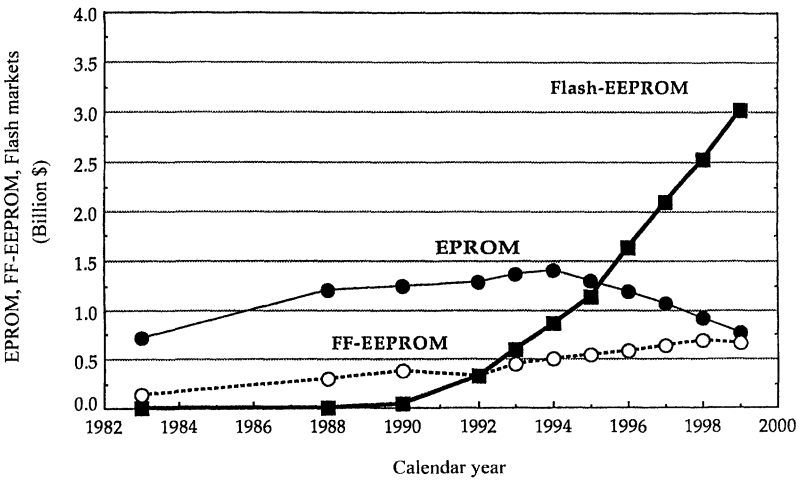


Figure 1.2 Comparison of the actual global sales (up to 1995) of EPROM, EEPROM, and Flash EEPROM, and forecasted market evolution (after 1995).

Petitioner Micron Ex-1024, 0003

(i.e., memory cards, small, compact, and portable memories). These Flash EEPROMs were themselves developed in the late 1980s and introduced around 1990 when manufacturers were searching for nonvolatile devices that were still electrically erasable, but that could become nearly as cost effective as EPROMs. They combined the best concepts of EPROM and traditional EEPROM into a single-transistor Flash EEPROM.

This chapter presents the basic concepts and physics of operation of all the nonvolatile semiconductor memory types and classes listed previously. It is intended as a solid introduction to all the following chapters in this book. We will first present the basic principles and history of nonvolatile memory (NVM) devices in Section 1.1. The different programming mechanisms used in the various devices are discussed in Section 1.2, and the basic NVM memory products are presented in Section 1.3. A review of the major NVM devices in use today is given in Section 1.4 and is concluded by a rather general comparison of the different types of memory concepts. The basic equations and models specific to these NVM devices are presented in Section 1.5, which is followed by a detailed discussion in Section 1.6 of the NVM device characteristics and reliability issues for the different types of devices. Finally, Section 1.8 discusses some specific radiation aspects of NVM devices.

1.1. BASIC PRINCIPLES AND HISTORY OF NVM DEVICES

1.1.1 Basic Operating Principle

The basic operating principle of nonvolatile semiconductor memory devices is the storage of charges in the gate insulator of a MOSFET, as illustrated in Fig. 1.3. If one can store charges in the insulator of a MOSFET, the threshold voltage of the transistor can be modified to switch between two distinct values, conventionally

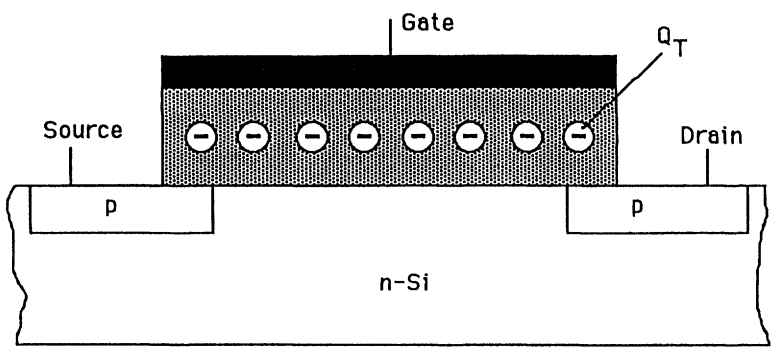


Figure 1.3 Basic operating principle of nonvolatile semiconductor memory: the storage of charges in the gate insulator of a MOSFET.

Petitioner Micron Ex-1024, 0004

defined as the "0" or erased state and the "1" or written (programmed) state, as illustrated in Fig. 1.4.

From the basic theory of the MOS transistor, the threshold voltage is given by

$$V_{TH} = 2\phi_F + \phi_{ms} - \frac{Q_I}{C_I} - \frac{Q_D}{C_I} - \frac{Q_T}{\epsilon_I} d_I \tag{1.1}$$

where ϕ_{MS} = the work function difference between the gate and the bulk material
ϕ_F = the Fermipotential of the semiconductor at the surface
Q_I = the fixed charge at the silicon/insulator interface
Q_D = the charge in the silicon depletion layer
Q_T = the charge stored in the gate insulator at a distance d_I from the gate
C_I = the capacitance of the insulator layer
ϵ_I = the dielectric constant of the insulator

Thus, the threshold voltage shift, caused by the storage of the charge Q_T is given by

$$\Delta V_{TH} = -\frac{Q_T}{\epsilon_I} d_I \tag{1.2}$$

The information content of the device is detected by applying a gate voltage V_{read} with a value between the two possible threshold voltages. In one state, the transistor is conducting current, while, in the other, the transistor is cut off. When the power supply is interrupted, the charge should, of course, remain stored in the gate insulator in order to provide a nonvolatile device.

The storage of charges in the gate insulator of a MOSFET can be realized in two ways, which has led to the subdivision of nonvolatile semiconductor memory devices into two main classes.

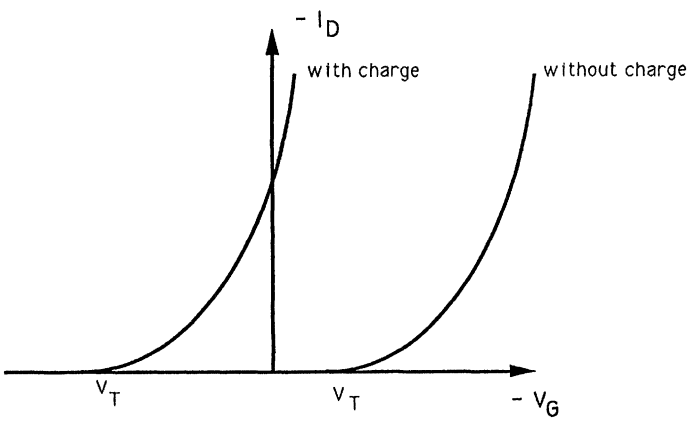


Figure 1.4 Influence of charge in the gate dielectric on the threshold of a p-channel transistor.

Petitioner Micron Ex-1024, 0005

The first class of devices is based on the storage of charge on a conducting or semiconducting layer that is completely surrounded by a dielectric, usually thermal oxide, as shown on Fig. 1.5*a*. Since this layer acts as a completely electrically isolated gate, this type of device is commonly referred to as a floating gate device [1.6, 1.7].

In the second class of devices, the charge is stored in discrete trapping centers of an appropriate dielectric layer. These devices are, therefore, usually referred to as charge-trapping devices. The most successful device in this category is the MNOS device (metal–nitride–oxide–semiconductor) structure [1.2, 1.8], in which the insulator consists of a silicon nitride layer on top of a very thin silicon oxide layer, as shown in Fig. 1.5*b*. Other possibilities, such as Al_2O_3 (MAOS) and Ta_2O_5 (MTOS) [1.9, 1.10], have never been successfully exploited.

Further details on the cell types, features, and new developments, as well as a comparison of these classes of nonvolatile memory cells, are given in Section 1.4.

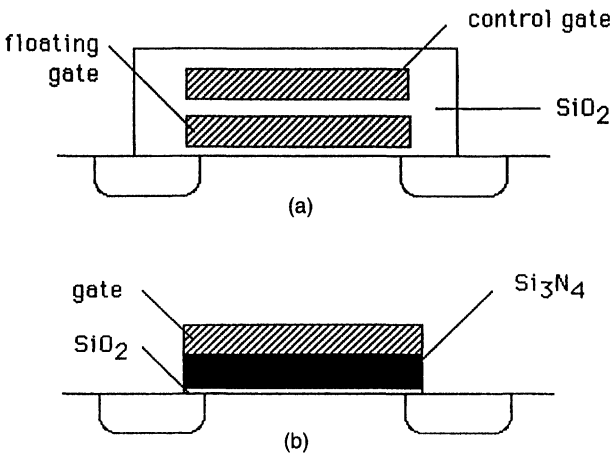


Figure 1.5 Two classes of nonvolatile semiconductor memory devices: (*a*) floating gate devices; (*b*) charge-trapping devices (MNOS device).

1.1.2 Short Historical Review

The idea of using a floating gate device to obtain a nonvolatile memory device was suggested for the first time in 1967 by D. Kahng and S. M. Sze [1.1]. This was also the first time that the possibility of nonvolatile MOS memory devices was recognized.

The memory transistor that they proposed started from a basic MOS structure, where the gate structure is replaced by a layered structure of a thin oxide I_1, a floating but conducting metal layer M_1, a thick oxide I_2, and an external metal gate M_2, as shown in Fig. 1.6. This device is referred to as the MIMIS (metal–insulator–metal–insulator–semiconductor) cell. The first dielectric I_1 has to be extremely thin in order to obtain a sufficiently high electric field to allow tunneling of electrons toward the floating gate. These electrons are then "captured" in the conduction band of the floating gate M_1, if the dielectric I_2 is thick enough to prevent discharging. When the gate voltage is removed, the field in I_1 is too small to allow

Petitioner Micron Ex-1024, 0006

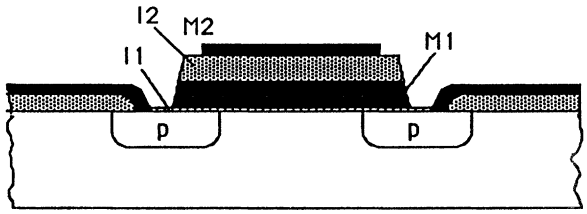


Figure 1.6 Introduction of the floating gate principle: the MIMIS structure, introduced by Kahng and Sze [1.1]. Writing and erasing the device is performed by direct tunneling of electrons through the thin oxide I1.

backtunneling. The injection mechanism to bring electrons to the floating gate is direct tunneling. To discharge the floating gate, a negative voltage pulse is applied at M_2, removing the electrons from the floating gate by the same direct tunneling mechanism.

The direct tunneling programming mechanism imposes the use of very thin oxide layers (< 5 nm), which are difficult to achieve without defects. Any pinhole in I_1 will cause all the charge stored on M_1 to leak off. Because of this technological constraint, the MIMIS cell could not be reliably built at that time. Therefore, the importance of this device is merely historical, not only because it introduced the basic concept of nonvolatile memory devices in general, but also because it contained several essential concepts that have led to the development of both classes of nonvolatile memory devices: the direct tunneling concept has been used in charge-trapping devices, while the floating gate concept has led to a whole range of floating gate memory types.

In order to solve the technological constraint of the MIMIS cell, two types of improvements are possible: (1) replace the conducting layer on top of I_1 by a dielectric layer without losing the capture possibilities, which is actually the approach utilized in charge-trapping devices, or (2) increase the thickness of the tunneling dielectric I_1, which implies the need for other injection mechanisms.

The first solution was used in the MNOS cell, introduced in 1967 by Wegener et al. [1.2], almost simultaneously with the MIMIS cell. In the MNOS cell, the M_1 and I_2 layers are replaced by a nitride layer, as shown in Fig. 1.5*b*, which contains a lot of trapping centers in which holes and electrons can be captured. These traps fulfill the storage function of M_1 with the important difference that an eventual pinhole in the thin tunneling oxide (I_1) will not lead to a complete discharge of the cell since the individual traps are isolated from each other by the nitride. The device is programmed by applying a high voltage to the gate such that electrons tunnel from the silicon conduction band to the nitride conduction band and are then trapped in the nitride traps. This results in a positive threshold voltage shift. Erasing is achieved by applying a high negative voltage to the gate, so that holes tunnel from the silicon valence band into the nitride traps, resulting in a negative threshold voltage shift. The MNOS device has the intrinsic advantage that both programming and erasing operations can be performed electrically. The concept has been used widely in several kinds of applications, specifically in a class of memory products called EEPROM, which are further discussed in Section 1.3. At present, however, this class of memory cells is used only for military and applications that must be resistant to radiation, and only marginally in commercial high-density nonvolatile memory circuits.

Petitioner Micron Ex-1024, 0007

The second solution has been used in a wide range of nonvolatile memory devices. The first operating floating gate device, shown in Fig. 1.7, was introduced in 1971 by Frohman–Bentchkowsky and is known as the *F*loating gate *A*valanche injection *MOS* (FAMOS) device [1.6–1.7, 1.11–1.12]. In the original p-channel FAMOS cell, a polysilicon floating gate is completely surrounded by a thick ($\approx$ 100 nm) oxide. Here, the problem of possible shorting paths is obviated, but, at the same time, direct tunneling is excluded as the programming mechanism. In the FAMOS cell, the charging mechanism is based on injection of highly energetic electrons from an avalanche plasma in the drain region underneath the gate. This avalanche plasma is created by applying a high negative voltage (> 30 V) at the drain. The injected electrons are drifted toward the floating gate by the positive field in the oxide induced by capacitive-coupling between the floating gate and the drain. The FAMOS device has found wide applications and was the first cell to reach volume manufacturing levels comparable to other semiconductor memory types. FAMOS devices have evolved into a class of memory products called EPROM, and are further discussed in Section 1.3. The original FAMOS device, however, had several drawbacks, with the inefficiency of the programming process as the most salient one. In addition, no mechanism for electrical erasure existed since no field emission is possible due to the lack of an external gate. Therefore, erasure was possible only by UV or X-ray irradiation.

The drawbacks of the FAMOS device were alleviated in several adapted concepts. In the *S*tacked gate *A*valanche injection *MOS* (SAMOS) [1.13, 1.14], an external gate is added, as shown in Fig. 1.8, in order to improve the writing efficiency, and thus, the programming speed by an increased drift velocity of the electrons in the oxide, a field-induced energy barrier lowering at the Si–SiO_2 interface, and a decreased drain breakdown voltage. Electrical erasure also became possible by field emission through the top dielectric due to polyoxide conduction. Consequently, EEPROM products became feasible.

These first floating gate memory devices were all p-channel devices. In n-channel devices, avalanching the drain yields hole injection, which is much less efficient. Several alternative injection mechanisms have been proposed, most of which, however, were not sufficiently adequate for large-volume applications. Out of the various proposed injection mechanisms, only a few have proven feasible in floating gate applications for large production volumes. These programming mechanisms are discussed in the next section, and the cells that have emerged are the subject of Section 1.4.

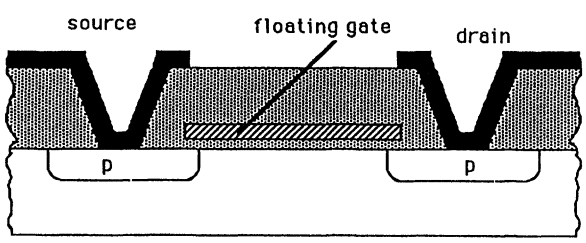


Figure 1.7 First operating floating gate device: the FAMOS (**F**loating gate **A**valanche injection **MOS**) device, introduced by Frohman–Bentchkowsky [1.6]. Writing the device is performed by injection of high energetic electrons created in the drain avalanche plasma. Erasure is possible by UV or X-ray radiation.

Petitioner Micron Ex-1024, 0008

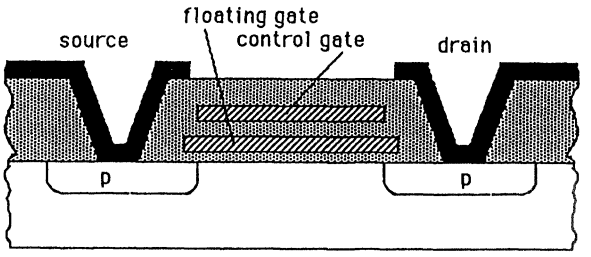


Figure 1.8 The SAMOS (Stacked gate Avalanche injection MOS) device [1.13]. The device is written like the FAMOS device. Several different erasure mechanisms are possible.

1.2. BASIC PROGRAMMING MECHANISMS

Electrical conduction through thin dielectric layers has been studied extensively in the past. It is generally understood that the electrical current behavior through dielectrics can be divided into two main classes: *bulk-limited conduction* and *electrode-limited conduction*. In the bulk-limited class, the current is determined mainly by the characteristics of the dielectric itself, and is independent of the electrodes from which the current originates. In the class of electrode-limited current, on the other hand, the conduction is determined by the characteristics of the electrodes, that is, the interface from which the current originates.

Many dielectrics, such as silicon nitride (Si_3N_4) or tantalum oxide (Ta_2O_5), belong to the bulk-limited conduction class. The current through silicon nitride is determined by Schottky emission from trapping centers in the nitride bulk and is commonly referred to as Poole–Frenkel conduction [1.15]. Thin nitride layers (< 30 nm), however, also show a strong electrode-limited contribution.

In silicon oxide, on the other hand, the current is determined mainly by the electrode characteristics, more specifically by the characteristics of the injection interface. This is due to the fact that SiO_2 has a large energy gap (about 9 eV compared to 5 eV for Si_3N_4) and a high energy barrier at its interface with aluminum or silicon. For example, the barrier of SiO_2 is about 3.2 eV for electrons in the conduction band of silicon and 4.8 eV for holes in the valence band, compared to 2 eV for holes and electrons in Si_3N_4, as shown in Fig. 1.9, which gives a comparison of the band structure for both materials. This means that conduction through SiO_2 will be determined primarily by electron injection, while, in Si_3N_4, both holes and electrons can contribute to the injection currents [1.16].

In both classes of nonvolatile memory devices, charge-trapping and floating gate devices, the charge needed to program the device has to be injected into an oxide layer, either to store it in the isolated traps in the nitride for the case of MNOS devices or to collect it at the floating gate in floating gate devices.

During the last two decades, various mechanisms for charge injection into the oxide have been considered. In order to change the charge content of floating gate devices, four mechanisms have been shown to be viable: Fowler–Nordheim tunneling (F–N) through thin oxides (< 12 nm) [1.5, 1.17], enhanced Fowler–Nordheim tunneling through polyoxides [1.18, 1.19], channel hot-electron injection (CHE)

Petitioner Micron Ex-1024, 0009

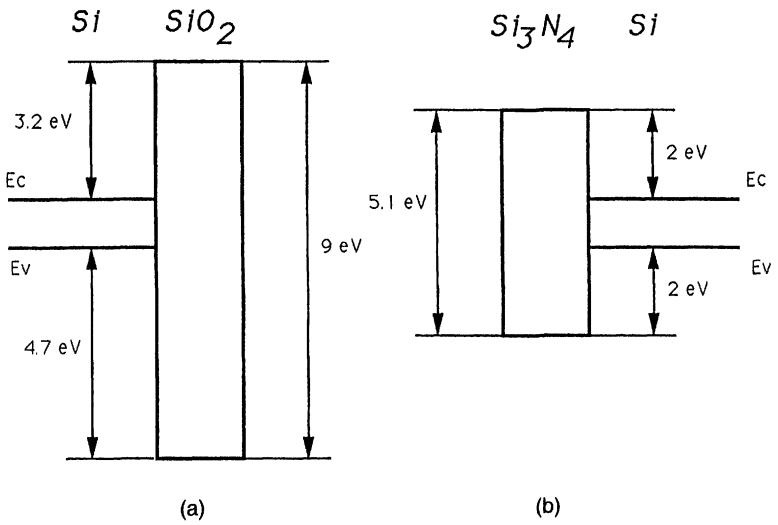


Figure 1.9 Energy band structures of (*a*) the Si–SiO_2 system and (*b*) the Si–Si_3N_4 system.

[1.20, 1.21], and source-side injection (SSI) [1.22, 1.23]. The first two are based on a quantum mechanical tunneling mechanism through an oxide layer, whereas the last two are based on injection of carriers that are heated in a large electric field in the silicon, followed by injection over the energy barrier of SiO_2. In order to change the charge content in charge-trapping devices, direct band-to-band tunneling and modified Fowler–Nordheim tunneling mechanisms are used. In the following sections, these six mechanisms are discussed briefly.

1.2.1 Fowler–Nordheim Tunneling

One of the most important injection mechanisms used in floating gate devices is the so-called Fowler–Nordheim tunneling, which, in fact, is a field-assisted electron tunneling mechanism [1.24]. When a large voltage is applied across a polysilicon–SiO_2–silicon structure, its band structure will be influenced as indicated in Fig. 1.10. Due to the high electrical field, electrons in the silicon conduction band see a triangular energy barrier with a width dependent on the applied field. The height of the barrier is determined by the electrode material and the band structure of SiO_2. At sufficiently high fields, the width of the barrier becomes small enough that electrons can tunnel through the barrier from the silicon conduction band into the oxide conduction band. This mechanism had already been identified by Fowler and Nordheim for the case of electrons tunneling through a vacuum barrier, and was later described by Lenzlinger and Snow for oxide tunneling. The Fowler–Nordheim current density is given by [1.24]:

Petitioner Micron Ex-1024, 0010

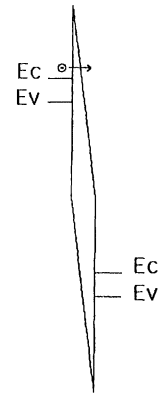


Figure 1.10 Energy band representation of Fowler–Nordheim tunneling through thin oxides: the injection field equals the average thin oxide field. Electrons in the silicon conduction band tunnel through the triangular energy barrier.

$$J = \alpha\, E_{inj}^2 \exp\left[\frac{-E_c}{E_{inj}}\right] \tag{1.3}$$

with

$$\alpha = \frac{q^3}{8\pi h \phi_b} \frac{m}{m^*} \tag{1.4}$$

and

$$E_c = 4\sqrt{2m^*}\, \frac{\phi_b^{3/2}}{3\hbar q} \tag{1.5}$$

where h = Planck's constant
ϕ_b = the energy barrier at the injecting interface (3.2 eV for Si–SiO_2)
E_{inj} = the electric field at the injecting interface
q = the charge of a single electron (1.6×10^{-19}C)
m = the mass of a free electron (9.1×10^{-31} kg)
m^* = the effective mass of an electron in the band gap of SiO_2 (0.42m [1.22])
$\hbar$ = $h/2\pi$

Equation (1.3) is the simplest form for the Fowler–Nordheim tunnel current density and is quite adequate for use with nonvolatile memory devices. A complete expression for the tunnel current density takes into account two second-order effects: image force barrier lowering and the influence of temperature.

The image force lowers the effective barrier height due to the electrostatic influence of an electron approaching the interface. Two correction factors $t(\Delta\phi_b)$ and $v(\Delta\phi_b)$, have to be introduced into Eq. (1.3), both of which are tabulated elliptic integrals and slowly varying functions. The reduction in energy barrier height ($\Delta\phi_b$) is given by [1.24]:

Petitioner Micron Ex-1024, 0011

$$\Delta\phi_b = \frac{1}{\phi_b}\sqrt{\frac{q^3\ E_{inj}}{4\pi\epsilon_{ox}}} \tag{1.6}$$

Although tunneling is essentially independent of temperature, the number of electrons in the conduction band, available for tunneling, is dependent on the temperature. This dependence can be taken into account by a correction factor f(T), given by [1.24]:

$$f(T) = \frac{\pi ckT}{\sin(\pi ckt)} \tag{1.7}$$

with

$$c = \frac{2\sqrt{2m^*}\ t(\Delta\phi_b)}{hqE_{inj}} \tag{1.8}$$

Taking these two corrections into account, we see that the expression for the Fowler–Nordheim tunnel current density becomes:

$$J = \alpha\ E_{inj}^2\ \frac{1}{t^2(\Delta\phi_b)}\ f(T)\ \exp\left[\frac{-E_c}{E_{inj}}\ v(\Delta\phi_b)\right] \tag{1.9}$$

The influence of the correction factors is small, however, and, for most practical calculations, the basic Eq. (1.3) is sufficiently accurate.

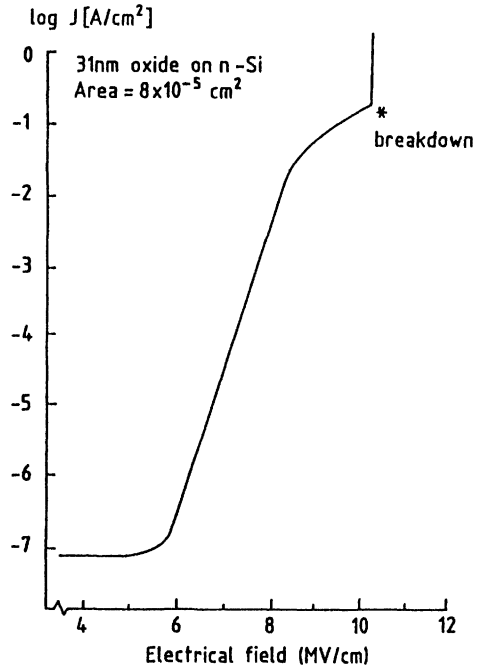
The Fowler–Nordheim tunnel current density is, thus, almost exponentially dependent on the applied field. This dependence is shown in Fig. 1.11*a* for the monocrystalline silicon–SiO_2 interface. The Fowler–Nordheim current is usually plotted as $\log(J/E^2)$ versus $1/E$, which should yield a straight line with a slope proportional to the oxide barrier, as shown in Fig. 1.11*b*. In this case, the numerical expression is

$$J\ [A/m^2] = 1.15\ 10^{-6}\ E_{inj}^2\ \exp\left[\frac{-2.54\ 10^{10}}{E_{inj}}\right] \tag{1.10}$$

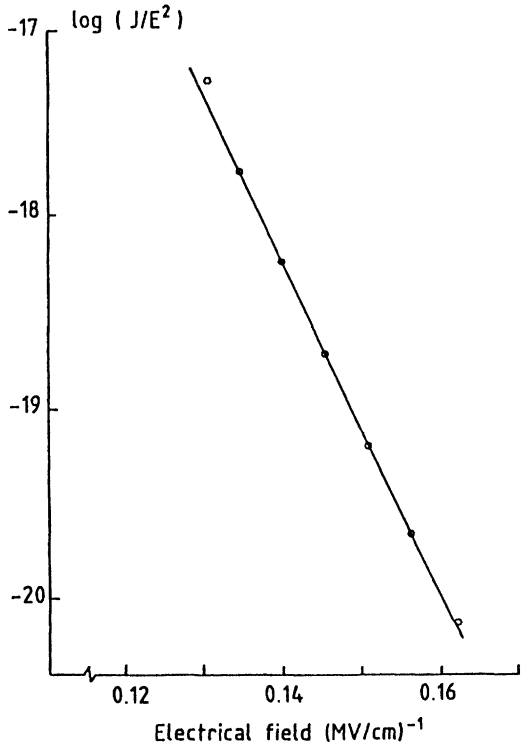
which, at an injection field of 10 MV/cm, leads to a current density of approximately $10^7\ A/m^2$ or $10^7\ pA/\mu m^2$. This high value of injection field is of the order of that needed across the oxide during the programming of a nonvolatile memory device. The breakdown field of these oxides should, of course, be significantly larger than this value. In order to reach these high-field values and limit the voltages needed during programming, very thin tunnel oxides are used; an injection field of 10 MV/cm is attained by applying a voltage of 10 V across an oxide of 10 nm thickness. In order to reduce the programming voltage, the tunnel oxide should become even thinner. A thickness of 6 nm, however, is the lower limit for good retention behavior. But these thin oxides are difficult to grow with low defect densities, as is required for floating gate devices. Moreover, below these values, other injection mechanisms, such as direct tunneling, can become important. Yield considerations now limit the usable oxide thicknesses to 8 to 10 nm [1.23].

It should be noted that the tunnel current density is totally controlled by the field at the injecting interface, and not by the characteristics of the bulk oxide. Once

Petitioner Micron Ex-1024, 0012



(a)



(b)

Figure 1.11 (*a*) Fowler–Nordheim tunneling current as a function of applied field across the oxide. The current is exponentially dependent on the field. Breakdown occurs around 10 MV/cm. (*b*) Fowler–Nordheim plot: J/E^2 as a function of $1/E$, extracted from the data (*a*). A straight line is obtained.

Petitioner Micron Ex-1024, 0013

the electrons have tunneled through the barrier, they are traveling in the conduction band of the oxide with a rather high saturated drift velocity of about 10^7 cm/s [1.26].

For the calculation of the injection field at a silicon–SiO_2 interface, however, the flatband voltage has to be taken into account as seen by

$$E_{inj} = \frac{V_{app} - V_{fb}}{t_{ox}} \tag{1.11}$$

where V_{app} = the voltage applied across the oxide
V_{fb} = the flatband voltage
t_{ox} = the thickness of the oxide

When voltages are applied so that the silicon is driven into depletion, a voltage drop in the induced depletion layer must be accounted for in the calculation of the oxide field.

The tunnel current for a given applied voltage can be calculated as the product of the tunnel current density and the injecting area only if the tunnel current density has the same value over the whole injecting surface—that is, if the injection occurs uniformly over the area of the tunnel oxide. This assumes a perfectly plane injecting interface which, in many practical devices, will not be the case. Special cases of nonuniform injection are discussed in Sections 1.2.2 and 1.2.3.

1.2.2 Polyoxide Conduction

Fowler–Nordheim tunneling requires injection fields on the order of 10 MV/cm to narrow the Si–SiO_2 energy barrier so that electrons can tunnel from the silicon into the SiO_2 conduction band, as discussed in the previous section.

