

Chapter 2

Basic Principle of Nonvolatile Memory

2.1 Introduction

There is a widespread variety of Nonvolatile Memory (NVM) devices, and they all show different characteristics according to the structure of the selected cell and the complexity of the array organization. A NVM memory cell has to commute from one state to the other and that can store the information independently of external conditions. There are several methods to achieve the NVM memory characteristic, such as transistor V_T shifts, charge displacements, and resistance change [2.1]. In this thesis, we focus on one solution that a transistor with a threshold voltage that can change repetitively from a high to a low state, corresponding to the two states of the memory cell. Most operations with a shift in the threshold voltage on novel nonvolatile memories, such as nanocrystal and SONOS memories are base on the concept of Flash memory. If a datum has to be stored in a bit of the memory, there are different procedures. The threshold voltage shift of a Flash transistor can be written as [2.2][2.3]:

$$\Delta V_T = -\frac{\bar{Q}}{C_{FC}}$$

where \bar{Q} is the charge weighted with respect to its position in the gate oxide, and the capacitances between the floating gate and control gate. The threshold voltage of the memory cell can be altered by changing the amount of charge present between the gate and the channel, corresponding to the two states of the memory cell, i.e., the binary values (“1” and “0”) of the stored bit. Figure 2-1 shows the threshold voltage shift between two states in a Flash memory. To a nonvolatile memory, it can be “written” into either state “1” or “0” by either “programming” or “erasing” methods, which are decided by the definition of memory cell itself. There are many solutions to achieve “programming” or “erasing”.

In this chapter, we will discuss program/erase mechanisms from the relation between bias and energy band bending. Tunneling injection, channel hot electron injection, and band to band assisted electron/hole injection will be discussed briefly. The reliability of nonvolatile memory and physical characteristic of nanocrystal NVM will be also discussed.

2.2 Basic Program and Erase Mechanisms

2.2.1 Energy band diagram during program and erase operation

Fig. 2-2 illustrates the program/erase physical operation of a SONOS memory device. In the write operation, a positive voltage is applied on gate electrode relative to the p-type substrate, which forms an electron channel. Then the electrons tunnel through the tunnel oxide into the silicon nitride film and can be stored in deep-level traps. Some electrons which are not trapped in the nitride film will tunnel through a blocking oxide into the gate electrode. The trapped electrons provide the electrostatic screening of the channel from the control gate, and result in a threshold voltage (V_T) shift. During the erase operation under a negative voltage bias on the gate electrode, the holes tunnel from the substrate into the silicon nitride and are partially trapped in a manner similar to electrons. And some holes “pile-up” at the blocking oxide interface because of the larger barrier height (5eV). Further, trapped electrons may be de-trapped into the nitride conduction band and then tunnel back to the channel. Thus, for SONOS memory device operation both carrier types are involved in the transport process.

The write and erase processes for an n-channel semiconductor nanocrystal memory device are illustrated schematically in Fig. 2-3. During the write process, a positive gate voltage is applied to inject channel inversion-layer electrons into the nanocrystals. During the erase process, a reverse gate bias is applied to cause the electrons to tunnel back into the channel and the accumulation layer holes to tunnel into the nanocrystals from the channel.

2.2.2 Carrier Injection Mechanisms

(a) Tunneling Injection

Tunneling is a quantum mechanical process akin to throwing a ball against a wall often enough that the ball goes through the wall without damaging the wall or the ball. It also loses no energy during the tunnel event. The tunneling probability, depending on electron barrier height ($\phi(x)$), tunnel dielectric thickness (d), and effective mass (m_e) inside the tunnel dielectric, is expressed as [2-4]

$$T = \exp\left(-2 \int_0^d \frac{\sqrt{\phi(x) * m_e}}{\hbar} dx\right)$$

Basically, tunneling injection must have available states on the other side of the barrier for the carriers to tunnel into. Tunneling through the oxide can be attributed to different carrier-injection mechanisms. Which process applies depends on the oxide thickness and the applied gate field or voltage. Direct tunneling (DT), Fowler-Nordheim tunneling (FN), modified Fowler-Nordheim tunneling (MFN) and trap assisted tunneling (TAT) are the main programming mechanisms employed in memory [2.5-2.7] as shown in Figure 2-4.

Direct Tunneling

Direct Tunneling is the flow of electrons through the full oxide thickness illustrated in Figure 2-4(a). For nanocrystal memories, the control-gate coupling ratio of nanocrystal memory devices is inherently small [2.8]. As a result, F-N tunneling cannot serve as an efficient write/erase mechanism when a relatively thick tunnel oxide is used, because the strong electric field cannot be confined in one oxide layer. The direct tunneling is employed in nanocrystal memories instead. In the direct-tunneling regime, a thin oxide with thickness less than 3 nm is used to separate the nanocrystals from the channel. During program/erase operations, electrons/holes can pass through the oxide by direct tunneling, which gives the advantages of fast write/erase and low operation voltage. In the other hand, the direct tunneling is more sensitive to the barrier width than barrier height, two to four orders of magnitude reduction in leakage current can still be achieved if large work function metals, such as Au or Pt [2.9].

