Advanced Structures and New Detection Methods for Future High Density Non-volatile Memory Technologies

Alvaro Padilla

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-9
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-9.html

January 29, 2008

Acknowledgement

I take this opportunity to express my gratitude to all those who supported me through my journey in both undergraduate and graduate school at Berkeley. First and foremost, I express my thanks to my advisor, Professor Tsu-Jae King Liu, whose unfailing support, encouragement and supervision has made my journey through graduate school possible. I also express my gratitude to all those who supported my application for admission into graduate school: Thanks to Professor Luke Lee for encouraging me to apply to the Berkeley Edge Program. Thanks to Professor Nathan Cheung for his guidance and support during the application process. Thanks to both Carla Trujillo and Professor Kristofer Pister for their unfailing support, common vision, and pivotal role in my admission to Berkeley.

Advanced Structures and New Detection Methods for Future High Density Non-Volatile Memory Technologies

by

Alvaro Padilla

B.S. (University of California at Berkeley, USA) 1997
M.S. (University of California at Berkeley, USA) 2005

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Chenming Hu
Professor Peter Yu

Fall 2007

The dissertation of Alvaro Padilla is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2007

Advanced Structures and New Detection Methods for Future High Density Non-Volatile Memory Technologies

by

Alvaro Padilla

ABSTRACT

Advanced Structures and New Detection Methods for Future High Density Non-Volatile Memory Technologies

by

Alvaro Padilla

Doctor of Philosophy in Engineering –
Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

The proliferation of portable electronic devices has spawned demand for ultra-high-density non-volatile semiconductor memory (NVM). Until recently, aggressive scaling of conventional (flash, SONOS) NVM cell structures (coupled with the use of algorithms that enable the storage of multiple bits of information within every cell) has resulted in a significant increase in NVM storage density. However, additional scaling of these technologies (beyond the 45nm node) is a major challenge due to both short-channel effects (SCE) and the enhanced cell-to-cell variation (in threshold voltage, V_T) that results from NVM cell structures with smaller dimensions.

This dissertation investigates the use of novel materials, charge detection methods and NVM Field Effect Transistor (FET) structures that (in principle) enhance the scalability of conventional semiconductor flash memory technologies. This assessment proposes solutions (based on materials and structures) that are compatible with

conventional CMOS process flows. Chapter 1 introduces the main challenges affecting the scalability of conventional NVM cell structures. Chapter 2 explores the use of high-k dielectrics within the gate-stack of a charge-trapping NVM cell and highlights both the limited benefits obtained with this approach and the need for a new charge detection method that mitigates variation and has reduced sensitivity to charge stored in the complementary bit(s) of the structure. Chapters 3 through 6 explore the use of double-gated Silicon-on-Insulator FET (DG-FET) structures as NVM cells. In Chapter 3, a dual-bit FinFET SONOS NVM cell structure is demonstrated. This structure can utilize either the conventional and/or a novel read method to independently distinguish the digital information stored at either bit. Since the novel read method is less sensitive to charge stored in the complementary bit, its use alone can enhance the scalability of multi-bit NVM cells. In Chapter 4 (5), a novel n-channel (p-channel) dual-bit FinFET-based NVM cell design with two separate gate-sidewall charge-storage sites is presented for the first time. This Gate-Sidewall Storage (GSS) cell design enhances the scalability of conventional SONOS cells since it can utilize a thinner gate-stack EOT and its charge-storage sites are physically separated (which suppresses sensitivity to charge stored in the complementary bit). Finally, Chapter 6 explores the use of (either SONOS or GSS) DG-FET's as 4-bit NVM cell structures. In terms of layout efficiency, the optimum practical implementation of these structures involves the use of the Back-Gated FET design, since its use most effectively reduces the size per bit of each unit cell within either NOR- or NAND-type array architectures.

_____

Professor Tsu-Jae King Liu, Committee Chair

To H.L.

Table of Contents

ACKNOWLEDGEMENTS

I take this opportunity to express my gratitude to all those who supported me through the peaks and valleys of my journey through both undergraduate and graduate school at Berkeley. First and foremost, I express my thanks to my advisor, Professor Tsu-Jae King Liu, whose unfailing support, encouragement and supervision has made my journey through graduate school possible. I also express my gratitude to all those who supported my application for admission into graduate school: Thanks to Professor Luke Lee (from the Bioengineering department) for encouraging me to apply to the Berkeley Edge Program. Thanks to Professor Nathan Cheung for his guidance and support during the application process. Thanks to both Carla Trujillo (former Direction of the Minority Engineering Program) and Professor Kristofer Pister for their unfailing support, common vision, and pivotal role in my admission to Berkeley.

I would like to thank Professor Chenming Hu, for chairing my qualifying examination committee and serving as a member of my dissertation committee. I also thank Professor Peter Yu for serving on both my qualifying examination committee as well my dissertation committee.

I also would like to thank all staff members of the Microfabrication Lab for their help and support in my projects. I am also grateful to Ruth Gjerde for her invariant kindness and timely information regarding department policy. I am thankful to many former and current members of the device group, and especially to Dr. Kyoungsub Shin and Noel Arellano for their assistance during the fabrication of FinFETs. I also thank David Carlton, Sunyeong Lee, and Chun Wing Yeung for their collaboration on my projects.

Finally (on a personal note), I would also like to thank all those who (either directly or indirectly) assisted me in my painful, difficult immigration to the United States and the eventual achievement of one of the most precious dreams of my life (the attainment of an undergraduate *and* graduate education at one of the most prestigious universities in this country). I started life in this country with absolutely nothing (but dreams and aspirations), and I can honestly say that I now have the life I always wanted to live. Thanks Mom for your constantly 'bugging' my father for those sporadic $50 dollar checks that eventually allowed me to finish high school in Mexico (thanks for those checks, Don Pepe). Thanks Maria, Marcos, and Don Pepe for assisting with affordable housing (in my first two years in the United States) so that I could go to community college in the evenings after work (and resume my education there). Thanks Heather (an angel sent from above to rescue me) for your unconditional support during our time together –your presence in my life has been quite meaningful; I could not have done *any* of this without you.

Thanks God for your blessing!

Chapter 1: Conventional Non-Volatile Memories

1.1 Introduction

Semiconductor memories are an essential component of all kinds of modern electronic systems (such as personal computers, digital cameras, cellular phones, or smart-media networks). These systems utilize various digital memory modules to store digital information (either permanently or temporarily), which can either be code (e.g. code used to run application software) or data collected over time by the electronic system.

Semiconductor memories can be classified into two broad categories: volatile memories, and Non-Volatile Memories. The former is a type of memory that retains data as long as the power supply is switched on. Overall, the main advantage of volatile memories is their ability to operate at very high speeds; hence, these memories are mainly used in the execution of code (where fast access to data is necessary). Their main disadvantages include their low storage density and their need of a power supply to operate.

Semiconductor memories that retain the information stored in them once the power supply is switched off are called “Non-Volatile” Memories (NVM). This feature (of non-volatility) has made these types of memories particularly well suited for portable electronic devices (such as cellular phones, digital cameras, USB keys, MP3 players, etcetera) since these devices require field updates of code or data and their portability

requires the user's ability to store the updated information in real time (and maintain it after the power is switched off).

In the last three decades, there has been an explosive growth of portable electronic devices. With this growth, several NVM technologies have been developed and introduced over the years. The first non-volatile memories that were introduced were read-only memories (ROM). In these memories, the stored information was encoded into the circuit topology (during the fabrication of the chip) through the addition or removal of diodes or transistors within the memory array layout [1.1]. Since the information stored in these memories is hardwired, the data stored cannot be modified; it can only be read. As a result, the use of ROM is only viable to a limited number of applications. To facilitate end-user programming of memories, Erasable Programmable Read Only Memories (EPROM) were introduced. EPROM can be electrically programmed; however, these memories have to be removed and exposed to ultra-violet (uv) light for a few minutes in order to erase their contents. Subsequently, Electrically Erasable Programmable ROMs (EEPROM) were introduced. These memories are electrically erasable and programmable; however, the original version of EEPROMs used larger areas than EPROMs, and had therefore higher manufacturing costs and lower memory densities. Eventually, all these technological innovations led to the technology now referred to as *flash* memory. This is a NVM technology in which a single cell can be electrically programmed and either a single cell or a large number of cells (a block, sector, or page) can be electrically erased at the same time. The word flash refers to the ability of this technology to erase an entire block of memory cells at the same time (similar to EPROM). This technology combines the high memory density of EPROM

(since it also utilizes a single transistor as its unit cell) with the electrical in-system ability to erase of EEPROMs. Because of these features, this technology is now considered the driver of the NVM industry and one of the drivers of the semiconductor industry.

The proliferation of portable electronic devices (and the ongoing tendency to converge various applications into a single portable electronic device, see **Figure 1.1**) has now spawned demand for ultra-high-density non-volatile semiconductor memory (NVM); consequently, research in this area is necessary (to meet this demand).

Figure 1.1: The convergence of different applications into a single portable electronic device has now spawned demand for ultra-high density NVM [1.2].

1.2 The Standard Flash NVM Cell

1.2.1 Cell Structure and Operating Principles

The standard flash NVM cell structure (**Figure 1.2**) is similar to the traditional transistor structure, except that its gate oxide is essentially modified to include a charge-trapping film (*i.e.*, poly-crystalline silicon for a floating-gate NVM cell) that is placed between a tunnel oxide (Tt_{ox}) and a Control Oxide (Ct_{ox}) film. Bit information is stored in this cell through the controlled placement of electrons onto its charge-trapping film. The charge stored in the charge-trapping region modulates the threshold voltage (V_T) of the transistor, and this modulation allows the identification of the presence of electrons, or lack thereof, stored within the charge-trapping region of the cell

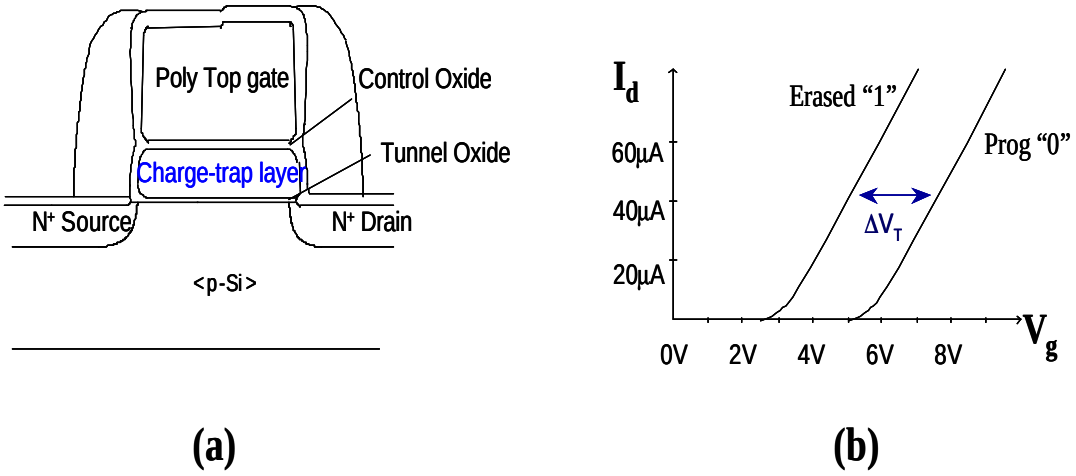


Figure 1.2: (a) Schematic cross-section of the conventional flash memory cell. (b) The storage of electrons within the charge-trapping layer modulates the cell's threshold voltage (V_T), and this modulation can thus be used to detect charge storage (and hence the information stored within the cell).

In a single-bit-per-cell structure, this V_T modulation results in two binary states: a state with a low V_T (binary state "1"), and a state with a high V_T (binary state "0"). For optimum performance, the separation between these two states (measured as ? V_T) needs to be maximized so that each state can be properly recognized; additionally, charge leakage from the charge-trapping film must be minimized so that the data stored in the cell is maintained as long as possible.

Three basic operations (read, program, and erase) are performed on a NVM cell in order to access or modify the information stored within it. A "read" operation refers to the (current or voltage) measurement performed on a NVM cell to distinguish the digital information stored within it. In the conventional NVM cell, this is achieved by sensing the transistor current in the on state, since its threshold voltage (V_T) will be affected by the presence of charge stored within the charge-trapping layer (particularly, those electrons placed in the region next to the source electrode). In practice, this is achieved by placing the n-channel NVM cell in inversion mode through application of large positive drain-to-source (V_{DS}), gate-to-source (V_{Read}) bias voltages (**Figure 1.3**). The charge state of the cell is then determined from the transistor's drain-to-source (I_{DS}) current: if electrons are stored, the threshold voltage will be high so that the cell's I_{DS} read current will be low; if no electrons are stored, the read current will be high. The resulting cell's I_{DS} current is compared (using a sense amplifier) against the current of a reference cell (whose V_T is normally placed halfway between the expected V_T of programmed and erased cells). Thus, since an erased cell has a lower V_T, it draws more current than the current of the reference cell; conversely, the programmed cell has a higher V_T, hence it draws less current than the current of the reference cell (I_{ref}). In

either case, the sense amplifier identifies the difference in current, and the binary state of the cell is thus determined.

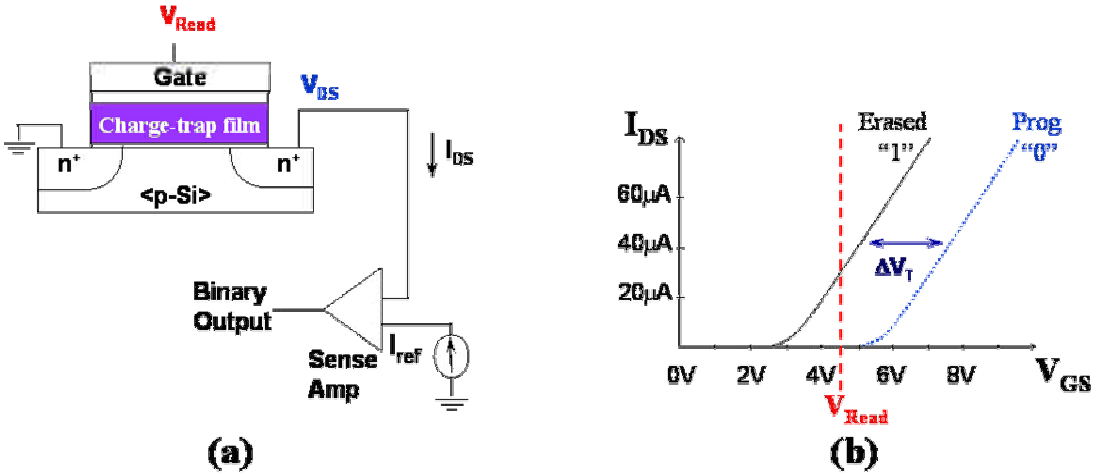


Figure 1.3: Read operation in a NVM cell: (a) Measurement setup. (b) A measure of the cell's on-state I_{DS} current (at a specific gate voltage, 'V_{Read}', and drain-to-source voltage, 'V_{DS}') permits identification of the binary state of the NVM cell.

Programming is the act of placing electrons onto the charge-trapping site of a NVM cell. To program the cell, two methods are predominantly employed: Hot Electron Injection ((HEI) [1.3], and Fowler-Nordheim (FN) tunneling [1.4]. In the HEI method (**Figure 1.4(a)**), electrons are first accelerated laterally through application of a large enough lateral field (i.e., a large V_{DS}), with a small portion of these electrons (the hot electrons) gaining enough kinetic energy to overcome the potential barrier (~3.1 eV) imposed by the silicon-oxide interface (i.e., the bottom interface of the tunnel silicon dioxide film). Some of these hot electrons (referred to as 'lucky' electrons in the "lucky electron" model [1.5] often used to model this programming mechanism) are then re-

directed towards the charge-trapping film (in the region near to the drain electrode) through application of a large transverse field (i.e., a large V_{GS}). With these settings, the lucky electrons get trapped within the charge-trapping film, thereby programming the cell.

In the FN Tunneling programming mechanism (**Figure 1.4(b)**), the Source and Drain (S/D) electrodes are either kept floating or grounded, and a large and positive transverse electric field (i.e., a large gate-to-bulk bias, V_{GB}) is applied to the structure such that an inversion channel exists under the gate and the electrons in this channel can quantum-mechanically tunnel through the tunnel oxide and become trapped within the charge-trapping region of the NVM cell. Since this programming mechanism utilizes tunneling currents (to introduce electrons onto the charge-trapping site of the cell), its programming rate is much slower (in the order of milliseconds) than that of the HEI mechanism (in the order of microseconds), which utilizes hot carriers and large currents to program the cell. Consequently, since the programming rate with this method is directly proportional to the tunnel oxide thickness of the structure, a thin tunnel oxide is preferred for a fast programming rate with this method.

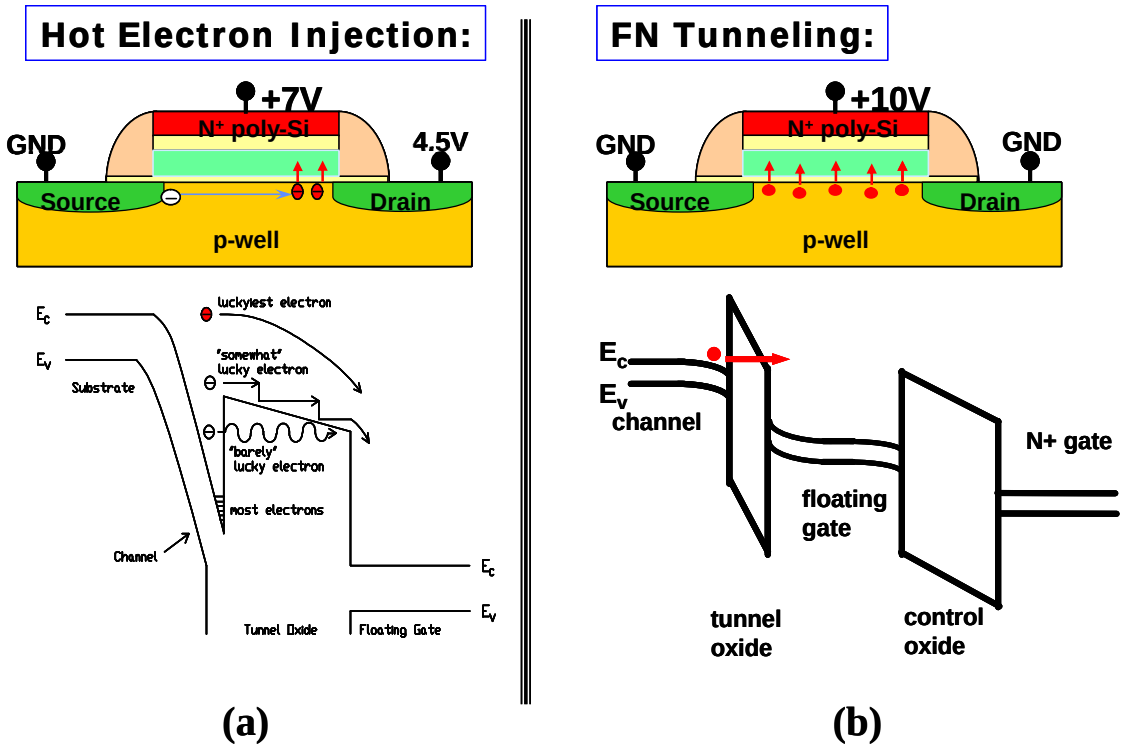


Figure 1.4: Predominant programming mechanisms: (a) Hot Electron Injection (HEI). (b) Fowler-Nordheim (FN) Tunneling.

Erasing is the act of removing electrons from the charge-trapping layer of the flash memory cell. To erase a cell, two methods are predominantly employed: Band-to-Band Tunneling Hot Hole Injection (HHI) [1.6][1.7], which is very similar in principle to the HEI mechanism, and FN tunneling. In the HHI method (**Figure 1.5(a)**), the drain is biased to a high voltage and the source is grounded to generate hot carriers (either via impact ionization or band-to-band tunneling) next to the drain electrode. To direct the generated holes towards the charge-trapping layer, the gate electrode is biased to a large negative voltage. With these settings, some of the generated holes drift towards the gate electrode (in the region near to the drain electrode), whereas the generated electrons drift

towards the drain electrode. A small portion of the generated holes that drift towards the gate electrode (the *hot* holes) have enough energy to overcome the potential barrier (~4.8 eV) imposed by the silicon-oxide interface. Some of these hot holes (the 'lucky' holes) are successfully re-directed towards the charge-trapping site (in the region near to the Drain electrode), where they get trapped. As a result, the holes trapped within a localized portion of the charge-trapping layer annihilate the electrons initially stored there, thereby erasing the cell.

The FN tunneling erase method (**Figure 1.5(b)**) consists of the application of a large & negative transverse field (i.e., a large $|V_{GB}|$) such that either electrons are able to tunnel from the floating gate onto the substrate or holes tunnel from the substrate onto the charge-trapping layer (since the device is placed in accumulation mode, holes accumulate at the surface of the channel and thus are able to tunnel due to the large transverse field). The voltage normally required for this operation is fairly large; additionally, since FN tunneling is exponentially dependent on the number of carriers present in the gate, there is a log-time erase dependence, hence, the erase operation is relatively long (in the order of 10's of milliseconds).

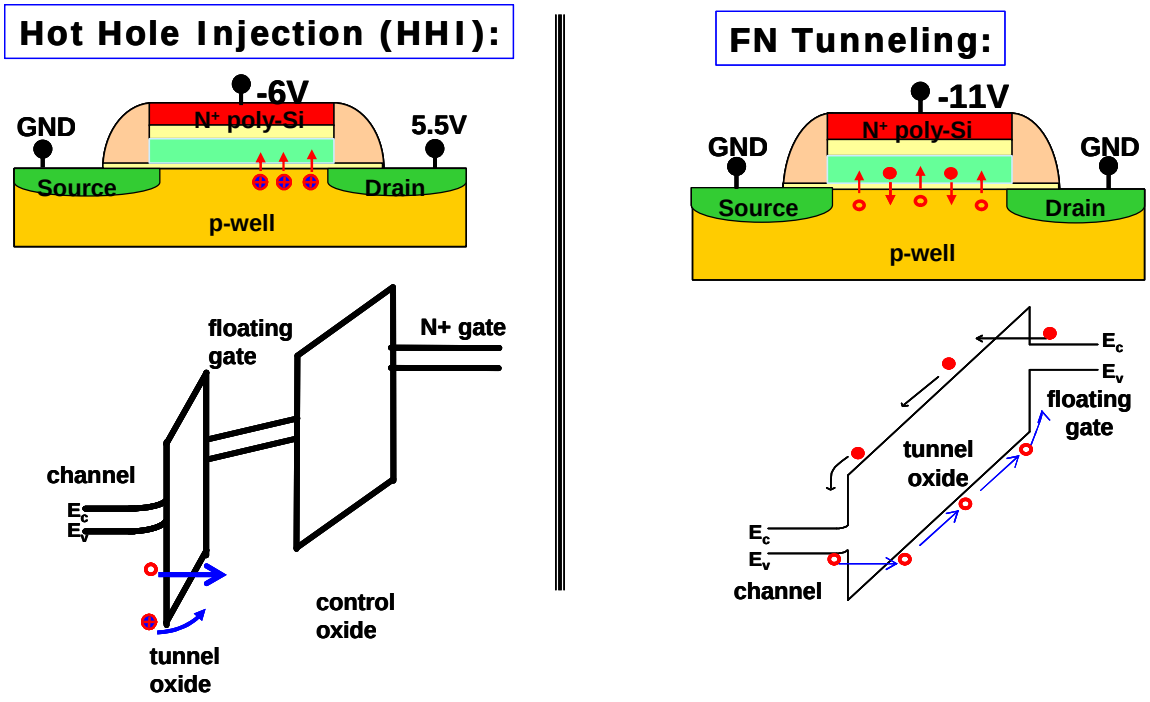


Figure 1.5: Predominant erase mechanisms: (a) Hot Hole Injection (HHI). (b) Fowler-Nordheim (FN) Tunneling.

1.2.2 Reliability Testing of Non-Volatile Memories

Manufacturers of flash memory products normally perform tests that assess the NVM cell's ability to sustain repeated program & erase cycles (*endurance test*) or the ability to retain the data stored within it (*charge retention test*). These tests are done to assure customers that the flash memory products are reliable and that the data stored in them is maintained. Each test is discussed is more detail below.

1.2.2.1 Endurance Test

Endurance is defined as the cell's ability to sustain many program/erase (P/E) cycles. Since fast program & erase operations require high voltages and currents through the gate stack, the cell's tunnel oxide layer is expected to degrade as the number of P/E cycles on the cell increases. However, a minimum level of product performance must be assured; therefore, an accelerated endurance test is normally performed either on single cells or arrays to assess product performance. The endurance test normally consists of a large number of program/read/erase/read operations (each instance referred to as 'P/E cycle') that are performed on a NVM cell. To assess the effect of these cycles on the performance characteristics (*i.e.*, the cell's ability to store electrons within its charge-trapping film) of the structure, the ΔV_T read window (*i.e.*, the difference in V_T between the programmed and erased state of the cell) is monitored (and plotted) as a function of the number of P/E cycles that are performed on the cell. Since the tunnel oxide and the ΔV_T read window are expected to degrade with increasing P/E cycles, a *minimum* ΔV_T read window must be maintained (after a specific number of P/E cycles is performed on the structure) to assure product performance.

1.2.2.2 Charge-Retention Test

By definition, non-volatile memories are designed (and expected) to maintain the stored information for a prolonged time (even with no power applied to the memory array). Charge retention (*i.e.*, the NVM cell's ability to retain the stored electrons for a prolonged time) is therefore a direct measure of *non-volatility* with these types of

memories. **Figure 1.6** shows the energy band diagram of a floating gate flash memory device during retention. The electrons are stored in the conduction band of the poly-silicon floating gate and have a non-zero probability of (quantum-mechanically) tunneling through the tunnel oxide and onto the channel (thereby losing the information stored there). Consequently, a rather thick tunnel oxide (>7nm) is required in floating-gate memories to reduce the tunneling probability of electron leakage and thus achieve the desired (normally, 10 year) retention time. Unfortunately, high operating voltages are required to program or erase cells with a thick tunnel oxide (which tend to degrade their endurance). Charge retention can be quantified by measuring the time it takes for the cell to discharge and thus lose the information stored within it. In practice, this is normally achieved by monitoring the V_T shift of a programmed cell over time when it is exposed to an elevated temperature (i.e., 85 ^oC).