In oxides thermally grown on monocrystalline silicon, the injection field is equal to the average field in the SiO_2; therefore, thin oxides have to be used to achieve large injection fields at moderate voltages. Oxides thermally grown on polysilicon, called polyoxides, however, show an interface covered with asperities due to the rough texture of the polysilicon surface [1.27, 1.28]. This has led to the name "textured polyoxide." These asperities give rise to a local field enhancement at the interface and an enhanced tunneling of electrons [1.29, 1.30]. In polyoxides, the field at the injecting interface is, therefore, much larger than the average oxide field. Consequently, the band diagram of a polysilicon–polyoxide interface is as shown schematically in Fig. 1.12. Average oxide fields of the order of 2 MV/cm are sufficient to yield injection fields of the order of 10 MV/cm. This has the big advantage that large injection fields at the interface can be obtained at moderate voltages using relatively thick oxides, which can be grown much more reliably than the thin oxides necessary for Fowler–Nordheim injection from monocrystalline silicon.

A quantitative analysis of the tunnel current–voltage relations for polyoxides is rather complex. Although the tunnel mechanism itself is described by the same formula (Eq. 1.3), discussed in a previous section, the difficulty lies in the accurate determination of the injection fields to be used. It is no longer possible to use a single value for this injection field because of the nonuniformity of the field enhancement

Petitioner Micron Ex-1024, 0014

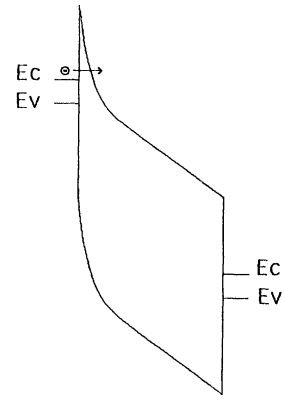


Figure 1.12 Energy band representation of Fowler–Nordheim tunneling through oxides thermally grown on polysilicon: the injection field is much higher than the average oxide field. The high injection field is due to local field enhancement at polysilicon–oxide interface asperities.

over the injecting interface. Indeed, the field enhancement factor is not uniform over the surface of one asperity bump [1.31–1.33]: the factor is maximum at the top of the asperity and decreases strongly down the slope on the bump surface. In addition, variations of the bump shape may be another cause for the nonuniformity.

In the past, attempts have been made to model the current through the polyoxide by use of some mean field enhancement factor [1.34], but as was proven in [1.32], this method always leads to incorrect results. A complete model for the current conduction must be based on [1.32, 1.33]:

the Fowler–Nordheim expression for tunnel current density

a model for the distribution of the field enhancement factors over the total injecting area

a model for the charge-trapping behavior of the oxide under current injection

The last-named model has to be taken into account because charge trapping is of much more importance in polyoxides than in oxides grown on monocrystalline material. This is again due to the strong nonuniform field enhancement. Initially, the injection current originates almost completely from the regions of maximum field enhancement. Extremely large current densities occur at these injection points, leading to strong local trapping of electrons near these sites. This trapping reduces the injection locally. Consequently, the current is taken over by regions with a slightly lower field enhancement. This process proceeds gradually so that the injection current, which initially is extremely localized, becomes more and more uniform and decreases continuously. Unlike conventional Fowler–Nordheim injection in which the trapping occurs only after some critical current level has been reached, charge trapping and current injection occur simultaneously over the whole current range during polyoxide injection. The nonuniform field enhancement at the polysilicon–polyoxide interface makes it impossible to use a closed analytical expression for polyoxide conduction. A complete model for polyoxide conduction, based on the principles indicated above, can be found in [1.32, 1.33].

Petitioner Micron Ex-1024, 0015

An example is given in Fig. 1.13 where the injection current is shown during a ramped voltage experiment for two consecutive runs on the same polyoxide capacitor. The dashed lines represent the experimental currents, while the solid lines are simulations based on the above referenced model [1.32]. As can be seen, during the first run the current is increasing less rapidly than expected from the conventional Fowler–Nordheim mechanism. This is due to the gradual decrease of the mean enhancement factor of the polyoxide surface, which is caused by the local electron trapping and accompanying shielding of the sites of maximum field enhancement. The second ramp shows a large shift with respect to the first one. Unlike the case of uniform Fowler–Nordheim injection, however, *this shift cannot be interpreted in terms of trapped charge only*, but is due mainly to a decrease in the mean enhancement factor after the first run.

Another consequence of the fast decrease in mean enhancement factor due to local trapping in polyoxides is the fast current decrease observed during measurement of the time behavior of the current after application of a voltage step across the polyoxide. This decrease can be described by a time power law ($I = C\ t^{-n}$), with a decay factor n. Whereas this decay factor is expected to be 1 for uniform trapping, the decay factor is found to be smaller than 1 for polyoxides [1.29]. Again, this is because the decay is due not only to charge trapping but also to the decrease in the mean enhancement factor of the polyoxide interface. An example is shown in Fig. 1.14, where, again, experimental results are compared with results from the above-mentioned model [1.32].

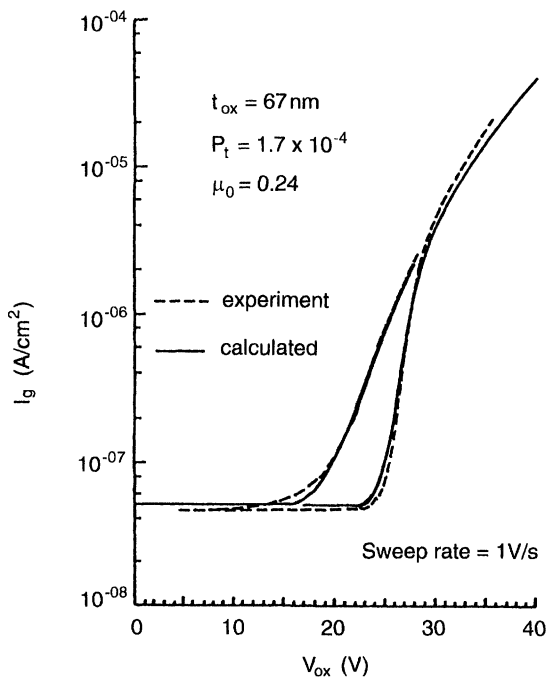


Figure 1.13 Injection current in polyoxides as a function of applied voltage during a ramped voltage experiment for two consecutive runs on the same polyoxide capacitor. The dashed lines represent experimental currents, while the solid lines are simulations based on the model by Groeseneken et al. [1.32].

Petitioner Micron Ex-1024, 0016

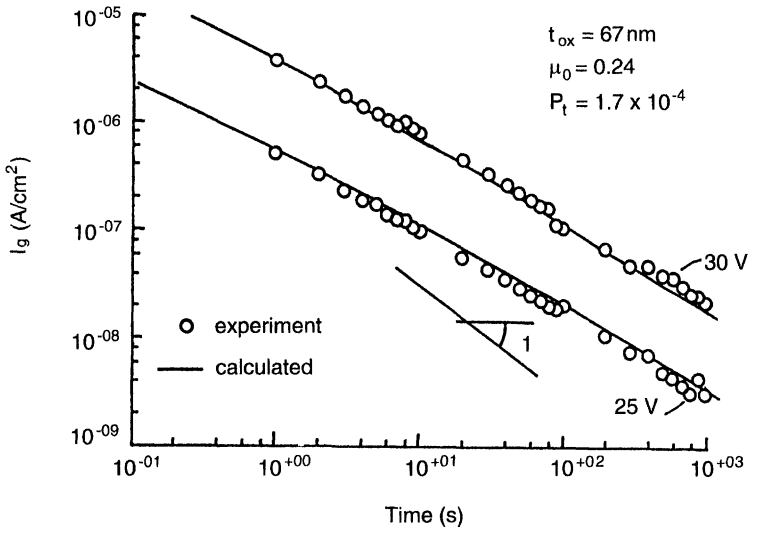


Figure 1.14 Time dependence of the polyoxide current for two values of applied voltage. The circles are experimental values, while the solid lines are results, based on the model of [1.32].

Polyoxide conduction has the advantage that considerable current levels can be attained at moderate average oxide fields, and thus, moderate applied voltages. The need for thin polyoxides is, therefore, not so stringent. From a reliability point of view, this is an advantage since the oxides are not stressed at large fields during programming so that dielectric breakdown failures are avoided [1.25]. On the other hand, the growth of textured polyoxides has to be carefully controlled in order to obtain the desired interface features (shape and size of the asperities) that determine the injection current and reliability characteristics. For this reason, reproducibility may be a problem for this kind of injection mechanism. Another disadvantage is that the injection is asymmetric with respect to polarity. For injection from a top polysilicon layer, the currents are much smaller. Finally, the strong change in injection currents due to a decrease in mean enhancement factor during current injection can pose severe constraints on the number of programming cycles that can be allowed if this mechanism is used for programming a memory cell [1.32]. In nonvolatile memories, polyoxides of 25 nm to 60 nm are used with programming voltages from 12 V up to 20 V.

1.2.3 Hot-Electron Injection

At large drain biases, the minority carriers that flow in the channel of a MOS transistor are heated by the large electric fields seen at the drain side of the channel

Petitioner Micron Ex-1024, 0017

and their energy distribution is shifted higher. This phenomenon gives rise to impact-ionization at the drain, by which both minority and majority carriers are generated. The highly energetic majority carriers are normally collected at the substrate contact and form the so-called substrate current. The minority carriers, on the other hand, are collected at the drain. A second consequence of carrier heating occurs when some of the minority carriers gain enough energy to allow them to surmount the SiO_2 energy barrier. If the oxide field favors injection, these carriers are injected over the barrier into the gate insulator and give rise to the so-called hot-carrier injection gate current [1.35, 1.36]. This mechanism is schematically represented for the case of an n-channel transistor in the energy band diagram shown in Fig. 1.15.

For nonvolatile memory applications, n-channel transistors are generally used, and therefore, the discussion here will be limited to n-channel devices. In case of an n-channel transistor, the gate current of the transistor consists of those channel hot electrons that actually reach the gate of the transistor. In a floating gate transistor, these electrons change the charge content of the floating gate. An important difference between hot-carrier injection and the two previously discussed injection mechanisms is that, with hot-electron injection, it is only possible to bring electrons onto the floating gate. They cannot be removed from the floating gate by the same mechanism. Although the use of hot-hole injection as a compensating programming mechanism has been tried [1.37], it has never found application due to the very small current levels that can be attained in this way.

In the past, several models have been used to describe the gate current due to channel hot-electron injection. In contrast to the Fowler–Nordheim tunneling case, no closed form analytical expression exists for the channel hot-electron injection current due to the complex two-dimensional nature of the phenomenon and many unknown physical parameters. Therefore, the models are merely qualitative. They can be divided into three main categories: the lucky electron models, the effective electron temperature models, and the physical models.

The lucky electron model [1.38, 1.39] assumes that an electron is injected into the gate insulator if it can gain enough energy in the large lateral electric field with-

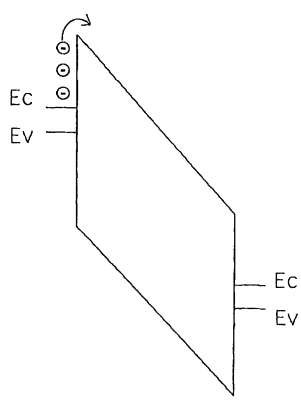


Figure 1.15 Energy band representation of hot-electron injection in the oxide; the oxide field is low, but the electrons are heated by the high lateral fields at the drain in the channel. Some of them acquire enough energy to overcome the interface energy barrier.

Petitioner Micron Ex-1024, 0018

out undergoing a collision, by which energy could be lost. By phonon scattering, the electrons are then redirected toward the Si–SiO_2 interface. If these electrons can reach the interface and still have enough energy to surmount the Si–SiO_2 energy barrier (and eventually also a repulsive field), they will be injected into the gate insulator.

The effective electron temperature model [1.40] assumes that the electrons forming the channel current are heated and become an electron gas with a Maxwellian distribution, with an effective temperature, T_e, that is dependent on the electric field. The gate current can then be calculated as the thermionic emission of heated electrons over the interface barrier energy.

The physical models [1.41] attempt to calculate the gate currents based on a more physical treatment and an accurate solution of the two-dimensional electric field distribution at the drain side of the channel. Then, the gate current is calculated based on an injection efficiency that is dependent on the interface barrier energy and the lateral electric fields.

For all the above-mentioned models, we always have to keep in mind that a difference exists between the number of injected electrons and the number of electrons actually reaching the gate. Indeed, due to a repulsive oxide field, all or part of the injected electrons can be repelled into the silicon [1.42].

Qualitatively, it can be stated that the gate current is determined on the one hand by the number of hot electrons and their energy distribution (which is largely dependent on the electric fields occurring in the channel of the transistor) and on the other hand by the oxide field (which determines the fraction of hot electrons that can actually reach the gate).

The magnitude of the gate current is dependent on both the applied gate and drain voltages. A characteristic gate current, as a function of the applied gate voltage and with the drain voltage as a parameter, is shown in Fig. 1.16 for an n-channel transistor. It is important to notice that the hot-electron gate current shows a maximum at approximately $V_g = V_d$, and thus, is not a monotonically increasing function of the applied gate voltage, as is the case for both Fowler–Nordheim and polyoxide conduction.

This typical shape is explained by both determining factors—the gate and drain voltages [1.36, 1.42]. For gate voltages greater than the drain voltage, the oxide field is always favorable for charge collection at the gate, which means that the gate current is limited by the number of hot electrons that are injected. The lateral electric field, and thus, the number of hot electrons that can be injected into the oxide, increases with decreasing gate voltage. Therefore, for $V_g > V_d$, the gate current increases with decreasing gate voltage. For gate voltages smaller than the drain voltage, however, the oxide field becomes repulsive for the injected electrons. Therefore, part of the injected electrons are repelled into the channel. Although the number of hot electrons that are available to be injected still increases with decreasing gate voltage, the gate current now drops rapidly with decreasing gate voltage. Due to this typical gate voltage dependence of the hot-electron gate current, the gate voltage during programming of a nonvolatile memory cell, using hot-electron injection, has to be chosen carefully in relation to the applied drain voltage.

Petitioner Micron Ex-1024, 0019

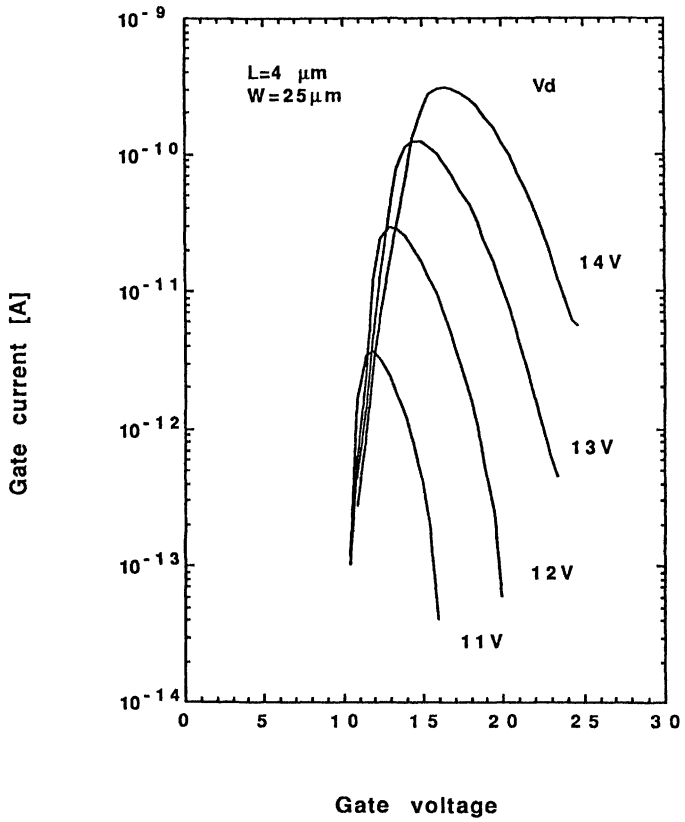


Figure 1.16 Hot-electron injection currents as a function of applied gate voltage with the drain voltage as a parameter. The maximum current occurs when $V_g = V_d$ and is exponentially dependent on the drain voltage.

In order to evaluate the dependence of the injection current on processing and geometrical parameters, a simplified expression for the lateral electric field in the channel can be used [1.43]:

$$E \approx \frac{V_d - V_{dsat}}{L} \tag{1.12}$$

with

$$L \approx 0.22\ t_{ox}^{1/3}\ x_j^{1/2} \tag{1.13}$$

and with V_{dsat} expressed as [1.44]:

$$V_{dsat} = \frac{(V_g - V_t)\ L_{eff}\ E_{sat}}{V_g - V_t + L_{eff}\ E_{sat}} \tag{1.14}$$

where E_{sat} is the electric field at which the electron mobility saturates.

Petitioner Micron Ex-1024, 0020

From these formulas, it can be concluded that the gate current increases with thinner gate oxides, shallower junctions, smaller effective channel lengths, and higher substrate doping levels (through the influence on the threshold voltage at the drain through the body effect).

1.2.4 Source-Side Injection

The main disadvantage of the conventional channel hot-electron injection mechanism for programming a nonvolatile element stems from its low injection efficiency, and consequently, its high power consumption. This is due to the incompatibility of having a high lateral field and a high vertical field, favorable for electron injection, at fixed bias conditions, as explained in the previous section. Indeed, the lateral field in a conventional MOS device is a decreasing function of the gate voltage, while the vertical field increases with the gate voltage. Therefore, in order to generate a large number of hot electrons, a low gate voltage is required, combined with a high drain voltage. However, for electron injection and collection on the floating gate of the memory device, a high gate voltage and a low drain voltage are required (see Fig. 1.17). In practice, both gate and drain voltages are kept high as a compromise. The main drawback is clearly the high drain current (on the order of mA's) and the correspondingly high power consumption.

Therefore, a novel injection scheme, now commonly referred to as source-side injection (SSI), has been proposed to overcome this problem [1.45]. In most cases, the MOS channel between the source and drain regions is split into two "subchannels" controlled by two different gates. The gate on the source side of the channel is biased at the condition for maximum hot-electron generation, that

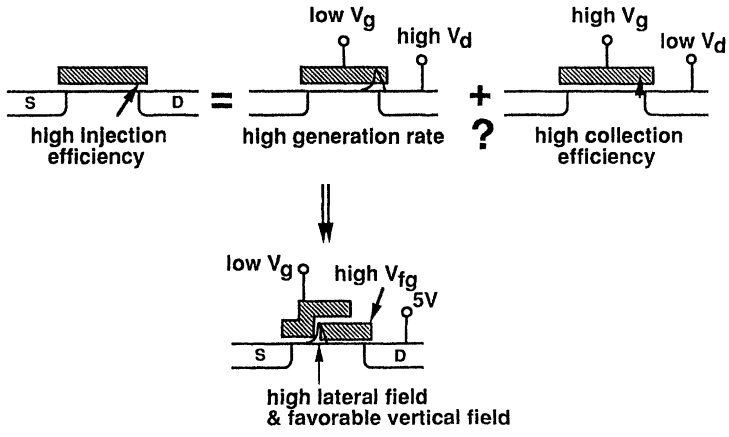


Figure 1.17 Schematic representation of the problem of low-injection efficiency for channel hot-electron injection and the principle of source-side injection.

Petitioner Micron Ex-1024, 0021

is, very close to the threshold voltage of this channel [1.42]. The gate at the drain side, which is the floating gate of the cell, is capacitively coupled to a potential that is comparable to or higher than the drain voltage in order to establish a vertical field component that is favorable for hot-electron injection in the direction of the floating gate. The latter condition can be accomplished either by implementing an additional gate with a high coupling ratio toward the floating gate [1.23, 1.45], or by using a high drain-coupling ratio [1.22]. As a result, the drain potential is entirely or partially extended toward the region between the gates that control the MOS channel. This effect is referred to as the virtual drain effect since the inversion layer under the floating gate merely acts as a drain extension, while the effective transistor channel is formed by the subchannel at the source side of the device [1.46]. Consequently, a high lateral field peak is obtained in the gap between both subchannels (Fig. 1.17). The hot electrons are thus generated inside the MOS channel and not at the drain junction of the cell. Because of the high floating gate potential, the vertical field at the injection point is favorable for electrons, and most of the generated hot electrons that overcome the potential barrier between the channel and the oxide layer are effectively collected on the floating gate.

The main advantage of this injection mechanism is the much higher injection efficiency (on the order of 10^{-3} and higher) which allows for fast 5 V-only and even 3.3 V-only operation combined with a low power consumption [1.23, 1.45]. This is illustrated in Fig. 1.18 where the gate currents of the conventional channel hot-electron injection and the source-side injection mechanisms are shown for comparable devices [same channel width to length ratio (W/L), same drain voltage, same technology]. It is clear that the SSI mechanism provides a gate current that is more than three orders of magnitude higher than conventional hot-electron injection [1.46]. At the same time, the drain current in the SSI case is also reduced by a factor

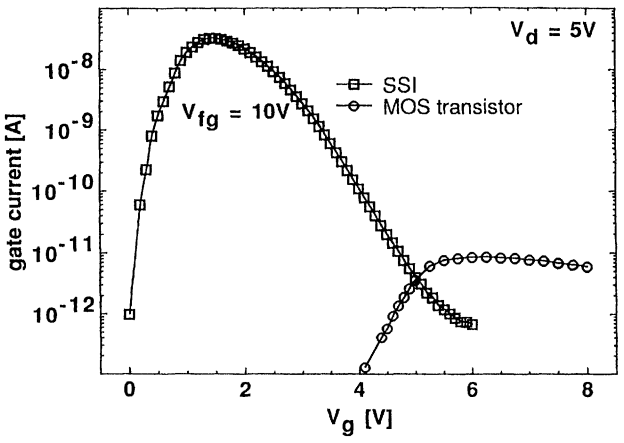


Figure 1.18 Comparison of injected gate current for source-side injection (SSI) and channel hot-electron injection, both measured at the same drain voltage of 5V.

Petitioner Micron Ex-1024, 0022

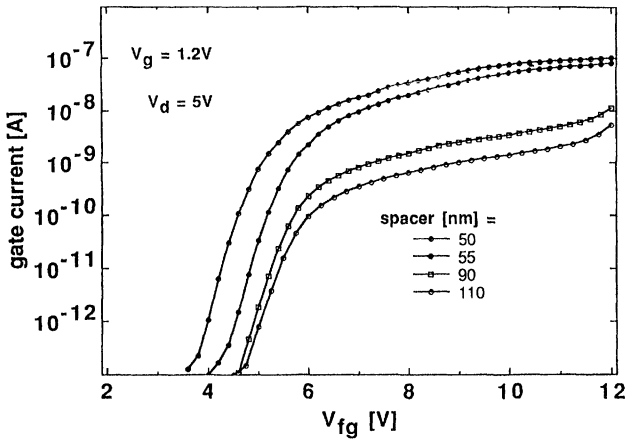


Figure 1.19 Source-side injection current versus floating gate voltage for various values of the interpoly spacer thickness. This characteristic also gives the evolution of the programming current during a programming operation.

of 40 with respect to the conventional case [1.45]. Figure 1.19 shows typical gate current characteristics for the SSI mechanism, but this time as a function of the floating gate voltage. The gate current tends to saturate with increasing floating gate voltage because the virtual drain potential (i.e., the channel potential at the injection point) approaches the externally applied drain voltage. Since the floating gate voltage changes during programming, the maximum observed in the conventional gate current characteristic (Fig. 1.16) is no longer relevant in the SSI case. The gate current decreases monotonically, and only slightly, while programming an SSI cell. Figure 1.19 also shows that the gate current is a strong function of the interpoly width between the gates that control the subchannels. Furthermore, the SSI mechanism is no function of the drain profile and is instead only a smooth linear function of the channel length of the device. This is in strong contrast to conventional hot-electron injection where the injection is strongly dependent on both drain profile and channel length.

1.2.5 Direct Band-to-Band Tunneling and Modified Fowler–Nordheim Tunneling

Section 1.2.1 treated Fowler–Nordheim tunneling, which is field-assisted electron tunneling from the silicon band into the silicon dioxide band through the triangular energy barrier. In MNOS devices with ultra-thin oxides (< 3 nm), the injection current can either be direct silicon band to nitride band tunneling only through the oxide barrier, as illustrated in Fig. 1.20*a*, or modified Fowler–Nordheim tunneling through the oxide barrier and a nitride barrier, as shown in

Petitioner Micron Ex-1024, 0023

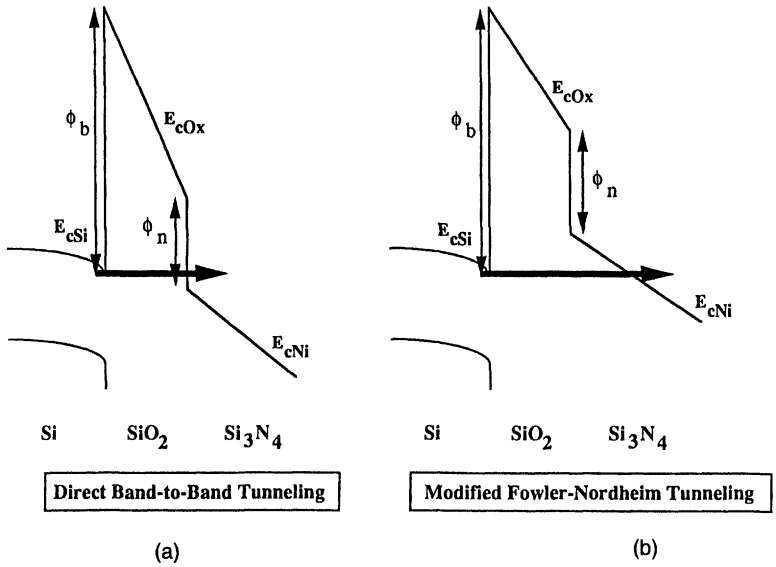


Figure 1.20 Energy band representation of (*a*) direct band-to-band tunneling and (*b*) modified Fowler–Nordheim tunneling between the silicon and the nitride conduction band.

Fig. 1.20*b*. Whether one or the other of these conditions applies depends strongly on the values of the oxide field and oxide thickness. These currents can be expressed as in [1.47]:

$$J = C_{FN}\, E_{ox}^2\, P_{ox}\, P_n \tag{1.15}$$

where C_{FN} is a constant with a similar meaning as α in Eq. (1.3), and P_{ox} and P_n represent the tunneling probabilities through the oxide and the nitride barriers, respectively, and are given by

$$P_{ox} = \exp\left\{-\frac{4}{3h}\sqrt{2qm_{ox}^*}\,\frac{\left[\phi_b^{3/2} - (\phi_b - E_{ox}t_{ox})^{3/2}\right]}{E_{ox}}\right\} \tag{1.16}$$

and

$$P_n = \exp\left\{\frac{4}{3h}\sqrt{2qm_n^*}\,\frac{(\phi_b - \phi_n - E_{ox}t_{ox})^{3/2}}{E_n}\right\} \tag{1.17}$$

where E_{ox} is the field in the oxide and E_n is the field in the nitride, m_{ox}^* and m_n^* are the effective masses in oxide and nitride, respectively, and ϕ_n is the oxide–nitride barrier. In Eqs. (1.16) and (1.17), a negative term within a radical must be replaced by zero.

If $(\phi_b - \phi_n - E_{ox}t_{ox}) < 0$, direct tunneling occurs only through the oxide potential barrier and $P_n = 1$ (see Fig. 1.20*a*). If $(\phi_b - E_{ox} \cdot t_{ox}) < 0$, Eq. (1.16) reduces to

Petitioner Micron Ex-1024, 0024

the expression for Fowler–Nordheim tunneling [Eq. (1.9)]. Since $E_n \approx (\epsilon_{ox}/\epsilon_n)E_{ox}$ with ϵ_{ox} and ϵ_n, the dielectric constants of oxide and nitride, respectively, Eq. (1.15) provides the current–oxide field relation for tunneling through these double potential barriers. The oxide field, E_{ox}, for a double insulator system (oxide thickness, t_{ox}, and nitride thickness, t_n) and for a given V_{app} and V_{fb} can be obtained from an expression similar to Eq. (1.11) for the case of a single insulating layer from

$$E_{ox} = \frac{V_{app} - V_{fb}}{t_{ox} + \frac{\epsilon_{ox}}{\epsilon_n} t_n} \tag{1.18}$$

1.3. BASIC NVSM MEMORY PRODUCTS

The nonvolatile memory cell concept has been used in several kinds of applications, and many different products have emerged in recent years. The core of the applications are high-volume stand-alone nonvolatile memories, but besides these, it has also been applied to other purposes such as electrically programmable logic devices (EPLD), application specific integrated circuits (ASIC) (embedded memories), and redundancy. Before discussing the different types of nonvolatile cells, this section first treats some of the important features of the main nonvolatile memory products. The aim is not to be complete, however, for the application fields and the product range for nonvolatile memory cells are so large that they cannot be covered within the limited focus of this chapter. A more general overview of the trends in nonvolatile memories can be found in [1.48, 1.49].

This section discusses the main features of EPROM and OTP, EEPROM, Flash EEPROM, and NOVRAM memory products, as well as a new type of concept, FRAM. These classes of nonvolatile memory products emerged under the influence of three main factors: (1) the limitations posed by the available cells (EPROM/OTP), (2) the requirements of the users (EEPROM, NOVRAM), and (3) market and price considerations (Flash EEPROM).

