

# Chapter 1

## Introduction

### 1.1 Overview of Nonvolatile Memory

Today, flash memory find wide applications and are considered as a technology driver for semiconductor industry in the next generation. It can be classified into two major markets: code storage application and data storage application. NOR type flash memory [1.1] is most suitable for code storage application, such as cellular phones, PC bios, and DVD player. NAND type flash memory [1.2] has been targeted at data storage market, which is an emerging application such as PDA, memory cards, MP3 audio players, digital cameras, and USB flash personal disc. These products all are based on flash memory that is nonvolatile and can keep stored information also when the power supply is switched off. Flash memory also has exhibited several advantages, such as the ability to be electrical programmed and fast simultaneous block electrical erased in a single-cell, smallest cell size to achieve highest chip density, and good flexibility [1.3-1.4]. In addition, the flash memory fabrication process is compatible with the current CMOS process and is a suitable solution for embedded memory applications. Therefore, flash memories are easily scalable replacements for EPROMs (Erasable Programmable Read Only Memory) and EEPROMs (Electrically Erasable Programmable Read Only Memory). Since flash memory possesses these key advantages, it has become the mainstream nonvolatile memory device nowadays.

In 1967, D. Kahng and S. M. Sze invented the floating-gate (FG) nonvolatile semiconductor memory (or flash memory) at Bell Labs [1.5]. The conventional floating-gate device structure is shown in Fig. 1-1. The FG acts as the storing electrode and is electrically governed by a capacitively coupled control gate (CG). Charge injected in the FG is maintained there, allowing the difference between threshold voltages of the cell transistor for nonvolatile memory application.

Recently, nonvolatile memory devices are moving toward high density memory array, low cost, low power consumption, high-speed operation, and good reliability.

Although conventional flash memory does not require refreshing and thus consumes less power and achieves much higher array density with a stacked floating gate structure. However, floating-gate flash memory is much slower to operation and has poor endurance. In order to improve the write/erase speed of a floating-gate device, the thickness of the tunnel oxide must be reduced. But conventional FG memory devices have limited potential for aggressive scaling of the tunnel oxide thickness. The tunnel oxide must be thin enough to allow quick and efficient charge transport to and from FG. On the other hand, the tunnel oxide needs to provide superior isolation under retention, endurance, and disturbed conditions in order to guarantee the data integrity for 10 years. For faster operation speed, thin tunnel oxide is desirable. However, it is desirable to increase the thickness of tunnel oxide for better isolation and reliability. So there is a trade-off between speed and reliability for the optimum tunnel oxide thickness. Currently, commercial flash memory devices use tunnel oxide thicker about 8-11 nm, which results in high programming voltage and slow programming speed [1.6].

To alleviate the tunnel oxide design trade-off for floating-gate memory devices, memory-cell structures employing discrete traps as charge storage media have been proposed. In the conventional floating gate flash memory, if there is one defect created in the tunnel oxide, all the charges stored on the floating-gate will leak back to the channel or the source/drain through the weak spots. Unlike conventional continuous floating gate, charges stored in discrete nodes cannot easily redistribute amongst themselves. Therefore, only a relatively small number of nodes near the oxide defects will be affected. Local charge storage in discrete nodes enables more aggressive scaling of the tunnel oxide by relieving the total charge loss concern. There are two promising candidates, SONOS [1.7-1.9] and nanocrystal nonvolatile memory devices [1.10-1.12], that have been demonstrated to lead to an improvement in retention time compared with conventional floating gate memory. Hence the tunnel oxide thickness can be reduced to allow faster programming and lower voltage operation.

### **1.1.1 SONOS Nonvolatile Memory Devices**

The first nitride-base devices is metal-gate nitride device MNOS (Metal/Nitride/Oxide/Silicon) which was reported in 1967 by Wegener et al [1.13].

However, it is well known that silicon nitride film contains many carrier traps which cause threshold voltage shift. Then the silicon nitride trap-based devices are extensively studied for charge storage device application in the early 70s. Fig. 1-2 illustrates the progression of device cross section, which has led to the present SONOS device structure. Initial device structures in the early 1970s were p-channel metal-nitride-oxide-silicon (MNOS) structures with aluminum gate electrodes and thick (45 nm) silicon nitride charge storage layers. Write/erase voltages were typically 25-30 V. In the late 1970s and early 1980s, scaling moved to n-channel SNOS devices with write/erase voltages of 14-18 V. In the late 1980s and early 1990s, n- and p-channel SONOS devices emerged with write/erase voltages of 5-12 V. The advantages of the ONO triple dielectric structure are: (1) lower programming voltage since the blocking action of the top oxide removes any limitation on the reduction of the nitride thickness; (2) charge injection from and to the gate electrode is minimized for both gate polarities, particularly for hole injection; (3) improved memory retention since there is minimal loss of charge to the gate electrode.