Fowler-Nordheim Tunneling

The Fowler–Nordheim (FN) tunneling is the flow of electrons through a triangular potential barrier illustrated in Figure 2-4(b). FN tunneling mechanism occurs when applying a strong electric field (in the range of 8–10 MV/cm) across a thin oxide. In these conditions, the energy band diagram of the oxide region is very steep. Therefore, there is a high probability of electrons' passing through the energy barrier itself. Using a free-electron gas model for the metal and the Wentzel–Kramers–Brillouin (WKB) approximation for the tunneling probability [2.10], one obtains the following expression for current density [2.11]:

$$J = \frac{q^3 F^2}{16\pi^2 h^2 \Phi_B} \exp\left[\frac{-4(2m_{ox}^*)^{1/2} \Phi_B^{3/2}}{3\hbar q F} \right]$$

Where Φ_B is the barrier height, m_{ox}^* is the effective mass of the electron in the forbidden gap of the dielectric, h is the Planck's constant, q is the electronic charge, and F is the electrical field through the oxide. The exponential dependence of tunnel current on the oxide-electric field causes some critical problems of process control because, for example, a very small variation of oxide thickness among the cells in a memory array produces a great difference in programming or erasing currents, thus spreading the threshold voltage distribution in both logical states.

Modified Fowler–Nordheim Tunneling

Modified Fowler–Nordheim tunneling (MFN) is similar to the traditional FN tunneling mechanism, yet the carriers enter the nitride at a distance further from the tunnel oxide-nitride interface. MFN mechanism is frequently observed in SONOS memories. The SONOS memory is designed for low-voltage operation (< 10V, depending on the Equivalent oxide thickness), a relatively weak electrical field couldn't inject charges by DT or FN mechanism.

Trap Assistant Tunneling

The charge storage mediums with many traps may cause another tunneling

mechanism. For example, the charges tunnel through a thin oxide and arrive to the traps of nitride layer at very low electrical field in SONOS systems. During trap assisted injection the traps are emptied with a smaller time constant then they are filled. The charge carriers are thus injected at the same distance into the nitride as for MFN injection. Because of the sufficient injection current, trap assistant tunneling may influence in retention [2.12].

(b) Channel Hot Electron Injection (CHEI)

The physical mechanism of HEI is relatively simple to understand qualitatively. An electron traveling from the source to the drain gains energy from the lateral electric field and loses energy to the lattice vibrations (acoustic and optical phonons). At low fields, this is a dynamic equilibrium condition, which holds until the field strength reaches approximately 100 kV/cm [2.13]. For fields exceeding this value, electrons are no longer in equilibrium with the lattice, and their energy relative to the conduction band edge begins to increase. Electrons are “heated” by the high lateral electric field, and a small fraction of them have enough energy to surmount the barrier between oxide and silicon conduction band edges (channel hot electron, CHE). Figure 2-5 shows schematic representation of CHEI MOSFET and the energy-distribution function with different fields. In the other hand, the effective mass of hole is heavier than one of electron. It is too hard to obtain enough energy to surmount oxide barrier. Therefore, hot-hole injection rarely is employed in nonvolatile memory operation.

Nevertheless, a description of the injection conditions can be accomplished with two different approaches. The HEI current is often explained and simulated following the “lucky electron” model [2.14]. This model is based on the probability of an electron’s being lucky enough to travel ballistically in the field for a distance several times the mean free path without scattering, eventually acquiring enough energy to cross the potential barrier if a collision pushes it toward the Si/SiO₂ interface. Consequently, the probability of injection is the lumped probability of the following events [2.15], which are depicted in Figure 2-6

- 1) The carrier has to be “lucky” enough to acquire enough energy from the lateral electric field to overcome the oxide barrier and to retain its energy after the collision

that redirects the electron toward the interface (P_{b}).

- 2) The carrier follows a collision-free path from the redirection point to the interface (P_{ED}).
- 3) The carrier can surmount the repulsive oxide field at the injection point, due to the Schottky barrier lowering effect, without suffering an energy-robbing collision in the oxide (P_{OC}).

(c) Band to Band Tunneling (BTBT)

Band to band tunneling application to nonvolatile memory was first proposed in 1989. I. C. Chen and et al. demonstrated a high injection efficiency (~1%) method to programming EPROM devices [2.16].

Band to Band Hot Electron Tunneling Injection

The injection is applied for n-type substrate nonvolatile memory device. Figure 2-7 shows the energy-band diagram and device operation during the band to band tunneling induced hot electron (BBHE) injection. When band-bending is higher than the energy gap of the semiconductor, the tunneling electron from the valence band to the conduction band becomes significant. The electrons are accelerated by a lateral electric field toward the channel region and some of the electrons with sufficient energy can surmount the potential barrier of SiO_2 like hot electron injection [2.16-2.18].

Band to Band Hot Hole Tunneling Injection

In p-type substrate, when a negative gate voltage and a positive drain voltage are applied to the cell, electron-hole pairs are generated by BTBT in the drain region, as shown in Figure 2-8. The holes are accelerated by a lateral electric field toward the channel region and some of them obtain high energy. The hot holes inject into charge trapping layer through the tunnel oxide and recombine the stored electrons. This injection is used for a new erase operation for nonvolatile memory device [2.19].

2.3 Basic Reliability of Nonvolatile Memory

For a nonvolatile memory, the important to concern is distinguishing the state in cell.

However, in many times operation and charges storage for a long term, the state is not obvious with charges loss. Endurance and retention experiments are performed to investigate Flash-cell reliability. In general, NVMs are required to withstand up to 10-100K program/erase cycles (endurance) with 10-year memory retention at temperatures as high as 85 °C.