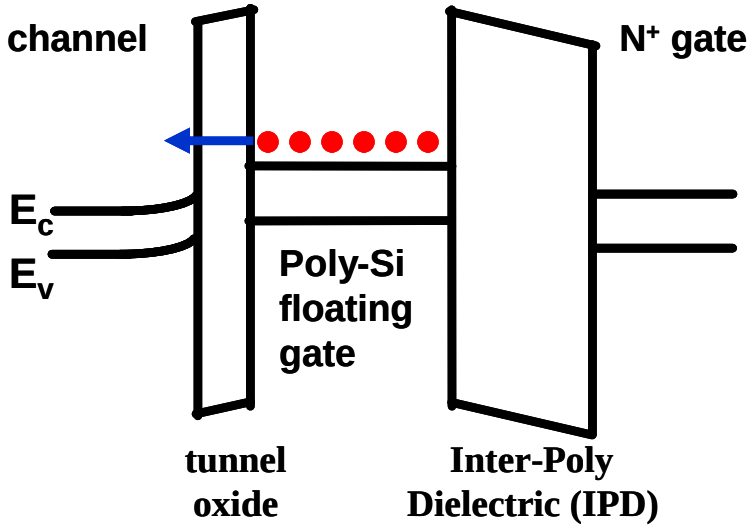


Figure 1.6: Energy band diagram of a floating-gate NVM cell during charge retention (i.e., with no voltages applied to a programmed cell). As shown, the stored electrons can (quantum-mechanically) tunnel back onto the channel (even with no bias applied to the structure).

1.2.3 Predominant NVM Array Architectures

Even though various memory architectures have been proposed, most NVM array architectures can be categorized as either NOR-type or NAND-type architectures [1.8]-[1.10]. **Figure 1.7(a)** illustrates a single-column circuit diagram of a common-source NOR-type array architecture that utilizes the conventional flash NVM FET structure as its unit cell. In this architecture, cells are arranged in a two-dimensional (2D) array, where the gates of all cells in the same row are connected to the same word-line (WL), and all the drain electrodes of all cells in the same column are connected to the same bit-line (BL), with the source electrode of each cell sharing common ground. In these architectures, each cell can be independently accessed or modified; consequently, these architectures have faster (individual cell) read access and thus are mainly used for embedded memory applications.

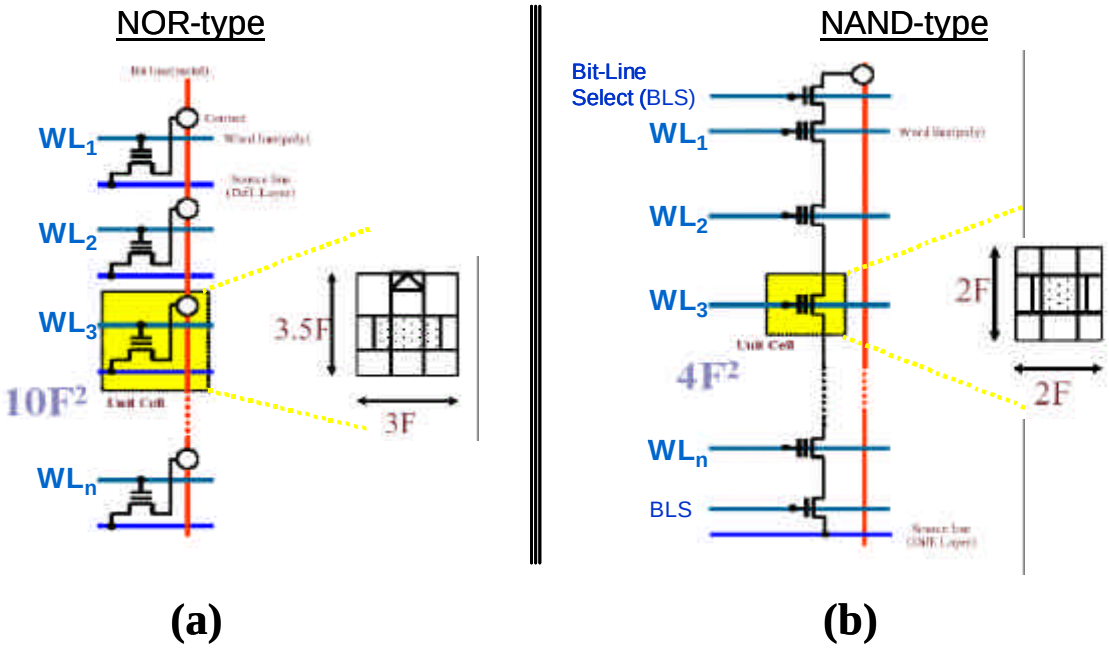


Figure 1.7: Predominant NVM array architectures [1.10]: (a) Common-Source NOR-type array architecture (mainly used for embedded applications). (b) NAND-type array architecture (mainly used for mass storage applications).

Figure 1.7(b) illustrates a single-column circuit diagram of a NAND-type array architecture. In this architecture, cells belonging to the same bit-line are connected in series between two bit-line select (BLS) transistors, and the gates of all cells in the same row share the same word-line. These architectures can utilize a more compact layout since adjacent cells in the same column share the same (source or drain) diffusion layer (which translates into higher memory densities). In addition, cells sharing the same word lines can be programmed (via FN tunneling) or read *simultaneously* (which translates into higher operating throughput). For these reasons, these architectures are mainly used for mass storage applications.

To further enhance NVM density, two different techniques (the Multi-Level Cell (MLC) and the Multi-Bit Cell (MBC) designs) have been widely utilized to store multiple bits of information within a single NVM cell, and thus attain higher NVM densities with the same process technology node [1.7] (**Figure 1.8**). In the MLC design (normally used in floating-gate NVM cells), various levels of charge storage are utilized to attain distinct threshold voltage (V_T) levels and thus distinct binary states within every cell of the array. In this approach, each V_T level corresponds to *one* binary state; consequently, the required number of distinct V_T levels increases substantially as the number of bits stored within every cell increases (for example, 4 distinct V_T levels are required to store 2 bits on every cell). The MBC design (e.g., Saifun's 2-bit NROM™ cell [1.11]) utilizes localized charge-trapping and the cell's symmetry (with respect to the Source and Drain electrodes) to attain two bits of storage in every cell. This symmetry allows treatment of each Bit-Line (BL) and source line within a (NOR-type) memory array layout as either a

Source or Drain electrode to *selectively* read (via the reverse read method [1.11]), program (via the Hot Electron Injection method), or erase (via the Hot Hole Injection method) the data stored at either side of the cell.

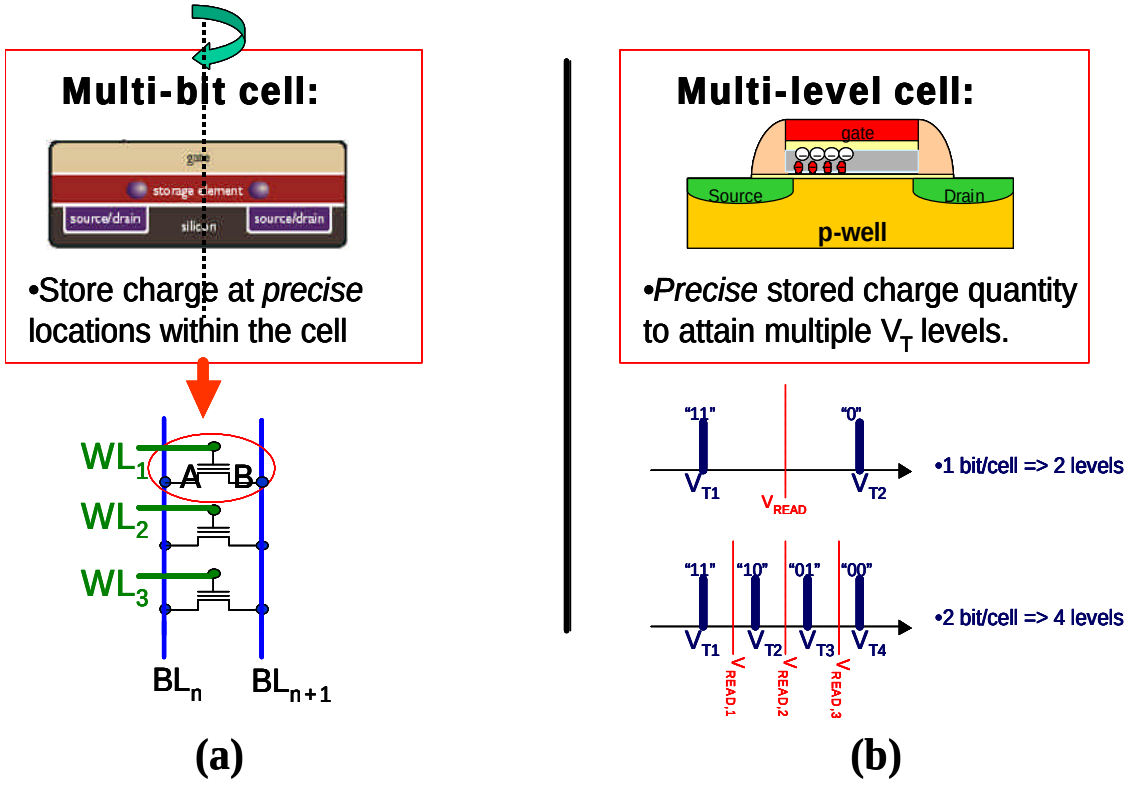


Figure 1.8: Approaches utilized to store multiple bits within every NVM cell [1.7]: (a) the multi-bit cell design. (b) The multi-level cell design.

1.3 Issues with Conventional NVM Technologies

As mentioned before, the proliferation of portable electronic devices has spawned demand for ultra-high-density NVM (for compact and low cost storage). Even though significant improvement in NVM density has been made in the last 2 decades (especially with the introduction of multi-bit architectures), conventional flash NVM technologies have limited scalability due to the inherent structure of the cell. Essentially, conventional floating-gate NVM devices are difficult to scale to gate lengths below 50nm because of their large gate-stack equivalent oxide thickness (EOT) [1.12]. The tunnel oxide thickness in a floating-gate memory device cannot be too thin (below ~8nm); otherwise stress-induced leakage current (due to program/erase operations) can cause charge to leak away from the floating gate, resulting in memory volatility. As an alternative, a charge-trapping NVM cell (such as the silicon-oxide-nitride-oxide-silicon ('SONOS) NVM cell design) has better scalability than a floating-gate device since charge is stored in discrete traps within a non-conductive nitride layer, thereby allowing for more aggressive scaling of its tunnel oxide. Still, a SONOS memory device has a much thicker EOT (~10nm) than a logic device (~2nm) and hence its electrostatic integrity (i.e., scalability) will be worse.

1.4 Research Objectives

In this dissertation, the use of novel material and structures that potentially enhance the scalability of conventional flash NVM devices is evaluated for future silicon-based NVM technologies. This assessment proposes solutions that are compatible with a conventional CMOS process flow. In addition, this assessment proposes the use of a novel charge detection algorithm that (in principle) enhances the scalability of conventional NVM cells and is compatible with the predominant NVM array architectures.

In chapter 2, the use of high-κ dielectrics (such as HfAlO) within the gate-stack of a charge-trapping (SONOS-like) NVM cell is investigated. The main benefit of this approach is the reduction in the EOT (and thus the enhanced scalability) of the proposed structure (as compared to the conventional SONOS NVM cell). Additionally, the use of a high-k dielectric as a new charge-trapping film should provide a significant improvement in the cell's retention time (due to the larger conductivity band offset of this material with respect to the tunnel oxide).

In Chapters 3 through 5, the use of the highly scalable silicon-on-insulator (SOI) double-gate "FinFET" structure [1.13] (modified to be a charge-trapping NVM cell) is evaluated for use as a dual-bit NVM cell. Chapter 3 explores the use of the conventional dual-bit SONOS FinFET NVM cell. The scalability of this structure is explored via 2-dimensional (2D) device simulations. In addition, SONOS FinFET NVM cells are fabricated (using a gate-first process flow similar to that used to fabricate SOI spacer FinFETs [1.14]) and tested to demonstrate the use of a new charge detection method that is proposed with this structure. This new charge detection method utilizes a

change in the cell's gate-induced drain leakage current (I_{GIDL}) [1.15] to detect charge stored on the bit next to the drain electrode. This method enhances the scalability of the SONOS FinFET NVM cell since it is essentially insensitive to charge stored in the complementary bit (next to the Source electrode). The scalability of the SONOS FinFET NVM cell (with its use of the new (ΔI_{GIDL}) read method) is also investigated.

In Chapters 4 and 5, two SOI FinFET-based NVM cell designs with two separate gate-sidewall charge-storage sites are presented for the first time. Chapter 4 (5) explores the use of an n-channel (p-channel) gate-sidewall-storage FinFET structure as a dual-bit NVM cell. The n-channel gate-sidewall-storage (GSS) FinFET NVM cell can utilize the conventional and/or the new (ΔI_{GIDL}) read method to detect the charge-storage state of each bit in the cell. The p-channel GSS FinFET NVM cell can utilize both the conventional read method to detect charge storage, and alternative program and erase methods (that make it compatible with NAND-type array architectures). Both GSS FinFET cell designs can in principle be used to achieve very high NVM storage density because of their high scalability and compatibility with standard CMOS process technology.

Chapter 6 explores the use of two NVM cell designs that utilize a double-gate FET structure (with either 2 or 4 physically separate charge-storage sites) to store 4 bits of information. The symmetry of these structures and the independent operation of each gate are leveraged to selectively access or modify each bit of these structures. The scalability of both cell designs is investigated via 2D device simulations. Read operation of these cells within NOR- and NAND-type array architectures is also discussed.

An overall summary of this dissertation is presented in Chapter 7. Key research findings (obtained in this thesis) and suggestions for future research (in semiconductor-based NVM technologies) are highlighted.

1.5 References

[1.1] Rabaey, J. M., "Digital Integrated Circuits -A Design Perspective", *Prentice Hall series in electronics and VLSI*, First Edition, p. 551-621 (1996).

[1.2] Cappelletti, P., "Memory Devices", *2004 International Electron Device Meeting Short Course* (2004).

[1.3] B. Eitan and D. Froham-Bentchkowsky, "Hot-electron injection onto the oxide in n-channel MOS devices", *IEEE Trans. Electron Devices*, Vol. 28, No. 3, p. 328 (1981).

[1.4] M. Lezlinger and E.H. Show, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Physics*, Vol. 40, No. 1, p. 278 (1969).

[1.5] S. Tam, P.-K. Ko, and C. Hu, "Lucky-electron model for Channel Hot-Electron Injection in MOSFET's", *IEEE Trans. Electron Devices*, Vol. 31, No. 9, p. 1116 (1984).

[1.6] K. Yoshikawa *et al.*, "Lucky-hole Injection Induced by Band-to-band Tunneling Leakage in Stacked Gate Transistors," *IEDM Technical Digest*, p. 580 (1990).

[1.7] R. Liu, *et al.*, "Memory Technologies for 45nm and Beyond," *2006 International Electron Device Meeting Short Course* (2006).

[1.8] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash Memory Cells –An Overview", *Proceedings of the IEEE*, Vol. 85, No. 8, p. 1248 (1997).

[1.9] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, "Flash Memories", *Kluwer Academic Publishers*, First Edition (1999).

[1.10] Rabaey, J. M., "Digital Integrated Circuits Course (Lecture Notes)", *http://inst.eecs.berkeley.edu/~ee141*, Lecture on Semiconductor Memories (2006).

[1.11] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NROM: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Letters*, Vol. 21, No. 11, pp. 543-545 (2000).

[1.12] M. L. French, Chun-Yu Chen, H. Sathianathan, and M. H. White, "Design and scaling of a SONOS multi-dielectric device for nonvolatile memory applications," *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, Part A, Vol. 17, No. 3, pp. 390-397 (1994).

[1.13] N. Lindert, L. Chang, Y.-K. Choi, E.H. Anderson, W.-C. Lee, T.-J. King, J. Bokor, and C. Hu, "Sub-60-nm quasi-planar FinFETs fabricated using a simplified process," *IEEE Electron Device Letters*, Vol. 22, No. 10, pp. 487-489 (2001).

[1.14] Y.-K. Choi, T.-J. King, and C. Hu, "Nanoscale CMOS spacer FinFET for the terabit era,", *IEEE Transactions on Electron Devices*, Vol. 23, No. 1, pp. 25-27 (2002).

[1.15] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters*, Vol. 8, No. 11, pp. 515-517 (1987).

Chapter 2: Use of High-k Dielectrics in Charge-trapping NVM Cells

2.1 Motivations

As discussed in chapter 1, the recent exponential growth of the portable electronics industry has placed a strong demand on the availability of very high-density Non-Volatile Memory (NVM) technologies. The conventional floating-gate flash NVM cell (**figure 2.1a**) faces scaling limitations beyond the 50nm node [2.1] because of its thick gate-stack effective oxide thickness (EOT). The gate-stack of a floating-gate NVM cell consists of a tunnel oxide film ("Tt_{ox}"), a highly doped PolySi film (where charge is stored), and an inter-poly dielectric film ("*IPD*", which normally consists of an oxide or an oxide-nitride-oxide gate-stack). In this structure, the tunnel oxide film cannot be scaled too aggressively; otherwise, stress-induced leakage current (SILC) can cause charge to leak away from the floating gate, resulting in memory volatility. Also, the IPD film, which isolates the floating gate from the control gate, must also be thick enough (and defect-free) to prevent charge leakage from the floating gate to the control gate (and vice versa) while the cell is either in retention, programming or erasing mode. Thus, the EOT of this gate-stack (which is proportional to the sum of the thicknesses of

these films) is by necessity very thick, and this limits the scalability of the cell due to its poor immunity against short-channel effects (SCE).

As an alternative, charge-trapping NVM cells, such as the SONOS (polySi-oxide-nitride-oxide-silicon) NVM device structure (**figure 2.1b**), utilize a thinner gate-stack EOT and consequently are more scalable in principle than the floating-gate NVM cell. The gate-stack of this cell consists of only 3 dielectric layers –a tunnel oxide, a charge-trapping layer (e.g., silicon-rich nitride for a SONOS cell), and a control oxide film ("Ct_{ox}")- stacked underneath the gate electrode. The SONOS structure stores electrons in discrete (localized) traps located below the nitride conduction band edge [2.2], which makes it more immune to defects within the tunnel oxide. As a result, this cell can utilize a thinner tunnel oxide, hence this cell design is more scalable and can utilize lower programming or erasing voltages (than a floating-gate cell). For these reasons, the SONOS NVM device structure has received lots of attention for high-density semiconductor non-volatile memory applications [2.3].

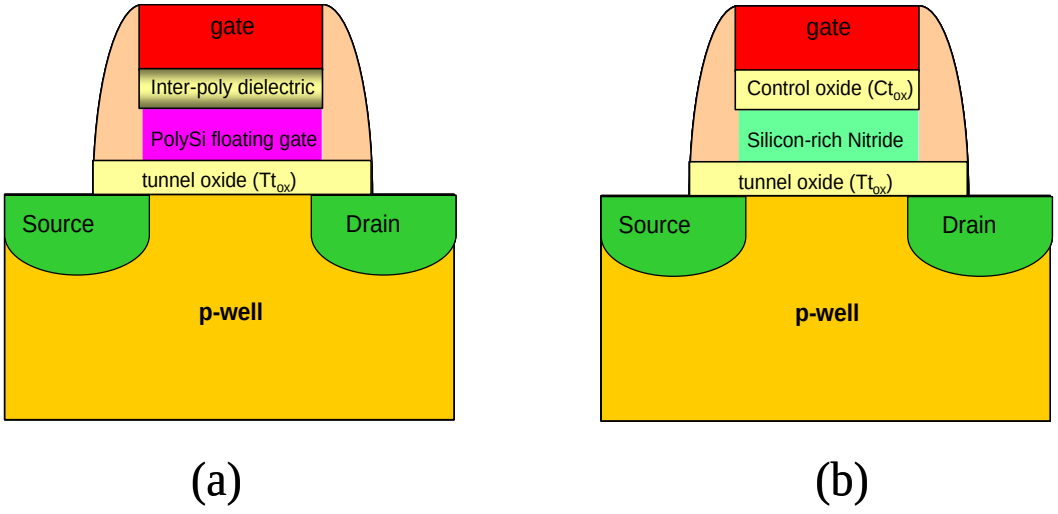


Figure 2.1: (a) Floating-gate versus (b) charge-trap-based (e.g., SONOS) NVM cell structures.

Nonetheless, the SONOS NVM cell still faces challenges for improvement. **Figure 2.2** shows the energy band diagram through the gate-stack of a SONOS cell while in charge-retention mode (i.e., with no voltages applied to a programmed structure). As shown, there are two charge-loss mechanisms: (1) direct tunneling, with an associated barrier height $\phi_o + E_t$; and (2) thermally assisted de-trapping into the nitride conduction band and subsequent tunneling through the tunnel oxide. In both cases, a high conduction band offset (ϕ_o) between the trapping layer and the tunnel oxide is essential for achieving long retention times (ϕ_o is only 1.03 eV for a nitride trap layer). Thus, it is desirable to use a charge-trapping material with a lower conduction band edge (larger electron affinity ?) to achieve a larger conduction band offset ϕ_o, and thus a significant improvement in charge retention times, as shown in **figure 2.3**.

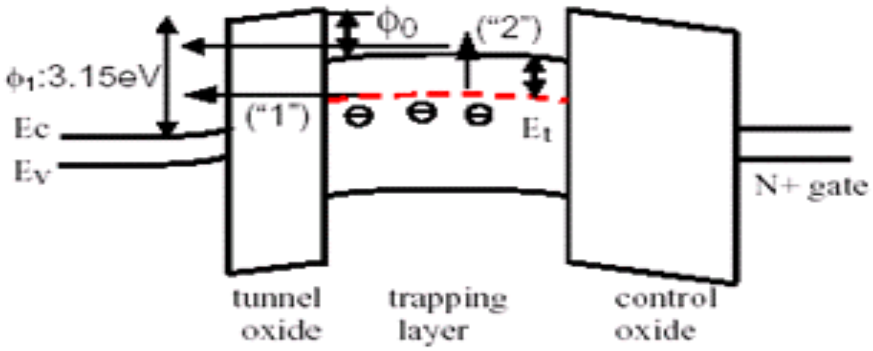


Figure 2.2: Energy band diagram through the gate-stack of a SONOS NVM cell while in retention mode (ϕ_o~1.03eV).

Recently, high-permittivity ("high-κ") dielectric materials such as HfO_2 and ZrO_2 have been investigated to replace thermal oxide as the MOSFET gate dielectric [2.4]. Such materials have a lower conduction band edge (and thus a higher ?) than does Si_3N_4. The conduction band offset ϕ_o for HfO_2 and TiO_2 are 1.65eV and 3.1eV, respectively, which are much larger than the 1.03eV offset associated with a nitride trapping layer. Thus, it should be advantageous to use a high-κ material as the charge-trapping layer in a SONOS-type NVM cell, provided that it contains a sufficient density of deep trap states. In addition, the electron trap energy levels (E_t) must be deep enough to assure good retention. The reported E_t values for ZrO_2 [2.5], Jet Vapor Deposited HfO_2 [2.6]and Si_3N_4 are 1.0eV, 1.5eV and 1.0eV, respectively.

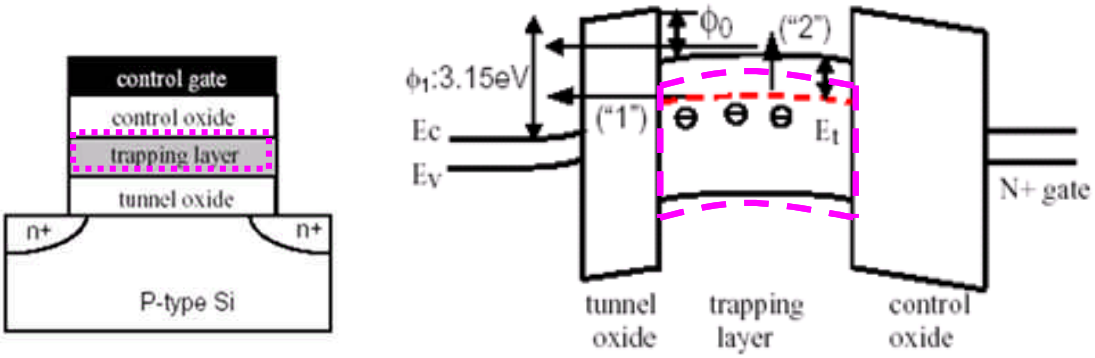
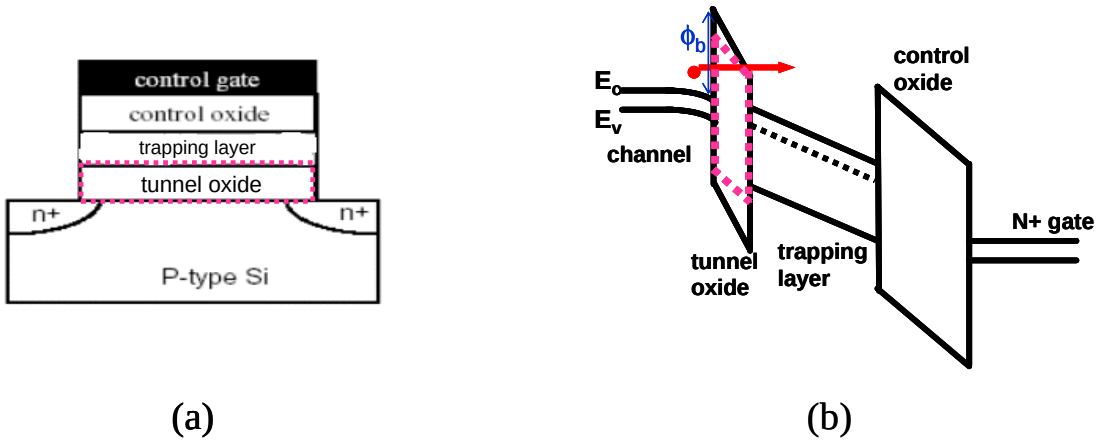


Figure 2.3: (a) Proposed structure, which utilizes a charge-trapping material with a larger electron affinity than Si_3N_4. (b) Energy band diagram, showing the improvement in charge retention with the proposed structure (due to its larger conductivity band offset, ϕ_o).