The SONOS (poly-Silicon-Oxide-Nitride-Oxide-Silicon) memory devices, as shown in Fig. 1-3, have attracted a lot of attention due to its advantages over the traditional floating-gate flash device. These include reduced process complexity, high speed operation, lower voltage operation, improved cycling endurance, and elimination of drain-induced turn-on [1.14-1.16]. The main difference between floating-gate and SONOS structure is the method of charge storage. The charge storage media in the floating-gate structure is the conducting polysilicon floating-gate electrode. In the SONOS memory structure, charges are stored in the physical discrete traps of silicon nitride dielectric. A typical trap has a density of the order  $10^{18}$ - $10^{19}$   $\text{cm}^{-3}$  according to Yang et al [1.17] and stores both electrons and holes injected from the channel. The charges cannot move freely between the discrete trap locations, hence the SONOS memory device is very robust against the defects inside the tunnel oxide and has good endurance.

The SONOS memory devices still face challenge in the future for high density nonvolatile memory application, which requires low voltage ( $< 5\text{V}$ ), low power consumption, long-term retention, and superior endurance. Various approaches have been proposed for improving the SONOS performance and reliability. Chen *et al.*

demonstrated a  $\text{Si}_3\text{N}_4$  bandgap engineering (BE) control method for better endurance and retention. A nitride with varied relative Si/N ratio throughout the film has increased the charge-trapping efficiency significantly [1.18]. Tan *et al.* showed that over-erase phenomenon in SONOS memory structures can be minimized by replacing silicon nitride with  $\text{HfO}_2$  as the charge storage layer. The charge retention and endurance performance is improved by the addition of 10%  $\text{Al}_2\text{O}_3$  in  $\text{HfO}_2$  to form  $\text{HfAlO}$ , while maintaining the over-erase resistance of  $\text{HfO}_2$  [1.19]. She *et al.* demonstrates that high-quality nitride is applied as the tunnel dielectric for a SONOS-type memory device. Compared to control devices with  $\text{SiO}_2$  tunnel dielectric, faster programming speed and better retention time are achieved with low programming voltage [1.20]. Lee *et al.* presents a device structure of  $\text{SiO}_2/\text{SiN}/\text{Al}_2\text{O}_3$  (SANOS) with TaN metal gate. It is demonstrated that the use of TaN metal gate blocks electron current through  $\text{Al}_2\text{O}_3$  layer more efficiently than a conventional polysilicon gate, resulting in faster program/erase speed and significant decrease of the saturation level of the erase  $V_T$  [1.21].

Chen *et al.* studies a polycrystalline silicon thin-film transistor (poly-Si TFT) with oxide/nitride/oxide (ONO) stack gate dielectrics and multiple nanowire channels for the applications of both nonvolatile silicon-oxide-nitride-oxide-silicon (SONOS) memory and switch transistor [1.22]. The proposed NW SONOS-TFT exhibits superior memory device characteristics with high program/erase efficiency and stable retention characteristics at high temperature. Such a SONOS-TFT is thereby highly promising for application in the future system-on-panel display applications.

New device structures are also indispensable in making flash memory more scalable. Since SONOS flash memory offers a thinner gate stack than floating gate flash memory, and a FinFET structure controls the short channel effect much better than a bulk structure. It has been demonstrated that the FinFET SONOS flash memory devices with a much smaller cell size can provide both excellent performance and reliability. Therefore, FinFET SONOS memory has potential to become the candidate for the next generation flash memory [1.23-1.24].