2.3.1 Retention

Retention describes the ability to the NVM to store and recover information after a number of program cycles at a specified temperature. In any nonvolatile memory technology, it is essential to retain data for over ten years. This means the loss of charge stored in the storage medium must be as minimal as possible. For example, in modern Flash cells, FG capacitance is approximately 1 fF. A loss of only 1 fC can cause a 1V threshold voltage shift. If we consider the constraints on data retention in ten years, this means that a loss of less than five electrons per day can be tolerated [2.20]. Possible causes of charge loss are: 1) by tunneling or thermionic emission mechanisms; 2) defects in the tunnel oxide; and 3) detrapping of charge from insulating layers surrounding the storage medium; 4) mobile ion contamination. Further, the retention capability of Flash memories has to be checked by using accelerated tests that usually adopt screening electric fields and hostile environments at high temperature.

2.3.2 Endurance

The term “endurance” refers to the ability of the nonvolatile memory to withstand repeated program cycles and still meet the specifications in the data sheet. In a conventional Flash memory the maximum number of erase/program cycles that the device must sustain is 10^5 .

A typical result of an endurance test on a single cell is shown in Figure 2-9. As the experiment was performed applying constant pulses, the variations of program and erase threshold voltage levels are described as “program/erase threshold voltage window closure” and give a measure of the tunnel oxide aging. In particular, the reduction of the programmed threshold with cycling is due to trap generation in the oxide and to interface state generation at the drain side of the channel, which are mechanisms specific to hot-electron degradation. The initial lowering of V_T the erase is due to a pile-up of

positive charge which enhances tunneling efficiency. While the long-term increase of V_T the erase is due to a generation of negative traps.

Actually, endurance problems are mostly given by single-cell failures, which present themselves like a retention problem after program/erase cycles. In fact, a high field stress on thin oxide is known to increase the current density at low electric field. The excess current component, which causes a significant deviation from the current–voltage curves from the theoretical FN characteristics at low field, is known as stress-induced leakage current (SILC). SILC is clearly attributed to stress-induced oxide defects and, as far as a conduction mechanism, it is attributed to a trap assisted tunneling, as shown in Figure 2-10. The main parameters controlling SILC are the stress field, the amount of charge injected during the stress, and the oxide thickness. For fixed stress conditions, the leakage current increases strongly with decreasing oxide thickness [2.21-2.23]

2.4 Basic Physical Characteristic of Nanocrystal NVM

2.4.1 Quantum Confinement Effect

The quantum dot, is quasi-zero-dimensional nanoscaled material, and is composed by small amount atoms. The quantum confinement energy dependence on nanocrystal size has been studied both experimentally and theoretically with the tight -binding model [2.24]. The quantum confinement effect becomes significant when the nanocrystal size shrinks to the nanometer range, which causes the conduction band in the nanocrystal to shift to higher energy compared with bulk material [2.25]. For example, a 3nm Ge nanocrystal can have a conduction band shift of 0.5eV as compared with bulk Ge, which is significant enough to affect the electrical performance of the nanocrystal memory cell.

2.4.2 Coulomb Blockade Effect

When one electron is stored, the nanocrystal potential energy is raised by the electrostatic charging energy $e^2/2C$, where C is the nanocrystal capacitance, which depends mainly on the nanocrystal size, though it also depends on tunnel oxide thickness and control oxide thickness. The capacitance is self-consistently calculated using an electrodynamics method [2.26]. The electron charge will raise the nanocrystal potential

energy and reduce the electric field across the tunnel oxide, resulting in reduction of the tunneling current density during the write process. It is more dominant at low programming voltages (< 3V). In a flash memory array, device cells often encounter disturbances with low gate voltage soft-programming. The Coulomb blockade effect can effectively inhibit the electron tunneling at low gate voltage and improve the flash memory array immunity to disturbance. However, the Coulomb blockade effect should be reduced by employing large nanocrystal if large tunneling current and fast programming speed were desired. The Coulomb blockade effect has a detrimental effect on the retention time, since the electrons in the nanocrystal have large tendency to tunnel back into the channel if the nanocrystal potential energy is high in retention mode. In the energy band diagram, the Coulomb blockade charging energy only raises the electrostatic potential of the nanocrystal; the quantum confinement energy shifts the nanocrystal conduction band edge upward so that the conduction band offset between the nanocrystal and the surrounding oxide is reduced.