1.3.1 EPROM/OTP

The electrically programmable read-only memory (EPROM) was, in fact, the first nonvolatile memory that could be electrically programmed by the user and that could be erased afterward. All EPROM products rely on the floating gate cell concept. Present EPROM devices all use channel hot-electron injection. Since this mechanism can only supply electrons to the floating gate, EPROM memories are not electrically erasable. UV light is used to erase the memory. For programming and reading, the memory is byte addressable, and each byte can be addressed separately. Obviously, the erase operation always affects the whole memory. It is the user who can perform both operations. For both operations, however, some additional tools are needed—an EPROM programmer and UV light for erasing.

The channel length of EPROM cells has been steadily decreasing, reaching values down to 0.8 μm [1.50]. Since the end of the 1980s, however, EPROMs have

Petitioner Micron Ex-1024, 0025

been gradually replaced by Flash EEPROM products. Channel hot-electron programming requires high currents and high voltages. Consequently, EPROM memory products require an external supply voltage of typically 12 V for programming. The programming time ranges from 1 ms down to 100 μs per byte of information. Erasing typically takes 20 minutes of UV light exposure. During erasure, the component is not powered on. Because the EPROM functionality does not need addressing down to byte level during an erase operation, the memory cell can be kept fairly simple. One floating gate transistor suffices to build an EPROM memory cell, as illustrated in Fig. 1.21, and is, therefore, called a single-transistor memory cell. This allows for small cell sizes in the range of 8 μm^2 for 0.8 μm technologies, and bit densities comparable to those of DRAMs. At present, the highest bit densities available are 4 and 8 Mbit. The evolution of EPROM cell size and bit density is shown in Figs. 1.22 and 1.23 [1.49].

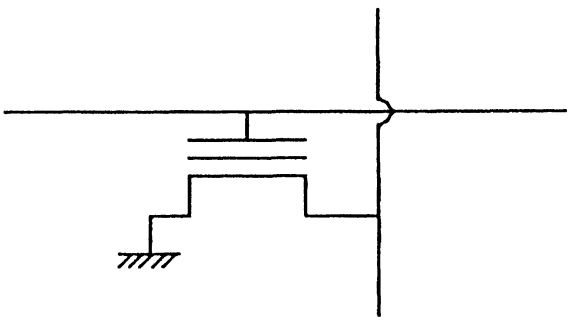


Figure 1.21 Single-transistor EPROM floating gate memory cell.

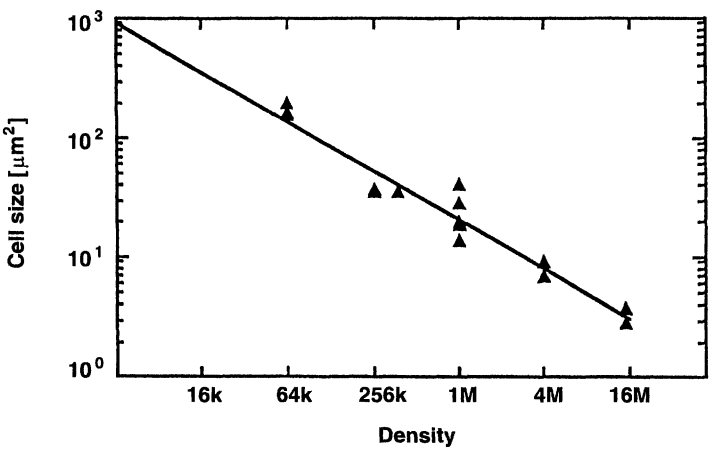


Figure 1.22 Evolution of the cell size of EPROM memory cells as a function of memory bit density.

Petitioner Micron Ex-1024, 0026

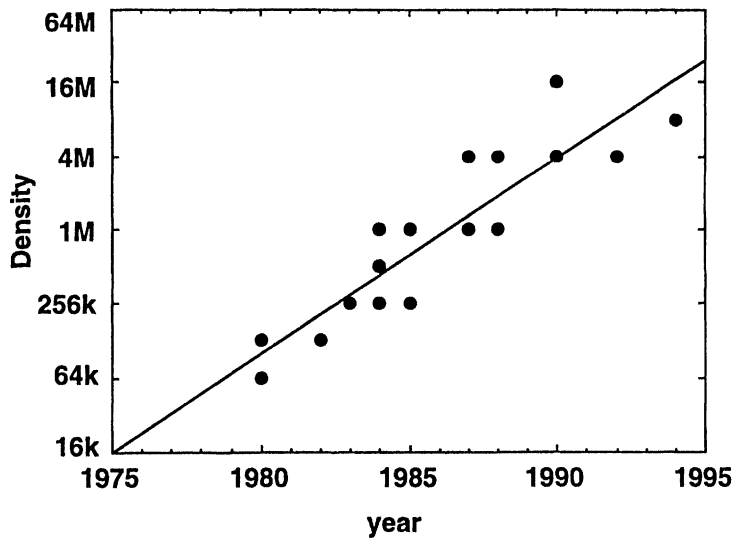


Figure 1.23 Evolution of bit density as a function of time showing a new generation of bit densities every three years.

Since UV light is used for erasure, a quartz window has to be provided in the EPROM package, which makes this package quite expensive. The package also has to be taken out of the circuit board in order to erase or reprogram the memory. In order to avoid these problems, a new product, which actually uses the same chips as the EPROM, called the one-time programmable (OTP) memory, was developed. This device can be written only once and is used like a PROM. Since no erasure of the memory is intended, the quartz window is not necessary and the device can be housed in a cheaper plastic package.

1.3.2 EEPROM

Although EPROM memories are reprogrammable, the reprogramming of the device is not user-friendly. The circuit has to be taken off the circuit board. The erase operation takes about 20 minutes, and then the whole memory circuit has to be reprogrammed byte by byte. This rather tedious erase procedure must be performed even if the content of a single byte has to be changed. These drawbacks have been obviated in the electrically erasable programmable read-only memory (EEPROM). In this type of nonvolatile memory circuit, all operations are controlled by electrical signals. The circuit can be reprogrammed while residing on the circuit board. Each operation, including erasing, can be performed in a byte-addressable way.

This higher level of functionality results in a larger memory cell. Since the EEPROM is byte addressable for reading and for all programming operations, the

Petitioner Micron Ex-1024, 0027

memory cell has to consist of a memory transistor and a select transistor [1.51] as shown in Fig. 1.24, thus leading to the so-called two-transistor memory cell. As a result, this memory cell is larger than the EPROM cell. Consequently, the densities of EEPROM products have always lagged behind EPROM densities by one to two generations, as illustrated in Fig. 1.25, with 1 Mbit to 2 Mbit memories as the present state-of-the-art EEPROM densities. Typical cell sizes of 1 Mbit parts range between 30 and 50 μm^2.

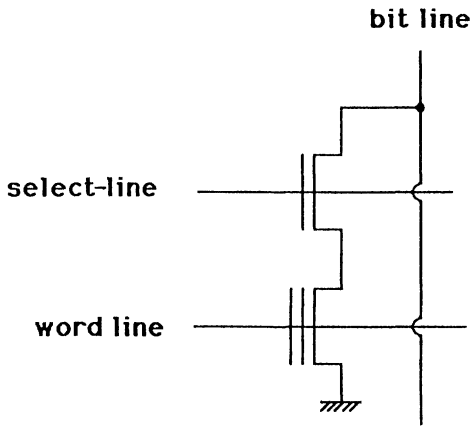


Figure 1.24 Two-transistor EEPROM memory cell. In order to allow byte selective write and erase, a select transistor is added to the memory cell.

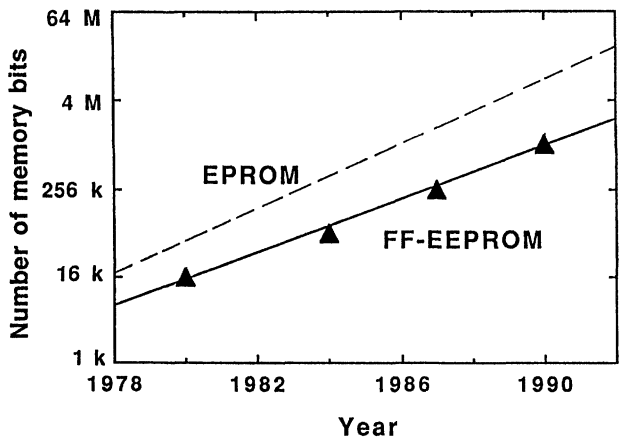


Figure 1.25 Evolution of the bit density as a function of time for EEPROM, and comparison with EPROM. EEPROM is lagging behind about one to two generations.

Petitioner Micron Ex-1024, 0028

Charge-trapping as well as floating gate cells are used for EEPROM products. As mentioned previously, the MNOS cell already is inherently electrically-erasable. The floating gate cells usually rely on Fowler–Nordheim tunneling or polyoxide conduction in order to achieve electrical programmability for both operations.

The continuing search for a nonvolatile memory part that is as easy to use as a RAM has led to the incorporation of more and more features on the memory chip, which had to be provided externally in earlier generations of EEPROMs. Considering the EEPROM evolution to this point in time, we can mention three generations. The first generation required an external, high-voltage power supply and wave-shaped write signals with critical specifications for rise, overshoot, and pulse times. In the second generation, the wave-shaping and high-voltage external power supply were eliminated, leading to the first 5 V-only EEPROM products [1.52–1.54]. The third generation is even more complex, with features such as data and address latching, internal timing for the programming operation, page-mode programming capabilities, on-chip pulse shaping, complete transistor-transistor logic (TTL) compatibility, power on/off protection circuitry, on-chip error checking and correcting circuits, data polling possibilities, and 5 V-only operation [1.55, 1.56]. This evolution has made the present EEPROM products completely compatible with other types of memories like SRAM and DRAM.

The EEPROM circuits have become 5 V-only by generating programming voltages on the chip by means of voltage multiplier circuits [1.52–1.54]. This is possible only because programming mechanisms rely on tunneling (direct, Fowler–Nordheim, or polyoxide tunneling), which does not require large programming currents. The erase operation that has to be performed before a byte can be written into a new state is made invisible to the user. Every write request is automatically preceded by the proper erase operation, which is totally controlled by circuits incorporated on the memory chip.

The properly shaped control signals for programming in all recent circuits are generated on chip. The memory just needs a TTL-compatible pulse to initiate a write operation. The timing and application of the different voltage levels are controlled by on-chip circuits [1.55]. Moreover, by using an intelligent data polling feature, the external circuitry can find out if the data have been successfully written into the memory, or if the internal write operation is still busy [1.55], which can be used to reduce the effective programming time.

In order to shorten the quite lengthy programming operation, so-called page-mode programming has been added [1.55, 1.56]. The user writes a whole page (typically 16 to 64 bytes) to the EEPROM as if it were a RAM. On the EEPROM chip, the information and the appropriate addresses are stored, and afterward, the whole page is written in parallel into the nonvolatile memory cells. This effectively reduces the programming time per byte by a factor of 16 to 64.

Devices incorporating all the above-mentioned features are called full-featured EEPROMs. Another class of EEPROM devices is high-speed EEPROMs. These circuits, though not as user-friendly, have a read access time in the range of 30 to 50 ns, comparable to SRAM devices and to bipolar products [1.57].

Petitioner Micron Ex-1024, 0029

1.3.3 Flash EEPROM

During the 1980s, a novel nonvolatile memory product was introduced, referred to as the Flash EEPROM [1.58]. The general idea was to combine the fast programming capability and high density of EPROMs with the electrical erasability of EEPROMs. The first products were merely the result of adapting EPROMs in such a way that the cell could be erased electrically. Consequently, these devices used channel hot-electron injection for programming and Fowler–Nordheim tunneling through a thin gate oxide or through a polyoxide for erasure.

All Flash EEPROM products are based on the floating gate concept. The memory can be erased electrically but not selectively. The content of the whole memory chip is always cleared in one step. The advantages over the EPROM are the faster (electrical) erasure and the in-circuit reprogrammability, which leads to a cheaper package. Its cost is lower than that of EEPROM devices, and the part was introduced partially to cope with the low volumes of the market that could be reached with the full-featured EEPROM, until recently.

In the 1990s, Flash memory has become the largest market in nonvolatile technology due to a highly competitive tradeoff between functionality and cost/bit. Since the cell size of Flash devices has the potential to track that of DRAM cells, competitively priced Flash concepts are expected to find a huge market and even to become one of the main technology drivers of the semiconductor industry. Apart from the replacement of EPROMs and EEPROMs, novel application fields have also arisen, such as solid-state disks for portable and handheld computers, and smart cards. Also, novel device structures have been proposed based on Fowler–Nordheim tunneling for both programming and erasure in order to allow operation from a single supply voltage. Moreover, source-side injection Flash devices (see Section 1.2.4) have gained considerable interest because of their unique combination of very fast programming capabilities with low power consumption.

Currently, Flash products up to 32 Mbit are commercially available. The cell size attained in a 32 Mbit product is on the order of 1.5 μm^2 [1.59–1.61]. Finally, there is a strong demand for embedded Flash memory on ASICs, digital signal processing (DSP) chips, microcontrollers, and so on. In the case of microcontrollers, process compatibility, development cost, and single-supply voltage operation are more stringent requirements than high density and high performance.

In 1995, Flash's cost/MB had already become smaller than DRAM's, and additional improvements are still to be expected because of its high scalability. Furthermore, due to the demand for ever higher Flash memory densities (also in embedded applications such as smart cards), the multilevel charge storage (MLCS) option has recently gained considerable interest. The MLCS principle is based on the relatively high stability of the charge level that can be stored inside the floating gate memory cell in a virtually continuous (or analog) manner. In this way, more than two levels, and hence, more than 1 bit, can be stored inside a single memory transistor. This further increases memory capacity without the need for considerable changes in die size or aggressive technology scaling, hence drastically decreasing the cost/MB even further.

Petitioner Micron Ex-1024, 0030

1.3.4 NOVRAM

The most complex nonvolatile memory device is the *NO*n *V*olatile *RAM* (NOVRAM) in which an EEPROM memory acts as a shadow memory for a static (or dynamic) RAM [1.62–1.64]. Each memory bit, therefore, consists of a RAM memory cell and an EEPROM element, as shown in Fig. 1.26. Some vendors provide nonvolatile RAM memories by using a battery backup included in the chip package. Battery backup NOVRAM parts up to 512 K are presently available. Other manufacturers provide inherent nonvolatility; that is, all data input/output (I/O) occurs through the RAM (thus allowing fast read and write operations). Data can be written into the EEPROM by copying the entire RAM content in parallel within 10 ms, the normal programming time of an EEPROM memory. Both charge-trapping and floating gate type cells have been used. The nonvolatile element can be based on polyoxide [1.64] or on thin oxide Fowler–Nordheim tunneling. This type of nonvolatile memory with a shadow SRAM is available in densities up to 16 Kbit. This density has not been increased, indicating that there is seemingly no need for larger devices, mainly because of its very high cost. Nevertheless, a 64 Kbit

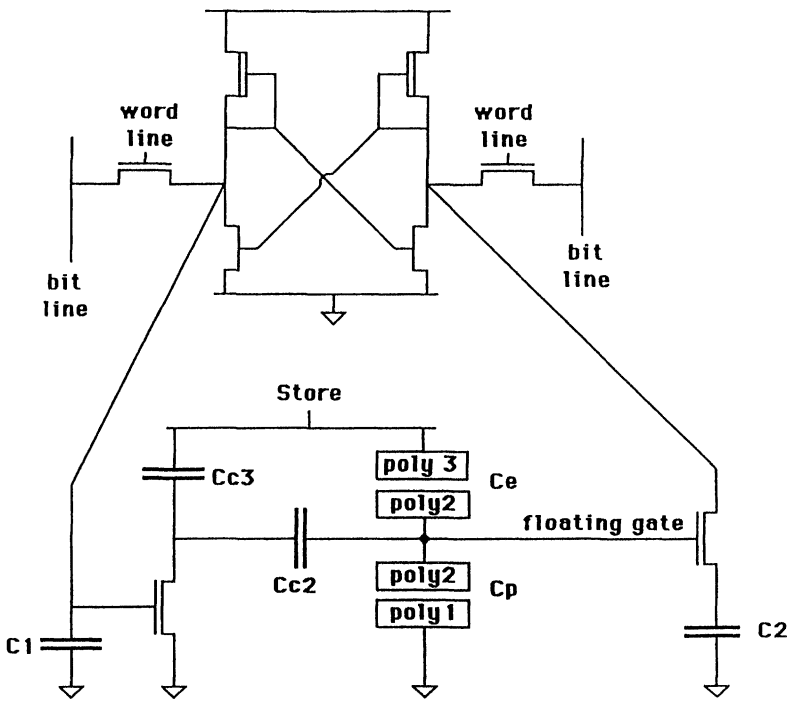


Figure 1.26 Example of a NOVRAM cell [1.64]. The cell couples a static RAM cell with a nonvolatile TPFG memory cell (see also Section 1.4.1.3).

Petitioner Micron Ex-1024, 0031

NOVRAM, making use of two SNOS transistors in combination with a static four-transistor RAM cell, has been reported [1.61]. A nonvolatile DRAM memory has also been reported which uses the combination of a conventional high-density DRAM cell and an EEPROM, leading to a shadow DRAM [1.66–1.68]. This results in a much smaller cell size than is obtained in commercially available products.

1.4. BASIC NVSM DEVICES PRESENTLY IN USE

Following our discussion of different programming mechanisms that have shown feasibility for use in nonvolatile memory cells, and of the different nonvolatile memory products, this section is concerned with the different basic nonvolatile memory devices. As already mentioned, nonvolatile memory devices can be subdivided into two main classes: floating gate devices and charge-trapping devices. Floating gate devices are used for EPROM as well as for EEPROM, Flash EEPROM, and NOVRAM, whereas charge-trapping devices are more suited for EEPROM or NOVRAM applications because of their inherent electrical erasability. The following subsections discuss the different basic floating gate device concepts, charge-trapping devices, and ferroelectric RAM device concepts, and presents a brief comparison between the various classes of devices. This discussion is concerned mainly with the basic configuration and operation principles of the different cells and with some of the most important variations and latest improvements. For a more detailed examination of the various types of cells and of more recent developments, the reader is referred to the corresponding chapters in this book.

1.4.1 Floating Gate Devices

Basically, all floating gate memory cells have the same generic cell structure. They consist of a stacked gate MOS transistor, as was shown in Fig. 1.5*a*. The first gate is the *floating gate*, since it is completely embedded inside the dielectric. The second gate, which is usually referred to as the *control gate*, acts as the external gate of the memory transistor. Between the floating gate and the substrate, and between the floating gate and the control gate, a dielectric layer is provided for isolating the floating gate from the external nodes. These dielectric layers can be oxides, nitrides, oxynitrides, or stacked layers of oxide and nitride (ONO). Furthermore, special features are provided in order to implement the selected programming mechanism inside the cell. In fact, the differences between the various classes of floating gate devices are based on the programming mechanisms that are used for writing or erasing the cell.

When channel hot-electron injection is used as the programming mechanism, the cell is referred to as a SIMOS (*S*tacked gate *I*njection *MOS*), and it is used mainly for EPROM purposes. When the programming mechanism is Fowler–Nordheim tunneling, the cell is often called FLOTOX (*FLO*ating gate *T*hin *OX*ide), which is used primarily in EEPROM, Flash EEPROM, and NOVRAM applications. When polyoxide conduction is used for writing and erasing the mem-

Petitioner Micron Ex-1024, 0032

ory, the cell is called TPFG (*T*extured *P*oly *F*loating *G*ate), which is used in both EEPROM and NOVRAM applications. Finally, there are cells in which a combination of two programming mechanisms is used to write or erase the device. In the following paragraphs, these different classes of nonvolatile memory cells are briefly discussed.

1.4.1.1 SIMOS (EPROM, FLASH EEPROM). The SIMOS (*S*tacked gate *I*njection *MOS*) cell is the n-channel version of the previously discussed SAMOS cell [1.20]. It consists of a double-polysilicon stacked gate device, as shown in Fig. 1.27, and is the basic cell configuration for almost all EPROM memories.

The SAMOS device, discussed in Section 1.1, is a p-channel device using drain avalanche electron injection as the programming mechanism, and consequently, a programmed device would behave as a normally-on device. Proper operation in a memory array, therefore, implies inclusion of an additional select transistor in each memory cell.

This is no longer the case for an n-channel transistor. Its threshold voltage increases due to electron injection, and it offers higher electron mobilities in comparison to p-channel devices. Avalanching the drain of an n-MOS transistor, however, only yields hole injection, which is even less efficient than drain avalanche electron injection in a p-MOS device, due to a higher oxide barrier for hole injection and a larger hole-trapping probability in silicon oxide.

Therefore, in the SIMOS device, channel hot-electron injection is used as the programming mechanism. However, as discussed in Section 1.2, this mechanism is inefficient, and therefore, programming the device is very power consuming, which has prevented the realization of 5 V-only devices. Therefore, all EPROM products require an external supply voltage, which is typically 12 V for programming, and the eventual use of on-chip high-voltage multipliers is excluded. Since channel hot-electron injection is only capable of putting electrons onto the floating gate, UV light is used for erasure.

Typical conditions during operation of the SIMOS cell are shown in Table 1.1. During the read operation, the control gate of the device, which is connected to the wordline of the memory array, is brought to V_{cc} (5 V), while the drain, connected to the bitline of the array, is held at 1 to 2 V, and the source is grounded. If the cell is programmed (high-threshold voltage), no current is detected, whereas an erased cell

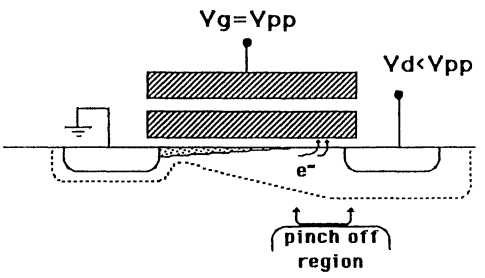


Figure 1.27 The SIMOS cell (Stacked-gate Injection MOS) is the n-channel version of the SAMOS cell [1.20]. Programming occurs through channel hot-electron injection. Erasing is done by UV irradiation or polyoxide conduction. The cell is used mainly for EPROM applications.

Petitioner Micron Ex-1024, 0033

TABLE 1.1 SIMOS OPERATION CONDITIONS

	V_{cg}	V_d	V_s
Read	V_{cc} (5 V)	2 V	GND
Write	V_{pp} (12 V)	8–9 V	GND

will conduct a high current. During programming, the control gate or wordline is brought to the programming voltage, V_{pp}, which is typically 12 V, while the drain is held at 8 to 9 V and the source is again grounded. It is important to notice that the reading and programming configurations are the same. The difference lies only in the voltage levels. The programming gate voltage can be generated on-chip since it is not consuming any current. The programming drain voltage, however, has to be supplied externally. It is also important to mention that selection of the bit during programming is automatically performed by raising the control-line and the bitline. Cells that are connected to the same wordline but to a different bitline, as well as cells connected to the same bitline but to a different wordline, will not be programmed. Consequently, no additional select transistor is needed, and a minimum-size one-transistor cell can be used.

The SIMOS cell has been and, in fact, still is the workhorse of almost all EPROM memories available on the market. Since it became clear that channel hot-electron (CHE) programmable transistors are to be used in high-density EPROMs, technology has been adapted in order to yield fast-programmable, high-density floating gate EPROM structures. The efficiency of the channel hot-electron programming process is largely dependent on substrate (and drain junction) doping level, effective channel length, and floating gate–drain junction overlap. High density can be achieved by using the self-aligned double-polysilicon stacked gate structure [1.69], in which the drain-source junctions are self-aligned with respect to the floating gate, and in which the double-polysilicon stacked gate structure is etched in one step, usually by means of an anisotropic dry etching process [1.69].

Although the conventional self-aligned stacked gate double-polysilicon transistor is still used at the 4 Mbit density level, several new architectures are reported to allow a further decrease in the effective cell size. A first approach is the contactless self-aligned EPROM cell [1.70, 1.71] which consists of a cross-point array configuration defined by continuous buried n^+ diffusions (forming the bitlines) and WSi_2/Poly control gate wordlines. Metal is used to contact the bitline every sixteenth wordline in order to reduce bitline resistance [1.72]. In fact, this is only one example of the use of a virtual ground array. Figure 1.28 shows this virtual ground array architecture in which the common ground line in the array, as well as the drain contact in each memory cell, is eliminated. This technique has been introduced to obtain small cell sizes [1.73, 1.74]. The array architecture relies on the use of asymmetrical floating gate transistors [1.73, 1.74], or on proper source and drain decoding [1.70].

Petitioner Micron Ex-1024, 0034

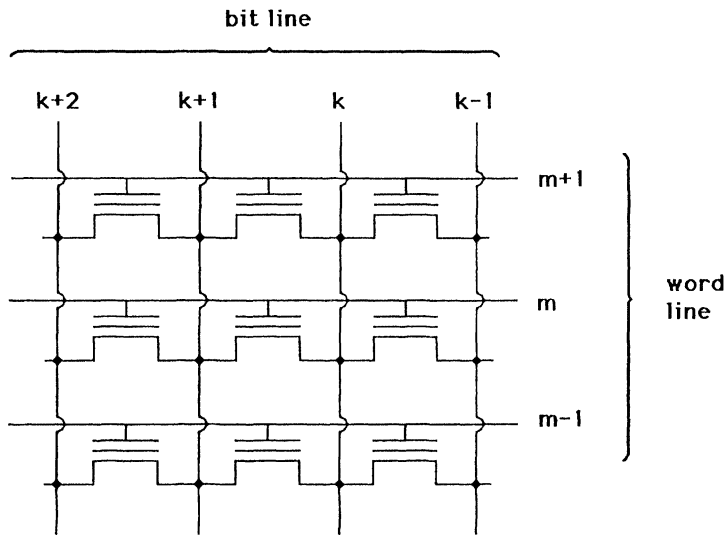


Figure 1.28 The virtual ground array architecture [1.70, 1.73, 1.74]. In this architecture, the common ground line in the array, as well as the drain contact in each memory cell, is eliminated.

Many new device structures, all relying on the original SIMOS concept, have been presented. These variations generally serve one common purpose—namely, to increase the injection current without increasing the programming voltage. The injection of hot electrons is a very inefficient process inasmuch as a proper biasing condition for hot-electron generation does not go together with a favorable condition for injection into the oxide. In practice, very high drain and gate voltages are needed, and the injection efficiency is very low. Alternatives should be based on creating a high hot-electron generating electric field in the channel, and simultaneously, a favorable oxide injection field at the site of the hot-electron generation. Several solutions that meet these requirements have been proposed.

One possibility is the use of a split gate EPROM cell [1.73, 1.75]. In this cell, shown in Fig. 1.29, the series transistor ensures a high immunity to drain turn-on, which otherwise can constitute a serious problem, and eventually, can put a limit on the minimum effective channel length of the EPROM transistor [1.76]. The series transistor also eliminates source-drain punch-through problems. This allows the use of a very small length for the floating gate, thus realizing a high programming speed and a high read current [1.73].

Alternative concepts rely on the source-side injection mechanism described in Section 1.2.4. In the dual gate structure [1.22], shown in Fig. 1.30*a*, a strong potential drop is induced in the center of the channel where neither of the gates is controlling the channel potential. The injection occurs at this site. The injection efficiency can be increased from 10^{-7}, for conventional hot-electron injection, to 10^{-3} for this cell.

Petitioner Micron Ex-1024, 0035

Figure 1.29 The split gate EPROM cell [1.73, 1.75]. The series transistor, incorporated in the memory transistor, allows the use of a floating gate with minimal effective length.

Another cell, based on source-side injection [1.23], is shown in Fig. 1.30*b*. It uses a side-wall gate and a conventional stacked gate structure. Under the spacer oxide between the side-wall gate and the stacked gate, a weak gate control region is formed. This creates a high channel field, located near the source, where the oxide field is highly favorable for injection. The injection efficiency of this cell is on the order of 10^{-5} to 10^{-6}. Other alternatives are the side-wall floating gate structure [1.77], shown in Fig. 1.30*c*, the trench gate–oxide structure [1.78], shown in Fig. 1.30*d*, and the focused ion-beam implantation cell [1.79], shown in Fig. 1.30*e*.

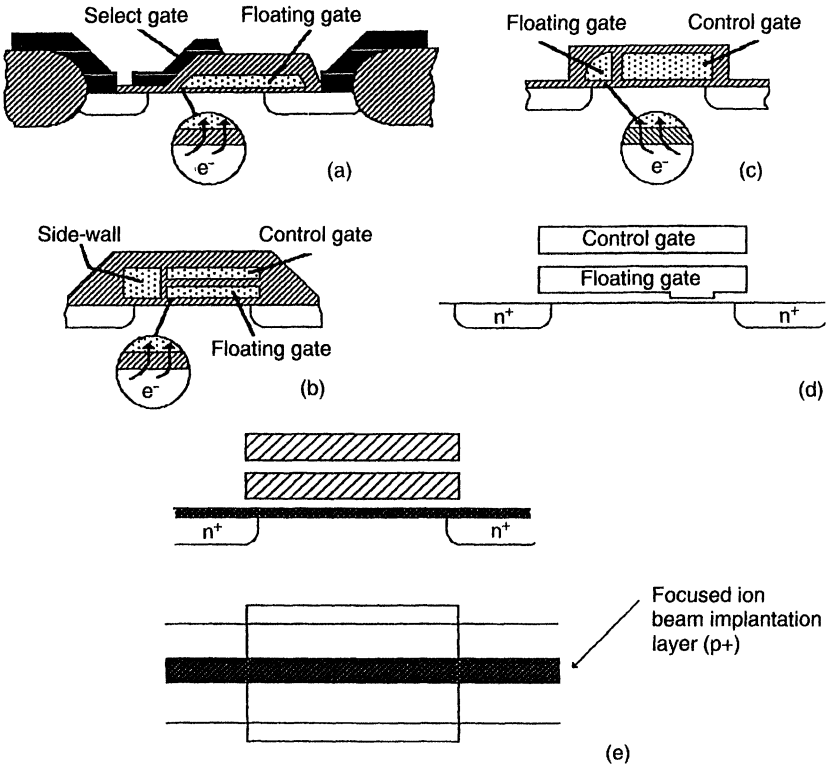


Figure 1.30 Five alternative EPROM cell structures for increased injection efficiency (*a*) PACMOS cell [1.22], (*b*) side-wall floating gate cell [1.23], (*c*) source-side injection cell [1.77], (*d*) trench–gate–oxide structure [1.78], and (*e*) focused ion beam implanted cell [1.79].