## 1.1.2 Nanocrystal Nonvolatile Memory Devices

Nanostructure nonvolatile memories are first introduced in the early 1990s. IBM

researchers first proposed flash memory with a granular floating gate made out of silicon nanocrystals [1.25]. Fig. 1-4 illustrates conventional nanocrystal nonvolatile memory (NVM) device structures. It is observed that the nanocrystals are separated from each other within the gate dielectric. The term “nanocrystal” refers to a crystalline structure with a nanoscale dimension and its electronic properties seem more similar to an atom or molecule rather than the bulk crystal. For a nanocrystal NVM device, the charge storage media is in the form of mutually isolated nanocrystals instead of the continuous polysilicon layer. The limited size and capacitance of nanocrystals limit the numbers of stored electron, collectively the stored charges screen the gate charge and control the channel conductivity of the memory transistor.

Nanocrystal-based NVM devices have recently received much attention due to their potential to overcome the limitations of conventional polysilicon -based flash memory. Using nanocrystals as charge storage media offers several advantages, the main one being the potential to use thinner tunnel oxide without sacrificing non -volatility. This is a quite attractive proposition since reducing the tunnel oxide thickness is a key to lowering operating voltages and/or increasing operating speeds. This claim of improved scalability results from the local charge storage in discrete nodes, which makes the storage more fault-tolerant and immune to the leakage caused by localized oxide defects. Further, the lateral charge migration effect between nanocrystals can be suppressed by the strongly isolation of surrounded dielectric. There are other important advantages though. First, nanocrystal memories use a more simplified fabrication process as compared to conventional stacked-gate FG NVM's by avoiding the fabrication complications and costs of a dual-poly process. Second, due to the absence of drain to FG coupling, nanocrystal memories suffer less from drain-induced-barrier-lowering (DIBL) and therefore have intrinsically better punch-through characteristics. One way to exploit this advantage is to use a higher drain bias during the read operation, thus improving memory access time [1.26]. Alternatively, it allows the use of shorter channel lengths and therefore smaller cell area. Finally, nanocrystal memories are characterized by excellent immunity to stress induced leakage current (SILC) and oxide defects due to the distributed nature of the charge storage in the nanocrystal layer.

Research in this regime has focused on the development of fabrication processes and

nanocrystal materials, and on the integration of nanocrystal -based storage layers in actual memory devices.

The fabrication of a nonvolatile memory cell requires a perfect control of four main parameters: (1) the tunnel oxide thickness, (2) the nanocrystal density, (3) the nanocrystal size, and (4) the control oxide thickness. An important consideration is the average size and aerial density of the nanocrystals. Larger-size nanocrystal array provides higher program/erase efficiency due to small quantum confinement and coulomb blockade effects, and hence larger tunneling probability. However, it is desirable to reduce the nanocrystal size for better reliability (stress induced leakage during retention). So there is a trade-off between programming speed and reliability in selecting the nanocrystal size. A typical target is a density of at least  $10^{12} \text{ cm}^{-2}$ , and requires nanocrystal size of 5 nm and below. Moreover, good process control is needed with regards to such nanocrystal features as: planar nanocrystal layer; inter-crystal interaction (lateral isolation); and crystal doping (type and level). Finally, it is preferred that the fabrication process is simple and that it uses standard semiconductor equipment.

After the first proposal of a memory transistor using silicon nanocrystals as floating gates. In order to improve the data retention in NVM, double layer Si nanocrystals memory has been investigated [1.27]. It seems interesting to use Ge nanocrystals rather than Si nanocrystals because of its smaller band gap. Indeed King and Hu have recently demonstrated the superior memory properties of Ge based nanocrystal memories over those based on Si [1.28]. Recently, germanium/silicon (Ge/Si) nanocrystals have been reported to possess superior charge retention capability than Ge or Si nanocrystals. This is due to the fact that Ge has a smaller band gap than Si and thus by introducing a Si interface around the Ge nanocrystal, it would create an additional barrier height at the Ge/Si interface which makes it harder for electrons to leak out of the nanocrystal [1.29,1.30]. However, semiconductor nanocrystal memory may not be the ultimate solution to nonvolatile memory scaling, although it still attracts a lot of attention now.