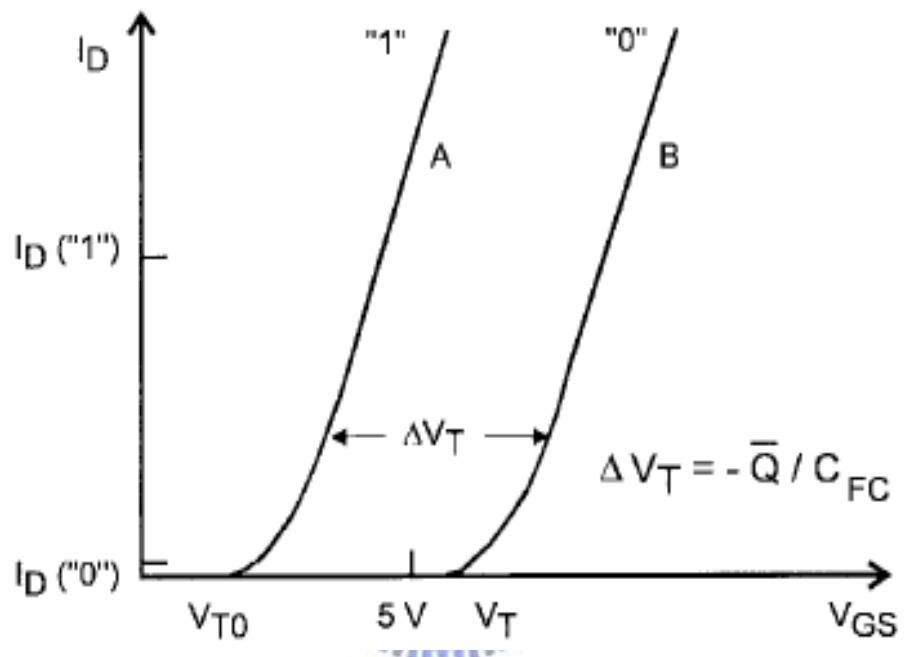
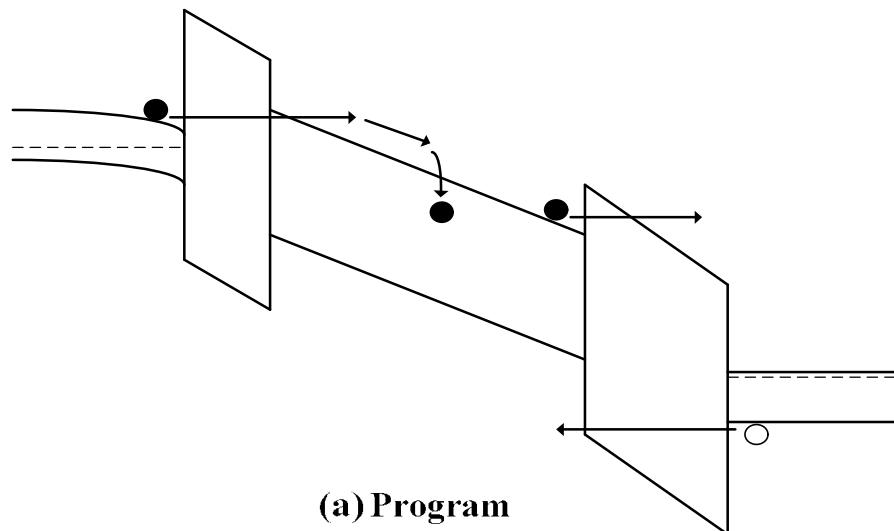


Figure 2-1 I-V curves of the floating-gate device when there is no charge stored in the floating-gate (curve A) and when a negative charge Q is stored in the floating-gate (curve B).

Si-sub / SiO₂ / Si₃N₄ / SiO₂ / N⁺ Poly gate



Si-sub / SiO₂ / Si₃N₄ / SiO₂ / N⁺ Poly gate

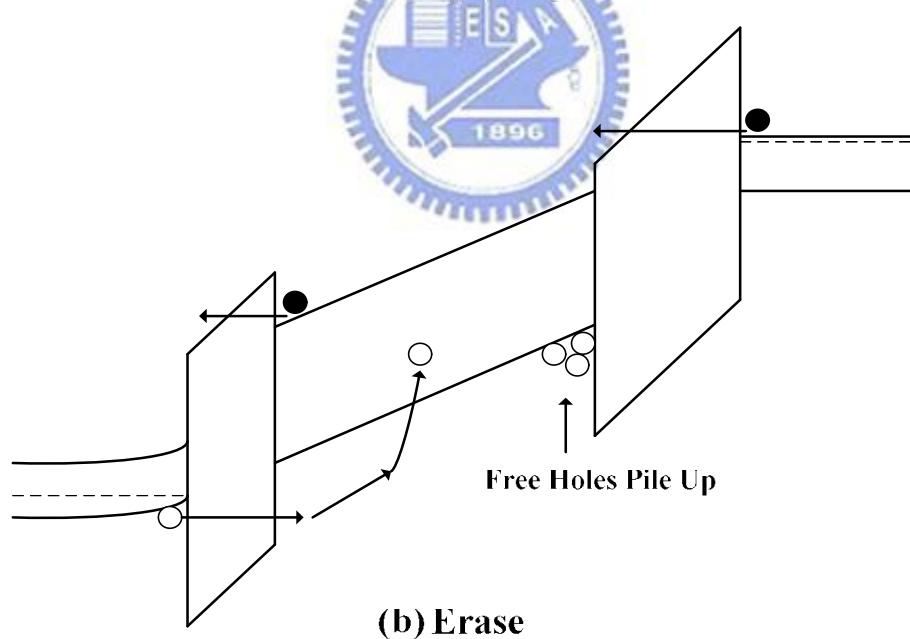


Figure 2-2 Energy band diagrams of the SONOS memory device under (a) program (b) erase operation. electrons, holes.