Additional modifications can be done to the traditional SONOS cell structure to improve its performance. For instance, the use of a high-k dielectric as a novel tunnel oxide film in a SONOS structure can improve both its programming speed and its retention time. The programming speed of most conventional programming methods (e.g., FN tunneling or Hot Electron Injection) depends on the height of the potential barrier (ϕ_b) between the silicon channel and the tunnel oxide, as shown in **figure 2.4**. Essentially, a larger ϕ_b requires larger programming voltage (or longer programming times) to program the cell. Consequently, when the barrier height of the tunnel oxide of a SONOS cell is lowered (through the use of a novel dielectric material, for instance), an improvement in the programming speed of the cell is expected. Furthermore, the use of a high-k dielectric material will improve the retention time of the cell since this film can be made physically thicker than a SiO_2 film (thereby reducing charge loss due to direct tunneling). These facts have been demonstrated before with the use of JVD nitride as a novel tunnel oxide in a charge-trapping NVM cell [2.7].



(a) (b)

Figure 2.4: (a) Proposed charge-trapping NVM structure, which utilizes a high-κ dielectric as the tunnel oxide. (b) Energy band diagram, showing the improvement in programming speed (as a result of the smaller ϕ_b with the use of the new tunnel oxide film).

In this chapter, the use of a high-κ dielectric material (HfO_2) is investigated for use as the new charge-trapping film in a charge-trapping (SONOS-like) NVM cell. The HfO_2 film under investigation is deposited via Rapid Thermal Chemical Vapor Deposition (RT-CVD) [2.8]. The overall performance of this structure is compared against that obtained in the conventional SONOS structure.

2.2 Experimental Approach

2.2.1 Device Fabrication

N^+ poly-si gated capacitors with tunnel-oxide/trapping-layer/control-oxide dielectric stacks were fabricated on p-type Si substrates. The devices with HfO_2 or Si_3N_4 as the charge-trapping layer are designated as "SOHOS" or "SONOS", respectively. In this experiment, the equivalent oxide thickness (EOT) was designed to be roughly the same for both devices. The 3.0 nm-thick tunnel oxide (Tt_{ox}) was grown on both devices at 800 ^oC in dilute O_2 ambient. Afterwards, a ~5 nm-thick charge-trapping Si_3N_4 film was deposited on the 'SONOS' device (via LPCVD @ 750 ^oC), whereas a ~15 nm-thick HfO_2 film was deposited on the 'SOHOS' device @ 450 ^oC (via Rapid-Thermal Chemical Vapor Deposition, "RT-CVD"). Note that the thickness of the HfO_2 film (in the SOHOS structure) is ~3 times thicker than that of the Si_3N_4 film (in the SONOS structure) to ensure that the EOT values of both structures are comparable. Then, the ~7.5 nm-thick high temperature oxide (HTO) film (Ct_{ox}) was then deposited on both devices via CVD @ 800 ^oC. After gate-stack formation, both devices were annealed @ 450 ^oC for 4 hours within an N_2 ambient.

2.2.2 Device Characterization

i) Programming Characteristics

Figure 2.5 shows the capacitance vs. gate-voltage (C-V) characteristics of both SOHOS, SONOS capacitors before and after programming via Fowler-Nordheim (FN) tunneling (through application of 8 volts at the gate with various programming pulses). As shown, there is clearly a shift (to the right) of the C-V curve (in each case) after each programming pulse, which results from electron storage within the charge-trapping film. FN tunneling programming characteristics (defined as the capacitor's flat-band voltage shift, ΔV_{FB}, as a function of programming time) for both devices are shown in **Figure 2.6**. As shown, the SONOS device shows a faster programming speed (i.e., a larger ΔV_{FB} for the same programming pulse) at programming times greater than 1msec. This behavior is most likely due to the fact that the HfO_2 film (in the SOHOS structure) contains a large amount of negative charge initially stored within it (perhaps due to its deposition process) [2.9].

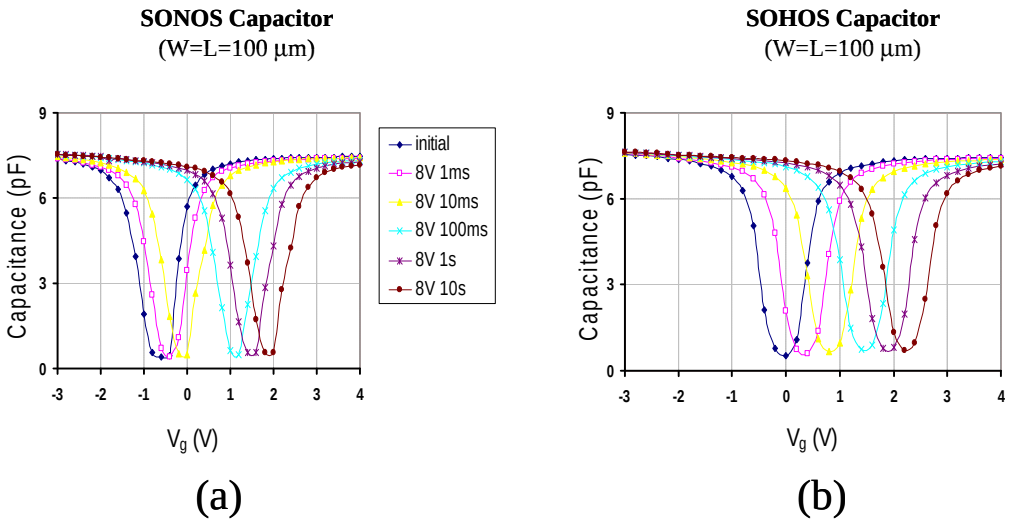


Figure 2.5: Capacitance vs. Voltage (C-V) Measurements (before and after application of various 8-volt programming pulses at the gate) of capacitors containing (a) the SONOS structure, and (b) the SOHOS structure.

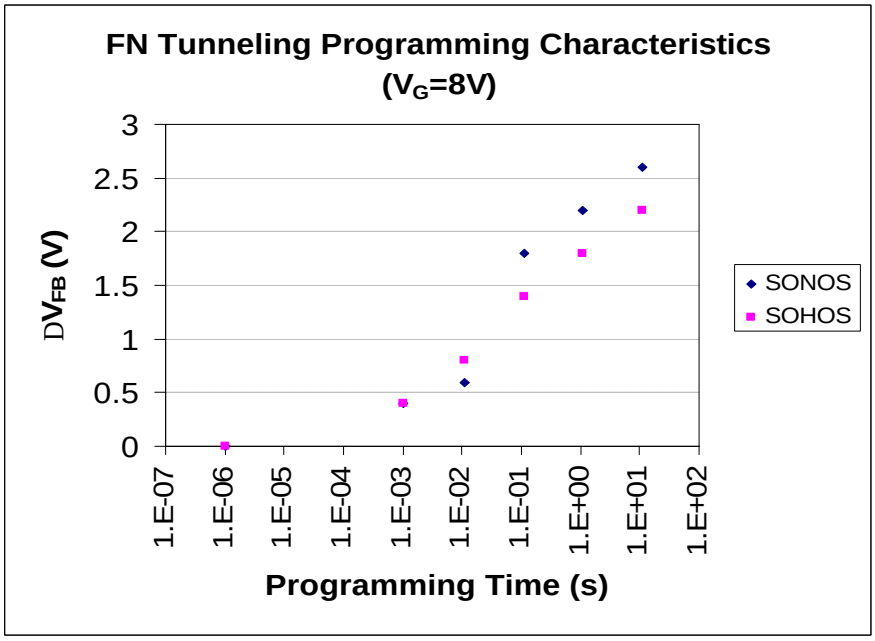


Figure 2.6: Programming Characteristics: Flat-band voltage shift (ΔV_{FB}) vs. programming time (at a gate voltage, V_G, of 8V) for both SONOS, SOHOS devices.

Figure 2.7 shows the C-V characteristics of both SOHOS, SONOS capacitors before and after erasing a programmed cell via FN tunneling (through application of V_G=-8V with various pulses). The resulting FN tunneling erase characteristics (ΔV_{FB} as a function of erase time) for both devices are shown in **Figure 2.8**. As shown in **Figure 2.7**, the C-V curve of the SONOS device shifts to the left (i.e., ΔV_{FB} becomes more negative, see **Figure 2.8**) with each erasing pulse. This is expected, since the holes that get injected (via FN tunneling) from the channel annihilate the negative charge initially stored there. On the other hand, the C-V curve of the SOHOS device shifts to the right (an indication that electrons, instead of holes, are injected and stored in the structure and thus increase ΔV_{FB}, see **Figure 2.8**) with each pulse. This shift results from the injection of electrons from the gate (and through the control oxide film) onto the charge-trapping layer. The observed shift also occurs when larger (in magnitude) gate voltages are utilized to erase the structure. This behavior, known as the 'erase saturation problem' [2.10], is common to all charge-trapping devices and places a lower limit on the thickness of the control oxide film (i.e., this film cannot be too thin; otherwise, electrons will be injected from the gate –and onto the charge trapping film- during the erase operation, thereby programming the structure). In this case, this problem arises only on the SOHOS device, and this is most likely due to the large negative charge that is initially stored within the HfO_2 film (which might effectively enhance in magnitude the transverse electric field through the control oxide film, thereby enhancing electron injection from the gate). Since the SONOS device does not contain negative charge within the nitride film [2.9], this behavior is not shown with this structure.

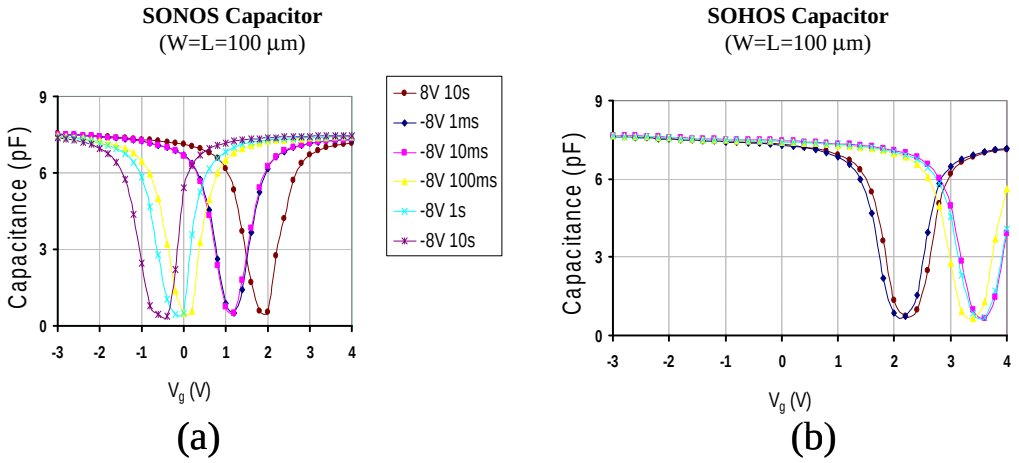


Figure 2.7: C-V Measurements (before and after application of various erasing pulses at the gate of a programmed structure) of capacitors containing (a) the SONOS structure, and (b) the SOHOS structure.

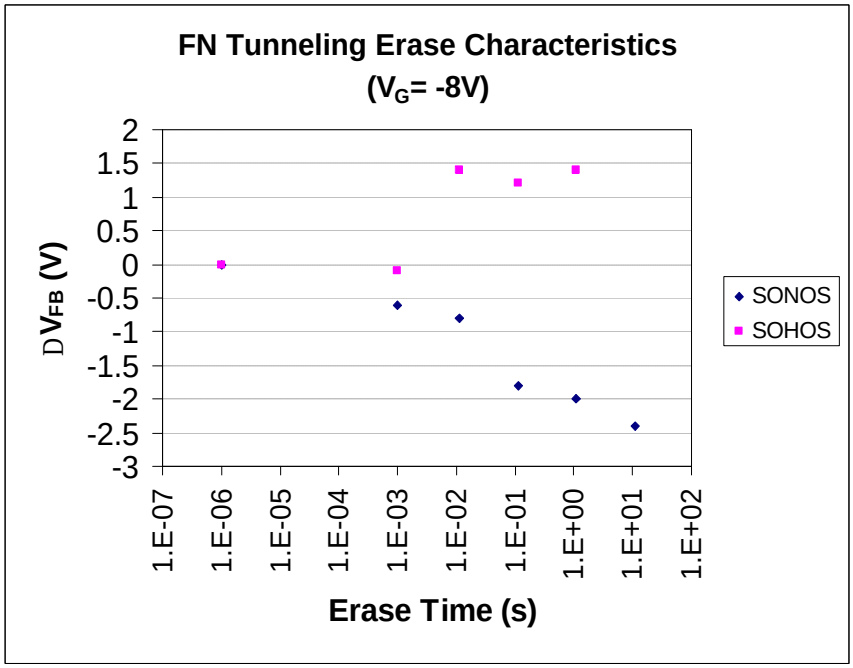


Figure 2.8: Erase Characteristics: Flat-band voltage shift (ΔV_{FB}) vs. erase time (at V_G =-8V) for both SONOS, SOHOS devices.

2.3 Scaling Issues of Charge-trapping NVM Cells

As already mentioned in chapter 1, the use of charge-trapping NVM cells with a thinner gate-dielectric EOT can enhance the scalability of these cells. Nonetheless, marginal benefits can only be obtained with this approach for two reasons: i) short channel effects (SCE), and ii) the enhanced variation in V_T that is attained with devices with smaller dimensions[2.11]. As previously shown in section 2.2, the erase saturation problem will place a lower limit on the thickness of the control oxide film. Also, the tunnel oxide film of these structures cannot be too thin (to mitigate charge loss due to stress-induced leakage current (SILC)). Furthermore, the thickness of the charge-trapping film cannot be too thin; otherwise, it will be difficult to program the structure and/or store enough electrons within it to induce a significant shift in V_T. For these reasons, the gate-stack's EOT of these structures remains larger than that of a logic device (*even* with the use of high-κ dielectrics), and this will ultimately limit the scalability of these structures (especially when multiple bits are stored within each cell). In addition, the enhanced variation in V_T that is attained in structures with smaller dimensions will reduce the V_T separation (ΔV_T) between programmed and erased states (and this will *also* limit the scalability of these structures).

2.4 Other Alternatives for Enhanced Scalability

2.4.1 The Need of a New Charge Detection Method

As already mentioned, the transistor's V_T is the metric normally used to identify the state information stored in a NVM cell. However, V_T (as normally defined) occurs in the linear region of the I_{ds}-V_{gs} curve, before saturation (see **figure 2.9**). The slope of linear region is highly sensitive to charge present *anywhere* within the gate stack (i.e., the charge-trapping layer, an interface, or the gate oxide). As a result, the cell's V_T is defined within a highly non-constant, variable region of the IV curve and therefore its measurement leads to significant cell-to-cell variation (especially in situations where some cells have endured more voltage stress than others). Ideally, a parameter that is more sensitive to charge stored within a specific region of the trapping layer (and less sensitive to charge stored elsewhere within the transistor's gate stack) *and* occurs in a saturated region of the IV curve should be used to identify the state information stored in the cell. In addition, the desired metric should provide a (current or voltage) signal that is large enough for easy detection.

A parameter that meets *most* of these requirements is the cell's Gate-Induced Drain Leakage ("GIDL") Current [2.12]. This parameter is *very* sensitive to charge stored next to the Drain electrode (and essentially insensitive to charge stored away from the Drain electrode) [2.13]. The enhancement in GIDL current in the charged state arises from the increase in the transverse electric field due to the charge stored at that site. Due to the symmetry of the NVM structure (with respect to the source and drain

electrodes), a change in the cell's GIDL current (due to charge stored next to the Drain electrode) could thus be used as a new charge detection method in both single- or multi-bit NVM cells.

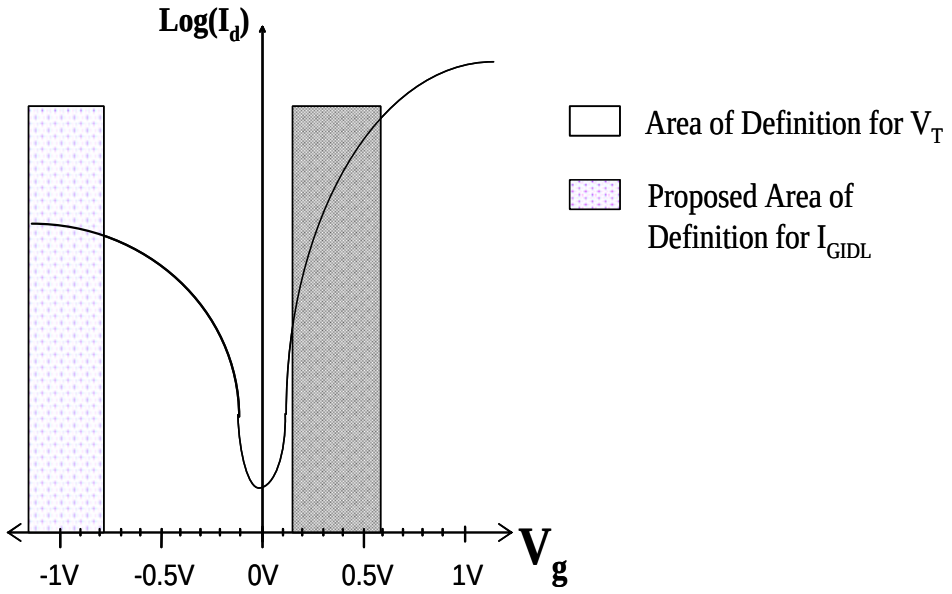


Figure 2.9: Illustrations that show the regions of the I_{DS}-V_{GS} curve where the transistor's threshold voltage (V_T) and 'saturated' GIDL current [2.12] would normally occur.

There is an issue with this approach though: the magnitude in GIDL current that could be obtained on single-gate NVM cells (in the order of a new nano-amperes) might *not* be large enough to allow easy detection. To fully leverage this new metric, the GIDL signal must somehow be amplified. This could be achieved in practice with the use of SOI Double-Gated Field Effect Transistors (DG-FETs) as NVM cells, since these transistors are able to provide larger GIDL currents with lower lateral and transverse fields applied to the structure [2.14]. Thus, the use of this new charge detection method (in SOI DG-FETs) is investigated in the chapters that follow.

2.5 Conclusions

In this work, HfO_2 was investigated as a new charge storage layer to improve SONOS-type flash memory performance. The fabricated ('SOHOS') structures did not show a significant improvement in programming performance (as compared with the conventional SONOS cell). In addition, the SOHOS structure is affected by the 'erase saturation problem' that is common to all charge-trapping NVM cells reported to date. Consequently, a comparison could not be made.

Nonetheless, trap-based NVM cell designs (with multi-bit or multi-level charge storage capability) are *still* the desired choice in the ongoing search for the silicon-based NVM cell structure that is most scalable. The optimum design of this structure might need to utilize a new charge detection algorithm that is less sensitive to charge stored in the complementary bit(s) *and* provides for a signal that is large enough for easy detection. Alternatively, the charge storage sites of the optimum cell structure might also need to be *physically* separated to further mitigate the complementary bit disturb (CBD) issue when the cell operates in multi-bit mode of operation.

2.6 References

[2.1] M. L. French, Chun-Yu Chen, H. Sathianathan, and M. H. White, "Design and scaling of a SONOS multi-dielectric device for nonvolatile memory applications," *IEEE Trans. Components, Packaging, and Manufacturing Technology*, Part A, Vol. 17, No. 3, p. 390 (1994).

[2.2] H. Aozasa, I. Fujiwara, A. Nakamura and Y. Komatsu, "Analysis of Carrier Traps in Si_3N_4 in Oxide/Nitride/Oxide for Metal/Oxide/Nitride/Oxide Silicon Nonvolatile Memory", *Japanese Journal of Applied Physics*, Vol. 38, Part 1, p. 1441 (1999).

[2.3] M.H. White, D.A. Adams, and J. Bu, "On the Go with SONOS", *IEEE Circuits and Devices*, Vol. 16, No. 4, p. 22 (2000).

[2.4] G. Bersuker, *et al.*, "Dielectrics for future transistors", International SEMATECH, January 2004.

[2.5] T. Yamaguchi, H. Satake, N. Fukushima, and A. Toriumi, "Band Diagram and Carrier Conduction Mechanism in ZrO_2/Zr-silicate/Si MIS Structure Fabricated by Pulsed-laser-ablation Deposition", *IEDM Technical Digest*, pp. 19-22 (2000).

[2.6] Zhu, W.; Ma, T.P.; Tamagawa, T.; Di, Y.; Kim, J.; Carruthers, R.; Gibson, M.; Furukawa, T.; "HfO2 and HfAlO for CMOS: thermal stability and current transport", *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 463 (2001).

[2.7] M. She, *et al.*, "JVD Silicon Nitride as Tunnel Dielectric in P-channel Flash Memory", *IEEE Electron Device Letters*, Vol. 23, No. 2, p. 91-93 (2002).

[2.8] S. Sayan, S. Aravamudhan, B.W. Busch, W.H. Schulte, F. Cosandey, G.D. Wilk, T. Gustafsson, and E. Garnfunkel, "Chemical vapor deposition of HfO_2 films on Si(100)", *J. Vac. Sci. Technol. A*, 20(2), pp. 507 (2002).

[2.9] M. She, H. Takeuchi and T.-J. King, "Improved SONOS-type flash memory using HfO_2 as trapping layer," presented at the *19th IEEE Non-Volatile Semiconductor Memory Workshop* (Monterey, California, USA), pp. 53-55, 2003.

[2.10] T. Mikolajic, M. Specht, N. Nagel, T. Mueller, S. Riedel, F. Beug, T. Melde, and K.-H. Kusters, "The Future of Charge Trapping Memories," *International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pp. 112-115 (2007).

[2.11] M. Pelgrom, *et al.*, *IEDM Technical Diges*t (1998).

[2.12] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters*, Vol. 8, No. 11, pp. 515-517, 1987.

[2.13] E. Lusky, *et al.*, "Investigation of channel hot electron injection by localized charge-trapping nonvolatile memory devices," *IEEE Electron Device Letters*, Vol. 51, No. 3, pp. 444-451, 2004.

[2.14] J. Chen, , *et al.*, "The enhancement of GIDL current in SOI MOSFET and its impact on SOI device scaling," *IEDM Technical Diges*t (1992).

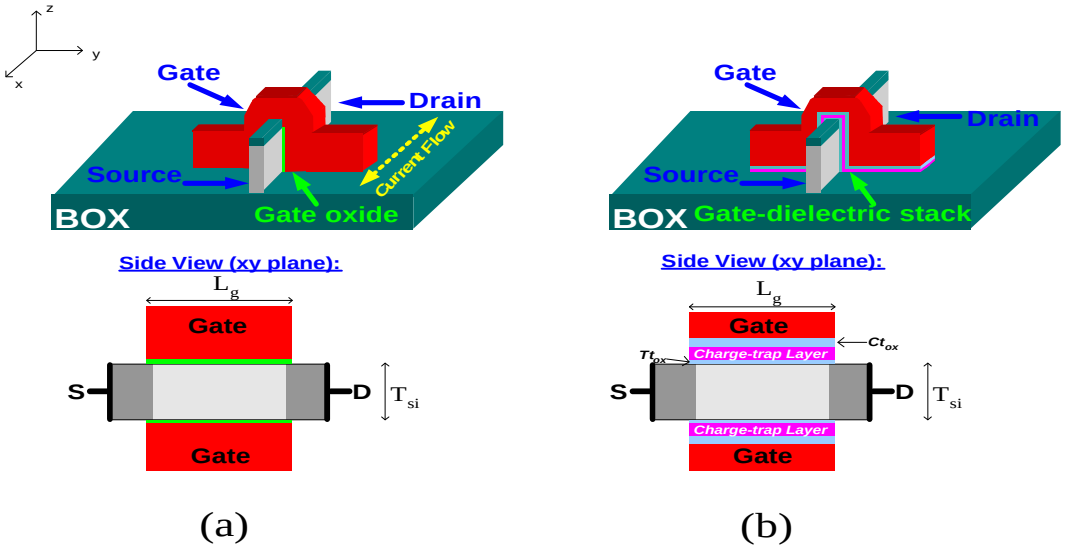
Chapter 3: Design of Dual-bit SONOS FinFET NVM Cells

3.1 Motivations

In this chapter (and the chapters that follow), the design of multiple-gate NVM cells is discussed. As already mentioned in previous chapters, conventional (floating-gate, SONOS) NVM cells are difficult to scale to gate lengths below 40nm because of their thick gate-stack effective oxide thickness (EOT), and this is partly because the metric used (the cell's threshold voltage, V_T) to detect charge storage is highly sensitive to short channel effects (SCE). One solution to this dilemma is to employ a novel charge detection method that is less sensitive to SCE, as discussed in Chapter 2. Another solution to this dilemma is to employ a transistor structure that *remains* scalable despite its thick gate-stack EOT. A structure with this quality is the double-gate thin-body FET structure [3.1], since it can achieve good electrostatic integrity without the need for a thin gate-dielectric stack [3.2] and attain the ideal subthreshold swing (60mV/decade at room temperature); as a result, this structure is more scalable than a conventional MOSFET [3.3]. The double-gated thin-body FET (DG-FET) structure can be implemented in practice with the FinFET structure [3.4]; hence, the FinFET structure

is a promising candidate for scaling conventional flash memory devices down to sub-40nm gate lengths (Lg) [3.5]- [3.7].

The FinFET device is a three-dimensional (3-D) transistor structure in which the gate electrode runs over a thin silicon-on-insulator (SOI) pillar (a "fin") structure (**Figure 3.1(a)**). In the dual-gate FinFET, current flows through two conducting channels that form along the fin's sidewalls –note that the top of the fin could also serve as a conductive channel if the oxide on top of it is thin enough; however, this particular tri-gate FinFET design is not considered here. Implementation of the conventional (floating-gate or SONOS) NVM cells with a FinFET structure is achieved in a straightforward manner by replacing its gate dielectric by a multi-layer stack, which essentially consists of a tunnel oxide film (Tt_{ox}, normally SiO_2), a charge-trapping film (doped PolySi for a floating-gate cell, or silicon-rich nitride for a SONOS cell), and a control oxide film (Ct_{ox}), see **Figure 3.1(b)**.