In optimizing nanocrystal NVM devices, the ideal goal is to achieve the fast write/erase of DRAM and the long retention time of Flash memories simultaneously. For this purpose we need to create an asymmetry in charge transport through the gate dielectric to maximize the  $I_{G, \text{Write/Erase}} / I_{G, \text{Retention}}$  ratio. One approach for achieving this

goal is to engineer the depth of the potential well at the storage nodes, thus creating a small barrier for writing and a large barrier for retention between the substrate and the storage nodes. This can be achieved if the storage nodes are made of metal nanocrystals by engineering the metal work function. The major advantages of metal nanocrystals over semiconductor nanocrystals include higher density of states around the Fermi level, scalability for the nanocrystal size, a wide range of available work functions, and smaller energy perturbation due to carrier confinement [1.31]. In addition, an electrostatic modeling from both an analytical formulation and numerical simulation is demonstrated that the metal nanocrystals will significantly enhance the electric field between the nanocrystal and the sensing channel set up by the control gate bias, and hence can achieve much higher efficiency in low-voltage P/E [1.32].

Toward better NVM device performance and reliability, numerous attempts have been made using metal nanocrystals. Liu *et al.* reported the growth of Au, Pt, and Ag nanocrystals on SiO<sub>2</sub> using an e-beam deposition method [1.31]. Lee *et al.* proposed a NVM structure using the Ni nanocrystals and high-*k* dielectrics [1.33]. Chen *et al.* present the stacked Ni silicide nanocrystal memory was fabricated by sputtering a comix target followed by a low temperature RTO process [1.34]. W nanocrystals on atomic-layer-deposited HfAlO/Al<sub>2</sub>O<sub>3</sub> tunnel oxide were presented for application in a memory device [1.35]. Using W nanocrystal double layers embedded in HfAlO to enhancement of memory window was demonstrated from the short channel devices down to 100nm [1.36]. Tang *et al* demonstrate that a chaperonin protein lattice can be used as a template to assemble PbSe and Co nanocrystal arrays for Flash memory fabrication. This provides a new approach to achieve a high density and good distribution uniformity nanocrystal array [1.37].

In the future, the primary drivers behind nanocrystal memories are the potential to scale the tunnel oxide thickness, resulting in lower operating voltages, and the simplicity of a single poly-silicon process. But there are still challenges await nanocrystal memories in the long road to commercialization. Nanocrystal memories have yet to deliver on most of their promises. In reality, part of the voltage gain is offset because of the poor control gate coupling. For fabrication processes, it is hard to control the uniformity of the nanocrystal size and their physical locations in the channel. It is not a surprise that

nanocrystal memories exhibit large device-to-device variation. Moreover, it has yet to be demonstrated that both the nominal and the statistical retention behavior are sufficient to meet true non-volatility requirements. Although single-dot memories have been demonstrated [1.38,1.39], but a more fundamental understanding of the scaling limits of nanocrystal memories is necessary, concentrating especially on the aspect of controlling channel conductance when relying on only a few discrete charge centers [1.40]. Finally, in order for that to happen, their claimed benefits will need to be more unambiguously substantiated, and a more appealing bundle of memory features will have to be demonstrated.

## 1.2 Organization of This Thesis

**In Chapter 1**, general background of flash, SONOS, and nanocrystal nonvolatile memory devices are introduced.

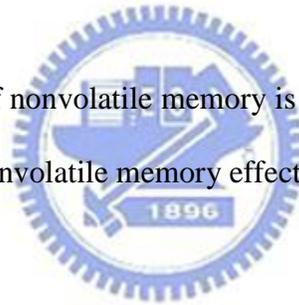
**In Chapter 2**, basic principle of nonvolatile memory is introduced.

**In Chapter 3**, formation and nonvolatile memory effect of Ni-Si-O and Ni-Si-N nanocrystal.

**In Chapter 4**, improved performance of nonvolatile Ni-Si-O and Ni-Si-N nanocrystal memory after thermal treatment.

**In Chapter 5**, multi-layer Ni silicide nanocrystal memory.

Finally, the conclusion is presented **in Chapter 6**.