Petitioner Micron Ex-1024, 0036

Although the SIMOS cell has been used mainly in EPROM devices, the hot-electron injection mechanism is also used in most Flash EEPROM devices. The cells that are used for these applications are primarily combinations of the SIMOS and the FLOTOX cell (to be discussed in the next section), in either a split gate or a stacked-cell configuration. This is discussed further in Section 1.4.1.4.

1.4.1.2 FLOTOX (EEPROM, Flash EEPROM, NOVRAM). The first nonvolatile memory device relying on Fowler–Nordheim tunneling for both writing and erasing was proposed by Harari et al. in 1978 [1.62], and was used in a nonvolatile RAM cell. This device incorporates a small, thin oxide region over the drain and has, in fact, exactly the same structure as the FLOTOX (*FLO*ating gate *T*hin *OX*ide) device, shown in Fig. 1.31. The FLOTOX approach [1.5] relies on Fowler–Nordheim tunneling through a thin oxide (8–10 nm) for both programming and erasure, and was introduced in 1980 as an EEPROM memory transistor [1.5].

In order to increase the threshold voltage of the cell, a high voltage is applied at the control gate of the device (typical 14 V), while source, drain, and substrate are grounded. This high voltage is capacitively coupled to the floating gate, by which the necessary high field appears across the thin oxide, leading to tunneling of electrons from the drain to the floating gate. When the threshold voltage has to be decreased, a large voltage is applied at the drain, while the control gate and substrate are grounded and the source is left open. By grounding the control gate and substrate, the floating gate is capacitively coupled to near ground, and again, a high field appears across the thin oxide, this time inducing electrons to tunnel from the floating gate to the drain.

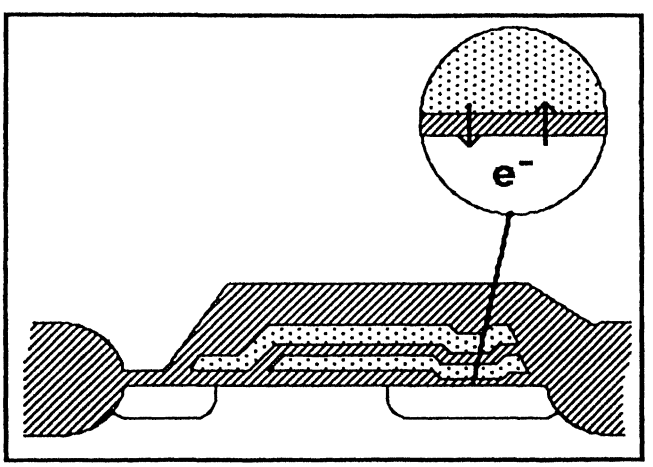


Figure 1.31 Cross section of the FLOTOX device (**FLO**ating gate **T**hin **OX**ide) [1.5]. The device is written and erased by Fowler–Nordheim tunneling of electrons through the thin oxide and is used for EEPROM and Flash EEPROM applications.

Petitioner Micron Ex-1024, 0037

Special attention should be paid when incorporating the FLOTOX cell into a memory array in order not to erase or write a cell on the same wordline or bitline when trying to program another cell. Indeed, unlike the SIMOS cell for which programming involves the application of two voltages (one at the wordline and one at the bitline), which is sufficient to make an appropriate selection of the cell to be programmed, programming of a FLOTOX-type cell involves the use of only one voltage, which is not sufficient to select one single cell. Cells that are connected to the same wordline, but at a different bitline, will also be programmed when the control-line is raised, while cells connected to the same bitline will lose charge from the floating gate when a high voltage is applied to the bitline.

Therefore, since the EEPROM should be byte addressable for read as well as for programming operations, a select transistor becomes necessary, as shown in Fig. 1.32. Each byte has its own erase/write control transistor. First, the control gates of the memory transistors and the select-line of the select transistor are raised to a high voltage while grounding the column lines, as shown in Fig. 1.32*a*. In this way, the floating gates are all charged with electrons. Then, the bits to be programmed are selectively purged of electrons by raising the drain to the programming voltage and grounding the control gates, while the select gates are raised to a high voltage, as shown in Fig. 1.32*b*.

During the read operation, the presence or absence of charge on the floating gate is detected by applying a positive voltage to the control gates and select gates, while biasing the column lines to about 2 V. Like the case of the SIMOS cell, the cells that have electrons on their floating gate do not conduct current, while the cells that are depleted of electrons on the floating gate are in a high conductive state.

For memory cells that use thin oxides, the injection field equals the average oxide field, as discussed in Section 1.2. Consequently, these cells need strong coupling between the floating gate and externally controlled terminals of the device in order to couple the high external voltages onto the floating gate, thereby inducing a high electric field across the thin oxide. The high gate capacitance of the floating gate to substrate transistor can be used for this purpose during the programming operation that lowers the threshold voltage (electrons tunneling off the floating gate). If both programming operations use F–N tunneling, large coupling-capacitance areas between the control gate and the floating gate are necessary.

The FLOTOX cell is used in many commercial products. One of the main reasons why is the low development-entry cost. The only additional process step required in standard double-poly processes is the growth of the thin oxide. It is even possible to realize this type of nonvolatile memory in single-poly processes. Hence, this cell is highly suitable for ASIC and logic applications [1.80, 1.81]. The drawback of the larger cell area is not so important for these applications. Scaling is often difficult because of the complex cell layouts used. To allow the programming voltages to decrease, the tunnel oxides should become even thinner. Thicknesses of 6 nm, however, are the limit for good retention behavior. But these oxides are hard to grow with low defect densities. Yield considerations now limit the oxide thickness to 8 to 10 nm [1.25]. Thinner tunnel oxides imply higher capacitances and thus, larger coupling areas. Really small floating gate tunnel oxide (FLOTOX) cells are, therefore,

Petitioner Micron Ex-1024, 0038

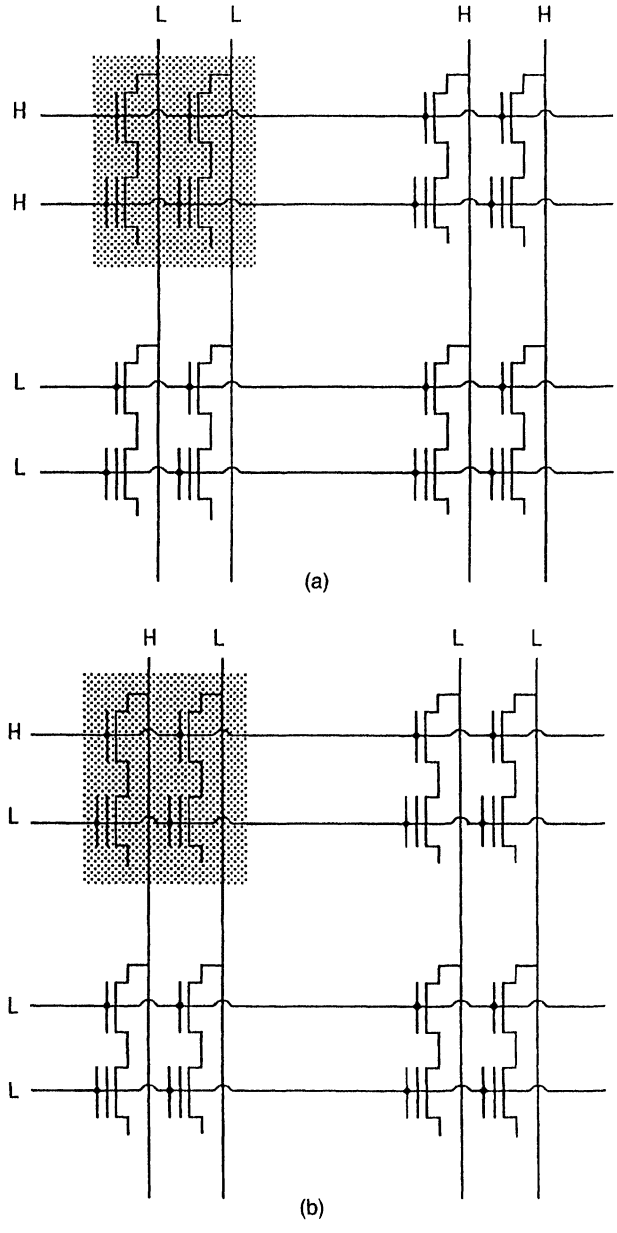


Figure 1.32 Configurations of the voltages in the EEPROM matrix during erase (*a*) and write (*b*) of a memory byte. The shaded parts in the figure are the locations that are being programmed.

hard to achieve. For large memories, however, further scaling and the use of thin oxides become mandatory [1.25, 1.82]. The use of new tunnel materials can obviate some problems. Oxynitrides or nitrided oxides offer better endurance [1.83, 1.84],

Petitioner Micron Ex-1024, 0039

while oxides grown on highly doped injection regions show higher tunnel current conductance [1.85, 1.86].

The advantage of using low power Fowler–Nordheim tunneling for both programming operations is the possibility of using on-chip high-voltage multipliers to generate the programming voltages. In this way, one of the main drawbacks of the SIMOS cell, namely, the need for an external power supply, can be avoided.

For the FLOTOX concept, several variations of the mechanisms themselves, or to the cell design, have been proposed in order to create alternative memory cells with improved performance. The thin oxide can cover the entire channel area of the floating gate transistor, as shown in Fig. 1.33. This type of nonvolatile memory transistor is called a *F*loating gate *E*lectron *T*unneling *MOS*, or FETMOS [1.17]. The device structure is very simple, and to realize it, only a few processing steps are required in addition to a standard *C*omplementary *M*etal *O*xide *S*emiconductor (CMOS) process. This makes small cell sizes impossible. The writing operation is done uniformly over the entire channel area by applying a high voltage at the gate (Fig. 1.33*a*), whereas the erasing operation can be done either uniformly, using a negative gate voltage, while grounding the other electrodes (Fig. 1.33*b*), or locally, at the drain, by grounding the gate and using a high voltage at the drain (Fig. 1.33*c*). The drawback of this structure is the large capacitance of the thin gate oxide, which

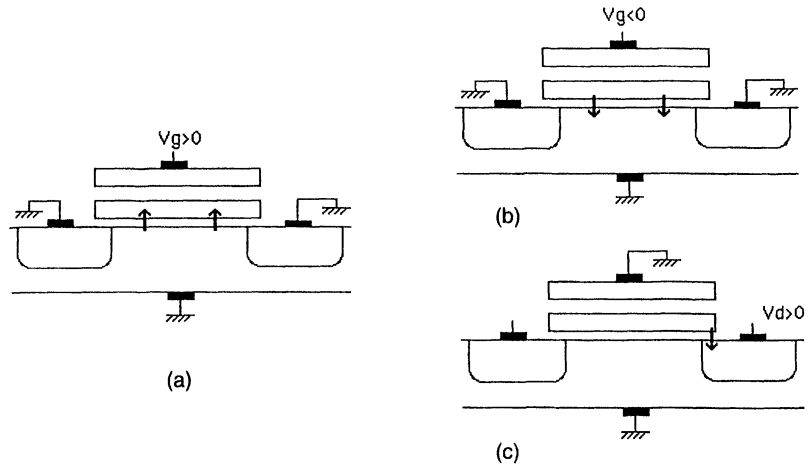


Figure 1.33 Cross section of the FETMOS device (**F**loating gate **E**lectron **T**unneling **MOS**) [1.17]. The thin oxide is covering the entire channel area. (*a*) When the control gate voltage is high, electrons tunnel from the channel to the floating gate. The opposite programming operation can be performed applying either (*b*) a negative voltage at the control gate (uniform) or (*c*) a positive voltage at the drain (nonuniform).

Petitioner Micron Ex-1024, 0040

counteracts the coupling between the control gate and floating gate, and which necessitates the use of large coupling capacitors.

Other variations of the FLOTOX device generally have one common purpose, namely, to increase the injection current without increasing the programming voltage. Whereas some of them are merely new ideas with a questionable chance for future application, others have already been implemented in real memory circuits.

For the tunneling mechanism, several possibilities can be considered to increase the injection efficiency without changing the programming voltage: (1) reducing the oxide thickness, (2) reducing the oxide injection barrier, or (3) increasing the coupling factor between the control and the floating gate without requiring cells with too large an area.

Reducing the tunnel oxide thickness, however, puts severe constraints on device reliability due to direct tunneling leakage problems and the enhanced problem of layer integrity. Reducing the effective tunnel oxide barrier was attempted in the past, for example, by using Si-rich SiO_2 [1.87] or nitrides [1.88]. In fact, the use of textured polyoxide (to be discussed in the next section) also falls within this category. Another possibility is to grow the thin tunnel oxide on a highly doped, n-type silicon substrate. It was reported that, in this way, the effective energy barrier for tunneling could be reduced from about 3.2 eV to 1.8 eV, allowing programming voltages of only 12 V with tunnel oxides of 14 nm [1.85, 1.86].

Increasing the coupling factor without sacrificing too much chip area can also be realized in several ways. One possibility is illustrated in Fig. 1.34 (*S*hielded *S*ubstrate *I*njection *MOS*). The floating gate is shielded from the substrate by the control gate [1.89]. This type of cell only takes about a quarter of the area of the larger cells (FLOTOX). Another version of this idea, the *S*tacked *S*elf-aligned *T*unnel *R*egion (SSTR) cell [1.90], is shown in Fig. 1.35. Another approach to increasing the coupling factor is to replace the coupling capacitor of the conventional stacked gate structure by one that is formed by a tunnel oxide MOS capacitor, as shown in Fig. 1.36 [1.80]. This cell can be realized in a single-poly process.

The NAND (Not AND) structure cell has also been proposed for reducing EEPROM cell size [1.91]. The main disadvantage of this approach is the very long read access time [1.92].

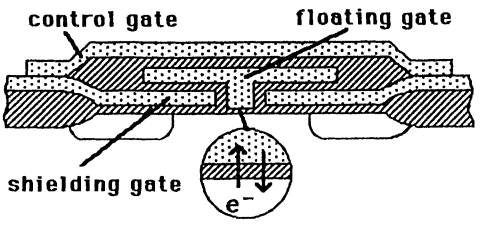


Figure 1.34 Cross section of the Shielded Substrate Injection MOS (SSIMOS) memory cell [1.89]. The floating gate is totally surrounded by the control gate.

Petitioner Micron Ex-1024, 0041

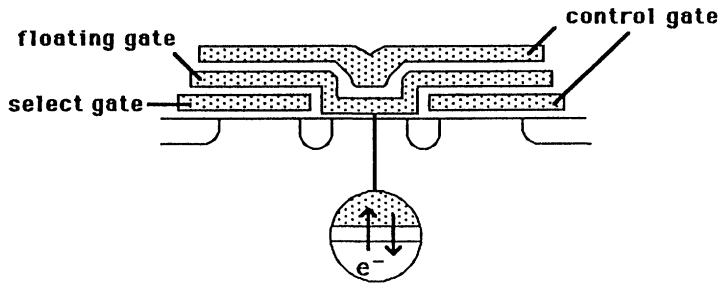


Figure 1.35 Cross section of the Stacked Self-aligned Tunnel Region (SSTR) cell [1.90]. The control gate shields the floating gate from the substrate, and the select transistor is incorporated in the cell.

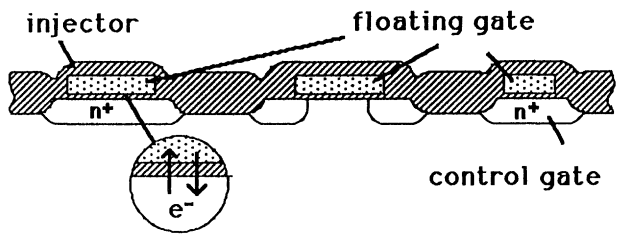


Figure 1.36 Cross section of the single-poly floating gate transistor [1.80]. The control gate is formed by a n+ doped region, which is coupled to the floating gate through a thin oxide region.

1.4.1.3 TPFG (EEPROM, NOVRAM). As discussed in Section 1.2, programming operations can also be based on tunneling through oxides grown thermally on polysilicon. These oxides feature current conduction at lower average oxide fields due to field enhancement at the asperities on the polysilicon surface. This allows the use of much thicker oxides for the same externally applied voltages during programming. Since injection will be enhanced in just one direction, a memory transistor, relying on this enhanced tunnel current for both programming operations, must incorporate two distinct injection regions.

A triple-poly nonvolatile memory transistor, called the *T*extured *P*oly *F*loating *G*ate (TPFG) transistor, was first reported in 1979 [1.18] when it was used in a nonvolatile RAM, and in 1980, as the key element of an EEPROM [1.19]. The device structure is shown in Fig. 1.37. Some other implementations of this enhanced

Petitioner Micron Ex-1024, 0042

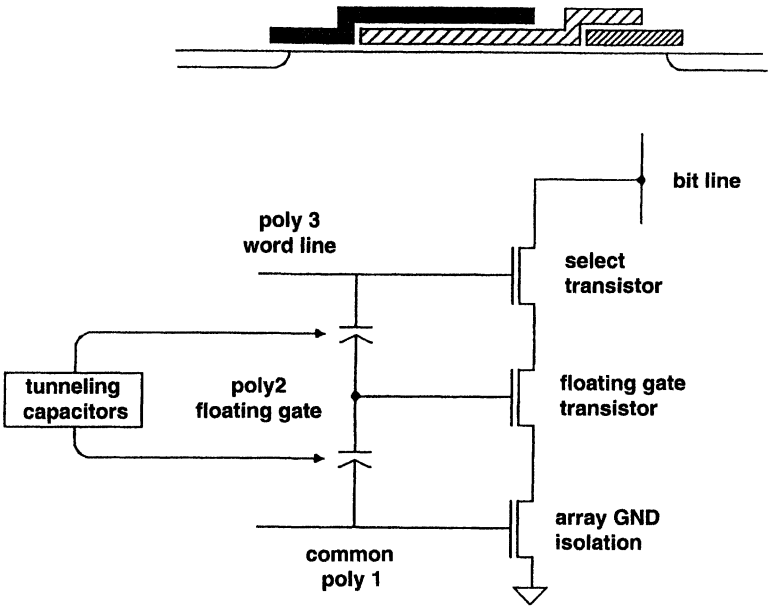


Figure 1.37 Cross section of the Textured Poly Floating Gate (TPFG) device [1.18, 1.19]. A select transistor is incorporated; writing and erasing the device is performed by means of Fowler–Nordheim tunneling through the polyoxide in two different areas. Therefore, a triple-polysilicon device is needed.

tunnel–current principle have been proposed. In the DEIS (*D*ual *E*lectron *I*njection *S*tructure) [1.87], tunnel–current enhancement was achieved by using Si-rich SiO_2 layers. Another approach is based on an improved technology for depositing polysilicon layers, resulting in symmetrical enhanced tunnel–current characteristics for both injection directions in polyoxide [1.93].

TPFG devices are used in large-density memory circuits. Again, as in the case of FLOTOX cells, no external power supply is needed, and on-chip high-voltage multiplier circuitry can be used. The main difficulty in this approach is the growth of polyoxides with desired interface features (shape and size of the asperities), which determine the injection current characteristics. Wearout features are quite dependent on the quality of the polysilicon–SiO_2 interface. The structure is rather complex. Three polysilicon layers [1.19], or two layers and an additional buried contact [1.94], are needed. The accurate alignment of these layers requires precise lithography. This has delayed use of this cell in ASIC or logic applications. The need for thin polyoxides is not that stringent. The injection current through these oxides is determined primarily by the shape and size of asperities, and not by the oxide thickness [1.95, 1.96]. The smaller number of injection points of a scaled TPFG memory cell can, however, aggravate intrinsic wearout as a result of trapping electrons in the polyoxide, a process known as trap-up.

Petitioner Micron Ex-1024, 0043

1.4.1.4 Combinations. As was previously discussed in Section 1.2, four different mechanisms are presently used to change the amount of charge on a floating gate: Fowler–Nordheim (F–N) tunneling through thin oxides (< 12 nm), enhanced Fowler–Nordheim tunneling through polyoxides, channel hot-electron injection (CHE), and source-side injection (SSI). Of these, only channel hot-electron injection and source-side injection can be used to bring electrons to the floating gate. The SSI mechanism is, in fact, a special case of channel hot-electron injection and will, therefore, be treated together with CHE in the following discussion. Therefore, a total of six combinations of main classes of floating gate cells can be defined.

Figure 1.38 shows these six possible combinations with an indication of how these devices are programmed. Each column corresponds to a mechanism that allows electrons to be brought to the floating gate. These are, from left to right, F–N tunneling, polyoxide conduction, and channel hot-electron injection (CHE or SSI). Each row corresponds to a mechanism that allows electrons to be removed from the floating gate. The upper row is for F–N tunneling, and the lower one is for polyoxide conduction. Table 1.2 summarizes the main advantages and drawbacks of the obtained cells.

Of these six combinations, two have already been discussed, namely, the FLOTOX cell (a) and the TPFG cell (e). The combinations (c) and (f) are, in fact, extensions of the SIMOS structure, where special features have been provided to allow an electrical erasure of the cell in order to obtain a Flash EEPROM cell. In the cell shown in (c), electrons are injected onto the floating gate by channel hot-electron injection, as in the SIMOS case. But in order to be able to remove the electrons from the floating gate electrically, either the gate oxide underneath the floating gate is kept thin, as in the case of the FETMOS cell, or a separate thin oxide is provided above the drain, as in the case of the FLOTOX [1.97]. This concept is the basic cell of most of the present Flash EEPROM technologies and is also referred to as electron tunneling oxide (ETOX).

For the cell shown in (f), again channel hot-electron injection is used for programming, but the electrons can now be removed from the floating gate by polyoxide conduction through the interpoly dielectric [1.98]. Because, for cells (c) and (f), electron injection toward the floating gate is performed by channel hot-electron injection, no large coupling areas between the control and floating gate are necessary. This leads to an electrically erasable cell but with a much smaller cell size. Therefore, these cells have been utilized primarily in Flash EEPROM applications [1.98]. The main disadvantage of both cells is still the high programming power, with the necessity of an external power supply, which again excludes the use of on-chip voltage multipliers.

Scaling structures that use channel hot-electron injection can lead to new opportunities for this kind of memory. At small channel lengths, a drain voltage of 5 V can be sufficient to generate hot electrons. Only the gate will then need a higher voltage for programming. Nonvolatile memories with CHE programming could thus operate from a 5 V supply voltage. But the high programming current remains, and therefore, these types of cells are not used in conventional EEPROMs (which are reprogrammed frequently). They are, however, the main cell for Flash EEPROM products.

Petitioner Micron Ex-1024, 0044

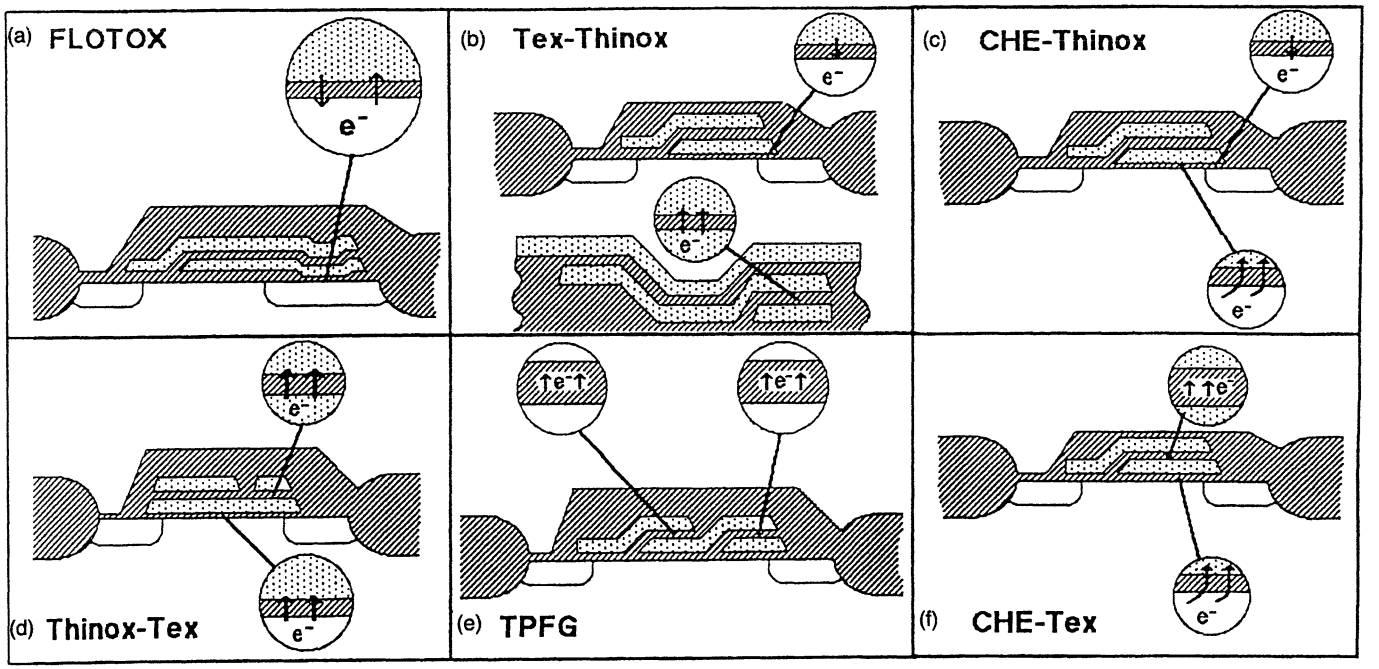


Figure 1.38 Six possible combinations of floating gate memory cells. The three columns represent the three mechanisms to bring electrons to the floating gate (thin oxide Fowler–Nordheim tunneling, polyoxide conduction, hot-electron injection), while the two rows represent the two mechanisms to remove electrons from the floating gate (thin oxide Fowler-Nordheim tunneling and polyoxide conduction). Hot-electron injection cannot be used to remove electrons from the floating gate.

Petitioner Micron Ex-1024, 0045

TABLE 1.2 COMPARISON OF VARIOUS POSSIBLE EEPROM CONCEPTS

	+	−
(a) FLOTOX EEPROM/ASIC/Logic	• compatibility • low development entry cost • possible with 1 poly layer	• large cell • difficult scaling • thinox defect density
(b) Tex-Thinox EEPROM/F-EEPROM	• small cell • easily scalable • low program power	• complex cell • 2 program mechanisms –thinox (defects) –polyoxide (wearout)
(c) CHE-Thinox EEPROM/F-EEPROM/ASIC	• small cell • relatively simple cell • easily scalable	• high program power • 2 program mechanisms –thinox (defects) –CHE (degradation) • ?? endurance ??
(d) Thinox-Tex EEPROM/F-EEPROM	• low program power	• large cell • scaling difficulties • 2 program mechanisms –thinox (defects) –polyoxide (wearout)
(e) TPFG EEPROM/ASIC	• thick oxide • small cell • low program power • direct write • easily scalable	• complex cell • higher program voltage –trap-up –window variation • critical tunnel oxide
(f) CHE-Tex EEPROM/EPROM/ASIC	• small cell • thick oxide • easily scalable	• high program power • 2 program mechanisms –polyoxide (wearout) –CHE (degradation)

Petitioner Micron Ex-1024, 0046

Of the six possible floating gate memory cells, two have never been used in commercial products, namely, the Thinox–Tex, shown as (d), and the Tex–Thinox, shown as (b). The Thinox–Tex approach seems to have only drawbacks (cf. Table 1.2). For example, large cell areas are needed for the thin oxide Fowler–Nordheim tunneling used to remove electrons from the floating gate. This also makes the cell difficult to scale. The two different programming mechanisms enhance the technological problems. The Tex–Thinox cell has a rather complex structure. Three polysilicon layers are needed, and the two different programming mechanisms have to be optimized separately. This concept has never been used in commercial products.

1.4.2 Charge-Trapping Devices

1.4.2.1 MNOS and SNOS devices (EEPROM, NOVRAM). MNOS (metal–nitride–oxide–silicon) devices were invented in 1967 [1.2] and were the first electrically alterable semiconductor (EAROM) devices. The nonvolatile function of these devices is based on the storage of charges in discrete traps in the nitride layer. These charges (electrons or holes) are injected from the channel region into the nitride by quantum mechanical tunneling through an ultra-thin oxide (UTO, typically 1.5 to 3 nm). These trapped charges cause a significant shift in the threshold voltage of the transistor [see Eq. (1.2) with Q_T the trapped charge in the nitride layer]. Although over time some of these charges will be lost after programming is completed and will, therefore, result in a gradual decrease of the threshold voltage, the programmed state of the device can typically be maintained for at least 10 years.