(a) (b)

Figure 3.1: Isometric view (and 2D planar cross-section) of (a) the dual-gate FinFET transistor structure [3.4], and (b) the modified charge-trapping NVM FinFET structure, which essentially replaces its gate dielectric by a multi-layer stack.

Even though both floating-gate and SONOS memory cells can be readily obtained with the FinFET structure (and their prototypes have already been demonstrated in [3.8], [3.9] respectively), the SONOS FinFET NVM cell is in principle more scalable than the floating-gate FinFET NVM cell. A SONOS cell has better immunity against SCE due to its thinner gate-stack EOT. In a floating-gate cell, the gate-stack EOT consists of the sum of the tunnel oxide thickness and the control oxide (or oxide-nitride-oxide) film thickness, whereas the gate-stack EOT of a SONOS cell consists only of the sum of its oxide-nitride-oxide (ONO) dielectric film thicknesses. Consequently, the SONOS cell is more scalable due to its thinner gate-stack EOT. In addition, the floating-gate NVM cell design suffers from significant coupling interference between adjacent cells due to its high coupling capacitance with neighboring cells [3.10], and this is considered a limiting factor for the realization of high-density NVM floating-gate cell arrays. In contrast, the SONOS cell design avoids floating-gate coupling interference (since the coupling capacitance between neighboring SONOS cells is very small). For these reasons, the FinFET SONOS cell design is a more appealing option for future high-density NVM technologies [3.7].

The dual-bit SONOS FinFET NVM cell design under assessment in this chapter consists of a double-gate transistor structure that contains the charge-trapping layer (silicon-rich nitride, “SirRN”) embedded within the gate dielectric stack underneath each gate electrode, as shown in **Figure 3.1(b)**. This is the conventional multi-gate SONOS NVM cell design, for which single-bit storage [3.5], multi-level charge storage [3.11], and dual-bit storage [3.6] have been demonstrated previously with the conventional charge detection method. The scalability of this cell is investigated in detail (first via

2D device simulations and subsequently through measurements on fabricated SONOS FinFET cells), for both the conventional read method and the new charge detection method (which utilizes a change in the cell's OFF-state current to determine whether charge is stored near to the drain electrode). Design tradeoffs are evaluated in terms of overall NVM cell performance.

3.2 SONOS FinFET NVM Cell Design and Operation

3.2.1 Operating Principles

The SONOS FinFET NVM cell is symmetric with respect to the Source (S) and Drain (D) electrodes (**figure 3.2(a)**). This symmetry can thus be utilized to store 2 bits within the structure, as already demonstrated in [3.6]. To perform dual-bit operation, each bit must be *independently* programmed and read *regardless* of the state of the complementary bit. In conventional dual-bit SONOS NVM cell structures, selective programming is achieved via the hot electron injection (HEI) mechanism, and the reverse-read method (which utilizes a change in the cell's threshold voltage, V_T, to detect charge storage) is used to selectively read the information stored in the bit next to the Source electrode (i.e., Bit 1) [3.17]. Needless to say, these methods can also be applied to the SONOS NVM cell to store and read 2 bits within it. These conventional methods (as well as other methods that may also be utilized with this dual-bit SONOS FinFET structure) are investigated in this chapter. The details of all these (read, programming) methods are further discussed in the sections that follow.

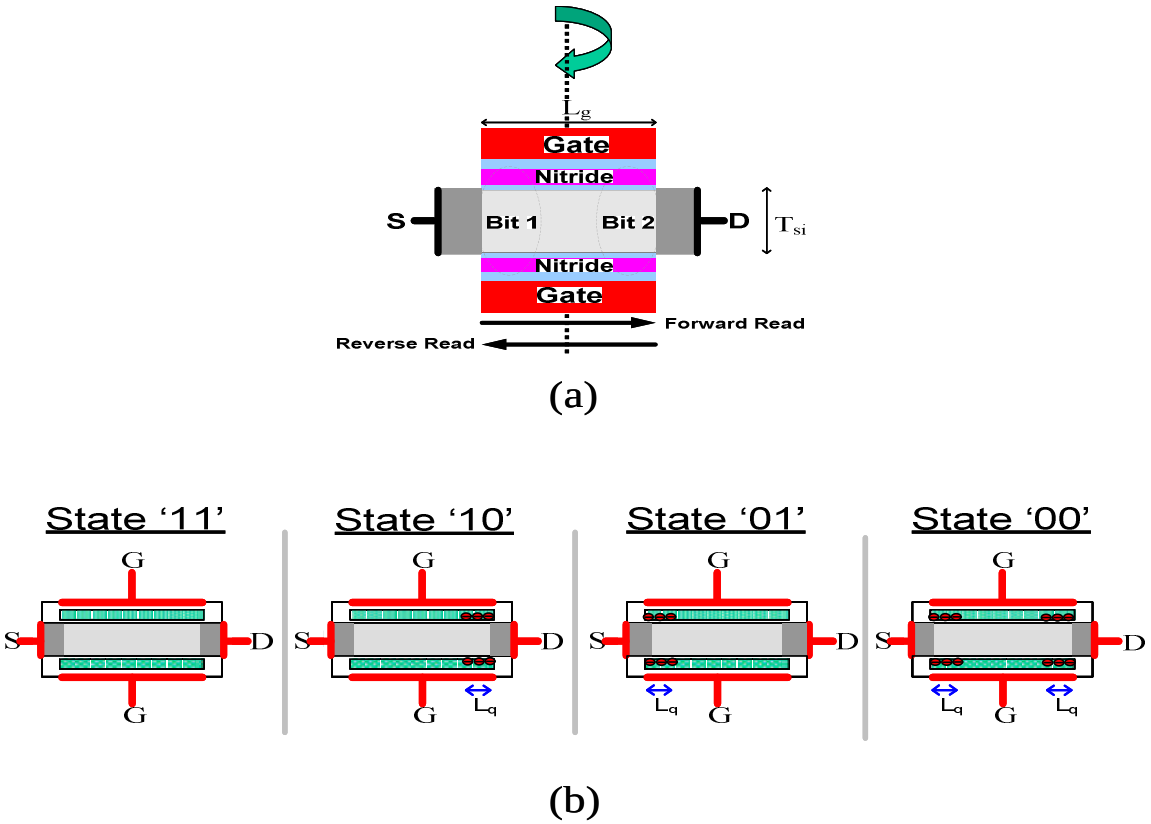


Figure 3.2: (a) the symmetry of the SONOS FinFET NVM cell (with respect to the S/D electrodes) allows for dual-bit operation. (b) Definition of the 4 binary states of this dual-bit NVM SONOS cell.

3.2.2 Programming Method

The conventional (HEI) method used to selectively program each bit of single-gate SONOS NVM cells [3.17] can also be used to program each bit of the SONOS FinFET NVM cell, as already demonstrated for the multi-gate SONOS FinFET NVM cell reported in [3.6]. The HEI method requires application of large lateral and transverse electric fields to generate hot electrons and then re-direct these onto the desired charge-trapping site, located next to the drain electrode (**figure 3.3**). In the SONOS FinFET

NVM cell, the bit next to the Drain electrode (i.e., Bit 2) may be selectively programmed via HEI by biasing the gate (V_G) and Drain (V_D) electrodes to a high voltage, while the source electrode (V_S) is kept grounded. To ensure that enough hot electrons are generated within the silicon body of the cell, the applied drain-to-source voltage (V_{DS}) should be larger than ~$V_{GS} - V_T$, so that the cell's channel is in the pinch-off condition.

With these settings, the following steps occur sequentially:

- Hot carriers are generated within both the front and back channels (next to the drain electrode) once a large lateral electric field, $E_{lateral}$ ($\geq 10^6$ V/cm), is applied to the cell. This is achieved through application of a large V_{DS}.
- The generated holes (electrons) drift towards the Source (Drain) electrode since the silicon fin is electrically floating and a large V_{DS} potential is applied to the cell. Some of the generated electrons gain enough kinetic energy by drift and thus become "hot".
- *Some* of the generated electrons are re-directed towards the charge-trapping film next to the drain electrode due to the large transverse electric field ($E_{transverse}$) that is applied to the structure. A small portion of the hot electrons (the 'luckiest' electrons) has enough energy to overcome the potential barrier of the tunnel oxide and thus *drift* (instead of tunnel) towards the SiRN charge-trapping site next to the Drain electrode. Eventually, these hot electrons get trapped within the SiRN charge-trapping film, thereby programming the bit.

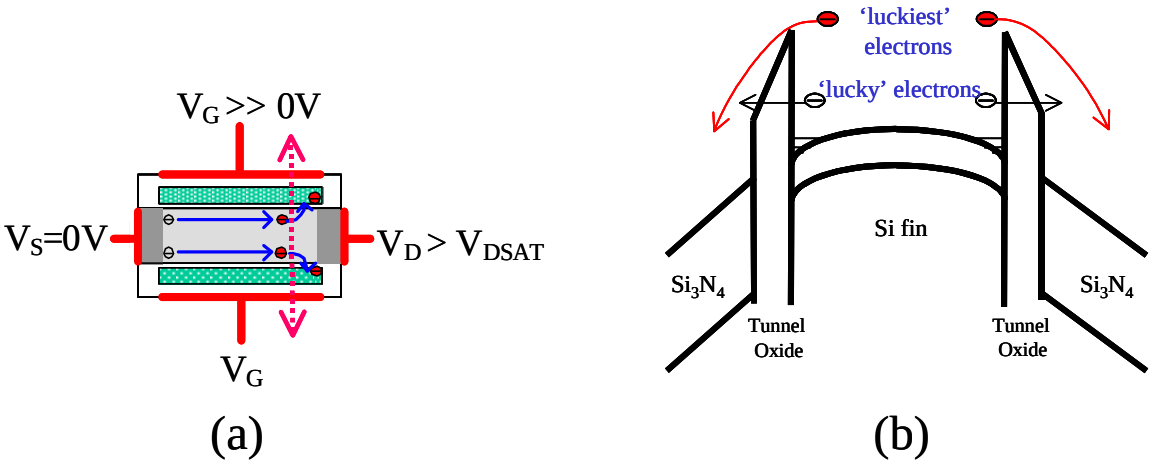


Figure 3.3: The HEI mechanism can be used to selectively program the bit next to the Drain electrode (Bit 2). (a) Bias conditions (b) Energy band diagram (through the dashed line in (a)), illustrating the HEI mechanism.

3.2.3 Charge Detection Methods

3.2.3.1 Read Simulation Setup

A schematic 2-Dimensional (2D) cross-section of the NVM cell design involved in this study is shown in **figure 3.4**, along with some of the parameter values used (in the Taurus [3.14] device simulations) on an optimized structure with L_g=100nm. This cell is comprised of an n-channel SONOS double-gate field-effect transistor (DG-FET) with the charge-trapping layers (made of any other material that is able to store charge, such as silicon-rich nitride, SiRN) embedded within the gate-dielectric stacks underneath the (N^+ PolySi) gate electrodes. In these simulations (and for optimum performance), the silicon fin is essentially undoped to both provide high carrier mobility and to minimize statistical doping fluctuation effects. Also, the silicon body thickness (T_{si}) is chosen to be ~0.45*L_{eff} or smaller (for instance, T_{si}=40nm for $L_{eff} \sim L_g \geq$80nm) to suppress SCE [3.15].

Furthermore, the thickness of the gate-stack underneath the gate electrodes includes a SiO_2 gate oxide thickness (T_{ox}) of 3nm, a (SiRN or doped PolySi) charge-trapping film with thickness (T_{trap}) of 6nm, and a SiO_2 control oxide film with thickness (Ct_{ox}) of 5nm. This setup allows for efficient tunneling or drift of electrons (holes) from the channel into the charge-trapping regions (and vice versa) during programming (erasing) of the structure at low programming (erasing) voltages since T_{ox} is thin enough. Also, a thick Ct_{ox} mitigates erase disturbance issues (due to tunneling of electrons from the gate electrodes onto the charge-trapping sites) when using large (negative) erasing voltages at both gates. In these read simulations, the areal density of charge at the bottom charge-trapping film interface (nearest to the channel) is set to a value of $5x10^{12}$ q/cm^2, (a value comparable to that used by other investigators [3.16]) and the length of the charge-trapping region ("L_q") is set to 25nm to mitigate the complementary bit disturb effect [3.17].

Parameter	Value
Gate length, L_g	100 nm
Fin thickness, T_{si}	40 nm
Fin doping (p-type)	1e13 /cm^3
Gate oxide thickness (T_{ox})	3 nm
Control oxide thickness (Ct_{ox})	5 nm
Trap-region thickness (T_{trap})	6 nm
Trap-region length (L_q)	25 nm
$Q_{ox,max}$ stored	-5e12 q/cm^2

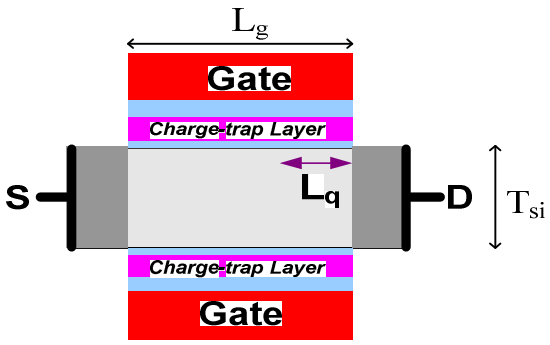


Figure 3.4: Parameter settings used in the 2-Dimentional (2D) Taurus device simulations on an optimized SONOS DG-FET structure with L_g=100nm.

3.2.3.2 The *Reverse Read* Method

The conventional "reverse read" charge detection method, normally used in single-gate SONOS NVM cells (as discussed in chapter 2), may be utilized to selectively distinguish the state of each bit of a SONOS FinFET NVM cell. In this method, charge storage is detected through a measure of a change in the cell's threshold voltage (V_T) when charge is stored on the bit next to the Source electrode (**figure 3.5(a)**). For example, to read the bit next to the Source electrode (*i.e.* Bit 1), the gates are biased to a positive voltage ("V_{GR}") while a moderate drain-to-source bias is applied (*e.g.* V_{DS} ~ 1.5V). The chosen value of V_{GR} lies roughly halfway between the V_T values of the programmed and erased states in order to maximize their separation and thus properly distinguish each state. The state of Bit 1 is determined from a measure of the cell's on-state current at the specified bias conditions: if electrons are stored in Bit 1 (binary states '01', '00'), the threshold voltage will be high so that the read current will be low; otherwise, if no electrons are stored in Bit 1 (binary states '10', '11'), the read current will be high (**figure 3.6(a)**). Due to the symmetry of the structure (with respect to the Source and Drain electrodes), charge storage in the complementary bit (i.e., Bit 2, next to the Drain electrode) may be detected in a similar manner upon interchange of the Source and Drain electrodes during the reverse read operation (to measure and thus identify the change in V_T due to charge storage on Bit 2).

The dual-bit charge storage scheme discussed above, based on charge storage within different regions of a *single* SiRN charge-trapping layer, has been implemented successfully in conventional single-gate (e.g. NROM™ and MirrorBit™) SONOS NVM cells (as discussed in Chapter 1), and has been demonstrated in multi-gated SOI

conventional SONOS FinFET cells [3.5]-[3.7]. These technologies are difficult to scale to sub-50nm L_g since V_T can be affected by charge stored next to the drain electrode, even for the FinFET SONOS NVM cell (which is in principle less susceptible to SCE). As shown in **figure 3.6(a)**, charge storage on Bit 2 increases the V_T of both Bit1 states (i.e. states '10' and '00'), and this decreases the *smallest* V_T separation (ΔV_T) between the programmed and erased Bit1 states (i.e. states '10' and '01'), which in turn limits the potential scalability of the structure. As an alternative, the novel DI_{GIDL} read method could also be utilized with this structure. This read method is less susceptible to SCE (since it is less sensitive to charge stored in the complementary bit) and thus in principle enhances the scalability of conventional NVM cells. The use of this novel read method in double-gated NVM cells is further discussed in the next section.

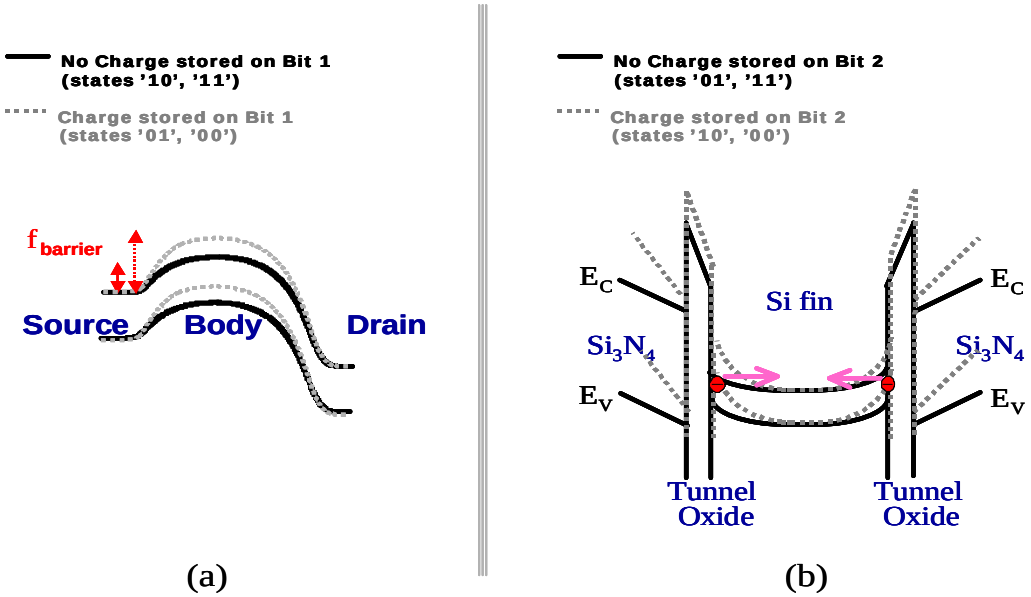


Figure 3.5: Energy band diagrams illustrating the charge detection methods that can be used with the FinFET SONOS NVM Cell: (a) charge stored on the bit next to the source electrode (i.e., Bit 1) affects the source-to-drain potential profile and thus the cell's V_T. (b) Charge stored on the bit next to the drain electrode (i.e., bit 2) affects the cell's transverse field (next to the Drain) and thus the cell's GIDL current.

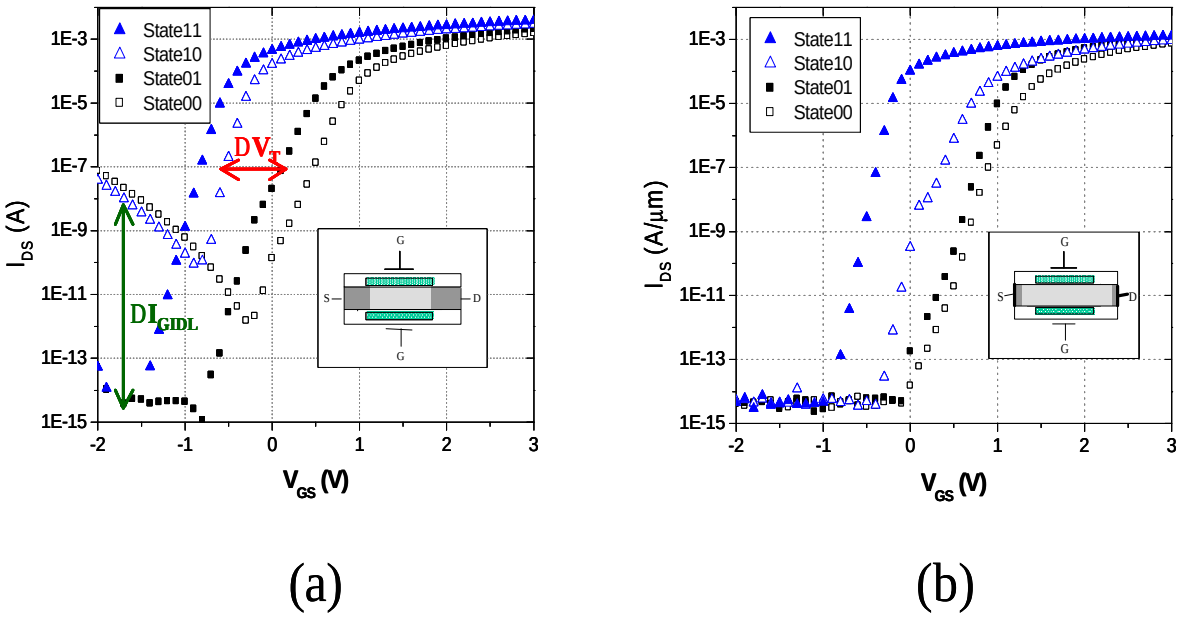


Figure 3.6: Simulated I_{DS}-V_{GS} curves for (a) a gate-aligned (with $L_g \sim L_{eff}$), and (b) a gate-underlapped (with $L_g < L_{eff}$) SONOS FinFET NVM cell, showing that GIDL can be only be used in the former case to distinguish the state of Bit 2 (L_g=100nm, T_{si}=40nm, L_q=25nm, V_{DS}=1.5V).

3.2.3.3 The Delta-GIDL (DI_{GIDL}) Read Method

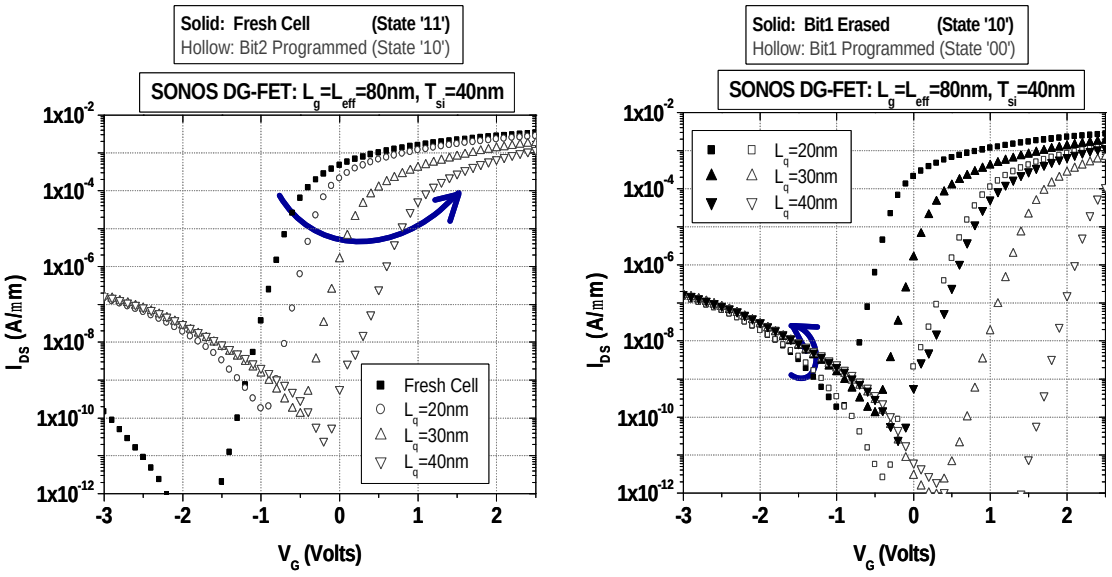
The dual-bit SONOS FinFET NVM cell structure may also utilize a change in the cell's Gate Induced Drain Leakage ("GIDL") current [3.18] to detect charge storage on the bit next to the Drain electrode (Bit 2) [3.19]. This change in off-state current arises from the change in the transverse electric field when charge is stored near the drain electrode, which significantly increases GIDL current due to band-to-band tunneling (**figure 3.5(b)**). Due to the symmetry of this structure, this novel method of charge detection may also be used to identify charge storage on the complementary bit (in this case, Bit 1) upon interchange of the Source and Drain electrodes, as normally done in the 'reverse-read' scheme discussed above. Alternatively, this novel charge detection

method may be used in conjunction with the conventional method –the former method is used to detect charge storage on Bit 2, while the latter is used to detect charge storage on Bit 1. This is confirmed by 2-D device simulations of an optimized SONOS FinFET device with gate-length L_g=100nm and silicon fin thickness T_{si}=40nm, shown in **Figure 3.6(a).** As shown, the shift in V_T with charge stored at Bit 1 (near the source) allows the state of this bit to be determined by applying small positive gate voltage (~0.1V) and moderate drain-to-source voltage (V_{DS} = 1.5V): low on-state current → electrons stored. Bit 2 (near the drain) can be read by applying a negative gate voltage ($V_{GS}\cong$-1.75V): high off-state current → electrons stored. With this approach, a reverse read operation is not required. Note that the change in GIDL current (ΔI_{GIDL}) with charge stored near the drain will not be significant if the cell has a gate-underlapped S/D structure, as shown in **Figure 3.6(b)**. Consequently, a gate-aligned structure design, where the source/drain (S/D) junction edges are perfectly aligned to the edges of the gate electrodes, is required for use (for optimum performance) with this charge detection method. This design minimizes the GIDL current of the erased Bit 2 states (since the gates do *not* overlap the S/D junction edges) and thus maximizes the GIDL current separation between the programmed and erased Bit 2 states.

3.3 Assessment of Short Channel Effects

3.3.1 The Complementary Bit Disturb Effect

To assess the complementary bit disturb (CBD) effect on both read methods, additional simulations were performed on a SONOS DG-FET structure that was properly scaled to suppress SCE, but with different amount of charge stored in the complementary bit. **Figure 3.7(a)** shows the simulated (I_{DS}-V_G) curves of a SONOS DG-FET structure with $L_g = L_{eff}$ =80nm, T_{si}=40nm (V_{DS}=1.5V) for the 2 erased Bit 1 states (*i.e.* binary states '*11*' and '*10*') for various trap-region lengths (L_q). As shown, the cell's V_T increases with the lateral extent of charge stored in the complementary bit ("L_q"). This effect enhances the variability of this metric and therefore limits the scalability of the cell, which is *not* desired. **Figure 3.7(b)** shows the simulated (I_{DS}-V_G) curves of a SONOS DG-FET structure with $L_g = L_{eff}$ =80nm, T_{si}=40nm (V_{DS}=1.5V) for the 2 programmed Bit 2 states (*i.e.* binary states '*10*' and '*00*') for various L_q. As shown, charge stored in the complementary bit (next to the Source electrode in this case) does *not* significantly affect the GIDL current of the cell, which indicates that the novel (off-state current sensing) method of charge detection is less sensitive to the CBD effect. Consequently, the use of the novel (ΔI_{GIDL}) charge detection method should (in principle) enhance the scalability of the SONOS DG-FET cells due to its improved immunity against the CBD effect.