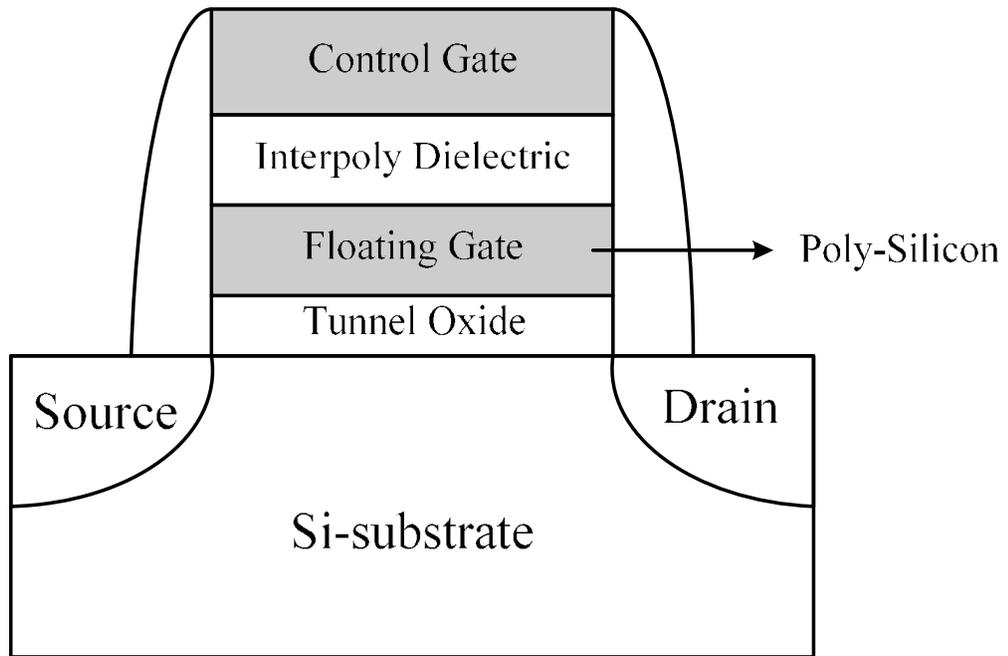


Figure 1-1 The structure of the conventional floating-gate nonvolatile memory device. Continuous poly-Si floating gate is used as the charge storage element.

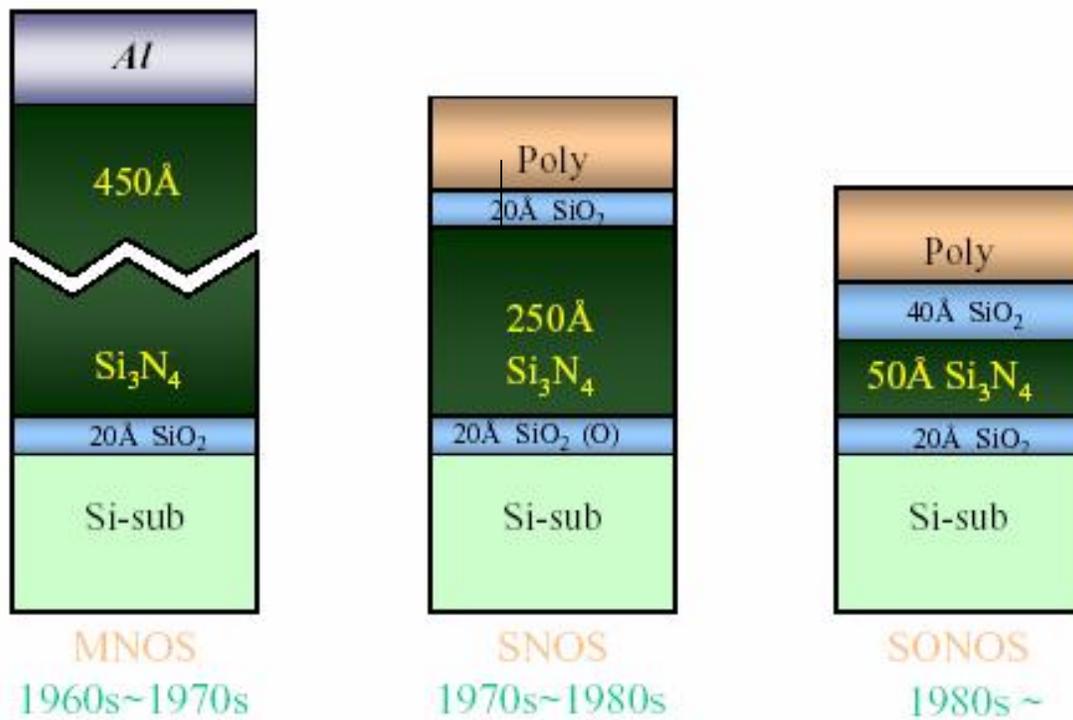


Figure 1-2 The development of the gate stack of SONOS EEPROM memory devices. The optimization of nitride and oxide films has been the main focus in recent years.