By 1975, metal gate, p-channel EAROM with densities up to 8 Kbit were available. They employed a 1 transistor per bit configuration based on the so-called trigate transistor cell concept [1.99]. In this transistor, shown in Fig. 1.39, only the center part of the channel contained the programmable UTO–nitride sandwich structure. At both drain and source, a thicker oxide–nitride sandwich was used, which induced a fixed threshold voltage in the erased state and prevented the device from going into the depletion mode. These memory devices suffered from low-speed, limited-density, inherent read disturbance (sensing the device-required application of a small read voltage to the gate), and the need for 2 to 3 voltage supplies to operate the memory.

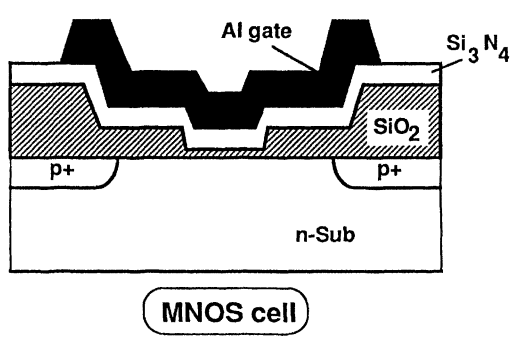


Figure 1.39 Cross section of the p-channel tri-gate MNOS device [1.99]. The thin tunneling oxide (1.5–3 nm) is present only at the center of the channel. At source and drain, a thicker oxide–nitride sandwich acts as a select transistor.

Petitioner Micron Ex-1024, 0047

An important breakthrough was achieved for MNOS in 1980 with the development of the Si-gate n-channel SNOS (silicon–nitride–oxide semiconductor) process [1.100], which resulted in the first 16 Kbit SNOS–EEPROM [1.4]. The reliability of the SNOS technology was based mainly on the use of *L*ow-*P*ressure *C*hemical *V*apor *D*eposition (LPCVD) silicon nitride and a pre-metallization high-temperature hydrogen anneal to improve the quality of the nitride–UTO–silicon interface [1.101–1.104]. A cross-sectional diagram of the SNOS cell, which is still used in today's commercially available 256 Kbit SNOS–EEPROM memories and the announced 1 Mbit parts, is shown in Fig. 1.40. A two transistor per bit configuration is used where the MOS transistor acts as the select device whose implementation has allowed the complete elimination of the problem of read disturbance [1.4]. The SNOS transistor consists of a silicon nitride layer (20–40 nm) on top of the UTO on silicon.

Because the SNOS transistor is, in fact, a two-polarity device, necessitating the application of memory bulk voltages, isolation of the memory bulk from the peripheral circuitry bulk is required. The most common approach is the use of separate p-wells for the peripheral MOS circuits and the memory array. Providing separate wells within the memory array then allows for full byte function. In LPCVD nitrides, net positive and negative charge can be stored in almost equal amounts.

The programming of the cell is as follows: during the write operation, a high (positive) voltage is applied to the gate with the well grounded. Electrons tunnel from the silicon conduction band into the silicon nitride conduction band through the modified Fowler–Nordheim tunneling process discussed in Section 1.2.5 and are trapped in the nitride traps, resulting in a positive threshold voltage shift. Erasing is achieved by grounding the gate and applying a high (positive) voltage to the well. This induces direct tunneling of holes from the silicon valence band into the nitride valence band, or the nitride traps [1.105, 1.106], resulting in a negative threshold voltage. During the off-state, the gate is grounded and the select transistor is required for proper operation within the array. Reading of the cell is accomplished by addressing the cell through the select transistor and by sensing the state of the SNOS transistor. Although the gate is grounded, the charge content within the nitride will be modified in time primarily by backtunneling charges through the UTO.

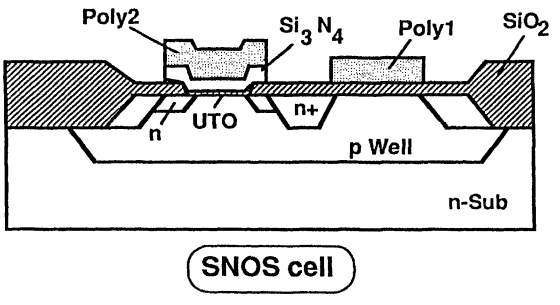


Figure 1.40 Cross section of the two-transistor n-channel SNOS memory cell [1.4, 1.100] consisting of a MOS select transistor and a SNOS memory transistor, both located in a p-well that allows full byte programmability.

Petitioner Micron Ex-1024, 0048

Hagiwara et al. [1.106] showed that the integrity of nitride layers can be guaranteed to thicknesses below 20 nm. Scaling of the devices, which must be carried out in parallel with the scaling of the peripheral MOS transistors, is straightforward up to memory densities of 1 Mbit. Yatsuda et al [1.105–1.107] proposed a scaling scheme for SNOS, which is based on a reduction in the dielectric layer thicknesses almost in proportion to the program voltage, except for the UTO. In order to conserve a constant programming time, the UTO thickness has to be reduced slightly. Minami et al. [1.108] showed that the written-state retentivity of these conventional SNOS devices is improved when the nitride thickness is further reduced. For a 1 Mbit SNOS–EEPROM, a nitride thickness of about 20 nm and a UTO thickness of 1.6 nm are used. The programming voltage is about 10 volts.

Hole injection from the gate will, however, limit the memory window [1.105, 1.109–1.111], a problem that becomes more severe for thinner nitride layers. An efficient way to cope with this problem is described in the next section (1.4.2.2). Another solution is the use of a silicon oxynitride layer instead of a pure silicon nitride layer [1.112, 1.113]. Although these layers require slightly larger programming voltages because of their higher energy barriers, it has been shown that retention and endurance properties of SNOS parts using oxynitrides with a moderate oxygen content are markedly better than their nitride counterparts. In particular, the improved endurance characteristics point to a significant reduction in gate injection. The optimum composition for the oxynitride layer has been found to be $[O]/([O] + [N]) \approx 0.17$ [1.113].

SNOS memories have two features that are worth mentioning and that have made this technology the first choice in military and space applications requiring nonvolatility. The first one is their inherent radiation hardness, which is discussed in Section 1.7. The second feature is the ability of SNOS devices to be adjusted to the envisaged application: very slow programming (1–100 ms) for long nonvolatile retention (years, EEPROM function) or fast programming (1–10 μs) for limited nonvolatile retention (hours, days, NOVRAM function) [1.114].

1.4.2.2 SONOS Devices. A reduction of the injection from the gate can be ensured by providing a thin oxide (2–3 nm) on top of the nitride, yielding the so-called SONOS (*S*ilicon–*O*xide–*N*itride–*O*xide *S*emiconductor) device [1.115], as illustrated in Fig. 1.41*a* (SONOS1). This oxide can be formed by steam oxidation of the nitride at the expense of the nitride thickness or by deposition. The blocking efficiency of this top layer has been proven for both types of oxide [1.105, 1.115, 1.116]. However, when thinning the nitride below 20 nm, another problem arises. It is known that the trapping length in nitrides is larger for holes than for electrons, that is, 15 to 20 nm for holes [1.117, 1.118] and 5 to 10 nm for electrons [1.118, 1.119]. If the nitride is reduced in thickness, holes will be trapped close to the gate electrode and will be mostly lost through the gate electrode, even in the presence of this thin oxide. This results in a significant reduction of the threshold voltage in the erased state.

Therefore, further scaling of the SONOS device for higher density memories and lower programming voltages requires a new concept. This concept was first

Petitioner Micron Ex-1024, 0049

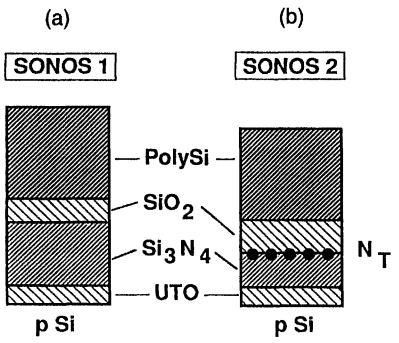


Figure 1.41 Schematical representation of the gate dielectric for two types of SONOS devices. In order to prevent hole injection from the gate, an oxide layer is added on top of the nitride layer. (*a*) The SONOS1 configuration [1.115] consists of a thin oxide (1–2 nm), a thicker nitride, and again a thin oxide (2–3 nm). (*b*) The SONOS2 configuration [1.120–1.122] consists of a thin oxide (1–2 nm), a thin nitride (< 10 nm), and a thicker oxide layer (> 3 nm).

proposed by Suzuki et al. [1.120] and was refined by others [1.121, 1.122]. This new SONOS concept is shown schematically in Fig. 1.41*b* (SONOS2). It consists of a UTO of the same thickness as before, a thin nitride layer (< 10 nm), and a thicker top oxide layer (> 3 nm). The aim of the top oxide is not only to inhibit gate injection, but also to block the charges injected from the silicon at the top oxide–nitride interface, resulting in a higher trapping efficiency and thus, a reduction in the problem related to nitride layer reduction. In this way, the total thickness of the insulator structure can be reduced, and consequently, the programming voltage can be reduced.

This device shows additional advantages. First, large oxygen-related electron trap densities are obtained at the nitride–top oxide interface due to the oxidation of the nitride [1.123]. This results in a larger memory window in spite of the decreased nitride thickness. For a constant top oxide layer thickness, this would eventually make the threshold of the written state independent of the nitride thickness [1.122]. Next, if pinholes are present in the thinner nitride layer, they can be filled with oxide afterward, during oxidation of the nitride. Finally, the retention and degradation behavior are improved, as is discussed in Section 1.6. Low-voltage operation, down to 5 V, has been demonstrated [1.124] for a nitride of 3 nm thickness and a blocking oxide thickness of 5.5 nm. Although optimization of the process and structure is still required, the application of this SONOS cell concept has allowed realization of memories with densities in the Mbit range.

1.4.3 Ferroelectric Devices

A new type of nonvolatile memory is emerging whose operation is based on the ferroelectric effect [1.125]. Certain crystalline materials show the tendency to polarize spontaneously under the influence of an external field and to remain polarized after the external field is removed. The polarization can simply be reversed by applying a field of opposite polarity. As such, a bistable nonvolatile capacitor is obtained in which stored information is based on polarization state rather than on stored charge.

The data stored in a capacitor can be read by sensing the interaction of a "read field" with the polarization state of the element. If a read voltage is applied to the ferroelectric capacitor of polarity opposite to the previous write voltage, the polar-

Petitioner Micron Ex-1024, 0050

ization state will switch, giving rise to a large displacement charge that can be sensed by proper circuitry.

The ferroelectric material used in nonvolatile memory applications is a lead–zirconate–titanate compound ($Pb[Zr,Ti]O_3$, PZT), which is a perovskite-type ceramic. Different configurations can be envisioned for a nonvolatile RAM which uses the polarizable medium as the storage element. These configurations can be divided into "backup"-type memories or "primary storage"-type memories.

For the first type, the configuration is, in fact, similar to that described for NOVRAMs in Section 1.3.4. It consists of a DRAM or SRAM configuration for which each memory element has a shadow ferroelectric capacitor backup cell. Only upon power failure or after an intentional store signal is the information present in the RAM transferred to the backup nonvolatile element. The cell itself does not affect the RAM performance during normal operation. When power comes up, or after a recall command, information stored in the ferroelectric capacitor is destructively read out and stored in the corresponding RAM cell. The advantages of this type of NOVRAM over the conventional concept discussed previously are first, the high programming speed of the ferroelectric elements, which allows a very fast transfer of the data content from the volatile to the nonvolatile part (typically well below 200 ns), and second, the high density that can be achieved since the nonvolatile capacitors are built above the conventional memory circuitry.

In order to achieve high-density, nonvolatile RAMs, however, the ferroelectric capacitor should be used as the primary storage element in an advanced DRAM-type configuration in which a single transistor and the ferroelectric capacitor make up the cell. This configuration is referred to as the *F*erroelectric *RAM*, or FRAM, and is the first true nonvolatile read/write memory. The FRAM configuration no longer needs a store or power-failure detection, but each write and access cycle is directed toward the capacitor. However, since each read operation is destructive and implies a rewrite, the FRAM concept can become successful only if very high endurance (more than 10^{15}) is assured and if program times are small enough, that is, below 50 ns. Figure 1.42 shows a schematic of a 2 capacitor/bit FRAM configuration. The memory bit consists of a wordline (WL) controlling two pass transistors, a

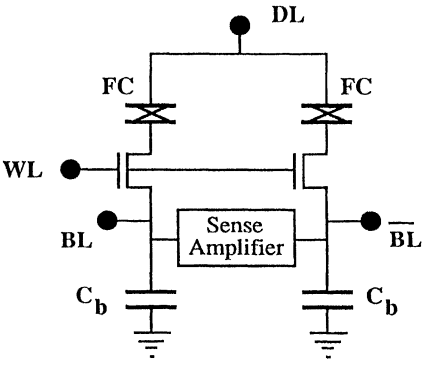


Figure 1.42 Schematic of a 2 capacitor/bit FRAM (Ferroelectric **RAM**) configuration.

Petitioner Micron Ex-1024, 0051

bitline (BL), a $\overline{\text{bit}}$line ($\overline{\text{BL}}$) to collect charge from the capacitors, and a driveline (DL) to drive the capacitors. A sense amplifier connects both bitlines.

The combination of fast write and read with nonvolatility, high-density, simple cell structure, better endurance, and radiation hardness makes this approach highly promising.

1.4.4 Comparison of the Floating Gate, Charge-Trapping, and Ferroelectric Devices

As the previous sections have made clear, the different nonvolatile technologies and approaches have their merits and drawbacks. The progress made in the physical understanding of the different programming mechanisms, the mastering of the technology, and the capability to adapt the technology and device to a specific application enable any technology or approach to be used or engineered for almost any application, provided sufficient effort is put into the development. However, other criteria, such as development-entry cost, compatibility with standard technology, experience with the technology, and environmental requirements, greatly influence and determine the selection of a particular technology or approach for a specific nonvolatile memory application. In Table 1.3, technologies or approaches (SNOS, TPFG, FLOTOX, Flash EEPROM, FRAM) are compared against a number of criteria.

TABLE 1.3 COMPARISON OF NONVOLATILE APPROACHES

Criteria	SNOS	TPFG	FLOTOX	Flash	FRAM
Scaled cell size	+	+	−	++	++
Voltage scaling	++	+	−	+	++
Complexity of technology	H	H	L	L	H
Compatibility	o	+	++	++	o
Complexity of cell	L	H	M	M	L
Retention	+	++	++	++	++
Endurance	+	+	+	o	++
Radiation hardness	++	−	−	−	++
Development entry cost	H	H	L	L	H

++: very good; +: good; o: medium; −: poor
H: high; M: medium; L: low

1.5. BASIC NVSM DEVICE EQUATIONS AND MODELS

Before discussing the most important nonvolatile memory device characteristics, special attention should be paid to some aspects that are typical for the understanding and modeling of floating gate memory devices. Indeed, apart from the dual dielectric layer that is used in charge-trapping devices, there are no fundamental

Petitioner Micron Ex-1024, 0052

differences between these devices and conventional MOSFETs from the point of view of typical MOSFET characteristics (I_d-V_g, I_d-V_d, etc.). For the modeling of their memory behavior, some basic considerations are made in Section 1.5.5. For floating gate devices, however, the presence of a floating, but conductive, gate inside the gate dielectric has some important consequences from a device modeling point of view. These consequences are the subject of the present section. First, the capacitor model is discussed, defining the important coupling factors for a floating gate cell. Then, the influence on I-V characteristics of a floating gate cell is examined, followed by the experimental procedures for determining the model parameters and the basic equations for modeling the memory behavior of floating gate (FG) devices.

1.5.1 The Capacitor Model

Obviously, in a floating gate memory cell, the floating gate itself cannot be accessed. Its voltage is controlled through capacitive-coupling with the external nodes of the device. Often, the floating gate transistor is modeled by a capacitor equivalent circuit [1.126–1.135] called the capacitor model.

In this model, shown in Fig. 1.43, all capacitors present in a typical double-poly floating gate transistor are represented. C_k is the total capacitance between the control and floating gates, while C_s and C_d are the floating gate source and drain capacitance, respectively. In the SIMOS device, C_d is determined by the floating gate–drain overlap, while, in the FLOTOX device, it is dominated by the thin oxide injection region capacitance. C_g and C_f are the floating gate channel and field region capacitance, respectively. C_t is then defined as the total capacitance of the floating gate: $C_t = C_k + C_d + C_s + C_g + C_f$.

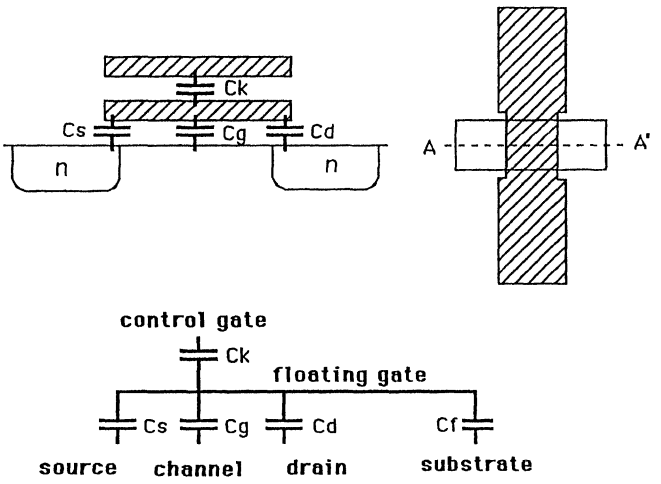


Figure 1.43 The capacitor model showing the various capacitances of the floating gate to the external nodes.

Petitioner Micron Ex-1024, 0053

Two important coupling ratios can now be defined: k, the control gate coupling factor, and d, the drain-coupling factor:

$$k = \frac{C_k}{C_t} \qquad d = \frac{C_d}{C_t} \tag{1.19}$$

These capacitances determine the fraction of the control gate and the drain voltage, respectively, that is capacitively-coupled to the floating gate. Therefore, they are important parameters in the design, modeling, and study of floating gate memory cells.

A capacitance that is often omitted in determining the coupling factors is the coupling capacitor C_f between the floating gate and the substrate through the field oxide. In a floating gate transistor, this capacitor can be of importance because of the large coupling area between the floating gate and the control gate located over the field oxide as, for example, in the case of a FLOTOX cell.

In order to estimate the error that is made by neglecting this field capacitance in the capacitor model, assume that t_k and t_f represent the effective oxide thickness of the coupling oxide and the field oxide, respectively, and k' and d' are the coupling ratios if the field oxide capacitor C_f is neglected. The correct coupling factors, k and d, can then be expressed as

$$\begin{cases} k = \dfrac{k'}{1 + k'\dfrac{t_k}{t_f}} \\ d = \dfrac{d'}{1 + k'\dfrac{t_k}{t_f}} \end{cases} \tag{1.20}$$

It can be concluded that, by neglecting the field oxide capacitance, there is an error in both d and k of the same relative importance. This error is larger if the coupling factor k' is large and if the ratio t_k/t_f is large. The error is the same for FLOTOX and SIMOS devices: the gate oxide thickness has no influence. Table 1.4 shows some examples that illustrate the error that can be made by neglecting the field oxide capacitance for the case when t_k/t_f is 1/12.

In most cases, the capacitor values are estimated by using a simple parallel-plate model. However, fringing capacitances should also be taken into account [1.128, 1.129]. These capacitances are due to coupling between the different terminals at the edges of the polysilicon layers and to the fringing fields existing in the substrate of the device. In thin oxide devices, normally the parallel-plate approximation is fairly good, and these fringing effects cause a deviation of only a few percentage points.

The capacitor model can be used to calculate the potential of the floating gate if the voltages at the external nodes and the charge on the floating gate are known. Once the floating gate potential is known, it can then be fitted into the conventional MOS models or equations as a replacement for the conventional gate voltage in order to describe the conventional MOS characteristics. Conversely, the capacitor model is also used to calculate the model parameters, such as the control gate and drain-coupling factors from the measured MOS characteristics.

Petitioner Micron Ex-1024, 0054

TABLE 1.4 INFLUENCE OF THE FIELD CAPACITOR ON THE COUPLING FACTORS

$\frac{t_k}{t_f} = \frac{1}{12}$			
k′ [%]	d′ [%]	k [%]	d [%]
50	10	48	9.60
60	10	57	9.52
70	10	66	9.45
80	10	75	9.38

In most cases, this procedure is applied under the assumption that the model's capacitors are formed by identical ideal conductors. In reality, however, the capacitor model, as depicted in Fig. 1.43, can only account for the electrostatics in the floating gate transistor. Therefore, the capacitor model, by itself, can lead to wrong conclusions. Indeed, in a MOS transistor, distinction should be made between the externally applied voltages and the internal electrostatic potential. In this section, the correct equations will be used, that is, the electrostatic potentials instead of the externally applied voltages.

Figure 1.44 shows an energy band diagram for a floating gate transistor at the onset of inversion—that is, when the band bending at the silicon surface equals twice the Fermipotential of the substrate. This situation clearly defines the threshold voltage as measured at the control gate.

The threshold voltage of the floating gate–oxide–substrate transistor is given by V_{t0}:

$$V_{t0} = \phi_{ms} + 2\phi_F - \frac{Q_{ox}}{C_{ox}} - \frac{Q_d}{C_{ox}} \tag{1.21}$$

where ϕ_F = the Fermipotential of the substrate (which is positive for the p-type substrate)

ϕ_{ms} = the work function difference between the gate material and the bulk material

Q_{ox} = the equivalent fixed oxide charge, located at the oxide substrate interface

Q_d = the charge in the depletion layer

First, it has to be noted that the influence of fixed oxide charges is taken into account by means of an equivalent oxide charge Q_{ox} located at the oxide–substrate interface. In order to account for the influence of this charge as sensed at the control gate, the exact distribution of the charges within the oxide must be known. Thus, charges located at the floating gate–oxide interface have no influence at all on the threshold voltage of the floating gate–oxide–substrate transistor, but will certainly be detected at the control gate. Since the exact distribution of the charges within the

Petitioner Micron Ex-1024, 0055

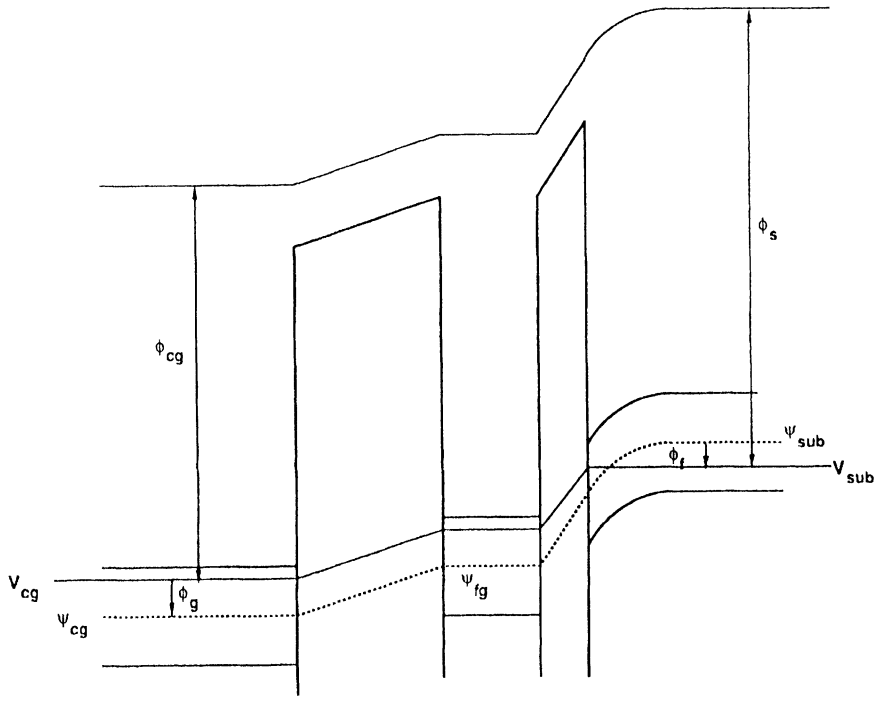


Figure 1.44 Energy band diagram for a floating gate transistor at the onset of inversion (with no charge on the floating gate).

oxide is not known, the assumption is made that the oxide-trapped charge is located at the oxide–substrate interface.

To define the threshold voltage at the control gate, the capacitor model [1.126] can be used to make the charge balance of the floating gate, using the electrostatic potential Ψ in the various regions of the cell (referred to as E_i; see Fig. 1.44):

$$Q_{fg} = (\Psi_{fg} - \Psi_{cg})\, C_k + (\Psi_{fg} - \Psi_{sub})\, C_g + (\Psi_{fg} - \Psi_d)\, C_d + (\Psi_{fg} - \Psi_f)\, C_f \qquad (1.22)$$

or

$$\Psi_{fg} = \frac{C_k}{C_t}\,\Psi_{cg} + \frac{C_g}{C_t}\,\Psi_{sub} + \frac{Q_{fg}}{C_t} + \frac{C_d}{C_t}\,\Psi_d + \frac{C_f}{C_t}\,\Psi_f \qquad (1.23)$$

with Q_{fg} being the charge on the floating gate. By taking $V_{sub} = 0$ as the reference, and with Eq. (1.19), this yields the floating gate voltage, V_{fg}:

$$V_{fg} + \phi_g = kV_{cg} + dV_d + \frac{Q_{fg}}{C_t} + k\,\phi_g + d\,\phi_d + \frac{C_g}{C_t}\,\phi_f - \frac{C_f}{C_t}\,\phi_f \qquad (1.24)$$

It is assumed that the substrate underneath the field capacitor is in accumulation, and thus, $\Psi_f = \phi_f$. Taking into account that the drain is n^+-doped like the gate,

Petitioner Micron Ex-1024, 0056

and thus, $\phi_g \approx \phi_d$, the threshold voltage at the control gate can be expressed as the control gate voltage for which $V_{fg} = V_{tO}$:

$$V_{tcg} = \frac{V_{tO}}{k} - \frac{Q_{fg}}{C_k} - \frac{d}{k}V_d - \frac{C_g}{C_k}(\phi_{ms} + 2\,\Phi_f) + \frac{C_f}{C_k}\,\phi_f \tag{1.25}$$

In most reports, only the first three terms of this formula are used, based on the capacitor model. The last two terms are usually omitted [1.127–1.132]. The fourth term accounts for the difference in work function between the p-doped substrate and the n^+-doped polysilicon (ϕ_{ms}) and band bending in the substrate ($2\phi_f$). The last term accounts for the influence of the field capacitor. Equation (1.25) is derived for the case of an n^+-doped polysilicon gate. If other gate materials are used (e.g., silicides), the fourth term of Eq. (1.25) will change due to a different work function.

The error introduced when neglecting the last terms in Eq. (1.25) can be significant. If we try to measure the coupling factor, k, by just dividing the measured threshold voltages at the floating gate and the control gate, as was proposed in [1.129] and based on the capacitor model, the resulting coupling factor for n-MOS floating gate devices with n^+-doped polysilicon gates is always too small. The error becomes relatively more important for floating gate transistors that have small coupling factors (as used in EPROM devices). This is shown in Table 1.5 where the ratio V_{tO}/V_{tcg} [Eq. (1.25)] is compared to the real coupling factor, k.

TABLE 1.5 COMPARISON OF k AND $\frac{V_{t0}}{V_{tcg}}$

$$V_{t0} = \phi_{ms} + 2.\phi_f - \frac{Q_{ox}}{C_{ox}} + \frac{Q_d}{C_{ox}}$$
$$= -0.9 + 0.6 + 1.1 = 0.8\,V$$

$d = 0.05 \quad k = 0.9.k'$

k	V_{tcg}	$\frac{V_{t0}}{V_{tcg}}$
0.5	1.87	0.43
0.6	1.51	0.53
0.7	1.25	0.64
0.8	1.06	0.75

1.5.2 I-V Characteristics of Floating Gate Devices

The floating gate forms an equipotential plane between the control gate and the substrate and is parallel to both of them. This kind of equipotential plane does not exist in an ordinary MOS transistor, and thus, there are some important consequences when examining the I-V characteristics of a floating gate memory cell [1.131]. Indeed, as discussed in the previous section, the floating gate voltage is determined by capacitive-coupling between the floating gate and the externally

Petitioner Micron Ex-1024, 0057

applied voltages, more specifically, the control gate and drain voltages. From the definitions of the coupling factors, the influence of external voltages on the I-V characteristics can be calculated using the capacitor model.

When calculating the I-V characteristics of the floating gate transistor, we can start from the I-V characteristics of the floating gate–oxide substrate transistor and substitute the floating gate voltage using Eq. (1.24). In the linear region, the current is calculated as

$$\begin{aligned} I_{ds} &= \frac{\mu C_{ox} W}{L}\left[V_{fg} - V_{tO} - \frac{V_d}{2}\right]V_d \\ &= \frac{\mu C_{ox} W}{L}\left[V_{cg} - V_{tcg} - \frac{V_d}{2k}\right]V_d \end{aligned} \tag{1.26}$$

In the saturation region, the current is

$$\begin{aligned} I_{ds} &= \frac{\mu C_{ox} W}{2L}(V_{fg} - V_{tO})^2 \\ &= k^2 \frac{\mu C_{ox} W}{2L}[V_{cg} - V_{tcg}]^2 \end{aligned} \tag{1.27}$$

In both formulas, which actually are valid for long channel transistors, of course, the correct value of V_{tcg}, taking into account the influence of the applied drain voltage [Eq. (1.25)], must be used.

As an example, in Fig. 1.45, the output characteristics of a conventional MOS transistor are compared to those of an equivalent floating gate transistor. The characteristics differ in three ways: the threshold voltage of the floating gate transistor is higher, while the transconductance and the output resistance are lower. The first two effects are due to capacitive-coupling between the control gate and the floating gate. The threshold voltage of the floating gate transistor is given by Eq. (1.25) and is, therefore, roughly a factor of 1/k higher than that of the corresponding MOS device if there is no charge on the floating gate. The transconductance decreases by the factor k, as given by Eqs. (1.26) and (1.27).