(a) (b)

Figure 3.7: The CBD effect on SONOS DG-FETs: (a) Simulated I_{DS} vs. V_G characteristics for the 2 erased Bit1 states (states '11' and '10'), illustrating the effect of stored charge in the complementary bit on the cell's V_T. As shown, the cell's V_T is highly sensitive to charge stored in the complementary bit (next to the drain electrode), which is not desired. (b) Simulated I_{DS} vs. V_G characteristics for the 2 programmed Bit2 states (states '10' and '00'), illustrating the effect of stored charge in the complementary bit on the cell's GIDL current. As shown, the cell's GIDL current is essentially insensitive to charge stored in the complementary bit (especially for V_G<-2V). These results illustrate the improved immunity against the CBD effect of the novel ΔI_{GIDL} read method.

3.3.2 Assessment of Scalability

To assess the scalability of the dual-bit SONOS DG-FET structure, additional simulations were performed on a structure that had a specific silicon body thickness (T_{si}) but different gate lengths (L_g). **Figure 3.8** shows the simulated (I_{DS}-V_G) curves of a SONOS DG-FET structure with (a) T_{si}=40nm, and (b) T_{si}=20nm (V_{DS}=1.5V) for the 2 binary states with *smallest* V_T separation ("DV_T") due to charge stored at Bit 1 (*i.e.* binary states '*10*' and '*01*'), for various L_g and a fixed charge-trapping length (L_q=T_{si}/2). As shown in **Figure 3.8(a)**, scaling of the cell's L_g degrades both its sub-threshold swing *and* DV_T (**Figure 3.8(c)**), which is not desired. Additionally (see **Figure 3.8(d)**), although scaling of the cell's L_g does *not* affect the separation in GIDL current due to charge storage on Bit 2, the severe swing degradation of the completely erased state of the cell (*i.e.* binary state '*11*', not shown in these graphs) detrimentally affects the separation in GIDL voltage "DV_{GIDL}" (where the GIDL voltage of each state is measured @ I_{DS}=10nA), which is also not desired. The use of a thinner silicon body thickness (T_{si}) exacerbates this behavior: As shown in **Figure 3.8(b)**, the observed swing degradation gets worse with a thinner T_{si} due to the enhanced gate-to-body coupling (and thus the enhanced sensitivity to charge stored underneath the gates) that exists with these structures [3.20]. Furthermore, the use of a thinner T_{si} enhances the GIDL current of the erased Bit2 state (*i.e.* binary state '10') and thus further reduces DV_{GIDL}, which is also not desired. These results highlight the fact that the SONOS DG-FET structure is quite susceptible to SCE (due to the placement of the charge-trapping sites *underneath* the gate electrodes), and this severely limits the scalability of this structure (when using either the conventional or the novel read method).

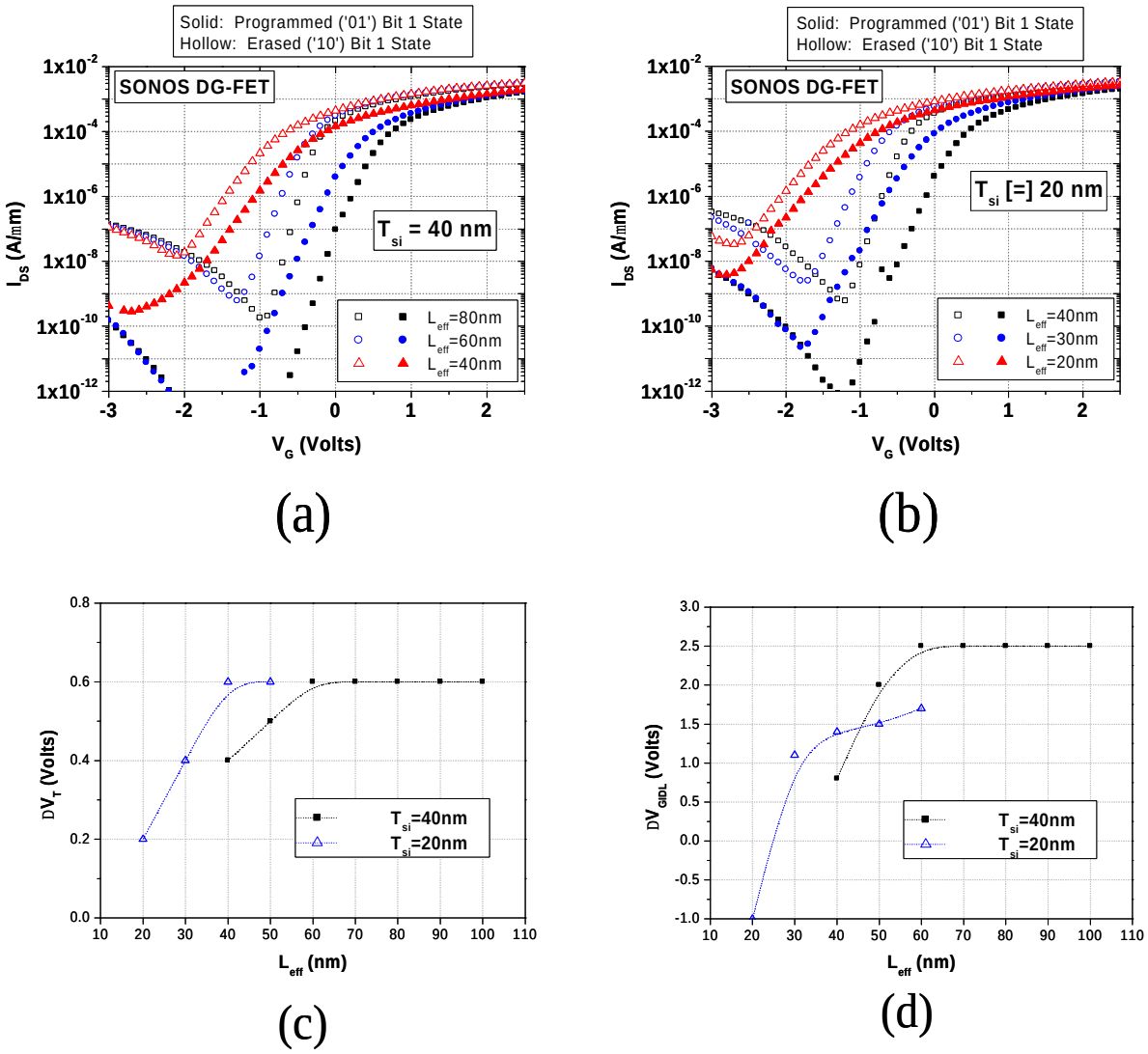


Figure 3.8: Simulated I_{DS} vs. V_G characteristics for the 2 binary Bit 1 states with smallest V_T separation (ΔV_T) of a SONOS DG-FET with (a) T_{si}=40nm, and (b) T_{si}=20nm (for various gate lengths (L_g)). As shown, scaling of the cell's L_g degrades both its sub-threshold swing and ΔV_T, as shown in figure (c). Additionally, the observed swing degradation detrimentally affects the separation in GIDL current (ΔV_{GIDL}) between the programmed and erased Bit 2 states (as shown in figure (d)). The observed behavior is exacerbated with smaller T_{si} (due to the enhanced gate-to-body coupling attained with these structures [3.20]). These effects ultimately limit the scalability of dual-bit SONOS DG-FETs.

3.4 Fabrication of SONOS FinFET NVM Cells

FinFET SONOS NVM cells were fabricated using a gate-first process flow similar to that used in the fabrication of spacer FinFETs [3.12] to investigate their dual-bit operation. **Figure 3.9** illustrates the process flow used to fabricate the FinFET SONOS NVM cells. A silicon-on-insulator (SOI) wafer is used as the starting substrate. The SOI layer is thinned to ~50nm by thermal oxidation. ~60nm SiO_2 is retained over the SOI to "deactivate" the channels at the top surfaces of the etched Si fins, so that channels are formed only along the sidewalls of the fins (**Figure 3.9(a)**).

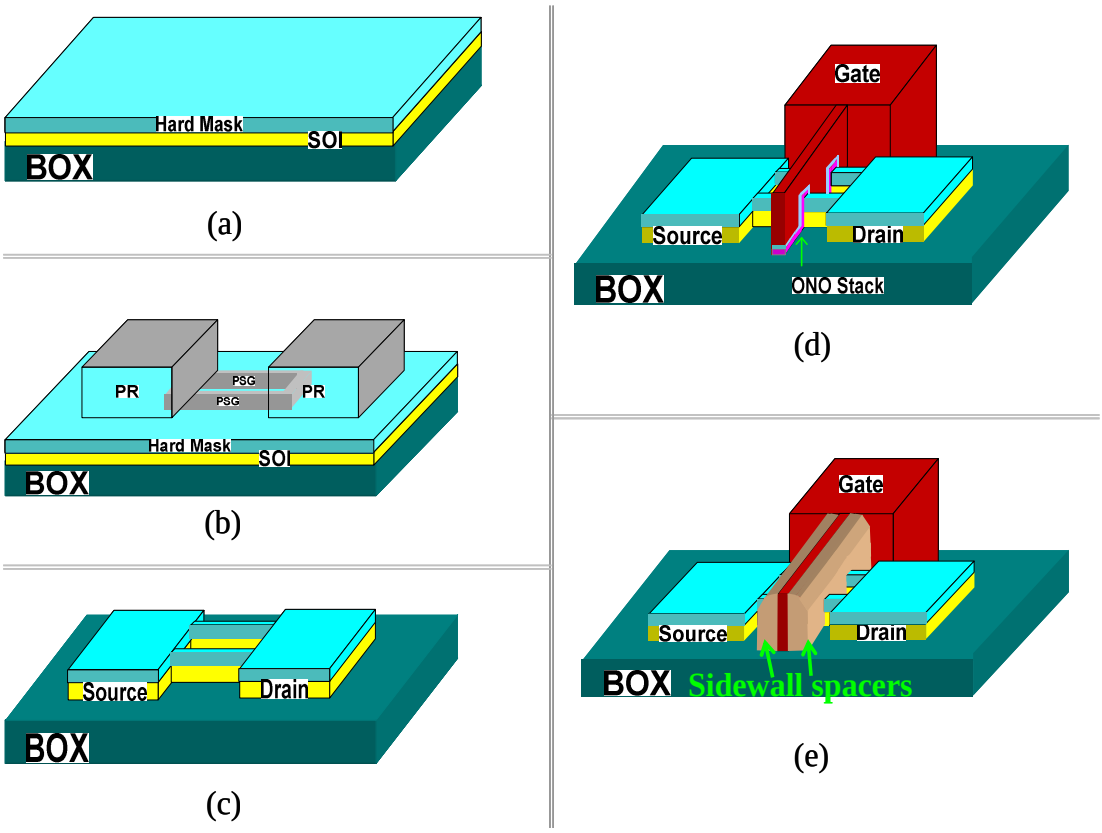
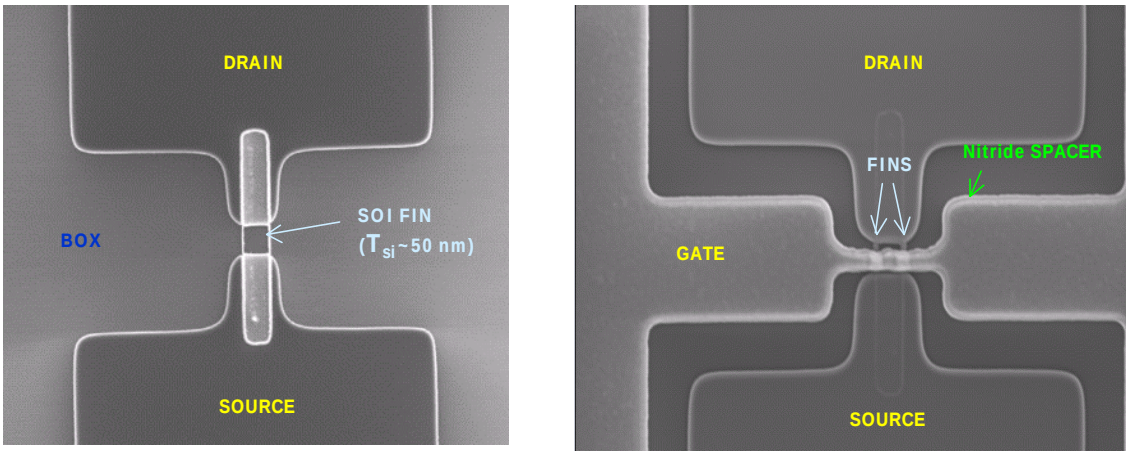


Figure 3.9: Process flow used to fabricate the dual-bit SONOS FinFET NVM cell.

Spacer lithography is then used to define the narrow Si fins (with width T_{si}~50nm), using amorphous silicon as the sacrificial material and ~50nm of phosphosilicate glass (PSG) as the spacer material. Optical lithography is then used to define the source and drain (S/D) contact regions. The composite active-area mask (comprised of PSG spacers + photoresist pads, as shown in **Figure 3.9(b)**) is then used to pattern the underlying SiO_2 hard mask and SOI layers by reactive ion etching (RIE), as shown in **Figure 3.9(c)**.

A sacrificial SiO_2 layer (~3nm) is thermally grown and then selectively removed (in dilute HF solution) to eliminate residual etch damage from the fin sidewalls [3.13]. The oxide-nitride-oxide (ONO) stack is then formed by thermal oxidation (~2.8nm SiO_2) followed by low-pressure chemical vapor deposition (LPCVD) of SiRN (~5.5nm) and SiO_2 (5nm). *In-situ* doped n-type polycrystalline silicon (N^+ Poly-Si) gate and low-temperature LPCVD oxide (LTO) gate hard mask layers, each ~150nm thick, are then deposited. The gate electrodes are patterned using optical lithography and RIE (**Figure 3.9(d)**). Si_3N_4 (~100nm) is then deposited and anisotropically etched back to form (~70nm) gate-sidewall spacers (**Figure 3.9(e)**). **Figure 3.10** shows a scanning electron micrograph (SEM) image of the FinFET SONOS NVM device after (a) active area formation and (b) nitride spacer definition. Next, ~30nm of LTO is deposited throughout, and phosphorus ion implantation (5E15 P^+/cm^2 at 40keV, 7^o tilt, R_P~40nm) is then used to dope the S/D regions. Optical lithography and wet etching are used to open probe/contact holes to the gate and S/D electrodes. Device fabrication is completed with thermal annealing (80s at 900^oC in N_2 ambient) to diffuse and activate the implanted

dopants, followed by a sintering step (30m at 450°C in forming gas) to improve Si/SiO_2 interface properties.



(a) (b)

Figure 3.10: Scanning Electron Micrograph (SEM) images of the FinFET SONOS NVM cell after (a) Active area formation, and (b) Nitride spacer formation.

3.5 Device Characterization

Figure 3.11 shows measured I_{DS}-V_{GS} characteristics for a SONOS FinFET NVM cell before and after Bit 2 was programmed via HEI. As expected, an increase in off-state GIDL current, and no increase in V_T, is seen for forward read operation with only Bit 2 programmed. For reverse read operation, an increase in V_T is evident, whereas no GIDL is seen. These results demonstrate that the new (ΔI_{GIDL}) read method can be used to selectively detect charge stored on the bit close to the drain. **Figure 3.12** shows measured I_{DS}-V_{GS} characteristics for each of the 4 states of a SONOS FinFET NVM cell,

which qualitatively match the simulations in f**igure 3.6(a)**: V_T is high when Bit 1 (near the source) is programmed (States '01' and '00'), so that on-state current can be used to determine the state of Bit 1; GIDL current is enhanced when Bit 2 (near the drain) is programmed (States '10' and '00'), so that off-state current can be used to determine the state of Bit 2.

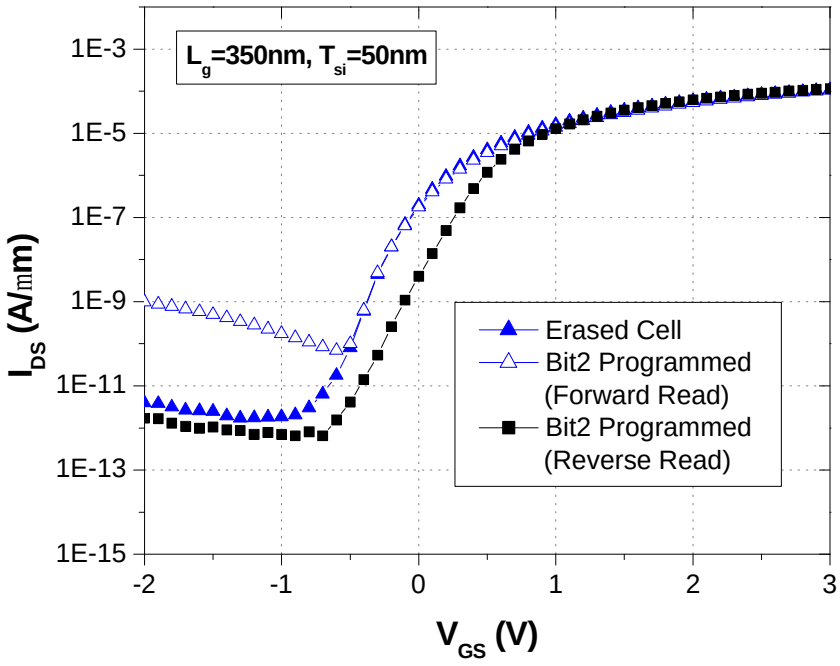


Figure 3.11: Measured I_{DS}-V_{GS} characteristics (V_{DS}=1.5V) of a SONOS FinFET cell before and after Bit 2 was programmed via HEI (V_{GS}=7.5V, V_{DS}=8.0V, 80μs).

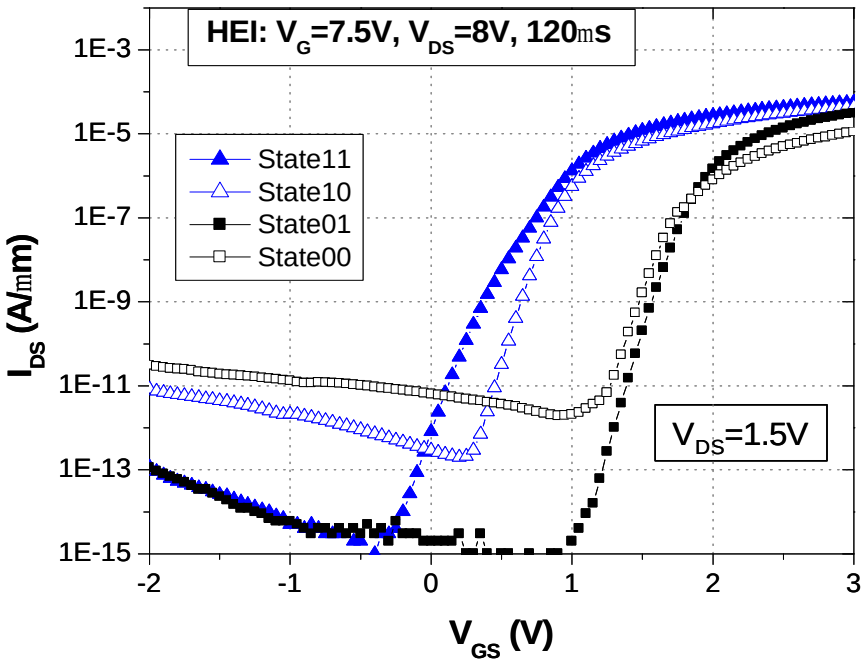
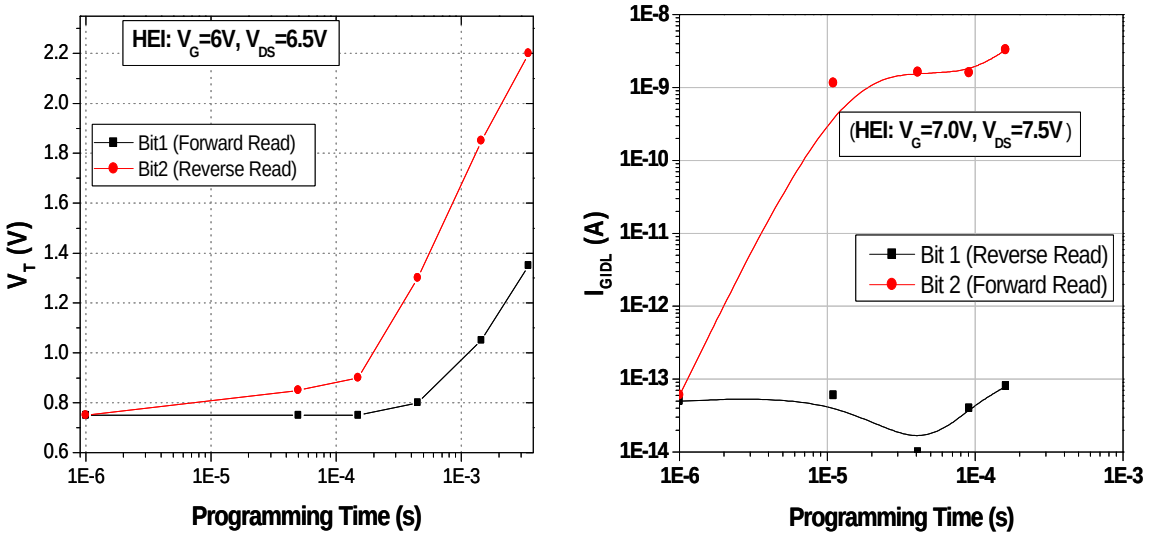


Figure 3.12: Measured SONOS FinFET cell I_{DS}-V_{GS} characteristics (V_{DS}=1.5V) for each of the 4 programmed states (L_g=350nm, T_{si}=50nm).

Figure 3.13(a) shows the change in V_T (defined at I_{DS}=0.1μA) *vs.* Bit 2 HEI programming time. As can be seen from this figure, a ~0.65V increase in reverse-mode V_T (indicative of the state of Bit 2) is achieved with a ~450 μs programming pulse; however, the forward-mode V_T (indicative of the state of Bit 1) also increases slightly. The disturbance of Bit 1 increases significantly with programming times longer than 500 μs. These results highlight a weakness of the conventional charge-detection method (that was highlighted previously with the simulation results of **figure 3.7**): V_T is sensitive not only to charge stored near to the source, but also to charge stored at the center of the channel; this makes it more difficult to achieve dual bit operation with high stored charge density and/or very short L_g. In contrast, GIDL current is less sensitive to charge stored

away from the drain electrode, as demonstrated by the measured HEI programming characteristics shown in **Figure 3.13(b)**.



(a) (b)

Figure 3.13: Bit2 HEI programming characteristics of a SONOS FinFET NVM cell, as measured by (a) the conventional (ΔV_T) and (b) the new (ΔI_{GIDL}) charge detection methods. In the latter case, the GIDL current of the programmed and erased states is measured @ V_{GS} =-1.5V, and V_{DS}=1.5V. The CBD effect in (a) is evident.

Charge stored in a SONOS FinFET NVM cell can be readily removed via BTBHHI, as evidenced by the measured erase characteristics in **Figure 3.14.** As shown, charge stored next to the Drain electrode can be effectively erased with this method without any detrimental effect on the uncharged state of the complementary bit.

The measurements obtained in this section corroborate the observations made earlier in the simulation study section of this chapter: As shown, the conventional (read, program, and erase) methods can readily be used on the NVM cell. The scalability of this structure is nonetheless limited (when using the conventional charge detection method) since V_T is highly sensitive to SCE. As an alternative, the new (ΔI_{GIDL}) charge detection method can enhance the scalability of this structure since it is less susceptible to SCE; however, placement of charge underneath the gate electrodes ultimately limits the scalability of this structure (even when the new charge detection method is used).