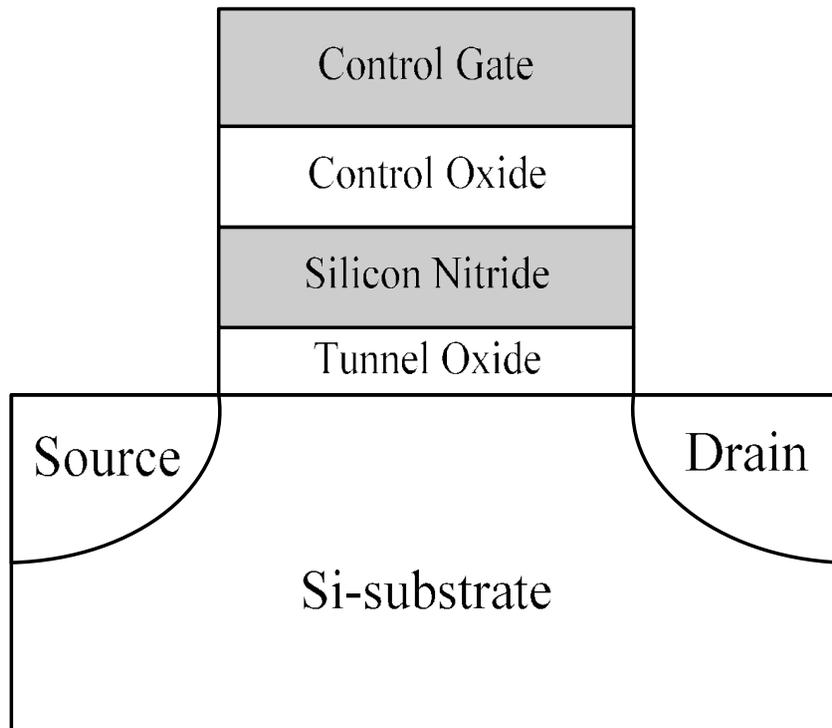


Figure 1-3 The structure of the SONOS nonvolatile memory device. The nitride layer is used as the charge trapping media.

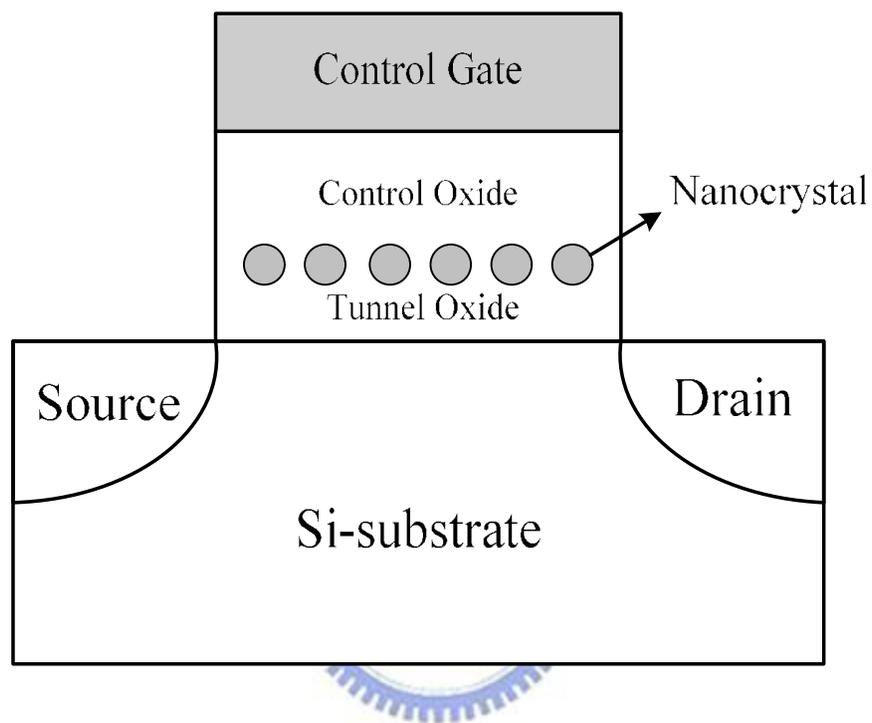


Figure 1-4 The structure of the nanocrystal nonvolatile memory device. The semiconductor nanocrystals or metal nanocrystals are used as the charge storage element instead of the continuous poly -Si floating gate.