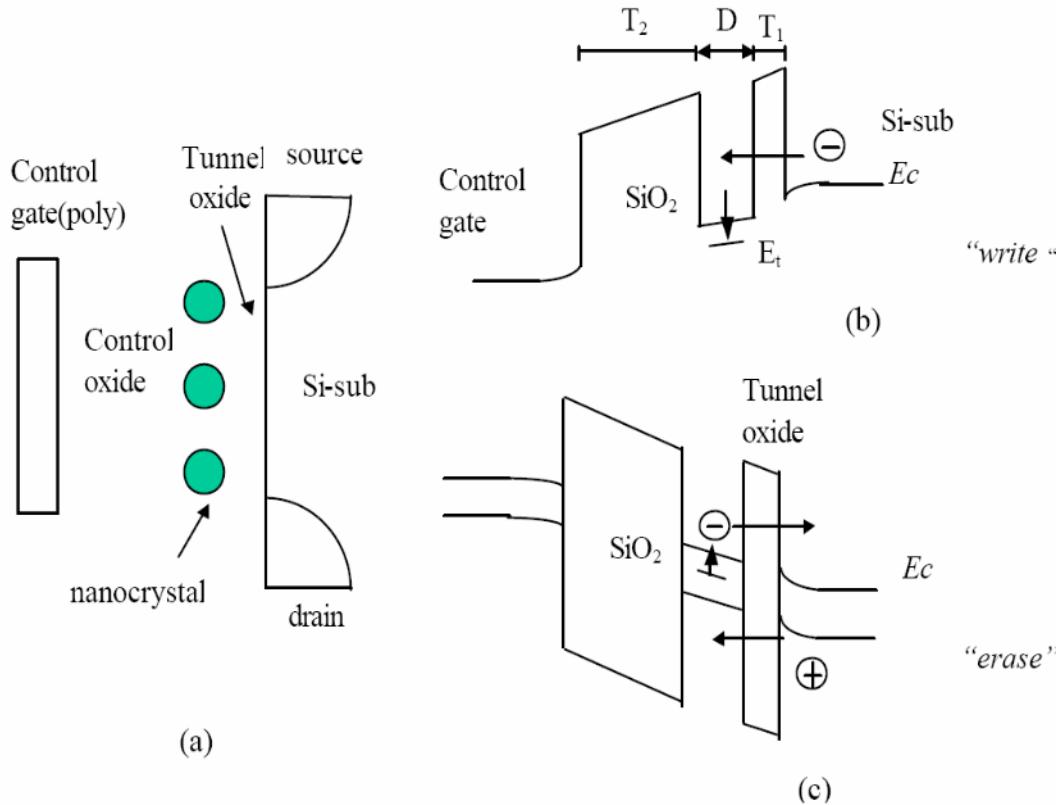
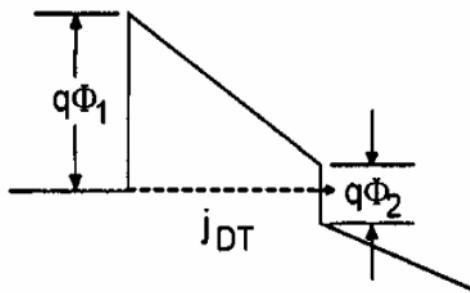


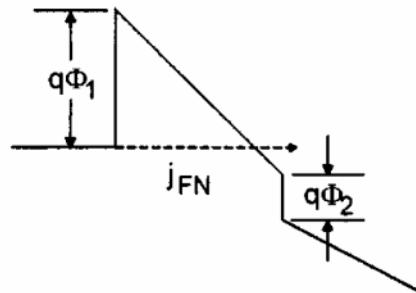
Figure 2-3 (a) Schematic cross-section of nanocrystal memory device structure; (b) illustration of write process: inversion-layer electrons tunnel into the nanocrystal; (c) illustration of erase process: accumulation layer holes tunnel into the nanocrystal, electron in nanocrystal can tunnel back to the channel.

(a)



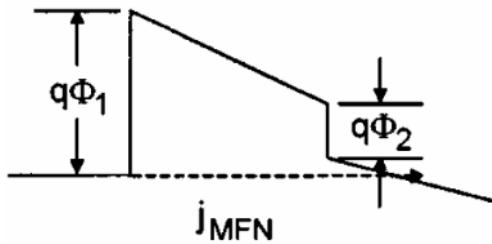
$$\text{DT occur when } \frac{\phi_1}{X_{OT}} > |E_{OT}| > \frac{\phi_1 - \phi_2}{X_{OT}}$$

(b)

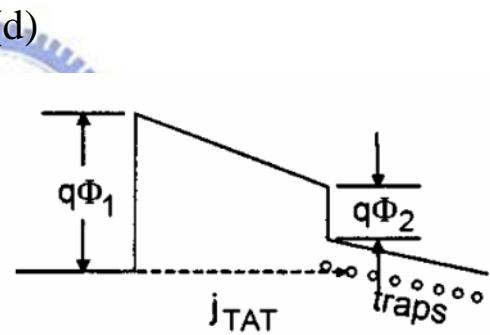
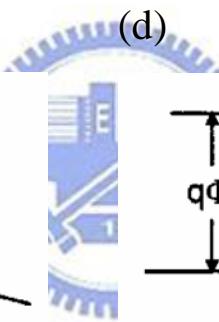


$$\text{FN occur when } |E_{OT}| > \frac{\phi_1}{X_{OT}}$$

(c)



$$\frac{\phi_1 - \phi_2}{X_{OT}} > |E_{OT}| > \frac{\phi_1 - \phi_2}{X_{OT} + \left(\frac{\epsilon_{OX}}{\epsilon_N}\right)X_N}$$



$$\frac{\phi_3}{X_{OT}} > |E_{OT}| > \frac{\phi_3}{X_{OT} + \left(\frac{\epsilon_{OX}}{\epsilon_N}\right)X_N}$$

$$\phi_3 = \phi_1 - \phi_2 - \phi_t$$

Figure 2-4 Fourth approaches to programming methods (a) Direct tunneling (DT) (b) Fowler-Nordheim (FN) tunneling (c) Modified Fowler-Nordheim (MFN) tunneling (d) Trap assistant tunneling (TAT).