The third distortion is the increase of the output current with drain voltage for the floating gate device. This is due to the capacitive-coupling between the drain and the floating gate, and is described by the drain voltage dependent V_{tcg}, as given by Eq. (1.25), in Eqs. (1.26) and (1.27). This effect can even lead to the turn-on of the transistor at high-drain voltages, even if the transistor is off at low-drain voltages.

1.5.3 Experimental Determination of the Coupling Factors k and d

The values of the different capacitors of the model can be calculated from the layout of the cell and the different oxide thicknesses, which can be determined from capacitor measurements. In these calculations, the effect of the field oxide capacitor and the influence of the stray capacitances must be taken into account. Therefore, this calculation will not be very accurate, and deviations from the experimentally obtained values can be as large as 10 to 15% [1.134].

Petitioner Micron Ex-1024, 0058

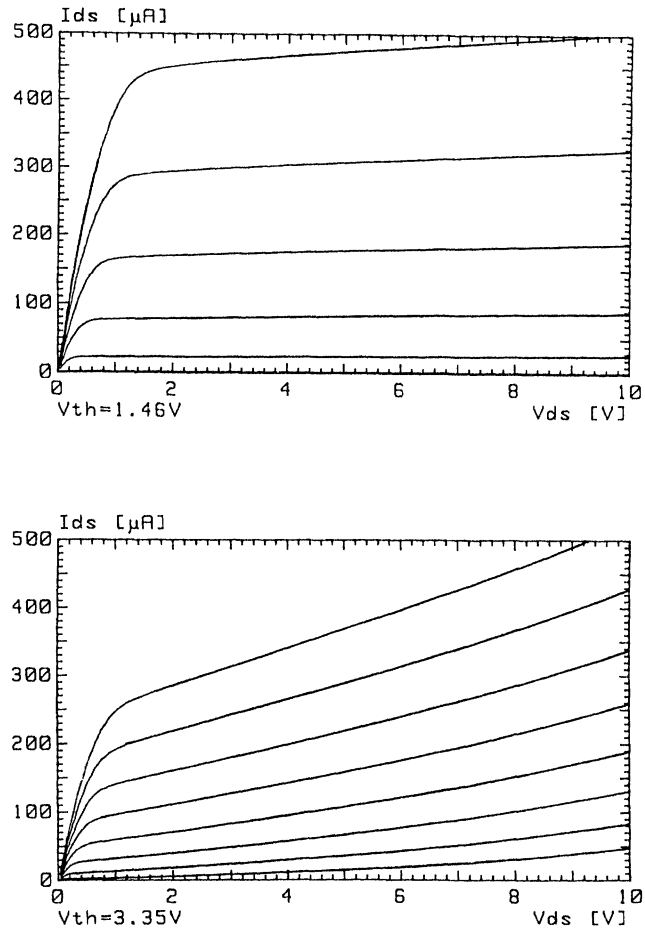


Figure 1.45 Comparison of the $I_{ds} - V_{ds}$ output characteristics of a conventional MOS transistor (upper) and its equivalent floating gate transistor with the same MOS geometry (lower). Upper figure: $V_{gs} = 2\,V$, $2.5\,V$, $3\,V$, $3.5\,V$, and $4\,V$. Lower figure: $V_{gs} = 3.5\,V$, $4\,V$, $4.5\,V$, $5\,V$, $5.5\,V$, $6\,V$, $6.5\,V$, and $7\,V$.

The value of the capacitors can be measured directly on the floating gate structures and on contacted floating gate structures [1.128]. The problem in this case is that the capacitance measurements on the small floating gate transistors are very difficult and limited in accuracy by the accuracy of the measurement equipment and setup.

A third method of determining the coupling factors is to compare the threshold voltage value of the floating gate device, V_{tcg} (measured at the control gate), to that of the equivalent contacted floating gate, V_{tO} (measured at the contacted floating gate). Apart from the uncertainty about charges stored on the floating gate, the conventionally used formulas [1.129] are also not valid because they only account

Petitioner Micron Ex-1024, 0059

for electrostatics in the structure, as was discussed previously. The use of the correct formulas introduces other unknown parameters, and the determination of the coupling factors again is not accurate.

The control gate coupling factor can also be determined from a comparison of the subthreshold I-V characteristics of floating gate devices and contacted floating gate devices [1.129]. But, as shown in reference [1.126], this will always result in a coupling factor k that is larger than the real one.

Another method is based on the comparison of the I-V characteristics of floating gate devices and contacted floating gate devices [1.128]. Equations (1.25–1.27) indeed show that the coupling factors k and d can be calculated from the ratio of the currents and from the dependence of the threshold voltage V_{tcg} on the drain voltage. This measurement also directly incorporates the influence of the field oxide capacitor. The most accurate methods for determining coupling factors therefore rely on the comparison of I-V characteristics of floating gate transistors and their contacted floating gate counterparts.

As an example, Fig. 1.46 illustrates the determination of k in both the linear (upper figure) and saturation (lower figure) region, where I_{ds}-V_{gs} curves are compared for a floating gate transistor and its equivalent MOS counterpart. In order to obtain an accurate result for k, it is mandatory that the structures indeed be identical. Not only must the layout of the transistors themselves be the same, but also the contacts made at the different terminals of the transistor and, in particular, the substrate contact, must be identical for the two structures that are being compared. Any difference in layout can cause different series resistances for the terminals of the transistors and can give rise to inaccuracies. If equivalent current ranges are considered and the same fitting method is used for the results from both the contacted floating gate transistor and the memory structure, the determined coupling factor is weakly dependent on these elements. For this example, k is found to be 0.487 in the linear region and 0.532 in the saturation region. As indicated in [1.128], the measured coupling factor k is dependent on the applied drain voltage by the dependence of the gate capacitance C_g on this drain voltage. The value of C_g is indeed smaller when the transistor operates in saturation than when it operates in the linear regime. The coupling factor k will, therefore, always be larger in the saturation regime than in the linear region.

The drain-coupling factor can be calculated from the dependence of the threshold voltage V_{tcg} on the drain voltage using Eq. (1.25). In principle, this dependence can be measured in both the linear and saturation regimes. But the range of drain voltages allowed to operate the transistor in the linear regime is so small that the drain-coupling factor cannot be determined accurately. Therefore, it is preferable that the I-V characteristics for different drain voltages be measured in the saturation regime and that the V_{tcg} values extracted from these results be used to calculate the drain-coupling factor d.

Figure 1.47 shows an example of the determination of the drain-coupling factor. In the upper figure, the $\sqrt{I_{ds}} - V_{gs}$ characteristics are shown, measured for several drain voltages in saturation (between 4 V and 10 V). It becomes clear from these curves, as expected from Eq. (1.27), that the transconductance remains constant, independent of the drain voltage, but that the threshold voltage decreased with drain

Petitioner Micron Ex-1024, 0060

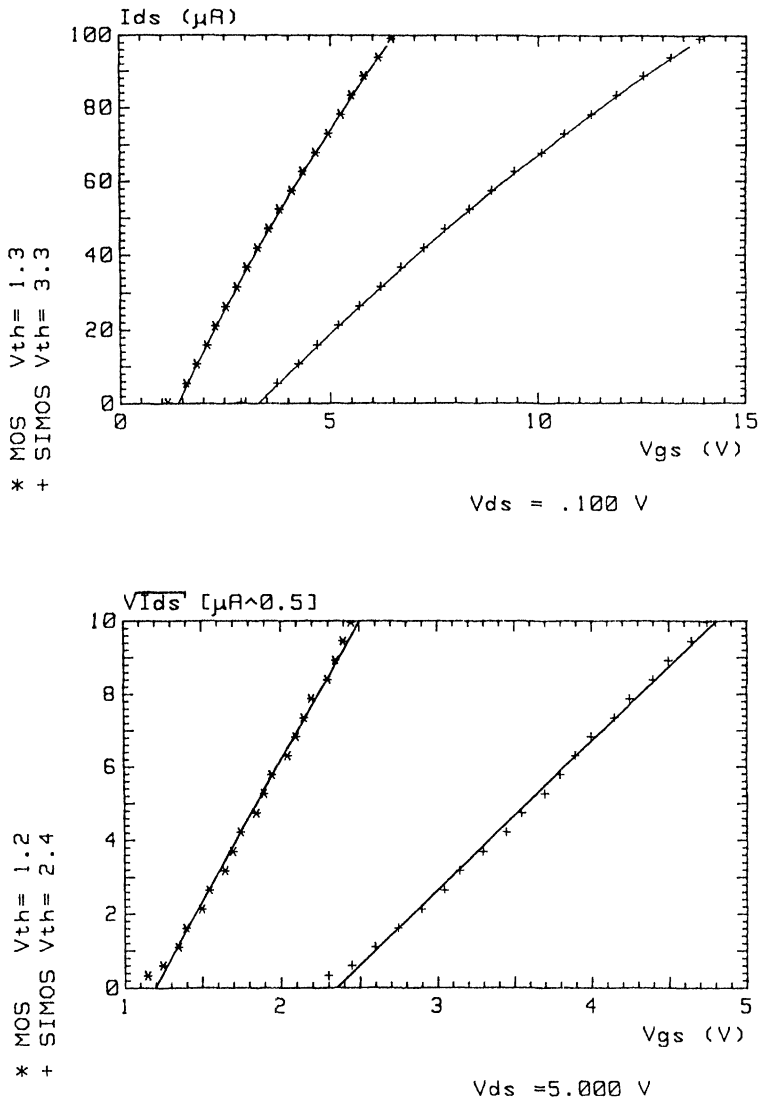


Figure 1.46 Determination of the control gate coupling factor k in the linear (upper) and in the saturation region (lower) from comparison of the $I_{ds} - V_{gs}$ characteristics of the floating gate device and its equivalent MOS counterpart.

voltage, as expected from Eq. (1.25). From Eq. (1.25), it can be concluded that the threshold voltage decreases linearly with drain voltage, with a slope of d/k. The extrapolated threshold voltage is plotted as a function of the drain voltage (lower figure), yielding a straight line. The slope of this line yields d/k, which is found to be 0.119 in this case.

Petitioner Micron Ex-1024, 0061

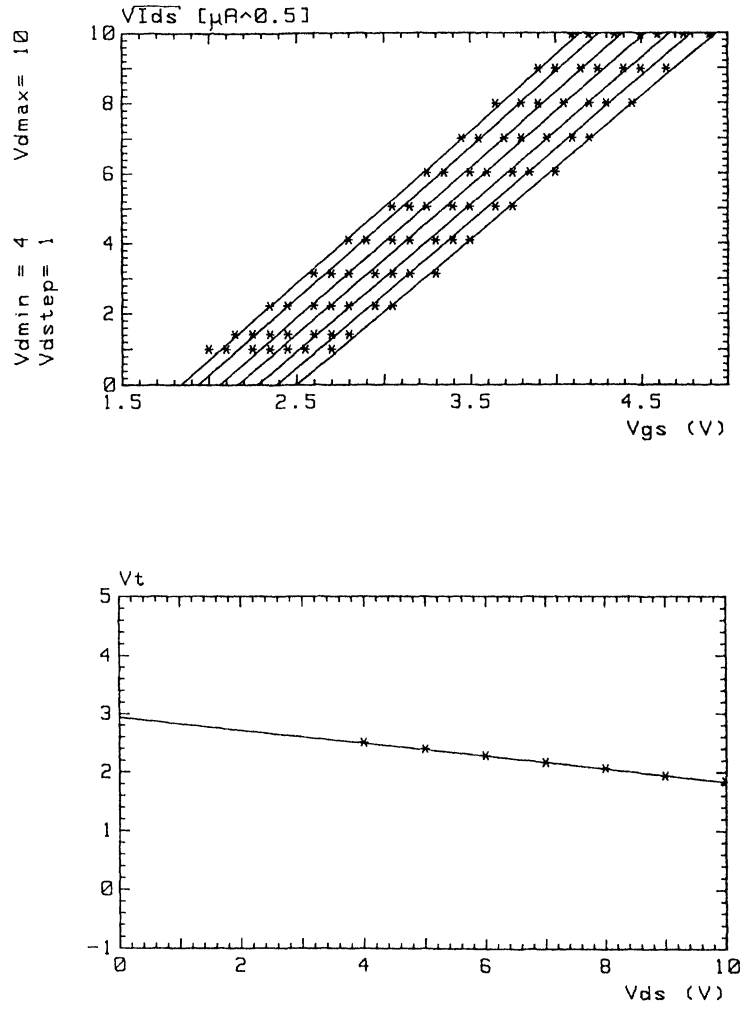


Figure 1.47 Determination of the drain-coupling factor d from the relationship between V_{tcg} and V_{ds}. The upper figure shows the influence of the drain voltage on the $\sqrt{I_{ds}} - V_{gs}$ characteristics. The lower figure shows the extrapolated threshold voltage as a function of V_{ds} yielding a straight line with slope d/k.

1.5.4 Modeling of the Memory Characteristics of FG Cells

Besides the use of the capacitor model and the I-V characteristics of floating gate devices for design and analysis purposes, they can also be applied to model the memory behavior of the cell. This model allows the calculation of the programming

Petitioner Micron Ex-1024, 0062

or transient characteristics, as well as the retention characteristics, which are discussed in Section 1.6. If a degradation model such as, for example, an electron trapping and generation or interface trap generation model is used, the endurance characteristics can also be calculated and predicted [1.135]. A comparison of measured and calculated characteristics reveals the validity of the assumptions made concerning the physical mechanisms governing the programming or degradation behavior. In this section, the basic equations that are needed for modeling the memory behavior of FG cells are discussed. The models themselves are strongly dependent on the type of cell (programming mechanism, geometry, and design) and are not treated here.

When charge is injected to or emitted from the floating gate, the floating gate potential will change, as given by Eq. (1.23). Due to the change of the floating gate potential, the electric fields in the surrounding oxides change as well, by which the injected currents increase or decrease. This process continues until a steady state is reached. Thus, it is important to remember that the injection currents during a programming operation are not constant, but change rather rapidly as a function of time. The modeling of the memory behavior of floating gate cells is, therefore, based on the following elements.

1. The basic memory model starts from the expression for the charge on the floating gate:

$$\frac{dQ_{fg}(t)}{dt} = \int_{A_{fg}} J_{fg}(t)dA \tag{1.28}$$

In this expression, the integral over the complete floating gate area can usually be replaced by a summation over the different oxides or dielectrics, "i", that surround the floating gate:

$$\frac{dQ_{fg}(t)}{dt} = \sum_{A_i} J_i(t) \tag{1.29}$$

where A_i is the injection area to the floating gate. The currents, J_i, through the different dielectrics are a function of time because the electric fields, E_i, in the dielectrics are changing with time. By integrating Eq. (1.29) with respect to time, an expression for the charge that is accumulated on the floating gate is obtained:

$$Q_{fg}(t) = \int_0^t \sum_{A_i} J_i[E_i(t)]dt \tag{1.30}$$

2. This expression cannot be solved as such because the electric fields, E_i, are dependent on the floating gate potential, and thus, also on the floating gate charge, $Q_{fg}(t)$. Therefore, a model is needed to allow the calculation of the fields occurring inside the device during programming. For floating gate transistors, this model is the capacitor model for the device described in Section 1.5.1.

Petitioner Micron Ex-1024, 0063

The relation between the floating gate potential and the charge on the floating gate is given by Eq. (1.24), which is often simplified by neglecting the correction factors due to the work function differences and the Fermipotentials:

$$V_{fg}(t) = k\ V_{cg}(t) + d\ V_d(t) + \frac{Q_{fg}(t)}{C_t} \tag{1.31}$$

where $V_{cg}(t)$ and $V_d(t)$ are the control gate and drain voltages, respectively, as a function of time during charge injection (e.g., programming). A more general expression can be obtained by also incorporating the other external potentials that are able to couple voltages to the floating gate such as, for example, the source voltage, V_s, and the channel potential, V_{ch}, together with their respective coupling factors. In most cases, however, these nodes are grounded, so they will not be considered further here.

From V_{fg} and the external potentials V_{cg}, V_d, V_s, and V_{ch}, the electrical fields, E_i, inside the respective dielectrics surrounding the floating gate can be calculated.

3. The next step is to implement a model for charge injection into the gate insulator under programming conditions based on the above calculated electrical fields. These models have been described in Section 1.2. As was discussed there, for some injection mechanisms like Fowler–Nordheim tunneling, a closed form expression for the current as a function of the electrical field exists. If the externally applied voltages are also time independent, Eq. (1.30) can be solved analytically. For other injection mechanisms, however, such as hot-electron injection or polyoxide conduction, as well as for time-dependent external voltages, no analytical expressions can be obtained. For these cases, Eq. (1.30) must be solved numerically.
4. If Eq. (1.30) has been solved, the external threshold voltage of the transistor cell can easily be found using expression (1.25), again given by the capacitor model. This expression links the threshold voltage of the nonvolatile memory transistor to the amount of charge stored on the floating gate. If the correction factors due to work function differences and Fermipotentials are neglected, this yields:

$$V_{tcg}(t) = \frac{V_{tO}}{k} - \frac{Q_{fg}(t)}{C_k} - \frac{d}{k}\ V_d \tag{1.32}$$

where, in this case, V_d is the drain voltage applied during sensing (i.e., reading) of the contents of the memory transistor.

This expression has to be used in combination with a clear definition of the threshold voltage as employed in the measurement of the transient programming characteristics. A commonly used definition is the voltage applied at the externally accessible gate of the transistor in order to allow a predefined current to flow from drain to source in the memory transistor. This imposes the problem that the I–V characteristics of the memory transistor have to be known. In most cases, the expression for the threshold voltage of

Petitioner Micron Ex-1024, 0064

the nonvolatile memory transistor is based on a physical criterion (like channel inversion) that must be related to the threshold voltage definition based on drain-source currents. In many cases, this distinction is not clearly made, but the error introduced by neglecting this difference is normally small since it is threshold voltage shifts that are considered in the first place, and the applied drain voltage during measurement of the threshold voltage and predefined current levels are kept small. If only the shift has to be predicted, Eq. (1.32) simplifies to

$$\Delta V_{tcg}(t) = -\frac{\Delta Q_{fg}(t)}{C_k} \tag{1.33}$$

5. In case of an eventual degradation of the oxide—such as electron trapping or interface trap generation, which will influence the injected currents J_i—an additional model that describes this degradation has to be implemented. This is needed not only to simulate the endurance characteristics but also in other cases since degradation of the injection currents can occur, even within the time needed for one programming operation. This is the case for programming operations based on polyoxide conduction, an injection mechanism that induces rapid degradation due to charge trapping in the polyoxide, as has already been discussed in Section 1.2.2. Numerous degradation models have been proposed in the past, but it is impossible to discuss these models within the focus of this introductory chapter. Instead, it is important to note that most of the degradation models used are based on experiments that were carried out on capacitors, based on either constant voltage or constant current injection experiments. In floating gate cells, however, neither constant voltage nor constant current conditions exist; therefore, care should be taken to extrapolate the results from capacitors to real situations in memory cells. These extrapolations can lead to erroneous models and conclusions concerning the degradation mechanisms that occur in transistor cells.

In principle, with the help of the above-described models, a complete transient programming characteristic (i.e., the shift of the threshold voltage of the nonvolatile memory transistor as a function of programming time during programming) can be calculated. This can be repeated for different programming conditions (e.g., applied voltages or voltage pulse shapes) and for different transistor geometries. In addition, parameters that cannot be measured directly can be calculated. It is important to know the fields occurring across the gate insulator during programming in order to avoid oxide breakdown, as well as the fields across the insulator between the floating gate and control gate in floating gate devices so that the leakage current through this insulator during programming can be calculated. The same equations can be used to predict or calculate the retention or endurance characteristics, provided that the correct physical mechanisms responsible for these characteristics (charge loss or gain mechanisms for retention and degradation mechanisms for endurance) are taken into account by an appropriate model.

Petitioner Micron Ex-1024, 0065

1.5.5 Modeling of the Memory Characteristics of SNOS and SONOS Devices

Modeling of the memory behavior (transient and programming characteristics) of SNOS and SONOS devices is based on numerical integration of the continuity equation which relates the current gradient to the rate of change of the local stored charge content in the nitride layer. The model, therefore, requires knowledge of the injection current mechanisms (at both interfaces and in both energy bands) and an appropriate trap model. Finally, the Poisson equation is used to determine the electric field distributions in all layers. Again, it is not the purpose of this chapter to treat these techniques and results in detail. The advantage of such a rigorous treatment is that charge distributions in the nitride layer can be obtained. The shape of the distribution seriously affects the retention behavior of these devices [1.108, 1.136]. This retention behavior can be computed numerically, as was done by Williams et al. [1.137], Heyns and Maes [1.138, 1.139], and Libsch et al. [1.140], once the trapped charge distribution is known. The effects of reducing the nitride thickness of SNOS or SONOS devices on the retention characteristics [1.108, 1.120] can then be studied and simulated [1.137], and will be useful in the further scaling of these devices. Numerous models have been proposed for the memory traps in silicon nitride. A review of these models is presented in [1.140]. Charge distributions obtained from computations based on two-carrier conduction and two or three trap-level models (electron trap, hole trap, and recombination center) were obtained by Remmerie et al. [1.116], whereas calculations based on two-carrier conduction and a single deep-level amphoteric trap model were presented by Libsch et al. [1.124, 1.140].

1.6. BASIC NVSM MEMORY CHARACTERISTICS

Besides conventional device characteristics, nonvolatile memory cells also have some important functional memory characteristics, which are used to evaluate the memory performance of the cell. These characteristics can be divided into three main classes.

1. The transient characteristics describe the time dependence of the threshold voltage during programming.
2. The endurance characteristics, which are meaningful only for EEPROM cells, give the memory threshold window, which is the difference between the threshold voltages in the written and the erased states, as a function of the number of programming cycles; they are characteristic for the intrinsic number of write/erase cycles that can be endured before both programmed states are no longer distinguishable.
3. Finally, the retention characteristics give the threshold voltage in either programmed state as a function of the time after programming, and indicate the intrinsic ability of the memory cell to retain its content over long periods of time.

Petitioner Micron Ex-1024, 0066

These basic memory characteristics are discussed in more detail in the following sections.

1.6.1 Transient Characteristics

The functioning of nonvolatile memory devices is based on the possibility of bringing charges onto the floating gate or into the gate insulator and removing them again in order to change the threshold voltage of the nonvolatile transistor. In the ideal nonvolatile cell, programming can be performed with externally applied voltages that are as low as possible. However, this programming operation should be as fast as possible.

The transient programming characteristic of a nonvolatile memory transistor is the shifting of the threshold voltage of the transistor as a function of time during programming. The exact knowledge of these characteristics allows the determination of the programming voltages and times needed to obtain a useful threshold voltage window.

Some remarks have been made in the preceding sections concerning the meanings of write, erase, program, and clear. Indeed, some confusing terminology is often used with respect to the different programming operations. In fact, two possible choices can be made:

- In a FLOTOX-type EEPROM memory matrix organization, 8 adjacent bits make up a byte. Due to the connection between these 8 bits (i.e., the sources of the memory transistors are connected), it is impossible for one programming operation to be performed selectively for the bits within one byte. Therefore, a byte in this EEPROM matrix architecture is programmed in two steps: first, the threshold voltages of the 8 bits are all brought to a high level simultaneously, and then some of the bits are selectively written to a low threshold voltage within one byte. Therefore, calling the operation that is performed simultaneously on all bits within one byte, "erase," and the selective operation, "write," can be defended. Then, "erasing" in this case means bringing the threshold voltage to a high (positive) level, while writing means bringing it to a low level. This convention is used in most EEPROM products.
- On the other hand, the output voltage on the data line of a memory device being equal to 0 V corresponds to the low-threshold voltage of the memory cell. Indeed, the memory chip will be designed so that the access time for output high and output low are the same. Because it takes more time for the output buffer to drive the data-line high, one will choose this to correspond to the memory cell threshold voltage that will be detected the quickest by the sense amplifier. For a memory cell with a high-threshold voltage, the bitline voltage (and thus, the input voltage of the sense amplifier) will not change, or will change only slightly, during read-out, and the sense amplifier does not need to switch. Therefore, a high-memory cell threshold voltage corresponds to an output voltage on the data-line $V_{out} = 5$ V, and the operation to achieve

Petitioner Micron Ex-1024, 0067

a high-memory threshold voltage is called "write." This choice also corresponds to the situation for EPROM memories where "programming" by means of hot carriers induces an increase in the threshold voltage. "Erasure" of EPROM memories is done with UV light and brings the threshold voltage low.

Since the transient programming characteristic consists of a shift of the threshold voltages during a programming operation, the transient programming measurement can be performed in two ways.

In the first approach, the programming operation is stopped before the programming is completed, and the threshold voltage is measured. Then, the memory transistor is erased until the initial threshold voltage is reached again. The programming operation is repeated under the same conditions as before, but with a longer time before the programming is halted. By repeating this procedure, a complete transient characteristic can be recorded. The drawback of this method of measurement is that it implicitly assumes that neither the programming operation nor the erase operation introduces any degradation. This degradation would then be superimposed on the real transient programming characteristic. A simple check on whether this assumption is valid consists of recording the transient programming characteristic twice: both transient programming characteristics should coincide. An advantage of this measurement method is that, in contrast with the second method, it can easily be implemented for different voltage pulse shapes.

A second measurement method for the transient programming characteristic divides the total considered programming time into short time periods. The programming voltages are applied during each small time period, and in between, the threshold voltage is measured. The simplest example is programming with constant voltages. During the measurement, voltage pulses of short duration are applied and the threshold voltage is measured after every pulse. In the transient characteristic, it is the cumulative effective programming time that is put on the x-axis. For this method, the assumption is made that the frequent interruption of the programming operation (in most cases, the application and disconnection of the programming voltages) has no influence on the results. This assumption can easily be checked by repeating the transient measurement where, this time, the programming conditions have to be applied during the whole programming time without interruption. Again, the two results should coincide. A drawback of this measurement method is that it cannot easily be implemented in the case where shaped programming voltage pulses are used.

A special sort of programming characteristic is the so-called soft-write. This terminology is used primarily for EPROM devices and is related to the conditions that occur at the different memory cells in a memory circuit during the programming of the whole memory. The soft-write characteristic describes the shift in threshold voltage of one cell under conditions that are present when another cell is being programmed. For EPROM (or Flash EEPROM) memories, it is mandatory that all cells of the circuit be written without disturbing the information

Petitioner Micron Ex-1024, 0068

content of any previously written cell. For all recent EEPROM circuits, the danger for soft-write does not exist because of the use of an isolating select transistor per memory cell.

Another special case of programming is the "read-disturb." During read-out of the information, voltages have to be applied to the nonvolatile memory transistor. These voltages can induce a threshold voltage shift in the cell that is addressed, as well as in some other cells. Read-disturb is mainly a concern for EPROM (and Flash EEPROM) memories for the same reason as noted above.

As an example, Fig. 1.48 shows the programming (both write and erase) characteristics of a FLOTOX-type nonvolatile memory transistor. For more details on the transient characteristics of the various cells, the reader is referred to the other chapters of this book.

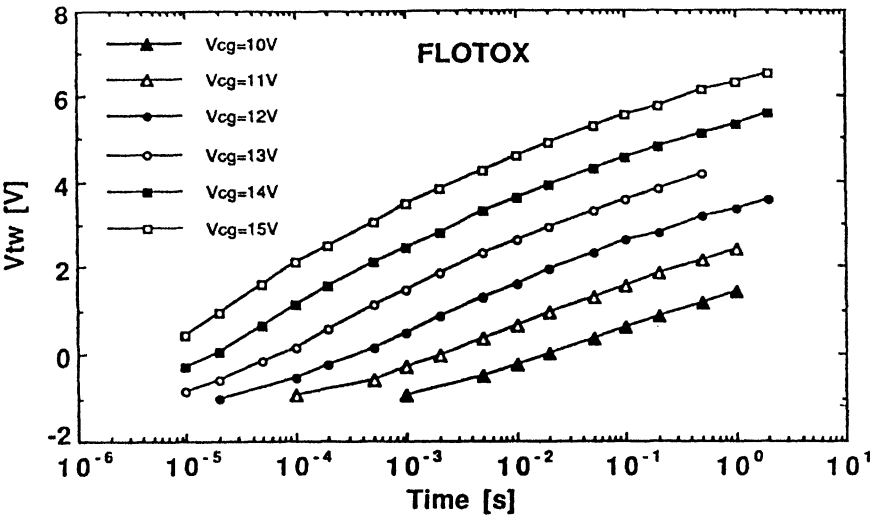
1.6.2 Endurance Characteristics

With respect to overall reliability, two different features of the nonvolatile memory have to be considered. Nonvolatile memories can be reprogrammed frequently, but, in contrast to RAM memories, each write operation introduces some sort of permanent damage. This implies that the total number of write operations is limited; for example, most commercially available EEPROM products are guaranteed to withstand, at most, 10^4 programming cycles. The damaging of the memory cell during cycling is normally referred to as "degradation" and the number of cycles the memory can withstand is normally called its "endurance." Another failure mode of the nonvolatile memory is retention failure. This failure mode is discussed in the next section.