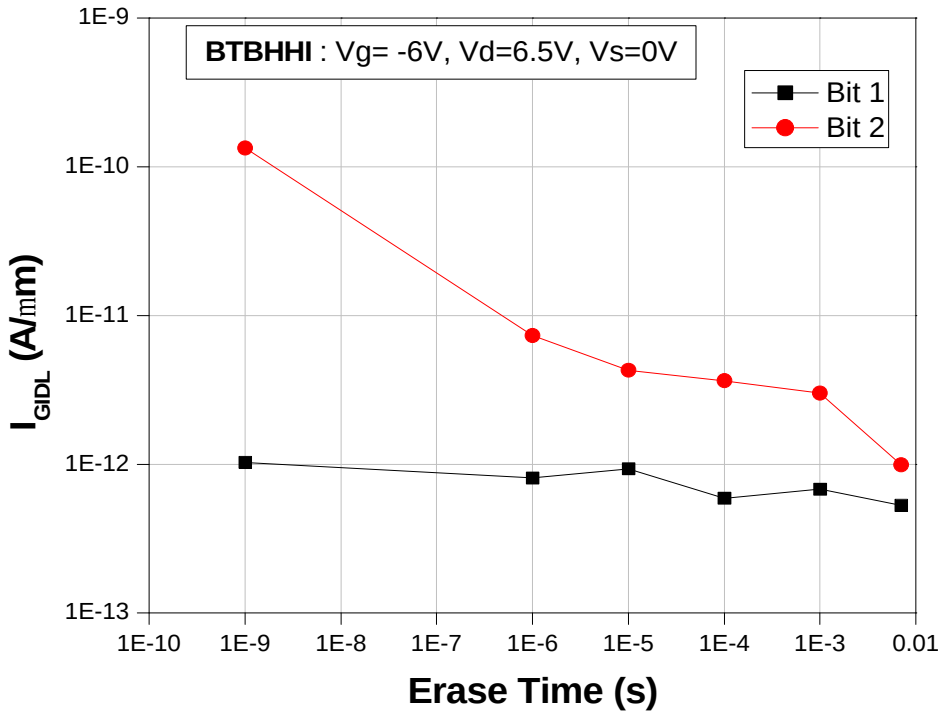


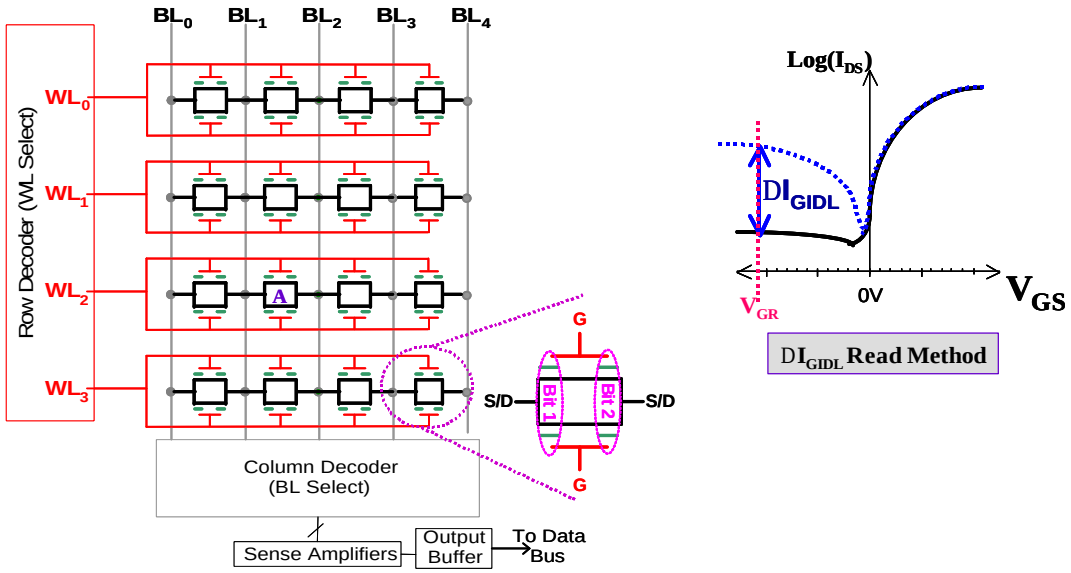
Figure 3.14: Bit2 BTBHHI erase characteristics of the SONOS FinFET NVM cell (measured @ V_{GS} =-1.5V, and V_{DS}=1.5V).

3.6 Memory Array Architectures

In this section, the use of a dual-bit DG-FET NVM cell (such as the cell discussed in this chapter or the gate-sidewall-storage FinFET cell discussed in the next chapter) within both a NOR-type and NAND-type architecture is discussed. The selective operation of each bit of a dual-bit DG-FET NVM cell within these array architectures is discussed, with particular emphasis on the selective read operation via the new DI_{GIDL} charge detection method that is proposed.

3.6.1 "Virtual Ground in SOI" NOR-Type Architecture

Figure 3.15 shows a circuit diagram of DG-FET NVM cells arranged in a (Virtual Ground) NOR array configuration, along with the word-line (WL) and bit-line (BL) bias conditions required to *selectively* read, program, and erase Bit 1 within Cell "A". In this architecture, cells are arranged in a 2-dimensional (2D) array, where the gates of all cells in the same row are connected to the same word-line, and all the source (or drain) electrodes of all cells in the same column are connected to the same Bit-line. This architecture utilizes the cell's symmetry to treat each Bit-Line (BL) as a source or drain electrode as necessary and thus attain 2 bits of storage in every cell. This approach has already successfully been used in the planar 'virtual ground' NROM architecture reported in [3.21] to attain 2 bits of storage per cell.



Operation on Cell 'A' (Bit 1)	V_{WL2}	All other WLs	V_{BL0}	V_{BL1}	V_{BL2}	V_{BL3}	V_{BL4}
Read (ΔV_T)	V_R (~0.5V)	0V	0V	0V	1.5V	1.5V	1.5V
Read (ΔI_{GIDL})	V_{GR} (~ -1.5V)	0V	1.5V	1.5V	0V	0V	0V
Program (HEI)	6.0V	0V	6.5V	6.5V	0V	0V	0V
Erase (HHI)	-6.0V	0V	6.5V	6.5V	0V	0V	0V

Figure 3.15: Schematic circuit diagram of 2-bit DG-FET NVM cells arranged in a NOR array architecture. Bias voltages used to selectively read, program, and erase Bit 1 of cell 'A' are indicated in the table.

In this architecture (as noted previously), each bit of the dual-bit DG-FET NVM cell can be *selectively* read, programmed or erased via the conventional reverse-read, HEI, and hot hole injection (HHI) methods, respectively. To read Bit 1 of Cell A via the DI_{GIDL} method, the WL connecting to its gates is biased to a negative voltage (e.g. V_{WL2} ~ -1.5V) with a moderate drain-to-source voltage (V_{DS}) applied (*e.g.* V_{BL1}-V_{BL2} = 1.5V) to mask the complementary bit (in this case, Bit 2). If no charge is stored at Bit 1, then no significant off-state current will flow in bit-lines BL_1 and BL_2. If charge is stored at Bit 1, the electric field in the transverse direction (near Bit 1) will be enhanced (**ref. Figure**

3.6(a)) so that significant off-state (GIDL) current will flow in bit-lines BL_1 and BL_2. Thus, the state of Bit 1 in Cell A can be distinguished by sensing the current flowing in either one of the bit-lines BL_1 and BL_2. To mitigate leakage current from all unselected cells, all of the other word lines are biased at 0V (or a small negative voltage) to ensure that non-selected cells (in different rows) are turned off so that they contribute negligible bit-line current. In addition, in order to prevent the non-selected cells along the same word line (i.e., WL_2) from contributing any bit-line current, the bit lines must be biased such that $V_{DS} = 0V$ for each of these non-selected cells.

Figure 3.16 shows a basic layout (not to scale) of an embodiment of the NOR-type array architecture shown (as a circuit diagram) in **figure 3.15**. As shown, this array layout utilizes SOI technology, and the unit cell in this architecture is the SONOS FinFET NVM cell (as an example). In this layout, the word lines (WLs) are n^+ Poly Si stripes, and the source or drain (S/D) bit-lines (which consist of n^+-Si stripes) are shared between adjacent cells to reduce space, as done with the *Virtual Ground* NOR-type NROM™ architecture [3.21]. Nonetheless, the use of this transistor design requires the placement of WLs both on the top *and* on the side of the active area of each cell, which increases slightly the size of the unit cell. Consequently, the size of the unit cell with this FET design is ~$7.5F^2$ (or ~$3.75F^2$ per bit), which is slightly larger than the value reported for the 2-bit NROM™ cell (~$2.5F^2$ per bit).

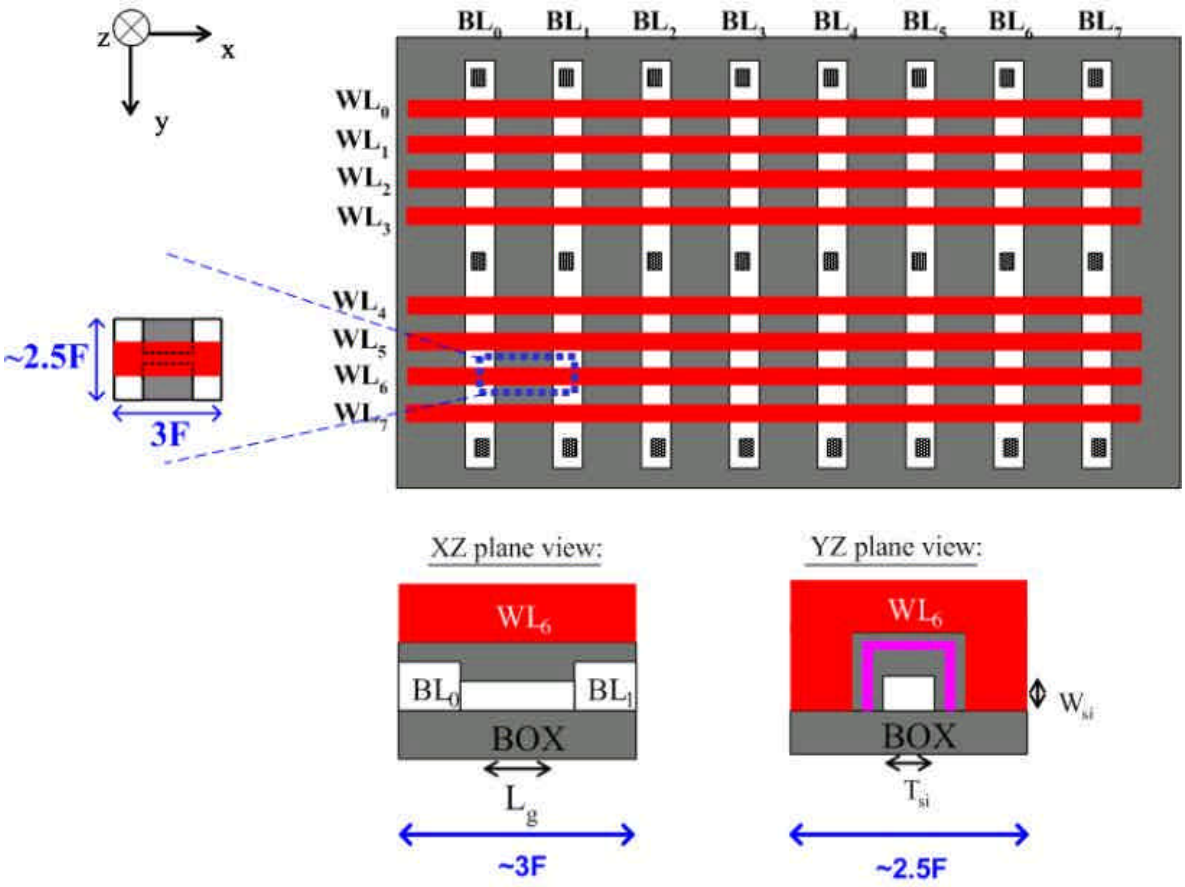
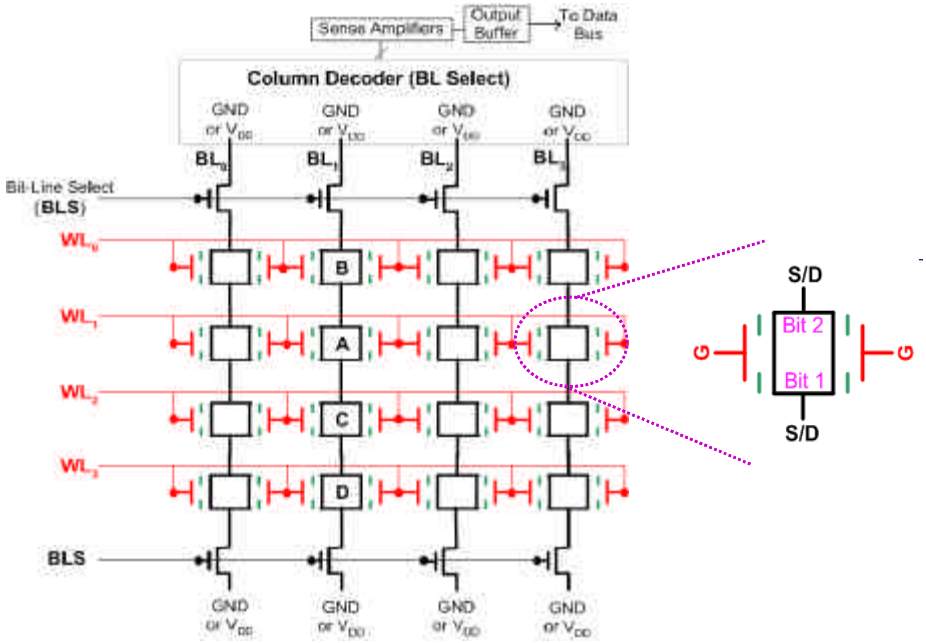


Figure 3.16: NOR-type array architecture layout (based on SOI technology) that utilizes the SONOS FinFET structure as its unit cell.

3.6.2 NAND-Type Array Architecture

Figure 3.17 shows a circuit diagram of a NAND-type array architecture that utilizes a double-gated NVM FET (such as the FinFET SONOS NVM cell) as its unit cell. In this architecture, cells belonging to the same bit-line are connected in series between 2 bit-line select (BLS) transistors, which must each be able to pass a high voltage (V_{DD}) as well as a low voltage (GND) to allow for forward-read as well as reverse-read operation of each cell.

Each bit of each 2-bit DG-FET NVM cell could be selectively programmed via the HEI method, though this method is not normally used with this architecture [3.17]. Alternatively, the Band-to-Band Tunneling Induced Hot Electron (BBHE) injection method could also be used to selectively program each bit within these structures. Though this method has only been demonstrated to date on p-channel SONOS-type NVM cells [3.22][3.23], its use has already been demonstrated with a NAND array architecture composed of dual-bit p-type Band-gap Engineered SONOS ('BE-SONOS') NVM cells [3.23].



Operation on Cell 'A' (Bit 1)	V_{WL1}	All other WLs	V_{BLS}	$V_{BL1(top)}$	$V_{BL1(bottom)}$	All other BLs
Read (ΔV_T)	$\mathbf{V_R}$ (~0.5V)	$\mathbf{V_P}$ (~ 2V)	~ 2V	$\mathbf{V_{DD}}$ (~1.5V)	0V	open
Read (ΔI_{GIDL})	V_{GR} (~ -1.5V)	$\mathbf{V_P}$ (~ 2V)	~ 2V	0V	$\mathbf{V_{DD}}$ (~1.5V)	open
Program (HEI)	6.0V	0V	~ 2V	0V	6.5V	open
Erase (FN Tunneling)	-9.0V	0V	~ 2V	0V	0V	open

Figure 3.17: Schematic circuit diagram of 2-bit DG-FET NVM cells arranged in a NAND array architecture. Bias voltages used to selectively read, program, and erase Bit 1 of cell 'A' are indicated in the table.

To read a specific cell in the array, each of the other cells that share the same bit-line are turned on strongly (by applying moderate word-line biases) so that these simply serve as pass transistors. Current flow through the bit-line is then determined by the state of the selected bit in the selected cell. Note that cells sharing the same word lines in a NAND array can be read simultaneously, in contrast to cells sharing the same word lines in a NOR array which must be read sequentially.

Figure 3.18 shows a basic layout (not to scale) of an embodiment of this NAND-type array architecture along with planar cross-sections of the unit cell. As shown, this architecture uses SOI technology, and the unit cell in this architecture is the dual-bit SONOS FinFET NVM cell structure (although the architecture also applies to other double-gated structures, such as gate-sidewall-storage FinFET structure discussed in chapter 4). As shown, the size of the unit cell within this architecture is very small (~$4F^2$, or ~$1F^2$ per bit), which indicates that this architecture is indeed the optimum array architecture design in terms of memory density (since it significantly reduces the size per bit.

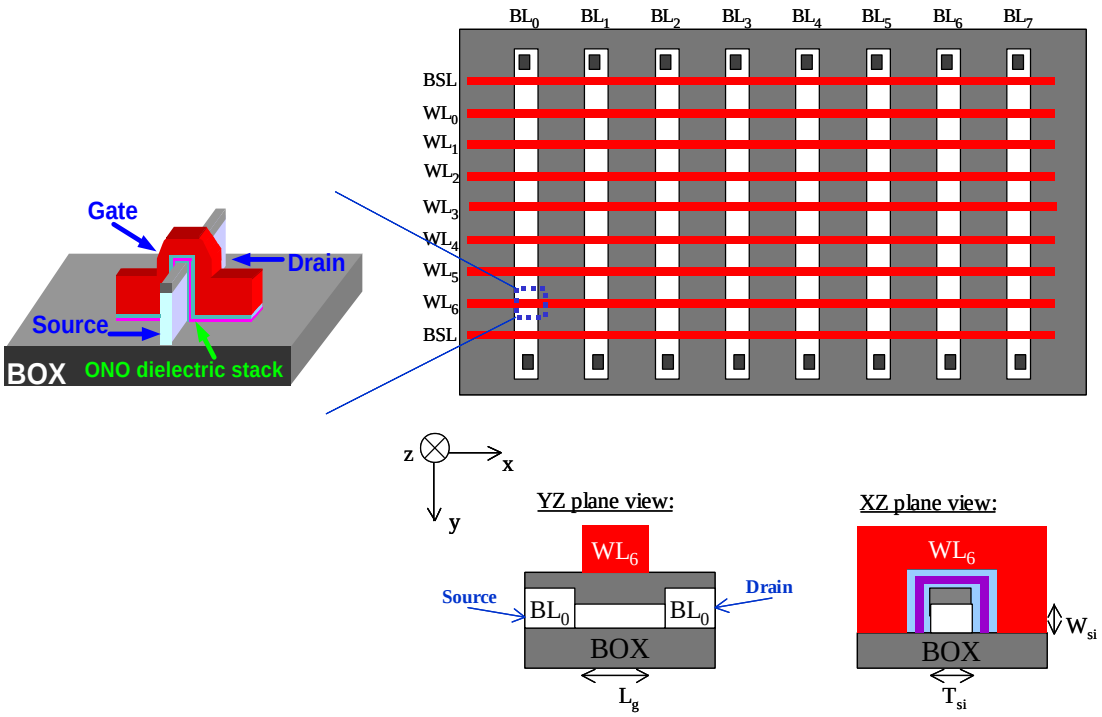


Figure 3.18: Basic schematic layout (not to scale) and cross-sections of the NAND-type array architecture that utilizes the dual-bit SONOS FinFET NVM cell as its unit cell (highlighted within the layout with a dashed line).

3.7 Conclusions

In this chapter, the scalability of a dual-bit SONOS FinFET NVM cell has been evaluated. The scalability of this structure is very limited (when the conventional charge detection method is used) since its threshold voltage is highly sensitive to short-channel effects. As an alternative, the new (ΔI_{GIDL}) read method enhances the scalability of this structure since this method is less sensitive to charge stored in the complementary bit. However, the placement of charge underneath the gates of this structure ultimately limits its scalability since the charge storage mechanism is intrinsically coupled with its electrostatic behavior (ideally, these two should be decoupled).

3.8 References

[3.1] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How far can Si go?," *IEDM Technical Digest*, p. 553 (1992).

[3.2] B. Yu, Y.-J. Tung, S. Tang, E. Hui, T.-J. King, and C. Hu, "Ultra-thin-body silicon-on-insulator MOSFETs for terabit-scale integration," *International Semiconductor Device Research Symposium*, p. 623 (1997).

[3.3] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," *Proc. IEEE*, Vol. 85, No. 4, p. 486 (1997).

[3.4] Lindert, N.; Chang, L.; Choi, Yang-Kyu; Anderson, E.H.; Wen-Chin Lee; Tsu-Jae King; Bokor, J.; and Chenming Hu, "Sub-60-nm quasi-planar FinFETs fabricated using a simplified process"; *IEEE Electron Device Letters*, Vol. 22, p. 487 (2001).

[3.5] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor, and T.-J. King, "FinFET SONOS Flash Memory for Embedded Applications", *IEDM Technical Digest*, p. 609 (2003).

[3.6] M. Specht, R. Kommling, F. Hofmann, V. Klandzievski, L. Dreeskornfeld, W. Weber, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rosner, M. Stadele, R.J. Luyken, H. Reisinger, A. Graham, E. Hartmann, and L. Risch, "Novel Dual Bit Tri-Gate Charge Trapping Memory Devices", *IEEE Electron Device Letters*, Vol. 25, p. 810 (2004).

[3.7] C.W. Oh, S. D. Suk, Y.K. Lee, S.K. Sung, J.-D. Choe, S.-Y. Lee, D.U. Choi, K.H. Yeo, M.S. Kim, S.-M. Kim, M. Li, S.H. Kim, E.-J. Yoon, D.-W. Kim, D. Park, K.

Kim, and B.-I. Ryu, “Damascene Gate FinFET SONOS Memory Implemented on Bulk Silicon Wafer”, *IEDM Technical Digest*, p. 893 (2004).

[3.8] E.S. Cho, T.Y. Kim, C.H. Lee, C. Lee, J.M. Yoon, H.J. Cho, H.S. Kang, Y.J. Ahn, D. Park, and K. Kim, “Optimized Cell Structure for FinFET Array Flash Memory”, *Solid-State Device Research Conference(ESSDERC)*,p. 289 (2004).

[3.9] S.-K. Sung, T.-Y. Kim, E.S. Cho, H. Y. Cho, B.Y. Choi, C.W. Oh, B.-K. Cho, C.-H. Lee, and D. Park, “Fully Integrated SONOS Flash Memory Cell Array with Body Tied FinFET Structure”, *IEEE Transactions on Nanotechnology*, p. 174 (2006).

[3.10] J.-D. Lee, S.-H. Hur, and J.-D. Choi, “Effects of Floating-gate Interference on NAND Flash Memory Cell Operation”, *IEEE Electron Device Letters*, p. 264 (2002).

[3.11] M. Specht, U. Dorda, L. Dreeskornfeld, J. Kretz, F. Hofmann, M. Stadele, R.J. Luyken, W. Rosner, H. Reisinger, E. Lnadgraf, T. Schulz, J. Hartwich, R. Kommling, and L. Risch, “20 nm Tri-gate SONOS Memory Cells with Multi-level Operation”, *IEDM Technical Digest*, p. 1083 (2004).

[3.12] Y.-K. Choi, T.-J. King, and C. Hu, “Nanoscale CMOS spacer FinFET for the terabit era”, *IEEE Transactions on Electron Devices*, Vol. 24, No. 7, p. 490 (2003).

[3.13] Y.-K. Choi, L. Chang, P. Ranade, J.-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T.-J. King, and J. Bokor, “FinFET Process Refinements for Improved Mobility and Gate Work Function Engineering”, *IEDM Technical Digest*, p. 259 (2002).

[3.14] Synopsys "Taurus Process & Device User Manual" 2003. http://www.synopsys.com.

[3.15] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "*Scaling Theory for Double-Gate SOI MOSFET's*," *IEEE Transactions on Electron Devices*, Vol. 40, No. 12, pp. 2326-2329 (1993).

[3.16] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan, "Characterization of Channel Hot Electron Injection by the Subthreshold Slope of NROM™ device," *IEEE Electron Device Letters*, Vol. 22, No. 11, pp. 556-558 (2001).

[3.17] R. Liu, *et al.*, "Memory Technologies for 45nm and Beyond", *2006 International Electron Device Meeting Short Course*, (2006).

[3.18] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters*, Vol. 8, No. 11, pp. 515-517, 1987.

[3.19] A. Padilla, and T.-J. King Liu, "Dual-bit SONOS FinFET Non-Volatile Memory Cell and New Method of Charge Detection," *VLSI-TSA*, paper T18 (2007).

[3.20] C.-H. Lin, X. Xi, J. He, L. Chang, R. Q. Williams, M. B. Ketchen, W. E. Haensch, M. Dunga, S. Balasubramanian, A. M. Niknejad, M. Chan, and C. Hu, "Compact Modeling of FinFETs Featuring Independent-Gate Operation Mode," *IEEE VLSI-TSA International Symposium on VLSI Technology Proceedings of Technical Papers*, pp. 120 (2005).

[3.21] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NROM: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Letters*, Vol. 21, No. 11, pp. 543-545 (2000).

[3.22] T. Ohnakado, K. Mitsunaga, M. Nunoshita, H. Onoda, K. Sakakibara, N. Tsuji, N. Ajika, M. Hatanaka, and H. Miyoshi, “Novel Electron Injection Method Using Band-to-Band Tunneling Induced Hot Electron (BBHE) for Flash Memory with a P-channel Cell,” *International Electron Devices Meeting Technical Digest*, pp. 279-282 (1995).

[3.23] H.-T. Lue, S.-Y. Wang, E.-K. Lai, M.-T. Wu, L.-W. Yang, K.-C. Chen, J. Ku, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “A Novel P-Channel NAND-Type Flash Memory with 2-bit/cell Operation and High Programming Throughput (>20 MB/sec),” *International Electron Devices Meeting Technical Digest*, pp. 331-334 (2005).

Chapter 4: Design of Dual-bit, n-channel Gate-Sidewall Storage (GSS) FinFET NVM Cells

4.1 Motivations

As discussed in the previous chapter, the conventional dual-bit SONOS FinFET NVM cell has limited scalability since it is highly susceptible to short channel effects (SCE). The threshold voltage (V_T) of this structure is sensitive not only to charge stored near the Source electrode, but also to charge stored near the center of its channel or even near to the Drain electrode –a phenomenon that is commonly known as the complementary bit disturb (CBD) effect [4.1]. As a result, it is difficult to achieve dual-bit operation (when using either the conventional DV_T or the DI_{GIDL} charge detection methods) with either high stored charge density and/or very short transistor gate lengths.

To ameliorate these effects, bit-to-bit interference can be reduced by physically separating the charge-trap sites. This can be done in practice, for instance, by adopting a NVM cell that utilizes the gate-sidewall spacers to store charge [4.2]. To employ the conventional read method with the gate-sidewall-storage structure, a gate-underlapped source/drain (S/D) design (where the channel length L_{eff} is larger than its physical gate length L_g) is necessary, which provides for lower read current.

In this chapter, the first n-channel FinFET-based dual-bit NVM cell with physically separated charge-storage sites (located along the sidewalls of the gate electrodes, see **Figure 4.1**) is presented [4.3]. The conventional read method, as well as the new ΔI_{GIDL} read method (which uses off-state current to determine whether charge is stored at the drain-side bit), can be used to determine the state of the cell. The new read method is much less susceptible to SCE than the conventional method and thus enhances the scalability of the proposed NVM cell. Additionally, the new read method is compatible with a gate-overlapped S/D transistor design (**Figure 4.1(c)**) for improved on-state current (desirable for a NAND array architecture), in contrast to the conventional method.