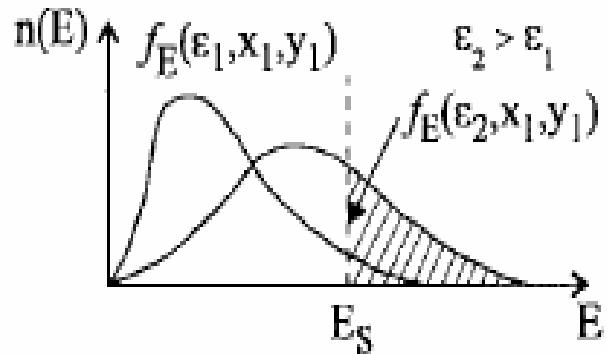
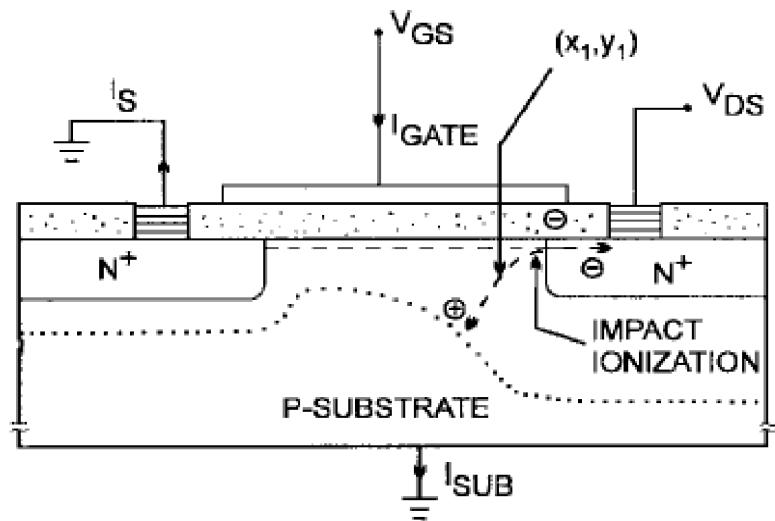


Figure 2-5 Schematics of channel hot electron injection (CHEI). The energy distribution function at point (X₁,Y₁) is also shown.

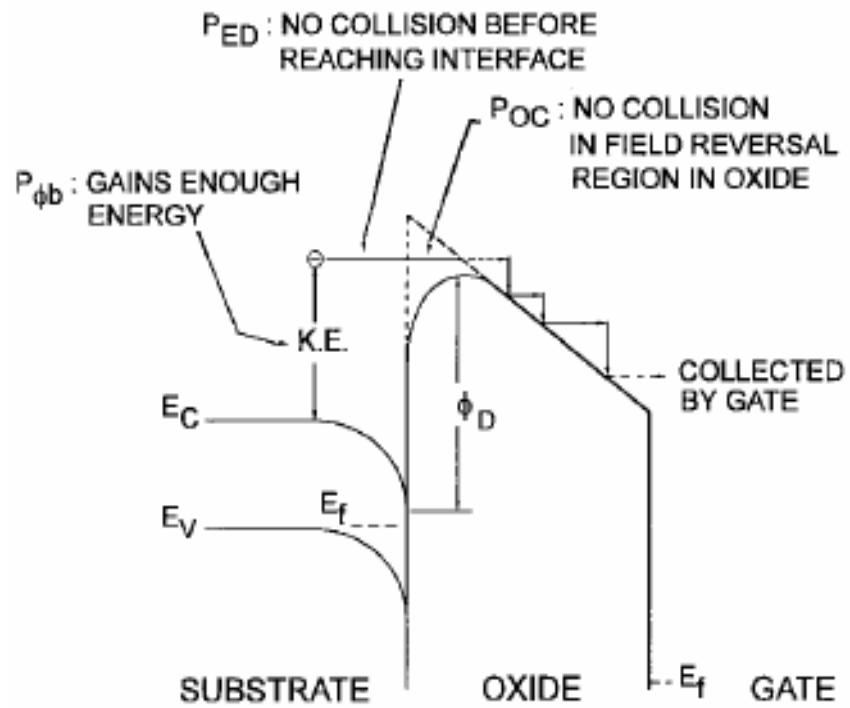


Figure 2-6 A schematic of energy band diagram describing the process involved in electron injection.