1.6.2.1 Floating Gate Devices. The program/erase endurance of floating gate devices is determined by four phenomena: tunnel oxide breakdown, gate oxide breakdown, trap-up, and degradation of the sense transistor characteristics. Whereas the first two are self-explanatory, trap-up is defined as the trapping of electrons in the oxide during programming operations. These trapped charges change the injection fields and thus, the amount of charge transferred to and from the floating gate during programming. This eventually leads to a situation where the difference in threshold voltage in the two possible memory states is so small that the sense circuit can no longer discern the two states. Degradation of the transistor characteristics occurs when CHE injection is used for programming. CHE injection is used primarily in EPROMs where endurance is restricted to a low number of cycles. Fowler–Nordheim injection also causes degradation of the transistor characteristics, but most devices make use of a separate tunnel area, leaving the sense transistor unaffected by programming operations.

As described by Mielke et al. [1.25], the main cause of endurance failure in TPFG devices is believed to be the trapping of electrons in the tunnel oxide (called trap-up), while thin oxide devices fail mainly because of thin oxide breakdown induced by very high oxide fields during programming. Trap-up also occurs in

Petitioner Micron Ex-1024, 0069



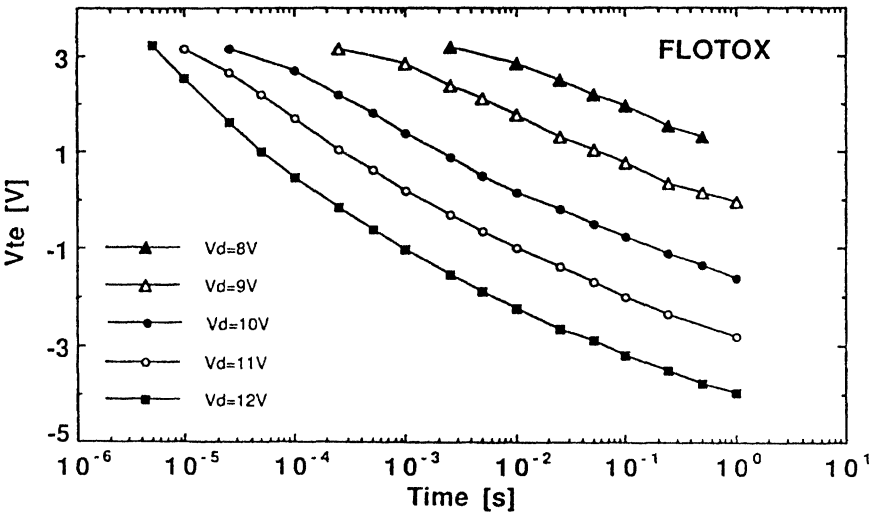


Figure 1.48 Typical programming characteristics of a FLOTOX-type memory cell showing the external threshold voltage as a function of the programming time. Upper figure = erasing, lower figure = writing.

thin oxide devices, but it is far less important than for TPFG transistors. Thin oxide devices are limited primarily by oxide breakdown.

The oxide breakdown phenomenon associated with thin oxide devices manifests itself slightly differently in device operation than during breakdown tests. As

Petitioner Micron Ex-1024, 0070

suggested in reference [1.25], the amount of energy available inside the memory transistor is too small to cause immediate and total breakdown. Therefore, this breakdown is generally regarded as being caused by a weak spot in the oxide that is "activated" by high electric fields and eventually becomes so leaky that the charges stored on the floating gate can no longer be retained. This is detected as a retention failure.

As explained above, thin oxide nonvolatile devices fail mainly because of the presence of defects (weak spots) in the thin oxide, which become shorts under the influence of high fields. As the total area of thin oxide per chip increases with increasing memory density, the probability that one memory cell will fail (and thus, that the circuit will fail) increases with increasing memory density. It was even predicted that thin oxide FLOTOX-type EEPROM circuits with densities larger than 16 KB would be less reliable than their TPFG counterparts, so that they would never become important [1.25]. As has been argued more than once [1.141], however, the density of weak spots in the thin oxide layer is largely dependent on processing conditions and is constantly decreasing by the use of advanced processing techniques. Recently, 1 Mbit EEPROM products, making use of thin oxide, have been achieved [1.142]. These circuits generally incorporate some sort of redundancy or error correcting circuitry, which can obviate the defect-related yield and reliability problems of thin oxide devices.

The fact that defects in the thin oxide limit the endurance features of thin oxide nonvolatile devices explains why little attention has been paid to the degradation behavior of these devices caused by charge trapping. This behavior, however, is important because it determines the endurance limit of the optimum memory cell (i.e., without defects) or its equivalent, constructed by means of redundancy. The study of degradation behavior also reveals some phenomena that would otherwise not be detected and that are responsible for some mismatches between experimentally obtained and calculated characteristics (e.g., some features of transient programming characteristics).

With respect to endurance, two different characteristics are important. The first one describes the degradation of the memory during cycling and gives the value of the threshold voltage for one memory transistor as a function of applied programming cycles, all with the same programming conditions. An example of such a characteristic for a FLOTOX cell is given in Fig. 1.49. A threshold voltage window opening in the first tens of cycles is observed, followed by a severe window closing after 10^5–10^6 cycles.

The results of threshold voltage degradation measurements are frequently presented in the literature. It is assumed that the results obtained on MOS capacitors can be used to predict the behavior of MOS transistors during high-field stressing. For stressing with alternating field polarities, as is the case for EEPROM devices, the degradation mechanisms have not been well understood until recently [1.25, 1.86, 1.126, 1.143, 1.144]. The explanation of observed memory degradation behavior is restricted to the general statement that threshold voltage window opening is caused by positive charge trapping, whereas window closing is caused by electron trapping in the oxide [1.25, 1.86, 1.143, 1.144].

Petitioner Micron Ex-1024, 0071

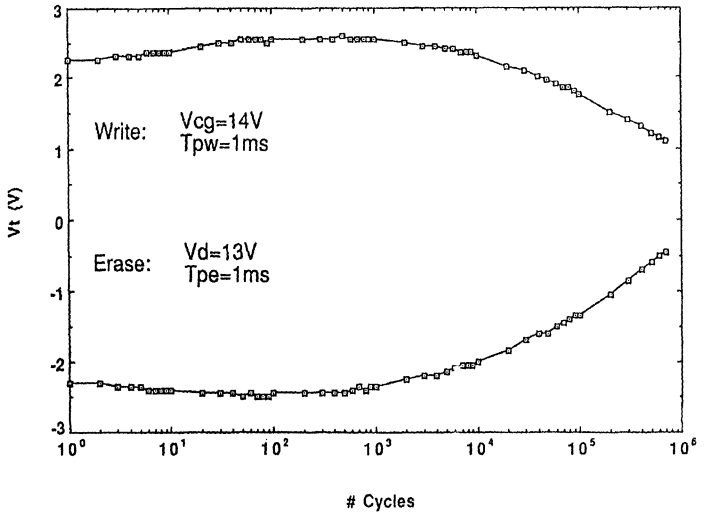


Figure 1.49 Typical endurance characteristics of a FLOTOX-type memory cell showing the threshold voltage in the written and the erased state as a function of the number of applied write/erase cycles. The threshold voltage shows a threshold voltage window opening during the first tens of cycles, followed by a window closure after 10^5–10^6 cycles.

For an understanding of the threshold voltage degradation characteristics, it is important that the threshold voltages be given from the first cycle on. In many cases, application of a number of cycles before the start of the degradation measurement in order to get a stable threshold voltage hides this valuable information.

A second concern about endurance is related to statistical information regarding the distribution of the cell characteristics. Specifically, the cumulative failure probability as a function of applied programming cycles for a large number of nonvolatile memory devices is an important indicator for the overall endurance reliability of memory circuits containing many memory cells. An example is shown in Fig. 1.50 [1.25]. For endurance failures originating from oxide breakdown, a broad random-life failure rate distribution with an almost flat failure rate is normally observed. For endurance failures caused by trap-up, which is a more intrinsic feature, a sharp wearout beyond some number of cycles is recorded.

In order to fully characterize the endurance behavior of a memory circuit, knowledge of the distribution of the endurance failures is not sufficient. The worst case endurance of any given cell in the array defines the endurance of the chip. Within a given chip, the endurances of individual bits have a small range of values, and it is pure chance whether a highly cycled bit has an endurance on the high end or low end of that distribution. Knowledge of the statistical distribution of the worst

Petitioner Micron Ex-1024, 0072

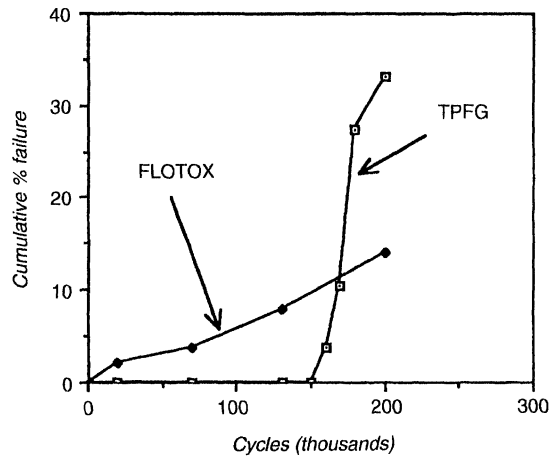


Figure 1.50 Cumulative failure distribution as a function of the number of cycles for two types of floating gate technologies (FLOTOX and TPFG) [1.25].

case bit from each of thousands of devices makes the endurance of a given chip statistically predictable.

1.6.2.2 Charge-Trapping Devices. For charge-trapping devices, program/erase endurance is determined by charge transport through the UTO layer. In conventional SNOS devices, it has been shown that the degradation is caused primarily by hole transport through the UTO layer [1.138, 1.145] by which hole traps are created in the oxide and interface traps are generated [1.112, 1.146], resulting in shifts of the threshold voltage and reduced retention. It was shown recently that hole transport toward the silicon is the most damaging [1.138]. This explains why a reduction in hole injection from the gate in SNOS devices with an oxynitride storage layer [1.113], or in SONOS devices, has given rise to improved endurance [1.105]. For the scaled-down SONOS structure, hole injection is almost eliminated, and consequently, even better endurance can be expected. This has been confirmed on ultra-thin SONOS structures (2 nm SiO_2, 8.5 nm Si_3N_4, 5 nm SiO_2), which showed no noticeable degradation after 10^7 10 V erase/write (E/W) pulse cycles [1.147].

1.6.3 Retention Characteristics

Another failure mode of the nonvolatile memory is retention failure. In this case, the memory loses its information and is, therefore, no longer nonvolatile. Most commercially available products are guaranteed to retain their information for at least 10 years after programming, either when operating or with power turned off. The retention feature of the memory device is influenced by degradation. It is possible that, after a number of programming cycles, the memory cell can still be reprogrammed but can no longer meet its retention limit.

Petitioner Micron Ex-1024, 0073

1.6.3.1 Floating Gate Devices. For floating gate devices, there is no intrinsic retention problem since retention is limited only by defect densities. Defects can be activated by the stress that the oxide layer undergoes in a high-temperature bake or from a large number of programming cycles. Thus, many endurance failures are actually retention failures.

Since the retention time to be guaranteed by the manufacturer is generally quite high, all retention tests use a combination of accelerating conditions. For example, higher voltages during read-out is one possible acceleration. The most commonly used retention test is storage of the programmed devices at a high temperature (up to 250°C). In any case, a complete reliability evaluation of a nonvolatile device must be a combination of endurance and retention tests.

1.6.3.2 Charge-Trapping Devices. For charge-trapping devices, there is an intrinsic retention problem. The threshold voltage of programmed SNOS devices decreases with time. This decrease is due either to *loss of charge* from the nitride layer by backtunneling to the silicon bands [1.136, 1.148] or by injection from the silicon into the nitride layer of carriers of the opposite type [1.139] or to *redistribution of charge* in the nitride layer [1.137]. The loss by backtunneling can be reduced through an appropriate hydrogen anneal step [1.103, 1.106], by which the interface trap density is reduced substantially [1.146]. For conventional SNOS and for SONOS devices, the threshold voltage decay is logarithmic in time, and, from extrapolation of the data taken over several decades of time, retention times of well over 10 years can be expected, even at elevated temperatures. Furthermore, a slowdown of the decay rate is observed for longer times [1.122, 1.139]. As an example, Fig. 1.51 shows the threshold voltage decay of a p-channel MNOS transistor and compares its behavior for annealed and nonannealed devices after storage at room temperature and at high temperature (125°C) [1.146]. The hydrogen anneal conditions used in this experiment were 800°C and 15 minutes [1.104]. As the figure shows, the decay is logarithmic in time for all cases and the hydrogen anneal results in a reduction of the decay rate by 25% [1.103]. Recently, Minami et al. [1.108] found that scaled-down SNOS devices show improved retention behavior: the written state decay rates do, in fact, decrease with nitride thickness, whereas the erased-state retentivity is almost independent of nitride thickness.

For the scaled-down SONOS device, it is not clear at present whether further scaling will improve the retention behavior as has been predicted [1.122]. If, in this scaled device, the major contribution to the threshold voltage shift is a result of charge stored at the top oxide–nitride interface, backtunneling can indeed be expected to decrease significantly. However, in this case, injection from the silicon and compensation of the stored charge could become a major issue of concern [1.139] in view of the high (> 2 MV/cm) fields that will be reached in the tunneling oxide for stored charges on the order of 10^{12}–10^{13} cm^{-2}.

Petitioner Micron Ex-1024, 0074

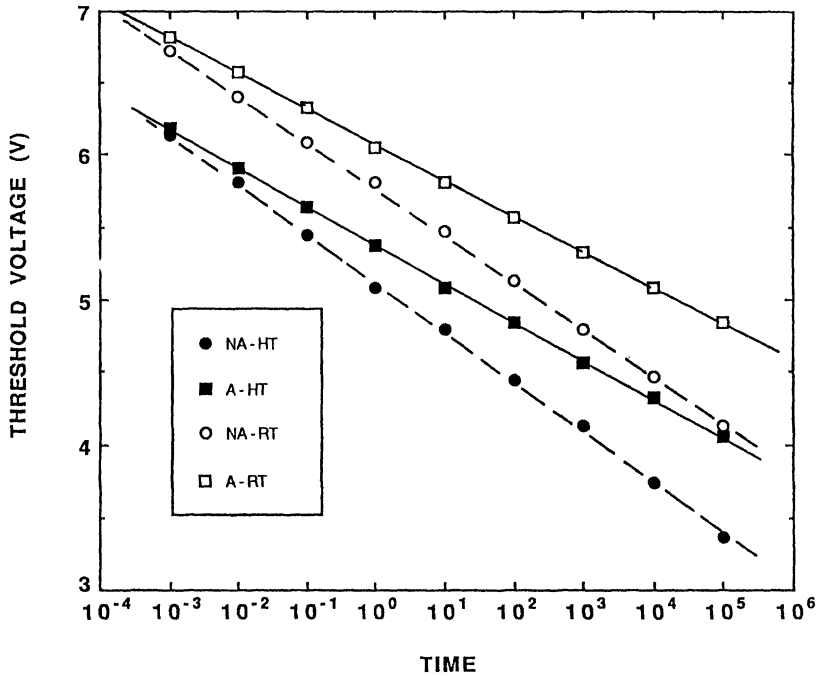


Figure 1.51 Threshold voltage decay in a p-channel MNOS transistor after a write pulse of 1 ms, $V_W = -20$ V, $t_{ox} = 2.1$ nm, $t_n = 52$ nm. Writing was always performed at room temperature. This figure shows the effect of a high-temperature hydrogen anneal (800°C, 15 min; A = annealed, NA = not annealed) and of the temperature during read (RT = room temperature, HT = high temperature = 125°C).

1.7. RADIATION ASPECTS OF NONVOLATILE MEMORIES

Finally, some issues related to the sensitivity of the different nonvolatile memory technologies to ionizing radiation are briefly discussed in this section.

1.7.1 SNOS Technology

SNOS memory devices are significantly harder than MOS structures because, unlike silicon dioxide, the mobilities of electrons and holes are not much different in nitrides. When exposed to ionizing radiation, both generated carriers can be rapidly swept out of the insulator, resulting in a negligible amount of trapped charge [1.149].

Petitioner Micron Ex-1024, 0075

Acceptable shifts for a total radiation dose of up to a Megarad (Si) at 77 K have been obtained for SNOS structures [1.150]. A reduction of the programming voltages and the absence of thick oxide parts in the newer SNOS devices have improved their radiation hardness. For SNOS/CMOS technologies, special techniques are, however, required to harden the peripheral circuitry [1.151, 1.152]. For scaled SONOS devices with a thick top oxide, radiation hardness might become a matter of concern in view of the increasing dominance of SiO_2 parts in the device.

1.7.2 Floating Gate Technology

Floating gate memories suffer from a lower failure rate due to α-particles than volatile memories since the charge is stored on a floating gate that is less susceptible to α-particle-induced electron-hole pairs [1.153]. As a result, the memory cell itself is not susceptible to α-particle upset. However, the peripheral circuitry is.

An important issue in radiation hardness of memory devices is their total dose characteristic. Since floating gate technology is essentially a MOS technology, the total dose radiation hardness of these technologies is expected to be rather weak if no special precautions are taken.

Recently, radiation characteristics of a floating gate memory technology were reported [1.154, 1.155]. Figure 1.52 shows an example of the memory threshold voltage window as a function of radiation dose [1.154]. The radiation was carried out with a Cobalt-60 source using a dose rate of 142 rad(Si)/sec. It can be seen that the high-V_t state decays with dose, whereas the low-V_t state remains nearly constant.

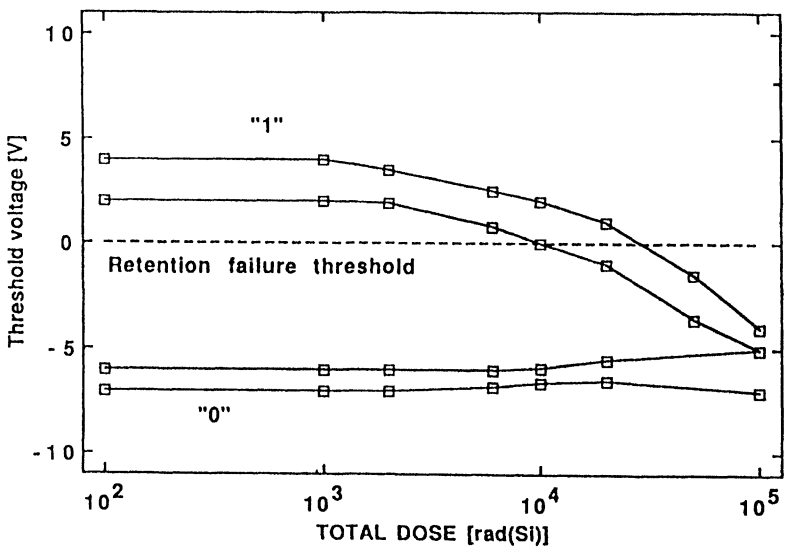


Figure 1.52 Memory threshold voltage window as a function of total radiation dose under Cobalt-60 radiation [1.154] for the floating gate technology.

Petitioner Micron Ex-1024, 0076

For a fixed reference sense amplifier, total dose radiation is limited to values of about 10 to 30 Krad(Si), depending on the initial (pre-radiation) high-V_t state of the memory cell. This indicates a tradeoff between the cell write time and the radiation retention failure. Moreover, by using a differential sense amplifier, maximum total dose values can be increased up to values of 100 Krad, at the cost, however, of memory density. It was shown [1.154] that the decay of the high-V_t state of the memory cell is caused by three main mechanisms, illustrated in Fig. 1.53: (1) holes (or electrons) generated in the oxide layers are injected onto the floating gate and decrease the stored charge (the oxide layers involved can be the tunnel oxide, the interpoly oxide, and, often the most important one, the field oxide [1.156]; (2) holes generated in the oxide can also be trapped in the oxide; and (3) electrons stored on the floating gate can be emitted over the energy barriers toward the substrate or the control gate.

Among the different cell types, floating gate memory cells manufactured in a single-polysilicon technology have recently been reported to show a much higher radiation tolerance than double-polysilicon cells [1.157]. Model calculations have indicated that the oxide thickness of the coupling capacitor of the cells appears to be the dominant parameter affecting their radiation response. The thinner oxide of this capacitor in single-polysilicon cells is the primary reason for their better hardness. Hardness levels of more than 110 Krad(SiO_2) have been achieved [1.157].

The radiation hardness of floating gate memory cells can be improved in different ways. The use of thinner oxides can increase the radiation hardness because the volume for carrier generation decreases [1.154, 1.157]. Another way to improve the radiation hardness is to perform an additional threshold voltage implant in order to increase the high-V_t state of the memory cell. The cost, however, is a reduction in speed of the cell. As already mentioned above, differential sensing can increase the maximum total dose to values of 100 to 200 Krad(Si). Finally, refreshing techniques, as in DRAMs, could be used to extend the hardness levels of the memory cells.

Another important point is that, in practical cases, it will not necessarily be the memory cell that causes the failure due to total dose radiation, but rather the peripheral circuits [1.155]. It was reported recently that, for a 256 K EEPROM, the total dose failure levels are limited to values of 10 to 30 Krad(Si) due to loss of drive

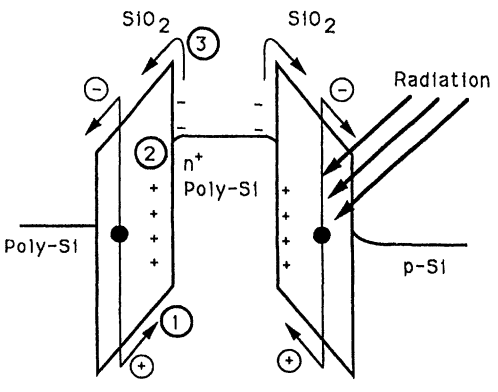


Figure 1.53 Schematic representation of the three main mechanisms that can cause the threshold voltage decay for the high V_t state of a floating gate memory cell: (1) hole injection onto the floating gate, (2) hole trapping in the oxides, and (3) electron emission from the floating gate.

Petitioner Micron Ex-1024, 0077

capability of the peripheral circuitry, and not charge loss from the memory transistors. Moreover, the operating mode during radiation (reading or writing) and the exposure time of radiation influence the total dose of memory circuits [1.155]. Finally, it was found that dose rates less than 10^8 rad(Si)/s can be allowed without generating dose-rate upset or latchup of the memory.

In conclusion, and based on the few reports that have been published, it can be stated that floating gate technologies can be used in low total dose applications [maximum 30 Krad(Si)] and with dose rates below 10^8 rad(Si)/s.

1.7.3 Ferroelectric Technology

Ferroelectric capacitors are much more radiation hard than MOS or SNOS capacitors. The data storage mechanism, polarization state, is the result of a net ionic displacement in the unit cells of the material. Thus, high energy particles, gamma radiation or neutrons, have to move ions in the lattice in order to distort the existing polarization states, which would require very high doses. Typically, these ferroelectric capacitors can function satisfactorily with little degradation to doses up to 10 Mrads, with or without applied bias. However, whereas the memory cell based on ferroelectric materials exhibits a high hardness, the ultimate hardness of the FRAM depends on that of the MOS parts in the memory cell and the control circuity.

References

[1.1] D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," *Bell Syst. Tech. J.*, vol. 46, p. 1288, 1967.

[1.2] H. A. R. Wegener, A. J. Lincoln, H. C. Pao, M. R. O'Connell, and R. E. Oleksiak, "The variable threshold transistor, a new electrically alterable, non-destructive read-only storage device," *IEEE IEDM Tech. Dig.*, Washington, D.C., 1967.

[1.3] H. E. Maes, "Recent developments in non-volatile semiconductor memories," *Digest of Technical Papers ESSCIRC83*, p. 1, 1983.

[1.4] T. Hagiwara, Y. Yatsuda, R. Kondo, S. Minami, T. Aoto, and Y. Itoh, "A 16kbit electrically erasable PROM using n-channel Si-gate MNOS technology," *IEEE J. Solid State Circuits*, vol. SC-15, p. 346, 1980.

[1.5] W. Johnson, G. Perlegos, A. Renninger, G. Kuhn, and T. Ranganath, "A 16Kb electrically erasable nonvolatile memory," *IEEE ISSCC Dig. Tech. Pap.*, p. 152, 1980.

[1.6] D. Frohman-Bentchkowsky, "A fully decoded 2048-bit electrically programmable MOS-ROM," *IEEE ISSCC Dig. Tech. Pap.*, p. 80, 1971.

[1.7] D. Frohman-Bentchkowsky, "Memory behaviour in a floating gate avalanche injection MOS (FAMOS) structure," *Appl. Phys. Lett.*, vol. 18, p. 332, 1971.

Petitioner Micron Ex-1024, 0078

[1.8] D. Frohman-Bentchkowsky, "The metal–nitride–oxide–silicon (MNOS)-transistor—Characteristics and applications," *Proc. IEEE*, vol. 58, p. 1207, 1970.

[1.9] S. Sato and T. Yamaguchi, "Study of charge behaviour in metal–alumina–silicon dioxide–silicon (MAOS) field effect transistor," *Solid State Electr.*, vol. 17, p. 367, 1974.

[1.10] R. L. Angle and H. E. Talley, "Electrical and charge storage characteristics of the tantalum oxide–silicon dioxide device," *IEEE Trans. Elect. Dev.*, vol. ED-25, p. 1277, 1978.

[1.11] D. Frohman-Bentchkowsky, "A fully decoded 2048 bit electrically programmable FAMOS read-only memory," *IEEE J. Sol. St. Circ.*, vol. SC-6, p. 301, 1971.

[1.12] D. Frohman-Bentchkowsky, "FAMOS—A new semiconductor charge storage device," *Sol. St. Electr.*, vol. 17, p. 517, 1974.

[1.13] H. Iizuka, T. Sato, F. Masuoka, K. Ohuchi, H. Hara, H. Tango, M. Ishikawa, and Y. Takeishi, "Stacked gate avalanche injection type MOS (SAMOS) memory," Proc. 4th Conf. Sol. St. Dev., Tokyo, 1972; *J. Japan. Soc. Appl. Phys.*, vol. 42, p. 158, 1973.

[1.14] H. Iizuka, F. Masuoka, T. Sato, and M. Ishikawa, "Electrically alterable avalanche injection type MOS read-only memory with stacked gate structure," *IEEE Trans. Elect. Dev.*, vol. ED-23, p. 379, 1976.

[1.15] S. M. Sze, "Current transport and maximum dielectric strength of silicon nitride films," *J. Appl. Phys.*, vol. 38, p. 2951, 1967.

[1.16] H. E. Maes and G. Heyns, "Two-carrier conduction in amorphous chemically vapour deposited (CVD) silicon nitride layers," *Proc. of the Int. Conference on Insulating Films on Semiconductors*, North Holland Publ. Co., p. 215, 1983.

[1.17] J. Yeargain and K. Kuo, "A high density floating gate EEPROM cell," *IEEE IEDM Tech. Dig.*, p. 24, 1981.

[1.18] R. Klein, W. Owen, R. Simko, and W. Tchon, "5-V-only, nonvolatile RAM owes it all to polysilicon," *Electronics*, October 11, p. 111, 1979.

[1.19] G. Landers, "5-V-only EEPROM mimics static RAM timing," *Electronics*, June 30, p. 127, 1980.

[1.20] B. Rössler and R. Müller, "Electrically erasable and reprogrammable read-only memory using the n-channel SIMOS one-transistor cell," *IEEE Trans. Elect. Dev.*, vol. ED-24, p. 806, 1977.

[1.21] D. Guterman, I. Rimawi, T. Chiu, R. Halvorson, and D. McElroy, "An electrically alterable nonvolatile memory cell using a floating gate structure," *IEEE Trans. Elect. Dev.*, vol. ED-26, p. 576, 1979.

[1.22] M. Kamiya, Y. Kojima, Y. Kato, K. Tanaka, and Y. Hayashi, "EPROM cell with high gate injection efficiency," *IEEE IEDM Tech. Dig.*, p. 741, 1981.

[1.23] A. Wu, T. Chan, P. Ko, and C. Hu, "A novel high-speed, 5-V programming EPROM structure with source-side injection," *IEEE IEDM Tech. Dig.*, p. 584, 1986.

Petitioner Micron Ex-1024, 0079

[1.24] M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling in thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, p. 278, 1969.

[1.25] N. Mielke, A. Fazio, and H.-C. Liou, "Reliability comparison of FLOTOX and textured polysilicon EEPROM's," *Proc. Int. Rel. Phys. Symp. (IRPS)*, p. 85, 1987.

[1.26] R. C. Hughes, "High field electronic properties of SiO_2," *Sol. St. Electr.*, vol. 21, p. 251, 1978.

[1.27] D. J. DiMaria and D. R. Kerr, "Interface effects and high conductivity in oxides grown from polycrystalline silicon," *Appl. Phys. Lett.*, vol. 27, p. 505, 1975.

[1.28] R. M. Anderson and D. R. Kerr, "Evidence for surface asperity mechanism of conductivity in oxide grown on polycrystalline silicon," *J. Appl. Phys.* vol. 48, p. 4834, 1977.

[1.29] H. R. Huff, R. D. Halvorson, T. L. Chiu, and D. Guterman, "Experimental observations on conduction through polysilicon oxide," *J. Electrochem. Soc.*, vol. 127, p. 2482, 1980.

[1.30] P. A. Heimann, S. P. Murarka, and T. T. Sheng, "Electrical conduction and breakdown in oxides of polycrystalline silicon and their correlation with interface texture," *J. Appl. Phys.*, vol. 53, p. 6240, 1982.

[1.31] R. K. Ellis, "Fowler–Nordheim emission from non-planar surfaces," *IEEE Elect. Dev. Lett.*, vol. EDL-3, p. 330, 1982.

[1.32] G. Groeseneken and H. E. Maes, "A quantitative model for the conduction in oxides thermally grown from polycrystalline silicon," *IEEE Trans. Elect. Dev.*, vol. ED-33, p. 1028, 1986.