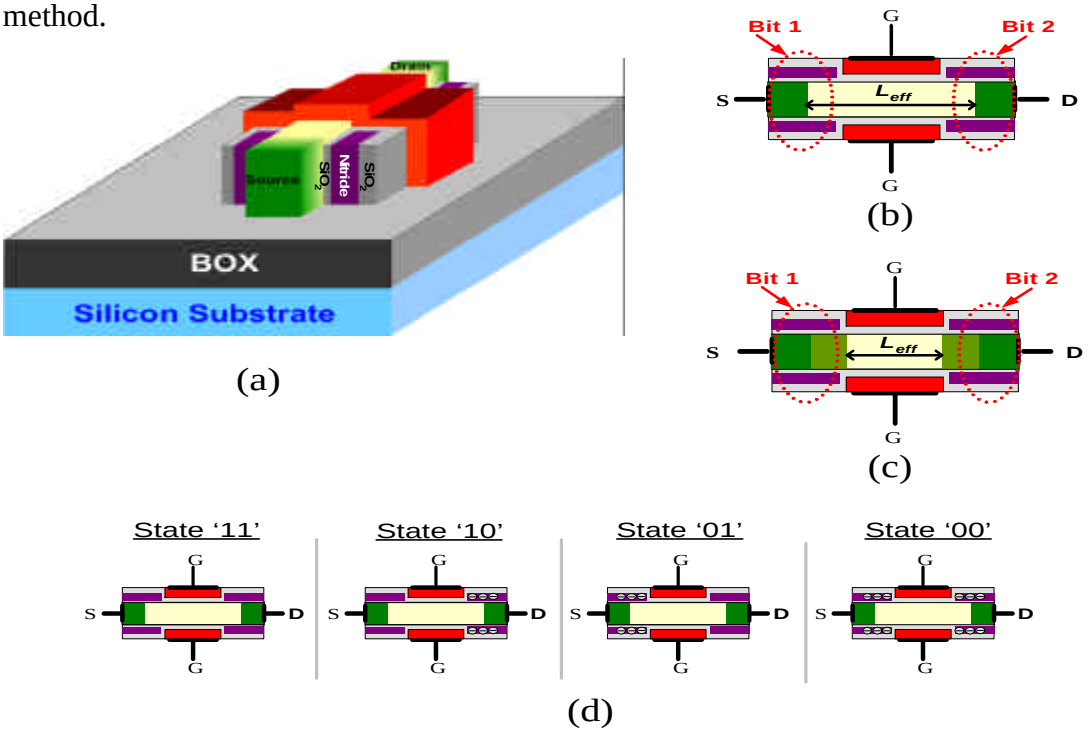


Figure 4.1: (a) Isometric view of the FinFET NVM cell with charge storage sites located at the gate sidewalls. The 2D schematic cross-sections of gate-underlapped ($L_{eff} > L_g$) and gate-overlapped ($L_{eff} \sim L_g$) cell structures are shown in (b) and (c), respectively. (d) Definition of the 4 charge-storage states of the dual-bit cell.

4.2 Cell Structure and Operation

The dual-bit NVM cell structure, illustrated in **Figure 4.1**, consists of an n-channel double-gate (DG) FinFET transistor with (Si_3N_4) charge-trap layers located along the sidewalls of each (p^+ poly-$Si_{0.8}Ge_{0.2}$ or n+ poly-Si) gate electrode. This cell design offers improved gate-length scalability because of its relatively thin gate-dielectric EOT. Each bit of this cell can be independently programmed (via hot electron injection, HEI), read (via the reverse read method), and erased (via hot hole injection, HHI) in a manner similar to that reported in [4.2] for a single-gate sidewall-charge-storage NVM cell. For example, the transistor on-state current can be used to determine the charge-storage state of the bit near to the source. For this conventional read method, however, the source junction must be located underneath the charge-storage site (**Figure 4.1(b)**) to maximize ΔV_T with charge storage at the source-side bit, *i.e.* a gate-underlapped S/D structure is required [4.2].

To allow for a gate-overlapped S/D structure, a new read method is proposed and demonstrated in this work. The structures shown in **Figure 4.1(b)** and **Figure 4.1(c)** were simulated using the 2-dimensional (2D) Taurus device simulator [4.4] using the parameter settings listed in **Figure 4.2**. For optimum performance, the simulated structures utilize a very lightly doped (1×10^{13} cm^{-3}) p-type silicon body (to provide high carrier mobilities and to minimize statistical dopant fluctuation effects) and a silicon body thickness T_{si} chosen to be less than ~0.45*L_{eff} to suppress SCE [4.5]-[4.6]. Additionally, these structures have a gate oxide thickness T_{ox}=6nm, tunnel oxide thickness T_{tox}=3nm, and gate-sidewall Si_3N_4 charge-trap layers with thickness T_{trap}=15nm and length L_{trap}=15nm. The thickness of the control oxide that isolates the charge-trap layer from

the gate electrode was CT_{ox}=5nm. The areal density of charge stored in the programmed charge-trap layer at its bottom interface (nearest to the Si) was set to a maximum value of $1\times10^{13}q/cm^2$, comparable to values used in other work [4.2],[4.7].

Parameter	Value
Gate length, L_g	32 nm
Fin thickness, T_{si}	16 nm
Fin doping (p-type)	1e13 /cm^3
Gate oxide thickness (T_{ox})	6 nm
Tunnel oxide thickness (Tt_{ox})	3 nm
Control oxide thickness (CT_{ox})	5 nm
Trap-region thickness (T_{trap})	15 nm
Trap-region length (L_q)	15 nm
$Q_{ox,max}$ stored	-1e13 q/cm^2

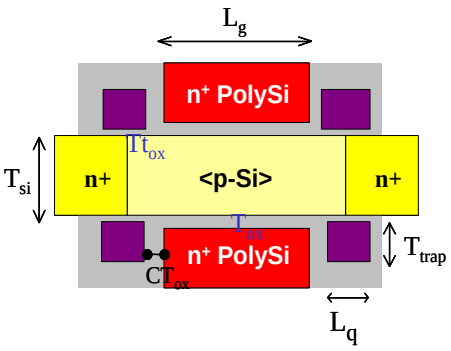


Figure 4.2: Parameter settings for Taurus 2-dimensional device simulations.

Simulated I_{DS}-V_{GS} characteristics of each of the 4 states for cells with a gate-underlapped *vs*. a gate-overlapped S/D structure (with L_g=32nm and T_{si}=16nm) are shown in **Figure 4.3**. For the gate-underlapped design, V_T is high when the source-side bit (Bit1) is programmed (States '01' and '00'), so that transistor on-state current can be used to determine the state of Bit1. Gate-induced drain leakage (GIDL) current [4.8] is enhanced when the drain-side bit (Bit2) is programmed (States '10' and '00'), so that transistor off-state current can be used to determine the state of Bit2. This enhancement in GIDL is due to an increase in the transverse electric field near the drain when charge is

stored in Bit2. The difference in on-state and off-state currents between programmed and erased states is large enough to allow the state of each bit to be determined by forward read operations. Note that V_T is affected by the state of the bit near to the drain due to SCE (specifically, drain-induced barrier lowering), thereby reducing the ΔV_T between programmed and erased states for the bit near to the source. In contrast, the off-state current is not affected by the state of the bit near to the source. This indicates that the new read method is less susceptible to SCE, so that it can be used for NVM cells with aggressively scaled gate lengths.

For the gate-overlapped structure (**Figure 4.3(b)**), which utilizes a lightly doped drain (LDD) extension to exclude the enhancement in GIDL current on the erased Bit2 states [4.9], there is no significant ΔV_T due to charge storage in Bit1, because the source potential barrier is *not* affected by charge stored in Bit1 in a gate-overlapped S/D structure, as discussed in [4.2]. However, there is a clear shift in GIDL current when charge is stored in Bit2. Thus, the state of each bit can be determined from the off-state currents under forward read operation and reverse read operation (with the roles of the S/D electrodes interchanged). Although there is significant off-state-current separation between programmed and erased Bit2 states, the magnitude of this current is not as large as the on-state current. GIDL current can be boosted by enhancing the transverse electric field near to the drain electrode [4.8]. This can be achieved in practice (for example) by applying a larger gate-to-drain (V_{GD}) bias voltage, by storing more electrons within the charge-trapping layers, by using a p+ poly-Si gate, or by enhancing the S/D doping concentration underneath the charge-trapping layers.

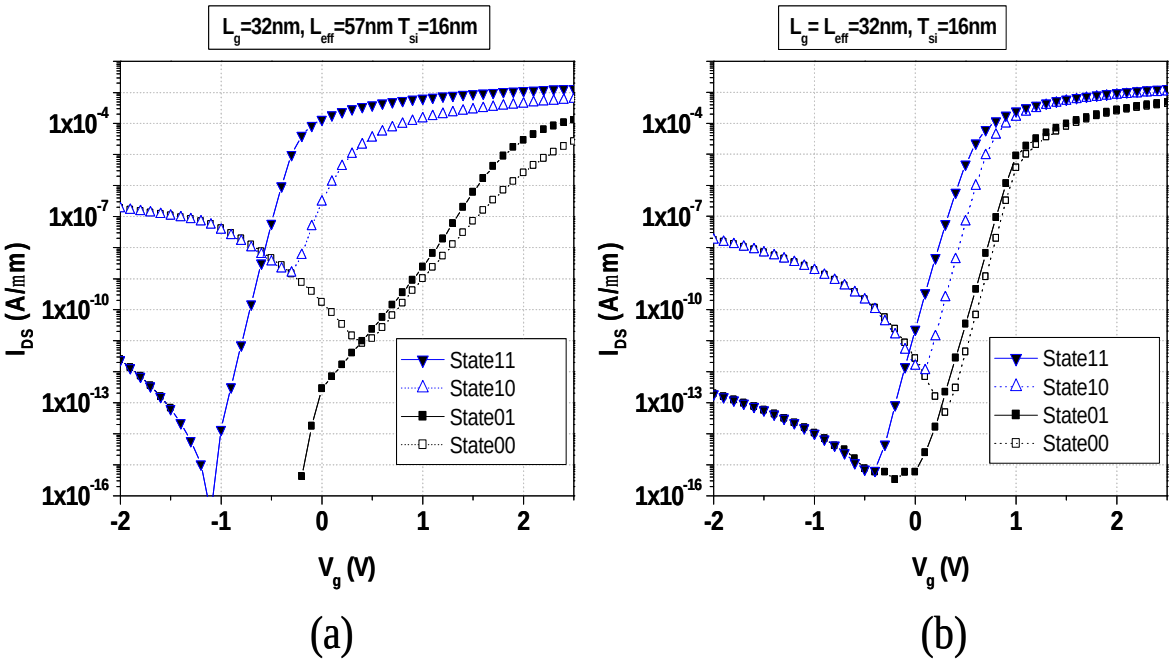


Figure 4.3: Simulated I_{DS}-V_{GS} (V_{DS}=1.5V) characteristics for each state of a gate-sidewall-storage FinFET NVM cell with (a) gate-underlapped source/drain regions and (b) gate-overlapped source/drain regions

4.3 Device Fabrication

This section describes the fabrication process flow of the first FinFET memory cell with physically separated charge-trap sites along the gate sidewalls. Prototype dual-bit NVM cells were fabricated using a gate-first process flow similar to that used to fabricate silicon-on-insulator (SOI) DG spacer FinFETs [4.10]. In the next sub-sections, the most relevant process steps are discussed in detail. Additional process flow details are included in Appendix I.

4.3.1 Active Area Definition

A SOI wafer was used as the starting substrate. The initial thickness of the SOI layer (100nm) was reduced to ~50nm by thermal oxidation (followed by wet etching as necessary), leaving a layer of ~60nm-thick SiO_2 on top of the thinned SOI layer (**Figure 4.4(a)**). This SiO_2 layer serves as a hard mask during the active area patterning process, and it also serves to "deactivate" the channel at the top surface of the patterned fins, so that channels are formed only along the fin sidewalls. Then, spacer lithography was used to define the narrow silicon fin regions: first, ~150nm-thick amorphous silicon (α-Si) sacrificial material was deposited, patterned into islands, and then thermally annealed (via solid phase crystallization); next, ~45nm (to ~75nm) of phosphosilicate glass (PSG) was conformally deposited and etched anisotropically using a reactive ion etch (RIE) process, leaving PSG spacers along the sidewalls of the sacrificial Si islands; then, the sacrificial Si was selectively removed with a dry etch process. Note that the width of the PSG spacers –which defines the width of the Si fins, T_{Si} – is determined by the thickness of the deposited PSG layer, and hence is not constrained by the resolution of the lithography process. Optical lithography was then used to define the source and drain (S/D) contact regions: photoresist was coated, exposed, and developed in the conventional manner, leaving photoresist only in the S/D contact regions (**Figure 4.4(b)**). **Figure 4.4(d)** shows a scanning electron micrograph (SEM) of a NVM cell structure after completion of this step. The combined pattern of PSG spacers and photoresist pads was then transferred to the underlying hard-mask SiO_2 and SOI layers by RIE (**Figure 4.4(c)**). **Figure 4.4(e)** shows a SEM of a NVM cell structure after completion of this step.

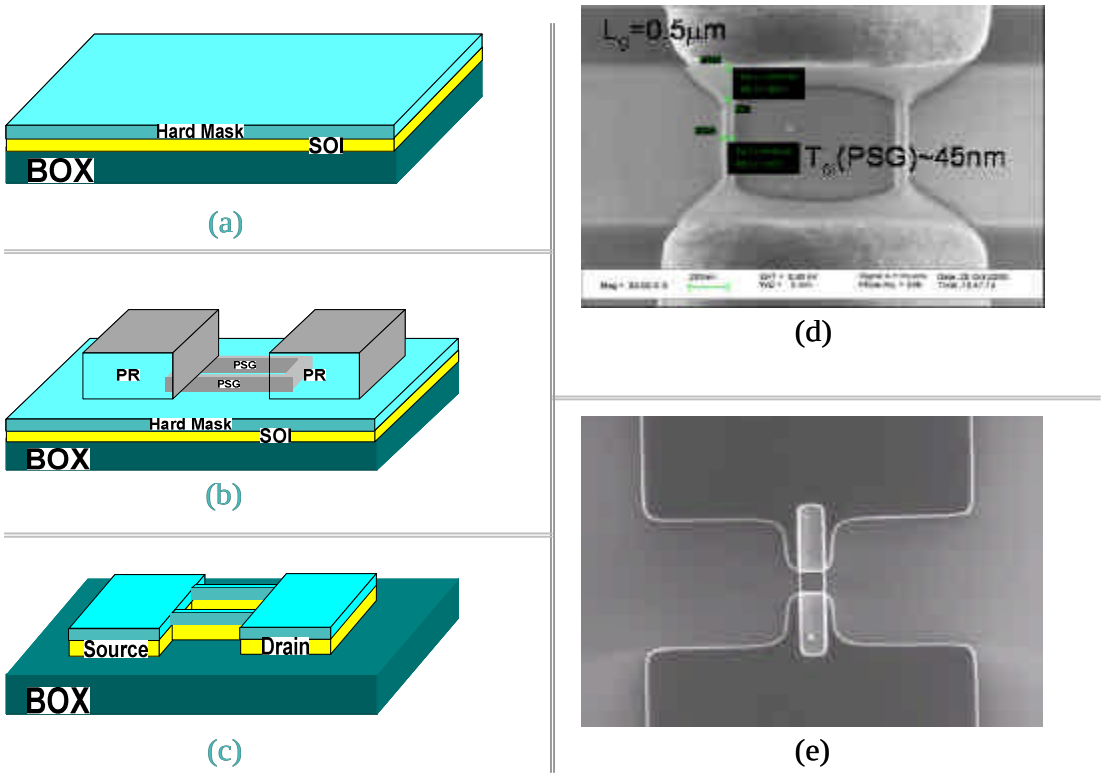


Figure 4.4: (a-c) Process flow used to fabricate gate-sidewall storage (GSS) FinFET NVM cells –ACTIVE Area Definition. (d-e) SEMs of a fabricated cell after completion of processing steps (b) and (c), respectively.

4.3.2 Gate Area Definition

After active-area patterning, a sacrificial oxide (~3nm thick) layer was grown and then selectively removed (in dilute HF solution) to eliminate residual etch damage from the fin sidewalls [4.11]. Then, a thermal oxidation step was performed to grow the gate oxide (~6nm). This was followed by deposition of an *in-situ* doped polycrystalline silicon (either n^+ poly-Si or p^+ poly-$Si_{0.8}Ge_{0.2}$) gate layer and low temperature oxide (LTO) gate hard mask layer, each ~100nm thick. Optical lithography and RIE was then used to pattern the gate electrodes (**Figure 4.3(b)**). **Figure 4.3(d)** shows a SEM of a NVM cell structure after completion of this step.

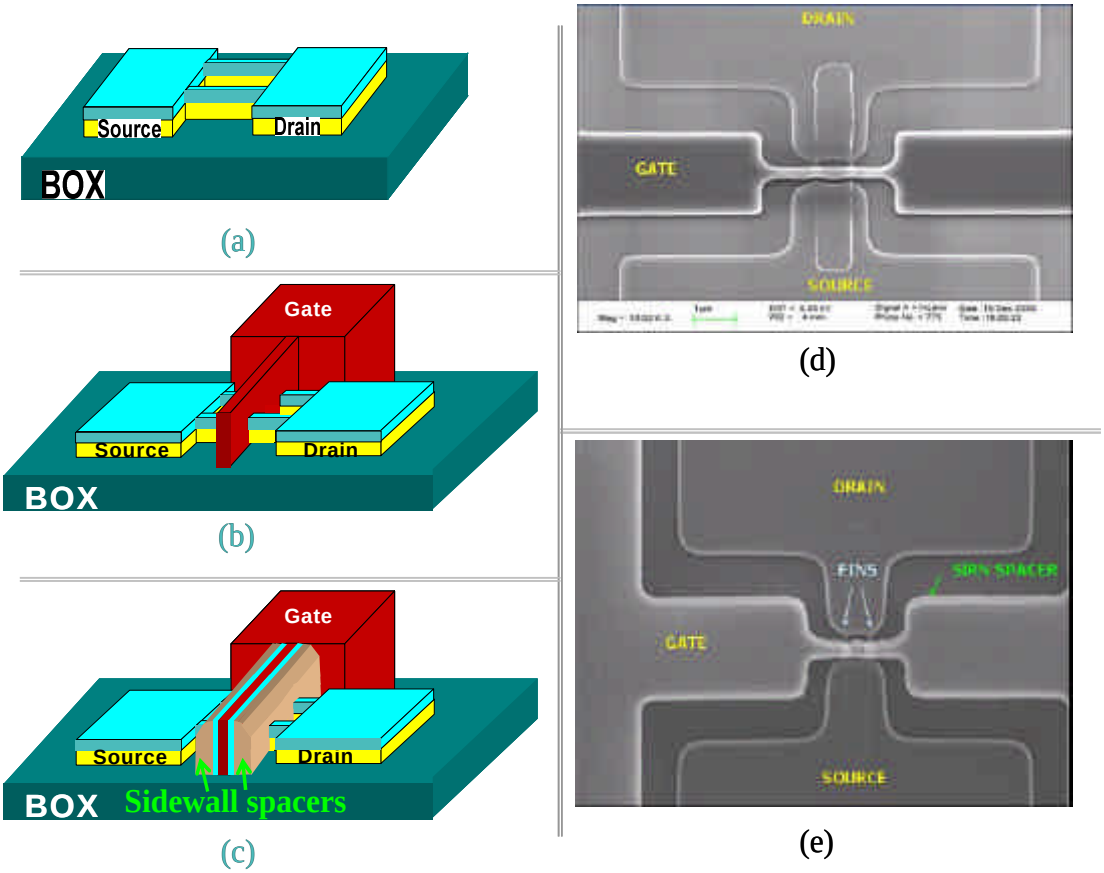
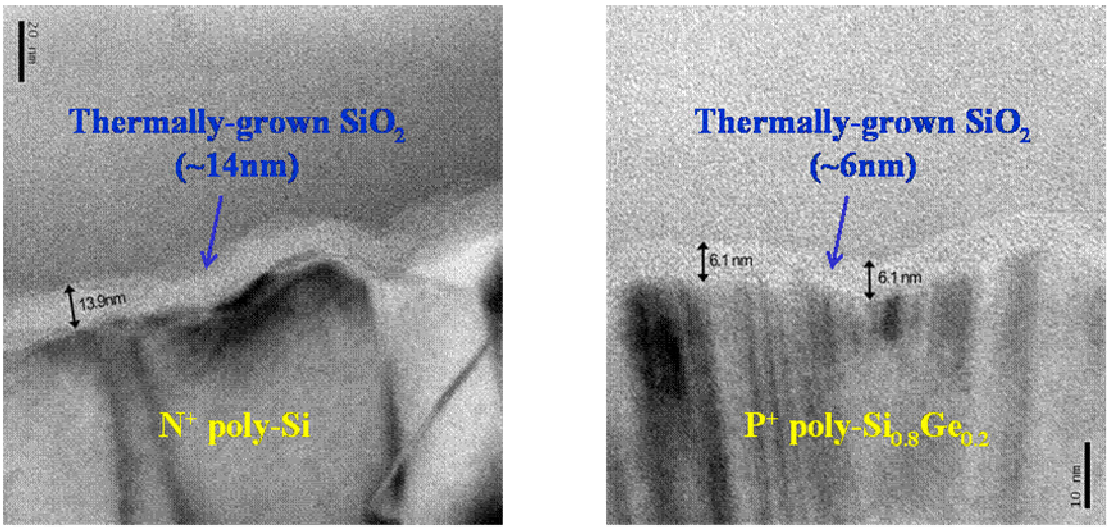


Figure 4.5: (a-c) Process flow used to fabricate gate-sidewall storage (GSS) FinFET NVM cells –Gate Area, Spacer Definition. (d-e) SEMs of a fabricated cell after completion of processing steps (b) and (c), respectively.

4.3.3 Spacer Definition

After the gate electrodes were defined by RIE, another sacrificial oxidation (~3nm SiO_2) was performed to improve the surface quality of the fin sidewalls, in preparation for growing a high-quality tunnel oxide layer. Thermal oxidation was then used to simultaneously grow the tunnel oxide (Tt_{ox}~3nm) and control oxide (CT_{ox}) spacers. Polycrystalline Si and $Si_{0.8}Ge_{0.2}$ each oxidize more quickly than crystalline Si, so that the

control oxide thickness along the gate sidewalls was much thicker than the tunnel oxide. This was confirmed by the cross-sectional transmission electron micrographs (TEM) of the oxidized gate films (**Figure 4.6**), showing that thickness of the control oxide film grown onto the n^+ poly-Si (p^+ poly-$Si_{0.8}Ge_{0.2}$) film is ~14nm (~6nm). The control oxide film serves to electrically isolate the charge-trapping layer from the gate electrodes; consequently, the thickness of this film must be thicker than that of the tunnel oxide (i.e., $CT_{ox} > Tt_{ox}$). Then, ~100nm of silicon-rich nitride (SiRN) was deposited via chemical vapor deposition. Anisotropic dry etching was then performed to form the SiRN spacers along the sidewalls of the gate electrodes (**Figure 4.5(c)**). **Figure 4.5(e)** shows an SEM after completion of this step.



(a) (b)

Figure 4.6: Cross-sectional TEM images that show the thickness of the control oxide film grown along the sidewalls of the in-situ doped (a) n^+ Poly-Si and (b) p^+ Poly-$Si_{0.8}Ge_{0.2}$ film.

4.3.4 Final Steps

After sidewall formation, phosphorous ion implantation (5E15 P^+/cm^2 at 40keV, 7^o tilt, R_P~40nm) was performed. Then, optical lithography and wet etching steps were performed to define the probe/contact regions. Device fabrication was completed with incremental rapid thermal annealing (RTA) steps to active the implanted dopants, followed by a sintering step to improve the Si/SiO_2 interface properties. The dopant activation step is highly critical: in order to maximize the current separation between programmed and erased states, the source and drain junction edges must be located somewhere underneath the width of the SiRN gate-sidewall spacers. This task is not trivial, since dopant diffusion in ultrathin Si films is not well characterized. Therefore, a conservative thermal budget (only 20s at 900^o C) was initially used. Additional RTA steps were performed (up to a cumulative time of 80s at 900^o C) as necessary to further laterally diffuse the source/drain dopants.

4.4 Device Characterization Results

4.4.1 Gate-underlapped S/D Cell Design

As already mentioned, each bit of the n-channel Gate-Sidewall Storage (GSS) FinFET NVM cell can be independently programmed (via HEI), read (via the reverse read method), and erased (via HHI) in a manner similar to that reported in [4.2] for a single-gate GSS NVM cell. The state of the bit next to the source (Bit 1) can be determined by the conventional method of charge detection, which relies on a shift in the cell's threshold voltage (V_T) with stored charge. To determine the state of the bit next to the drain (Bit 2), a "reverse read" operation (in which the roles of the source and drain are interchanged) can be performed. Alternatively, the state of Bit 2 can be determined by sensing the cell's GIDL current, which increases with stored charge near the drain (as already discussed in the last chapter for the conventional SONOS FinFET NVM cell).

Figure 4.7 shows measured I_{DS}-V_{GS} characteristics for a fabricated n-channel GSS FinFET NVM cell (L_g=240nm, W=100nm, T_{si}~75nm) before and after Bit 2 was programmed via HEI. As expected, an increase in GIDL current, and no increase in V_T, is seen for forward read operation with only Bit 2 programmed. For reverse read operation, an increase in V_T is evident, whereas no GIDL is seen.