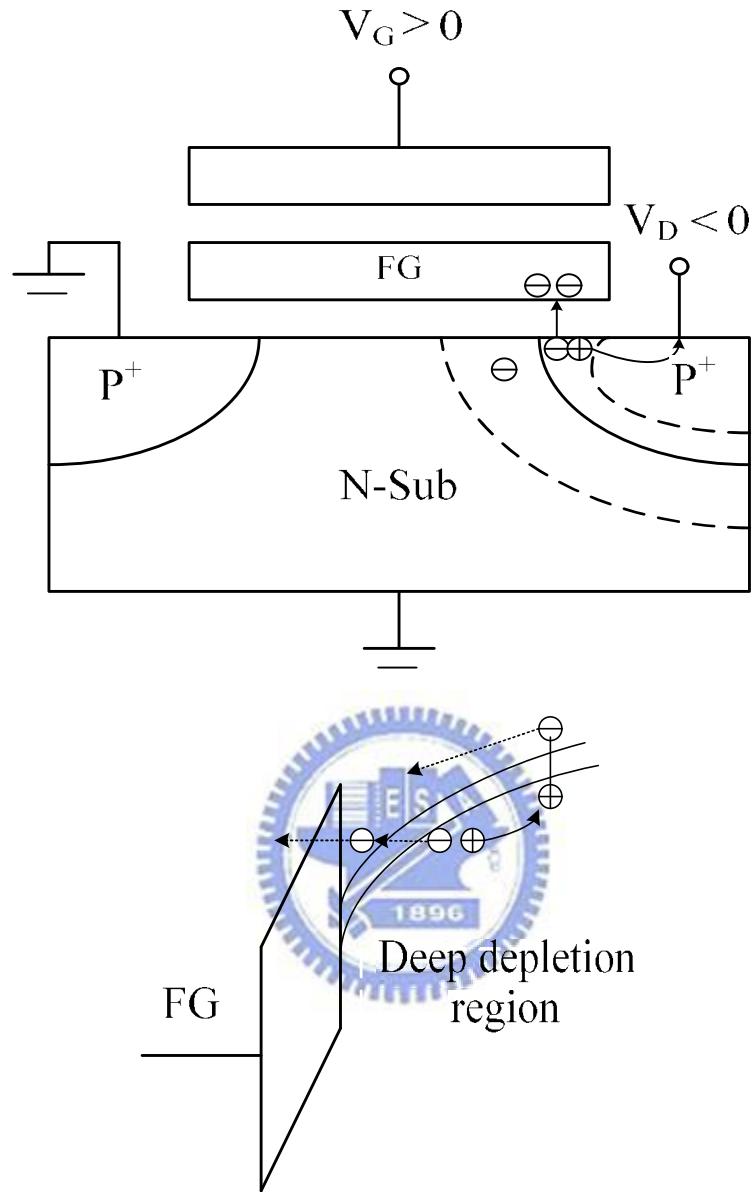


Figure 2-7 Energy-band diagram for the proposed band to band induce hot electron injection mechanism and schematic illustration cross of the Flash memory with p-channel cell. Due to the positive bias to the control gate, holes are not injected into the tunnel oxide.

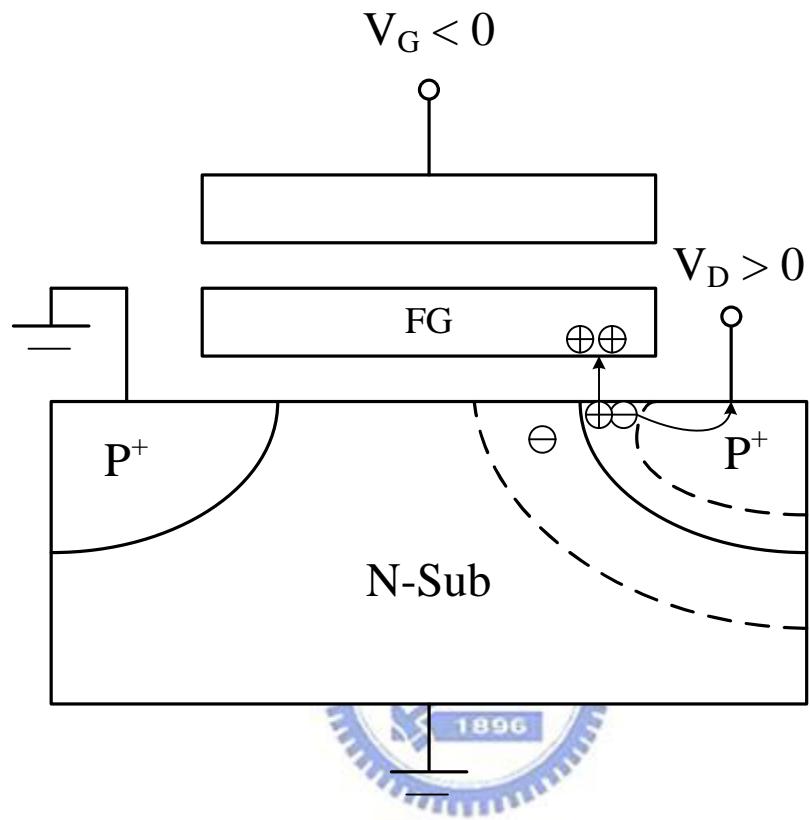


Figure 2-8 Band to band induce hot hole injection mechanism and schematic illustration cross of the Flash memory with p -channel cell.