[1.33] J. Bisschop, E. J. Korma, E. F. F. Botta, and J. F. Verwey, "A model for the electrical conduction in polysilicon oxide," *IEEE Trans. Elect. Dev.*, vol. ED-33, p. 1809, 1986.

[1.34] R. D. Jolly, H. R. Grinolds, and R. Groth, "A model for conduction in floating gate EEPROM's," *IEEE Trans. Elect. Dev.*, vol. ED-31, p. 767, 1984.

[1.35] P. E. Cottrell, R. R. Troutman, and T. H. Ning, "Hot electron emission in n-channel IGFET's," *IEEE J. Sol. St. Circ.*, vol. SC-14, p. 442, 1979.

[1.36] B. Eitan and D. Frohman-Bentchkowsky, "Hot electron injection into the oxide in n-channel MOS-devices," *IEEE Trans. Elect. Dev.*, vol. ED-28, p. 328, 1981.

[1.37] Y. Tarui, Y. Hayashi, and K. Nagai, "Electrically reprogrammable non-volatile semiconductor memory," *IEEE J. Sol. St. Circ.*, vol. SC-7, p. 369, 1972.

[1.38] C. Hu, "Lucky electron model of hot electron emission," *IEEE IEDM Tech. Dig.*, p. 22, 1979.

[1.39] P. Ko, R. Müller, and C. Hu, "A unified model for hot electron currents in MOSFET's," *IEEE IEDM Tech. Dig.*, p. 600, 1981.

[1.40] E. Takeda, H. Kume, T. Toyabe, and S. Asai, "Submicrometer MOSFET structure for minimizing hot carrier generation," *IEEE Trans. Elect. Dev.*, vol. ED-29, p. 611, 1982.

Petitioner Micron Ex-1024, 0080

[1.41] S. Tanaka and M. Ishikawa, "One-dimensional writing model of n-channel floating gate ionization-injection MOS (FIMOS)," *IEEE Trans. Elect. Dev.*, vol. ED-28, p. 1190, 1981.

[1.42] K. R. Hoffmann, C. Werner, W. Weber, and G. Dorda, "Hot-electron and hole emission effects in short n-channel MOSFET's," *IEEE Trans. Elect. Dev.*, vol. ED-32, p. 691, 1985.

[1.43] T. Y. Chan, P. K. Ko, and C. Hu, "Dependence of channel electric field on device scaling," *IEEE Elect. Dev. Lett.*, vol. EDL-6, p. 551, 1985.

[1.44] C. Hu, "Hot electron effects in MOSFET's," *IEEE IEDM Tech. Dig.*, p. 176, 1983.

[1.45] J. Van Houdt, L. Haspeslagh, D. Wellekens, L. Deferm, G. Groeseneken, and H. E. Maes, "HIMOS—a high efficiency Flash EEPROM cell for embedded memory applications," *IEEE Trans. Elect. Dev.*, vol. ED-40, p. 2255, 1993.

[1.46] J. Van Houdt, P. Heremans, L. Deferm, G. Groeseneken, and H. E. Maes, "Analysis of the enhanced hot-electron injection in split-gate transistors useful for EEPROM applications," *IEEE Trans. Elect. Dev.*, vol. ED-39, p. 1150, 1992.

[1.47] K. I. Lundström and C. M. Svensson, "Properties of MNOS structures," *IEEE Trans. Elect. Dev.*, ED-19, p. 826, 1972.

[1.48] H. E. Maes, J. Witters, and G. Groeseneken, "Trends in non-volatile memory devices and technologies," *Proc. 17th European Solid State Device Research Conference*, p. 743, 1987.

[1.49] H. E. Maes, G. Groeseneken, H. Lebon, and J. Witters, "Trends in semiconductor memories," *Microelectr. J.*, vol. 20, p. 9, 1989.

[1.50] S. Atsumi, S. Tanaka, S. Saito, N. Ohtsuka, N. Matsukawa, S. Mori, Y. Kaneko, K. Yoshikawa, J. Matsunaga, and T. Iizuka, "A 120ns 4Mb CMOS EPROM," *IEEE ISSCC Dig. Tech. Pap.*, p. 74, 1987.

[1.51] G. Yaron, S. Prasad, M. Ebel, and B. Leong, "A 16K EEPROM employing new array architecture and designed-in reliability features," *IEEE J. Sol. St. Circ.*, vol. SC-17, no. 5, p. 833, 1982.

[1.52] G. Landers, "5-V only EEPROM mimics static RAM timing," *Electronics*, June 30, p. 127, 1980.

[1.53] B. Gerber and J. Fellrath, "Low voltage single supply CMOS electrically erasable read-only memory," *IEEE Trans. Elect. Dev.*, vol. ED-27, p. 1211, 1980.

[1.54] D. H. Oto, V. K. Dham, K. H. Gudger, M. Reitsma, G. S. Gongwer, Y. W. Hu, J. Olund, H. Jones, and S. Nieh, "High-voltage regulation and process considerations for high-density 5V-only EEPROM's," *IEEE J. Sol. St. Circ.*, vol. SC-18, p. 532, 1983.

[1.55] D. Cioaca, T. Lin, A. Chan, L. Chen, and A. Milhnea, "A million-cycles CMOS 256K EEPROM," *IEEE J. Sol. St. Circ.*, vol. SC-22, p. 684, 1987.

[1.56] P. I. Suciu, M. Briner, C. S. Bill, and D. Rinerson, "A 64K EEPROM with extended temperature range and page mode operation," *IEEE ISSCC Dig. Tech. Pap.*, p. 170, 1985.

Petitioner Micron Ex-1024, 0081

[1.57] R. Vancu, L. Chen, R. L. Wan, T. Nguyen, C.-Y. Yang, W.-P. Lai, K.-F. Tang, A. Mihnea, A. Renninger, and G. Smarandoiu, "A 35ns 256K CMOS EEPROM with error correcting circuitry," *IEEE ISSCC Dig. Tech. Pap.*, p. 64, 1990.

[1.58] F. Masuoka, M. Asano, H. Iwahashi, and T. Komuro, "A new Flash EEPROM cell using triple polysilicon technology," *IEEE IEDM Tech. Dig.*, p. 464, 1984.

[1.59] K. Imamiya, Y. Iwata, Y. Sugiura, H. Nakamura, H. Oodaira, M. Momodomi, Y. Ito, T. Watanabe, H. Araki, K. Narita, K. Masuda, and J. Miyamoto, "A 35 ns-cycle-time 3.3V-only 32 Mb NAND Flash EEPROM," *IEEE ISSCC Dig. Tech.* Pap., p. 130, 1995.

[1.60] A. Nozoe, T. Yamazaki, H. Sato, H. Kotani, S. Kubono, K. Manita, T. Tanaka, T. Kawahara, M. Kato, K. Kimura, H. Kume, R. Hori, T. Nishimoto, S. Shukuri, A. Ahba, Y. Kouro, O. Sakamoto, A. Fukumoto, and M. Nakajima, "A 3.3V high density AND Flash memory with 1 ms/512B Erase and Program Time," *IEEE ISSCC Dig. Tech. Pap.*, p. 124, 1995.

[1.61] K.-D. Suh, B.-H Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.-S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, and H.-K. Lim, "A 3.3V 32 Mb NAND Flash memory with incremental step pulse programming scheme," *IEEE ISSCC Dig. Tech. Pap.*, p. 128, 1995.

[1.62] E. Harari, L. Schmitz, B. Troutman, and S. Wang, "A 256bit non-volatile static RAM," *IEEE ISSCC Dig. Tech. Pap.*, p. 108, 1978.

[1.63] R. Klein, W. Owen, R. Simko, and W. Tchon, "5-V-only, nonvolatile RAM owes it all to polysilicon," *Electronics*, October 11, p. 111, 1979.

[1.64] J. Drori, S. Jewell-Larsen, R. Klein, and W. Owen, "A single 5V supply non-volatile static RAM," *IEEE ISSCC Dig. Tech. Pap.*, p. 148, 1981.

[1.65] A. Weiner, C. Herdt, J. Tiede, and B. Geston, "A 64K Non-volatile SRAM," Non-volatile Semiconductor Memory Workshop, Vail, Colo., 1989.

[1.66] Y. Terada, K. Kobayashi, T. Nakayama, H. Arima, and T. Yoshihara, "A new architecture for the NVRAM—An EEPROM backed-up dynamic RAM," *IEEE J. Sol. St. Circ.*, vol. SC-23, no. 1, p. 86, 1988.

[1.67] Y. Yamauchi, K. Tanaka, and K. Sakiyama, "A novel NVRAM cell technology for high density applications," *IEEE IEDM Tech. Dig.*, p. 416, 1988.

[1.68] Y. Yamauchi, H. Ishihara, K. Tanaka, K. Sakiyama, and R. Miyake, "A versatile stacked storage capacitor on FLOTOX cell for megabit NVRAM applications," *IEEE IEDM Tech. Dig.*, p. 595, 1989.

[1.69] G. Gerosa, C. Hart, S. Harris, R. Kung, J. Weihmeir, and J. Yeargain, "A high performance CMOS technology for 256K/1Mb EPROMs," *IEEE IEDM Tech. Dig.*, p. 631, 1985.

[1.70] J. Esquivel, A. Mitchell, J. Paterson, B. Riemenschneider, H. Tiegelaar, T. Coffman, D. Dolby, M. Gill, R. Lahiry, S. Lin, D. McElroy, J. Schreck, and P. Shah, "High density contactless self-aligned EPROM cell array technology," *IEDM Tech. Dig.*, p. 592, 1986.

[1.71] T. Coffman, D. Boyd, D. Dolby, M. Gill, S. Kady, R. Lahiry, S. Lin, D. McElroy, A. Mitchell, J. Paterson, J. Schreck, P. Shah, and F. Takeda, "A

Petitioner Micron Ex-1024, 0082

1M CMOS EPROM with a 13.5μ^2 cell," *IEEE ISSCC Dig. Tech. Pap.*, p. 72, 1987.

[1.72] A. Esquivel, B. Riemenschneider, J. Paterson, H. Tiegelaar, J. Mitchell, R. Lahiry, M. Gill, J. Schreck, D. Dolby, T. Coffman, S. Lin, and P. Shah, "A 8.6 μm^2 cell technology for a 35.5mm^2 Megabit EPROM," *IEEE IEDM Tech. Dig.*, p. 859, 1987.

[1.73] B. Eitan, Y. Ma, C. Hu, R. Kazerounian, K. Sinai, and S. Ali, "A self-aligned split-gate EPROM," presented at the 9th NVSM Workshop, Monterey, Calif. 1988.

[1.74] K. Yoshikawa, S. Mori, K. Narita, N. Arai, Y. Oshima, Y. Kaneko, and H. Araki, "An asymmetrical lightly-doped source (ALDS) cell for virtual ground high density EPROMs," *IEEE IEDM Tech. Dig.*, p. 432, 1988.

[1.75] S. Ali, B. Sani, A. Shubat, K. Sinai, R. Kazerounian, C.-J. Hu, Y. Ma, and B. Eitan, "A 50ns 256K CMOS split-gate EPROM," *IEEE J. Sol. St. Circ.*, vol. SC-23, no. 1, p. 79, 1988.

[1.76] K. Prall, W. Kinney, and J. Macro, "Characterization and suppression of drain coupling in submicrometer EPROM cells," *IEEE Trans. Elect. Dev.*, vol. ED-34, no. 12, p. 2463, 1987.

[1.77] Y. Mizutani and K. Makita, "A new EPROM cell with a side-wall floating gate for high density and high performance device," *IEEE IEDM Tech. Dig.*, p. 63, 1985.

[1.78] S. Chu and A. Steckl, "The effect of trench-gate-oxide structure on EPROM device operation," *IEEE Elect. Dev. Lett.*, vol. EDL-9, no. 6, p. 284, 1988.

[1.79] S. Shukiri, Y. Wada, T. Hagiwara, K. Komori, and M. Tamura, "A novel EPROM device fabricated using focused Boron ion beam implantation," *IEEE Trans. Elect. Dev.*, vol. ED-34, p. 1264, 1987.

[1.80] R. Cuppens, C. Hartgring, J. Verwey, H. Peek, F. Vollebregt, E. Devens, and I. Sens, "A EEPROM for microprocessors and custom logic," *IEEE J. Sol. St. Circ.*, vol. SC-20, no. 2, p. 603, 1985.

[1.81] R. Yoshikawa, and N. Matsukawa, "EPROM and EEPROM cell structures for EPLD's compatible with single poly gate process," presented at the 8th NVSM Workshop, Vail, Colo., 1986.

[1.82] N. Mielke, L. Purvis, and H. Wegener, "Reliability comparison of FLOTOX and textured-poly EEPROMs," presented at the 8th NVSM Workshop, Vail, Colo., 1986.

[1.83] S. Lai, J. Lee, and V. Dham, "Electrical properties of nitrided-oxide systems for use in gate dielectrics and EEPROM," *IEEE IEDM Tech. Dig.*, p. 190, 1983.

[1.84] C. Jenq, T. Chiu, B. Joshi, and J. Hu, "Properties of thin oxynitride films used as floating gate tunneling dielectrics," *IEEE IEDM Tech. Dig.*, p. 309, 1982.

[1.85] N. Matsukawa, S. Morita, and H. Nozawa, "High performance EEPROM using low barrier height tunnel oxide," Ext. Abstr. Int. Conf. on Sol. St. Dev. and Mat., p. 261, 1984.

Petitioner Micron Ex-1024, 0083

[1.86] H. Nozawa, N. Matsukawa, and S. Morita, "An EEPROM cell using a low barrier height tunnel oxide," *IEEE Trans. Elect. Dev.*, vol. ED-33, no. 2, p. 275, 1986.

[1.87] D. DiMaria, K. De Meyer, C. Serrano, and D. Dong, "Electrically alterable read-only-memory using Si-rich SiO2 injectors and a polycrystalline silicon storage layer," *J. Appl. Phys.*, vol. 52, p. 4825, 1982.

[1.88] T. Ito, S. Hijiya, T. Nozaki, H. Arakawa, H. Ishikawa, and M. Shinoda, "Low voltage alterable EAROM cells with nitride barrier avalanche injection MIS (NAMIS)," *IEEE Trans. Elect. Dev.*, vol. ED-26, p. 906, 1979.

[1.89] R. Stewart, A. Ipri, L. Faraone, J. Cartwright, and K. Schlesier, "A shielded substrate injector MOS (SSIMOS) EEPROM cell," *IEEE IEDM Tech. Dig.*, p. 472, 1984.

[1.90] H. Arima, N. Ajika, H. Morita, T. Shibano, and T. Matsukawa, "A novel process technology and cell structure for megabit EEPROM," *IEEE IEDM Tech. Dig.*, p. 420, 1988.

[1.91] M. Momodomi, R. Kirasawa, R. Nakayama, S. Aritome, T. Endoh, Y. Itoh, Y. Iwata, H. Oodaira, T. Tanaka, M. Chiba, R. Shirota, and F. Masuoka, "New device technologies for 5V-only 4Mb EEPROM with NAND structure cell," *IEEE IEDM Tech. Dig.*, p. 412, 1988.

[1.92] Y. Itoh, M. Momodomi, R. Shirota, Y. Iwata, R. Nakayama, R. Kirasawa, T. Tanaka, K. Toita, S. Inoue, and F. Masuoka, "An experimental 4Mb CMOS EEPROM with a NAND structure cell," *IEEE ISSCC Dig. Tech. Pap.*, p. 134, 1989.

[1.93] R. Stewart, A. Ipri, L. Faraone, and J. Cartwright, "A low voltage high density EEPROM memory cell using symmetrical poly-to-poly conduction," presented at the 6th NVSM Workshop, Vail, Colo., August 1983.

[1.94] K. Sarma, A. Owens, D. Pan, B. Rosier, and L. Yeh, "Double poly EEPROM cell utilizing interpoly tunneling," presented at the 8th NVSM Workshop, Vail, Colo., 1986.

[1.95] H. Maes and G. Groeseneken, "Conduction in thermal oxides grown on polysilicon and its influence on floating gate EEPROM degradation," *IEEE IEDM Tech. Dig.*, p. 476, 1984.

[1.96] H. Wegener, "Build up of trapped charge in textured poly tunnel structures," presented at the 8th NVSM Workshop, Vail, Colo., 1986.

[1.97] G. Verma, and N. Mielke, "Reliability of ETOX based flash Memories," Proc. IRPS, p. 158, 1988.

[1.98] R. Kazerounian, S. Ali, Y. Ma, and B. Eitan, "A 5 volt high density poly-poly erase flash EEPROM," *IEEE IEDM Tech. Dig.*, p. 436, 1988.

[1.99] J. R. Cricchi, F. C. Blaha, and M. D. Fitzpatrick, "The drain-source protected MNOS memory device and memory endurance," *IEEE IEDM Tech. Dig.*, p. 126, 1973.

[1.100] Y. Yatsuda, T. Hagiwara, R. Kondo, S. Minami, and Y. Itoh, "N-channel Si-gate MNOS device for high speed EAROM," *Proc. 10th Conf. Solid State Devices*, p. 11, 1979.

Petitioner Micron Ex-1024, 0084

[1.101] G. Schols, H. E. Maes, G. Declerck, and R. Van Overstraeten, "High temperature hydrogen anneal of MNOS structures," *Revue de Phys. Appl. 13*, p. 825, 1978.

[1.102] Y. Yatsuda, S. Minami, R. Kondo, T. Hagiwara, and Y. Itoh, "Effects of high temperature annealing on n-channel Si-gate MNOS devices," *Jap. J. Appl. Phys.*, vol. 19, S19-1, p. 219, 1980.

[1.103] H. E. Maes and G. Heyns, "Influence of a high temperature hydrogen anneal on the memory characteristics of p-channel MNOS transistors," *J. Appl. Phys.*, vol. 51, p. 2706, 1980.

[1.104] G. Schols and H. E. Maes, "High temperature hydrogen anneal of MNOS structures," Proc. ECS, vol. 83-8, Eds. V. J. Kapoor and H. Stein, p. 94, 1983.

[1.105] Y. Yatsuda, T. Hagiwara, S. Minami, R. Kondo, K. Uchida, and K. Uchiumi, "Scaling down MNOS nonvolatile memory devices," *Jap. J. Appl. Phys.*, vol. 21, S21-1, p. 85, 1982.

[1.106] T. Hagiwara, Y. Yatsuda, S. Minami, S. Naketani, K. Uchida, and T. Yasui, "A 5V only 64k MNOS EEPROM," 6th NVSM, Vail, Colo., 1983.

[1.107] Y. Yatsuda, S. Minami, T. Hagiwara, T. Toyabe, S. Asai, and K. Uchida, "An advanced MNOS memory device for highly integrated byte erasable 5V only EEPROMs," *IEEE IEDM Tech. Dig.*, p. 733, 1982.

[1.108] S. Minami, Y. Kamigaki, K. Uchida, K. Furusawa, and T. Hagiwara, "Improvement of written-state retentivity by scaling down MNOS memory devices," *Jap. J. Appl. Phys.*, vol. 27, p. L2168, 1988.

[1.109] D. K. Schroder and M. H. White, "Characterization of current transport in MNOS structures with complementary tunneling emitter bipolar transistors," *IEEE Trans. Elect. Dev.*, vol. ED-26, p. 899, 1979.

[1.110] H. E. Maes and G. Heyns, "Carrier transport in LPCVD silicon nitride," 4th NVSM, Vail, Colo., 1980.

[1.111] H. E Maes and G. L. Heyns, "Two-carrier conduction in amorphous chemically vapour deposited (CVD) silicon nitride layers," Insulating Films on Semiconductor 83, Eds. J. Verwey and D. Wolters, p. 215, 1983.

[1.112] H. E. Maes and E. Vandekerckhove, "Non-volatile memory characteristics of polysilicon-oxynitride-oxide-silicon devices and circuits," *Proc. ECS*, vol. 87–10, Eds. V. J. Kapoor and K. T. Hankins, p. 28, 1987.

[1.113] H. E. Maes, "The use of oxynitride layers in non-volatile S-OxN-OS (silicon–oxynitride–oxide–silicon) memory devices," Chapter 6 in *LPCVD Silicon Nitride and Oxynitride Films*, pp. 127–146, Springer-Verlag, Berlin, Ed. F. H. Habraken, 1991.

[1.114] W. D. Brown, R. V. Jones, and R. D. Nasby, "The MONOS memory transistor: application in a radiation-hard nonvolatile RAM," *Sol. St. Electr.*, vol. 29, p. 877, 1985.

[1.115] P. C. Chen, "Threshold-alterable Si-gate MOS devices," *IEEE Trans. Elect. Dev.*, ED-24, p. 584, 1977.

[1.116] J. Remmerie, H. E. Maes, J. Witters, and W. Beullens, "Two carrier transport in MNOS devices," *Proc. ECS*, vol. 87-10, Eds. V. J. Kapoor and K. T. Hankins, p. 93, 1987.

Petitioner Micron Ex-1024, 0085

[1.117] H. E. Maes and R. Van Overstraeten, "Simple technique for determination of the centroid of nitride charge in MNOS structures," *Appl. Phys. Lett.*, vol. 27, p. 282, 1975.

[1.118] F. L. Hampton and J. R. Cricchi, "Space charge distribution limitations on scale down of MNOS memory devices," *IEEE IEDM Tech. Dig.*, p. 374, 1979.

[1.119] B. H. Yun, "Electron and hole transport in CVD nitride films," *Appl. Phys. Lett.*, vol. 27, p. 256, 1975.

[1.120] E. Suzuki, H. Hiraishi, K. Ishi, and Y. Hayashi, "A low voltage alterable EEPROM with metal–oxide–nitride–oxide–semiconductor (MONOS) structure," *IEEE Trans. Elect. Dev.*, vol. ED-30, p. 122, 1983.

[1.121] C. C. Chao and M. H. White, "Characterization of charge injection and trapping in scaled SONOS/MONOS memory devices," *Sol. St. Electr.*, vol. 30, p. 307, 1987.

[1.122] T. A. Dellin and P. J. McWhorter, "Scaling of MONOS nonvolatile memory transistors," *Proc. ECS*, vol. 87-10, Eds. V. J. Kapoor and K. T. Hankins, p. 3, 1987.

[1.123] V. J. Kapoor and S. B. Bibyk, "Energy distribution of electron trapping defects in thick-oxide MNOS structures," *Physics of MOS Insulators*, Eds. G. Lucovsky, S. Pantelides, and G. Galeener, p. 117, 1980.

[1.124] F. R. Libsch, A. Roy, and M. H. White, "Amphoteric trap modeling of multidielectric scaled SONOS nonvolatile memory structures," 8th NVSM, Vail, Colo., 1986.

[1.125] J. Evans and R. Womack, "An experimental 512-bit nonvolatile memory with ferroelectric storage cell," *IEEE J. Sol. St. Circ.*, vol. SC-23, p. 1171, 1998.

[1.126] J. Witters, "Characteristics and reliability of thin oxide floating gate memory transistors and their supporting programming circuits," Ph.D. Thesis, K. U. Leuven, 1989.

[1.127] A. Bhattacharyya, "Modelling of write/erase and charge retention characteristics of floating gate EEPROM devices," *Sol. St. Electr.*, p. 899, 1984.

[1.128] G. Groeseneken, "Programming behaviour and degradation phenomena in electrically erasable programmable floating gate memory devices," Ph.D. Thesis, K. U. Leuven, 1986.

[1.129] K. Prall, W. Kinney, and J. Macro, "Characterization and suppression of drain coupling in submicron EPROM cells," *IEEE Trans. Elect. Dev.*, vol. ED-34, p. 2463, 1987.

[1.130] P. I. Suciu, B. P. Cox, D. D. Rinerson, and S. F. Cagnina, "Cell model for EEPROM floating-gate memories," *IEEE IEDM Tech. Dig.*, p. 737, 1982.

[1.131] S. T. Wang, "On the I-V characteristics of floating gate MOS transistors," *IEEE Trans. Elect. Dev.*, vol. ED-26, p. 1292, 1979.

[1.132] S. T. Wang, "Charge retention of floating gate transistors under applied bias conditions," *IEEE Trans. Elect. Dev.*, vol. ED-27, p. 297, 1980.

[1.133] A. Kolodny, S. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of floating gate EEPROM cells," *IEEE Trans. Elect. Dev.*, vol. ED-33, p. 835, 1986.

Petitioner Micron Ex-1024, 0086

[1.134] R. Bez, D. Cantarelli, P. Cappelletti, and F. Maggione, "SPICE model for transient analysis of EEPROM cells," in *Proc. of ESSDERC88*, p. 677, 1988.

[1.135] J. T. Mantey, "Degradation of thin silicon dioxide films and EEPROM cells," Ph.D. Thesis, Ecole Polytechnique Federale de Lausanne, 1990.

[1.136] L. Lundkvist, I. Lundström, and C. Svensson, "Discharge of MNOS structures," *Sol. St. Electr.*, vol. 16, p. 811, 1973.

[1.137] R. A. Williams and M. E. Beguwala, "The effect of electrical conduction of nitride on the discharge of MNOS memory transistors," *IEEE Trans. Elect. Dev.*, vol. ED-25, p. 1019, 1978.

[1.138] G. Heyns, "Bijdrage von Electronen en Gaten tot de Conductie, Retentie en de Degradatie von MNOS Niet–Vluchtige Geheugens," Ph.D. Thesis, K. U. Leuven, Belgium, 1986, unpublished.

[1.139] G. Heyns and H. E. Maes, "A new model for the discharge behaviour of MNOS non-volatile memory devices," *Appl. Surf. Sci.*, vol. 30, p. 153, 1987.

[1.140] F. R. Libsch, A. Roy, and M. H. White, "A computer simulation program for erase/write characterization of ultra-thin nitride scaled SONOS/MONOS memory transistors," Ext. Abstr. Fall Meeting ECS 1986, vol. 86–2, Abstract 560, 1986.

[1.141] L. Blauner, "Study in reliability improvement of tunnel oxide for FLOTOX cell," presented at the 8th NVSM Workshop, Vail, Colo., August 1986.

[1.142] Y. Terada, K. Kobayashi, T. Nakayama, M. Hayashikoshi, Y. Miyawaki, N. Ajika, H. Arima, T. Matsukawa, and T. Yoshihara, "120ns 128kx8b / 64kx 16b CMOS EEPROM's," *IEEE ISSCC Dig. Tech. Pap.*, p. 136, 1989.

[1.143] B. Euzent, N. Boruta, J. Lee, and C. Jenq, "Reliability aspects of a floating gate EEPROM," *Procs. IRPS*, p. 11, 1981.

[1.144] K. Hieda, M. Wada, T. Shibata, I. Inoue, M. Momodomi, and H. Iizuka, "Optimum design of dual control gate cell for high density EEPROM's," *IEEE IEDM Tech. Dig.*, p. 593, 1983.

[1.145] E. Suzuki and Y. Hayashi, "Degradation properties of MNOS structures," *J. Appl. Phys.*, vol. 52, p. 6377, 1981.

[1.146] H. E. Maes and S. Usmani, "Charge pumping measurements on stepped-gate MNOS memory transistors," *J. Appl. Phys.*, vol. 53, p. 7106, 1981.

[1.147] A. Roy, F. R. Libsch, and M. H. White, "Investigations on ultrathin silicon nitride and silicon dioxide films in nonvolatile semiconductor memory transistors," *Proc. ECS*, vol. 87-10, Eds. V. J. Kapoor and K. T. Hankins, p. 38, 1987.

[1.148] H. E. Maes and R. J. Van Overstraeten, "Memory loss in MNOS capacitors," *J. Appl. Phys.*, vol. 47, p. 667, 1976.

[1.149] J. R. Cricchi, M. D. Fitzpatrick, F. C. Blaha, and B. T. Ahlport, "1 MRad hard MNOS structures," *IEEE Trans. Nucl. Sci.*, vol. NS-24, p. 2185, 1977.

[1.150] M. C. Peckerar and N. Bluzer, "Hydrogen annealed nitride/oxide dielectric structures for radiation hardness," *IEEE Trans. Nucl. Sci.*, vol. NS-27, p. 1753, 1980.

[1.151] P. Vail, "Radiation hardened MNOS : a review," *Proc. ECS*, vol. 83-8, p. 207, 1983.

Petitioner Micron Ex-1024, 0087

[1.152] M. G. Knoll, T. A. Dellin, and R. V. Jones, "A radiation-hardened 16 kbit MNOS EAROM," *IEEE Trans. Nucl. Sci.*, vol. NS-30, p. 4224, 1983.

[1.153] J. M. Caywood and B. L. Prickett, "Radiation induced soft errors in floating gate memories," Xicor Inc. Report.

[1.154] E. S. Snyder et al., "Radiation response of floating gate EEPROM memory cells," *IEEE Trans. Nucl. Sci.*, vol. NS-36, p. 2131, 1989.

[1.155] T. F. Wrobel, "Radiation characterization of a 28C256 EEPROM," *IEEE Trans. Nucl. Sci.*, vol. NS-36, p. 2247, 1989.

[1.156] D. Wellekens, G. Groeseneken, J. Van Houdt, and H. E. Maes, "On the total dose radiation hardness of floating gate EEPROM cells," Proc. NVMTR, IEEE Cat. #93TH0547-0, p. 54, 1993.

[1.157] D. Wellekens, G. Groeseneken, J. Van Houdt, and H. E. Maes, "Single poly floating gate cell as the best choice for radiation-hard EEPROM technology," *IEEE Trans. Nucl. Sci.*, vol. NS-40, p. 1619, 1993.

Petitioner Micron Ex-1024, 0088