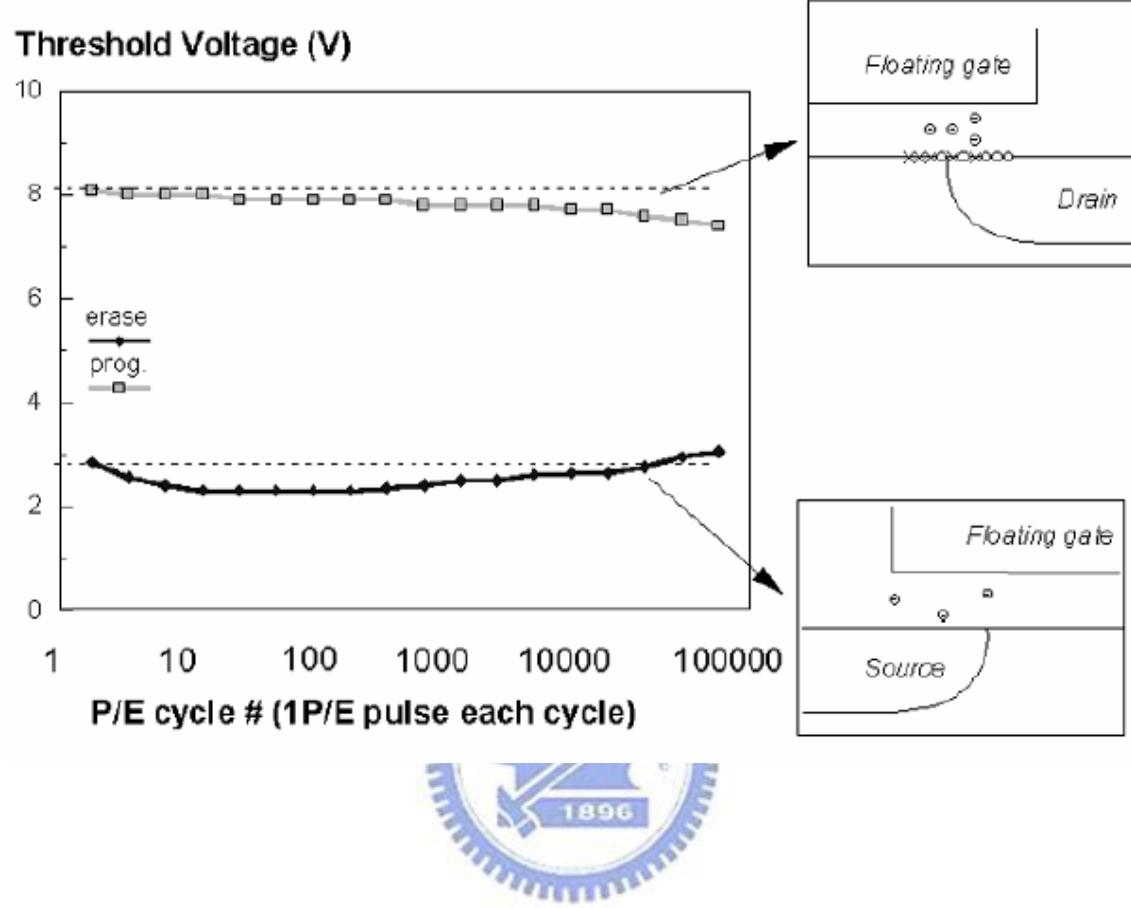


Figure 2-9 A typical result of an endurance test on a single cell. Threshold voltage window closure as a function of program / erase cycles.

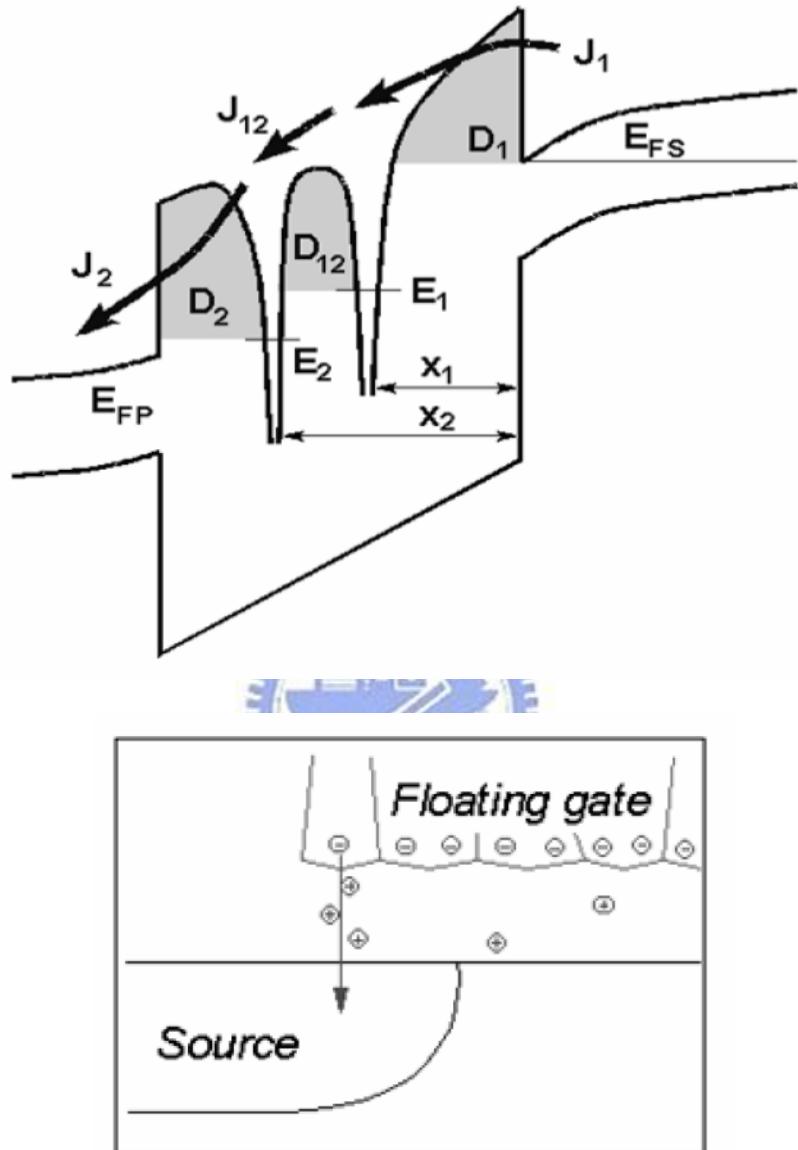


Figure 2-10 Anomalous SILC modeling. The leakage is caused by a cluster of positive charge generated in the oxide during erase. The multitrap assisted tunneling is used to model SILC: trap parameters are energy and position.