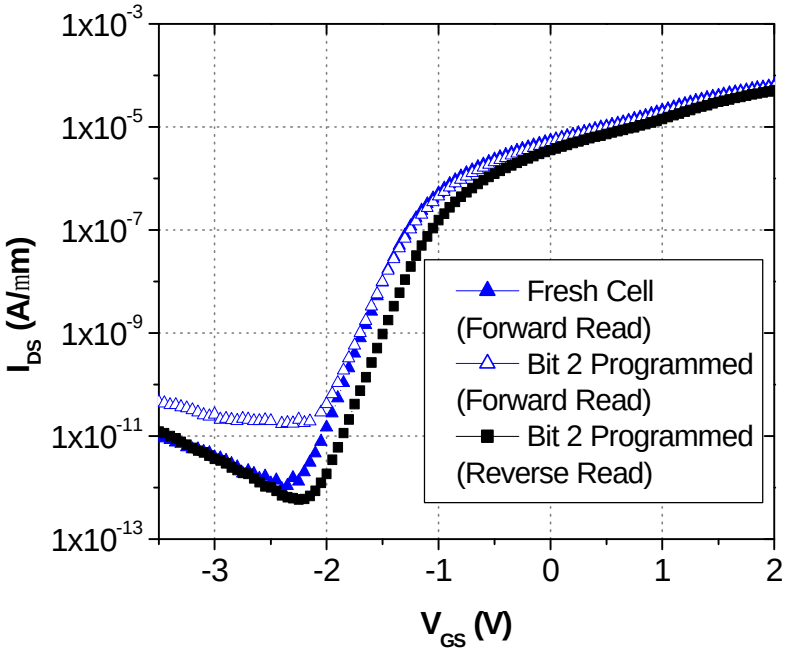


Figure 4.7: Measured I_{DS}-V_{GS} curves of an n-channel GSS FinFET NVM cell (p^+ Poly-$Si_{0.8}Ge_{0.2}$ gate) before and after Bit2 was programmed via HEI (V_{GS}=7V, V_{DS}=7.5V, 500μs). (L_g=240nm, W=100nm, T_{si}=75nm, V_{DS}=1.2V).

Figure 4.8 shows the change in V_T (defined at I_{DS}=0.1μA, V_{DS}=1.2V) *vs.* Bit 2 HEI programming time. Negligible programming disturbance on the complementary bit is seen, despite the long programming time, which shows that the GSS design is less susceptible to the CBD effect than the conventional SONOS design.

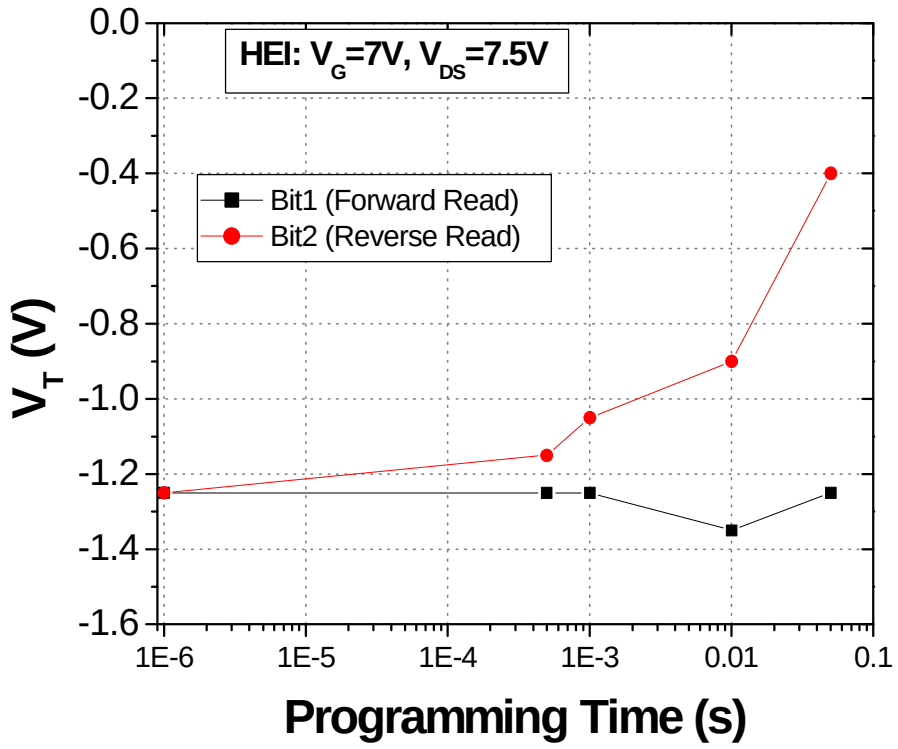


Figure 4.8: Measured HEI programming characteristics of an n-channel GSS FinFET NVM cell (p^+ Poly-$Si_{0.8}Ge_{0.2}$ gate). (L_g=240nm, W=100nm, T_{si}=75nm, V_{DS}=1.2V).

4.4.2 Gate-overlapped S/D Cell Design

As already mentioned, the dopant activation step is a very critical step in the process flow of GSS FinFET NVM cells. In this cell design, the source and drain junction edges must be located somewhere underneath the width of the SiRN gate-sidewall spacers; otherwise, charge stored next to the source electrode will not affect the cell's V_T (and this will reduce the V_T separation between programmed and erased Bit 1 states when the conventional read method is used). Additionally, it will be very difficult (if not impossible) to store electrons next to the Drain electrode (via the conventional HEI programming method) in a gate-overlapped S/D cell design.

I_{DS}-V_{GS} measurements were performed on multi-fin GSS n-channel FinFET structures (with gate lengths L_g=5μm, 10μm) that underwent RTA for 80s at 900 ^oC. To selectively program each bit of these structures, the band-to-band tunneling induced hot electron injection (BBHE) [4.12] and HEI programming mechanisms were attempted. Measured I_{DS}-V_{GS} characteristics after programming Bit 2 of the fabricated GSS FinFET NVM cell (via HEI) are shown in **figure 4.9**. As shown, no significant change in the cell's V_T or GIDL current were detected after the cell was programmed via HEI. These results are likely due to the fact that the S/D junction edges completely overlap the charge-trapping SiRN gate-sidewall spacers; as a result, most generated hot electrons are *not* able to tunnel towards the charge-trapping layers, and thus programming of Bit 2 does not occur. This indicates that HEI cannot be readily used to selectively program each bit of a gate-overlapped GSS FinFET structure.

Measured I_{DS}-V_{GS} characteristics are shown in **Figure 4.10(a)** for a cell in which Bit2 (next to the drain) was selectively programmed via BBHE (programming conditions: V_G = 7V, V_{SD} = -3V, 100ms), *i.e.*, State '10'. There is markedly higher GIDL current, but negligible change in V_T (ΔV_T), for the forward I_{DS}-V_{GS} characteristic (State '10') as compared with the reverse I_{DS}-V_{GS} characteristic (State '01'). Simulated I_{DS}-V_{GS} characteristics on a structure with similar dimensions (T_{si}=40nm, L_g=10μm) as those of the measured cell, shown in **Figure 4.3b**, show similar behavior (*i.e.* a large change in GIDL current and no significant ΔV_T), though the simulated results *under-estimate* the GIDL current of the programmed Bit2 state by an order of magnitude. This indicates that the fabricated cell has perhaps a gate-overlapped S/D structure (due to excessive

thermal annealing used to activate the implanted dopants). Nonetheless, these results demonstrate that each bit can be selectively programmed and read with the novel charge detection method.

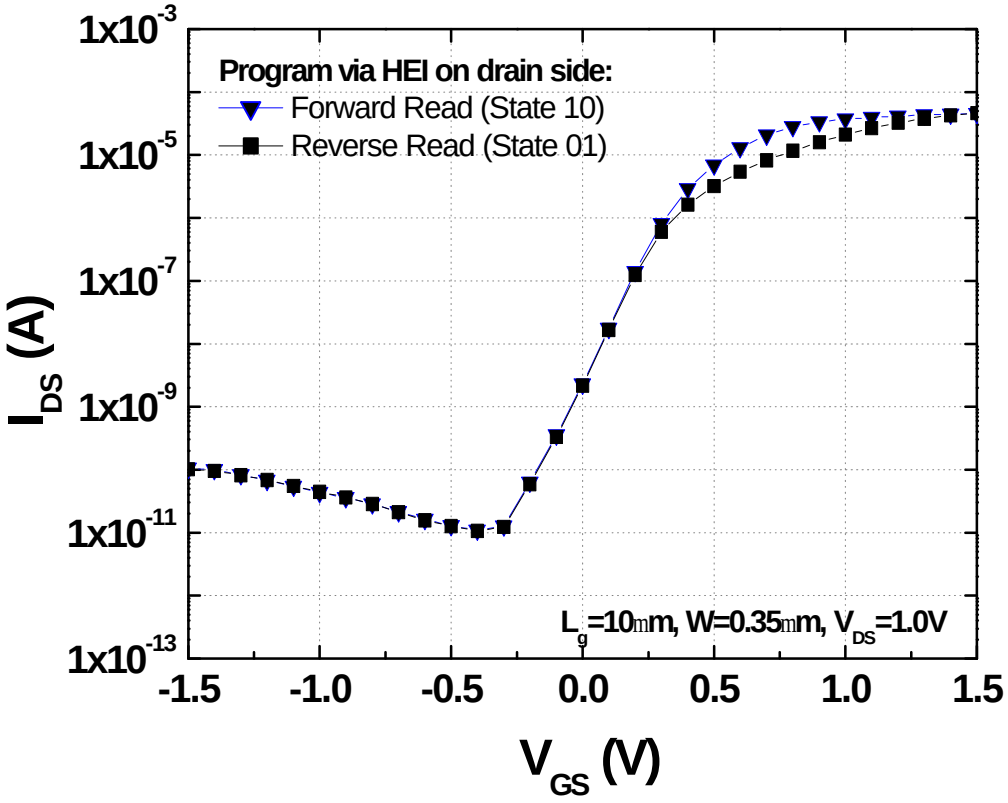


Figure 4.9: (a) Measured I_{DS}-V_{GS} characteristics of a dual-bit GSS n-channel FinFET NVM cell (n^+ Poly-Si gate) with gate-overlapped source/drain regions (L_g=10μm, W=1μm, V_{DS}=1.0V), showing a negligible change in V_T (and no change in GIDL current) when charge is stored on the drain-side bit via HEI (programming conditions: V_G = 7V, V_{DS} = 7V, 10ms).

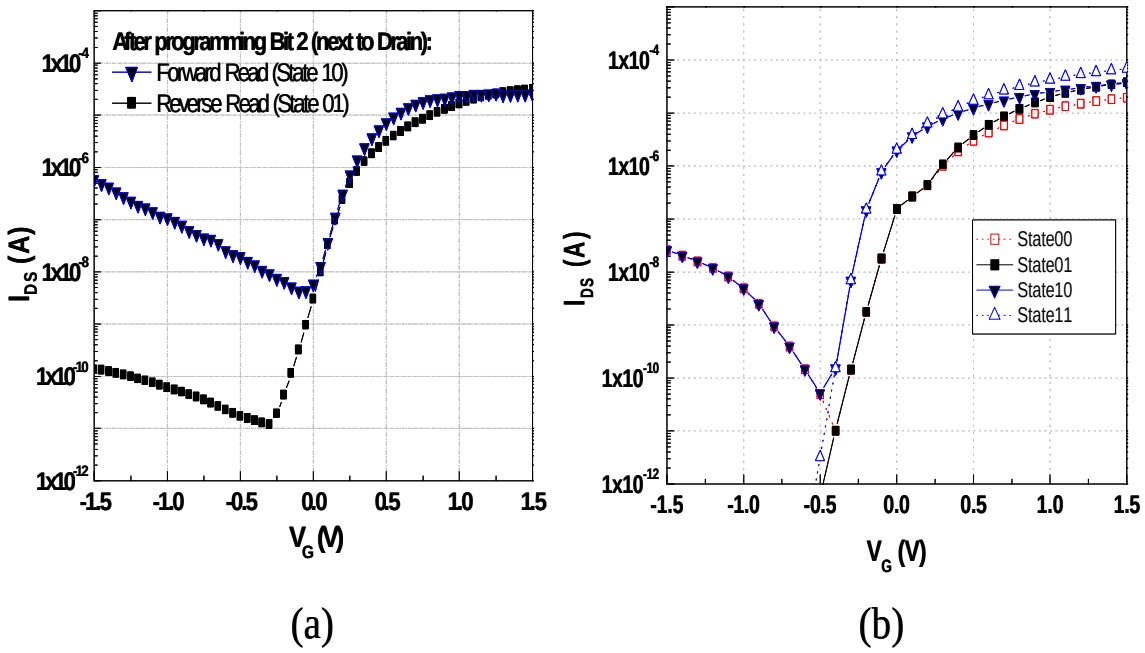


Figure 4.10: (a) Measured I_{DS}-V_{GS} characteristics of a dual-bit GSS n-channel FinFET NVM cell (n^+ Poly-Si gate) with gate-overlapped source/drain regions (L_g=10μm, W=1μm, V_{DS}=1.0V), showing a marked increase in GIDL current when charge is stored on the drain-side bit (programming conditions: V_G = 7V, V_{SD} = -3V, 100ms). (b) Simulated I_{DS}-V_{GS} characteristics for each state of a cell with similar dimensions.

4.5 Conclusions

An n-channel FinFET-based dual-bit NVM cell with gate-sidewall charge-storage sites is demonstrated for the first time. This multi-bit cell design is more scalable and thus offers improved storage density as compared with a more conventional FinFET-based SONOS cell design. The new read method may be utilized in conjunction with the conventional read method to fully determine the state of a NVM cell with gate-underlapped S/D regions using forward read operations. The new read method may also be utilized independently to fully determine the state of a NVM cell with gate-overlapped S/D regions, for which the conventional read method cannot be used.

4.6 References

[4.1] R. Liu, *et al.*, "Memory Technologies for 45nm and Beyond", *2006 International Electron Device Meeting Short Course*, (2006).

[4.2] M. Fukuda, T. Nakanishi, and Y. Nara, "New nonvolatile memory with charge-trapping sidewall," *IEEE Electron Device Letters*, Vol. 24, No. 7, pp. 490-492 (2003).

[4.3] A. Padilla, K. Shin, T.-J. King Liu, J. W. Hyun, I. Yoo, and Y. Park, "Dual-Bit Gate-Sidewall Storage FinFET NVM and New Method of Charge Detection," *IEEE Electron Device Letters*, Vol. 28, No. 6, pp. 502-505 (2007).

[4.4] Taurus Process & Device User Manual, 2003 (Synopsys, Inc.).

[4.5] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T-J. King, J. Bokor, C. Hu, M-R. Lin and D. Kyser, "FinFET scaling to 10nm gate length," *IEDM Technical Digest*, p. 251 (2002).

[4.6] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling Theory for Double-Gate SOI MOSFET's," *IEEE Transactions on Electron Devices*, Vol. 40, No. 12, pp. 2326 (1993).

[4.7] Y. Kamigaki, S. Minami, and H. Kato, "A new portrayal of electron and hole traps in amorphous silicon nitride," *Journal of Applied Physics*, Vol. 68, No. 5, pp. 2211-2215, 1990.

[4.8] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Letters*, Vol. 8, No. 11, pp. 515-517, 1987.

[4.9] S. Parke, J. Moon, P. Nee, J. Huang, C. Hu, and P. K. Ko, “Gate-induced drain leakage in LDD and fully-overlapped LDD MOSFETs,” *VLSI Technical Symp.,* pp. 49-50, 1991.

[4.10] Y.-K. Choi, T.-J. King, and C. Hu, “Nanoscale CMOS spacer FinFET for the terabit era,”, *IEEE Transactions on Electron Devices*, Vol. 23, No. 1, pp. 25-27, 2002.

[4.11] Y.-K. Choi, L. Chang, P. Ranade, J.-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T.-J. King, and J. Bokor, “FinFET process refinements for improved mobility and gate work function engineering,” *Electron Devices Meeting Technical Digest,* pp. 259-262, 2002.

[4.12] T. Ohnakado, K. Mitsunaga, M. Nunoshita, H. Onoda, K. Sakakibara, N. Tsuji, N. Ajika, M. Hatanaka and H. Miyoshi, “Novel Electron Injection Method Using Band-to-Band Tunneling Induced Hot Electron (BBHE) for Flash Memory with a P-channel Cell,” *Electron Devices Meeting Technical Digest,* pp. 279-282, 1995.

Chapter 5: Design of Dual-bit, Gate-Sidewall Storage p-channel FinFET NVM Cells

5.1 Motivations

FinFET SONOS (silicon-oxide-nitride-oxide-silicon) non-volatile memory (NVM) cells have been investigated recently to enhance the scalability of conventional flash NVM cells [5.1]-[5.4]. This is because the FinFET structure [5.5] is highly scalable [5.6], and SONOS cells avoid coupling interference between cells [5.7] and allow for a thinner tunnel dielectric, hence more aggressive gate-stack equivalent oxide thickness (EOT) scaling, and thus are preferred for future high-density flash memory technologies [5.8]. However, the erase saturation problem (affecting n-channel SONOS-type NVM cells) and slow Fowler-Nordheim (FN) tunneling programming speeds present serious challenges for implementation within a NAND-type array architecture [5.9].

To further enhance storage density, multiple bits can be stored within a single cell. The multi-bit storage scheme based on charge trapping in different regions of a single charge-trapping layer (*e.g.* used in NROM™ [5.10] and MirrorBit™ technologies) is difficult to scale to sub-50nm L_g since the cell's threshold voltage (V_T) can be affected by charge stored next to the drain electrode due to short-channel effects (SCE). To achieve stable multi-bit storage in the nanoscale regime, bit-to-bit interference can be

avoided by employing physically separated charge-trap sites, e.g. located along the gate sidewalls [5.11].

In this chapter, p-channel dual-bit gate-sidewall-storage (GSS) SOI FinFET NVM cell designs are investigated. The first p-channel GSS FinFET memory cells have been successfully fabricated, and are shown to be compatible with alternative program and erase methods (including the band-to-band tunneling induced hot electron injection (BBHE) programming method [5.12] and the band-to-band tunneling induced hole injection (BBHI) erase method), for potential use within a NAND-type array architecture.

5.2 Dual-bit GSS Cell Designs and Fabrication

The cell designs (**Figure 5.1**) studied in this work each utilize the vertical double-gate FinFET structure with charge-trapping layers (silion-rich nitride, "SiRN") located along the sidewalls of each gate electrode. In principle, these designs are more scalable and can achieve higher layout density than a conventional FinFET SONOS NVM cell because they do not utilize a thick gate dielectric stack. Furthermore, since the charge storage sites are physically separated, the complementary bit disturb (CBD) effect, normally seen in multi-bit SONOS NVM cells, should be mitigated.

GSS FinFET NVM cells were fabricated using the process flow described in the previous chapter (and in [5.13]) and illustrated in **Figure 5.2**. The starting silicon-on-insulator (SOI) layer is essentially undoped, and the gate dielectric stack consists of a thermally grown SiO_2 film (6nm) and ~100nm of *in-situ* doped p-type poly-$Si_{0.8}Ge_{0.2}$ (**Figure 5.2(a)**) or n-type poly-Si (**Figure 5.2(b)**). After gate stack formation, thermal

oxidation was used to simultaneously grow the tunnel oxide (~3nm) and control oxide. Polycrystalline silicon oxidizes more quickly than crystalline silicon, so that the control oxide thickness along the gate sidewalls is much thicker than the tunnel oxide. SiRN was then deposited (~100nm) and anisotropically etched back to form gate-sidewall spacers.

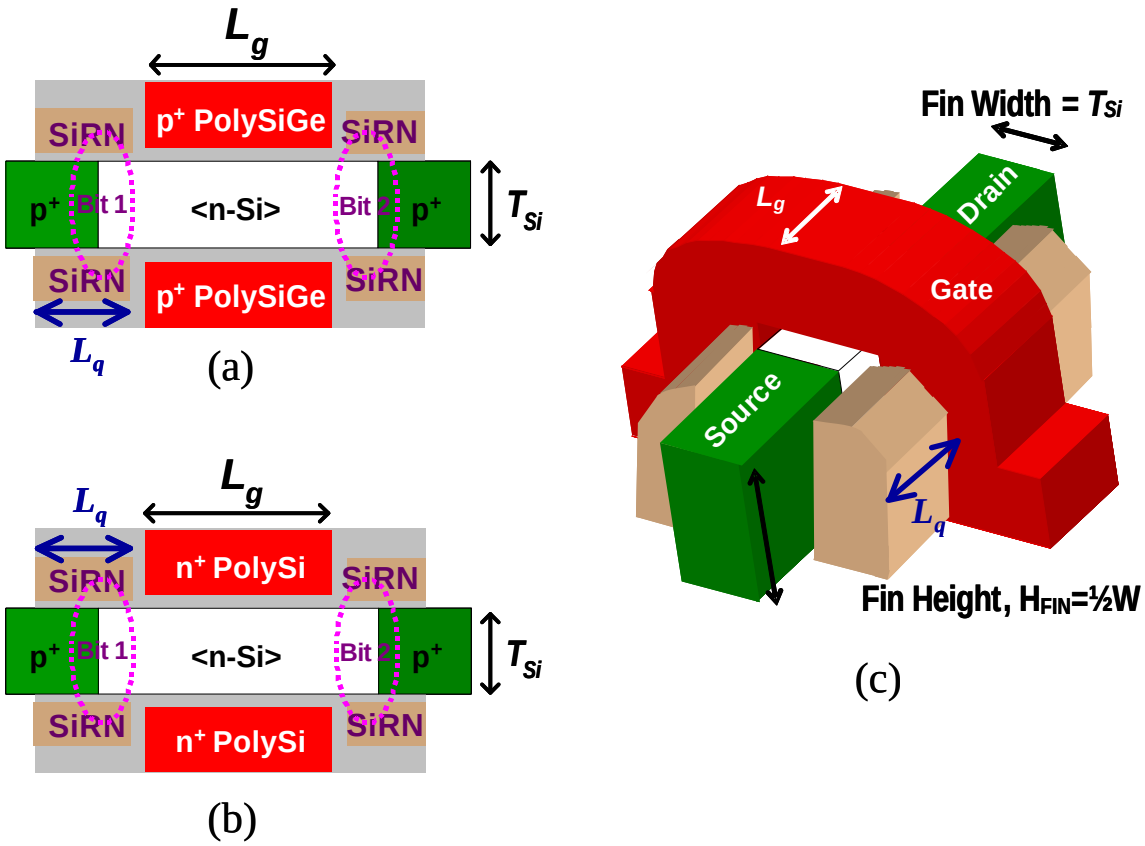
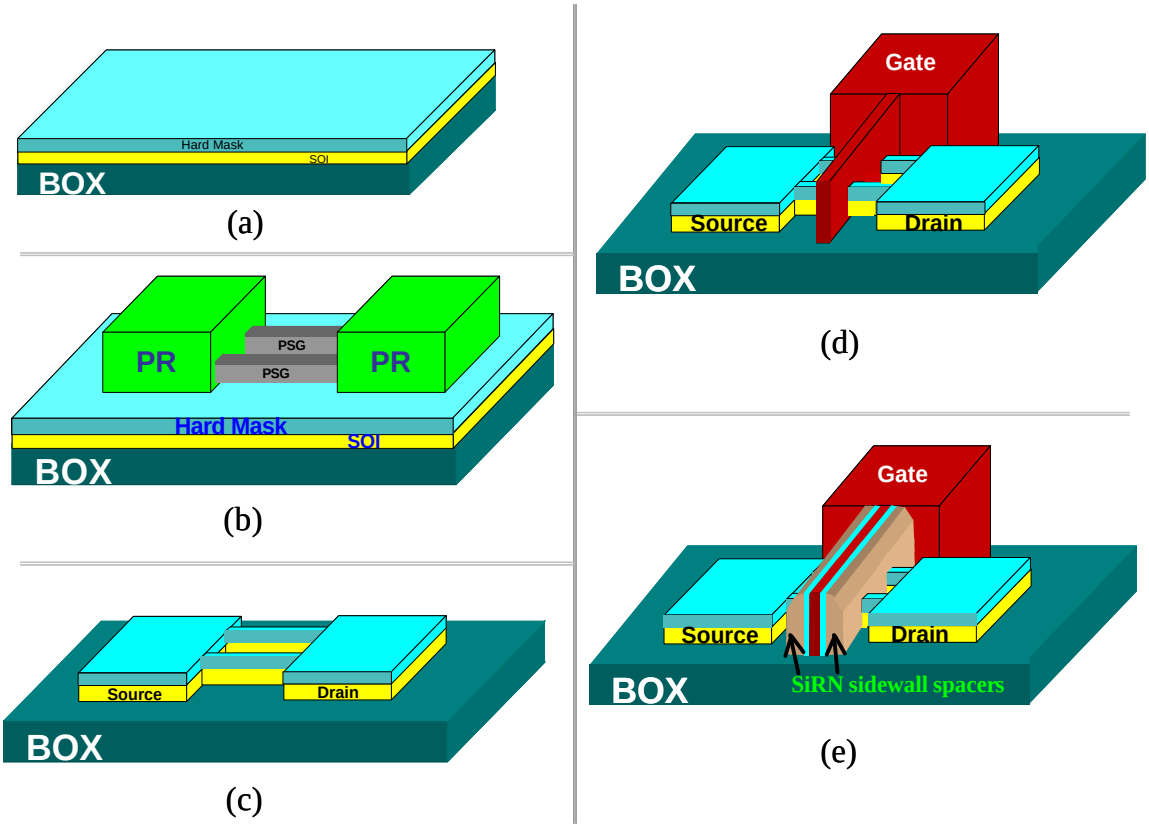


Figure 5.1: (a-b) Schematics of the two p-channel NVM cell designs studied in this chapter. (c) Isometric view of a dual-bit GSS FinFET NVM cell [5.13].



Process Flow
1)Active Area Formation (figures (a) through (c)): (a) Fin formation (PSG spacer lithography), T_{si}~75nm. (b) Optical lithography (define S/D pads) (c) Dry Etching
2) Gate Stack Formation (figure (d): (a) Gate oxidation (~6nm) (b) In-situ doped n^+ PolySi / p^+ PolySiGe CVD (~100nm) (c) LTO CVD (~150nm) (d) Gate lithography & dry etching, various gate-lengths (L_g)
3) Spacer Formation and Final Steps: (a) Thermal oxidation (~3nm in single-crystal silicon) (b) SiRN CVD (~100nm) and spacer dry etch (c) p+ S/D implantation (d) RTA Activation (~52sec @ 900 °C) and FGA

Figure 5.2: Process flow used to fabricate dual-bit GSS FinFET NVM cells.

Figure 5.3 shows a scanning electron micrograph (SEM) of a GSS FinFET NVM device after nitride spacer definition, and cross-sectional transmission electron micrograph (TEM) images of the oxidized gate film, showing that the thickness of the control oxide layer is indeed thicker than the tunnel oxide in each case (~6nm in the p^+ poly-$Si_{0.8}Ge_{0.2}$ film, and ~14nm in the n^+ poly-Si film).

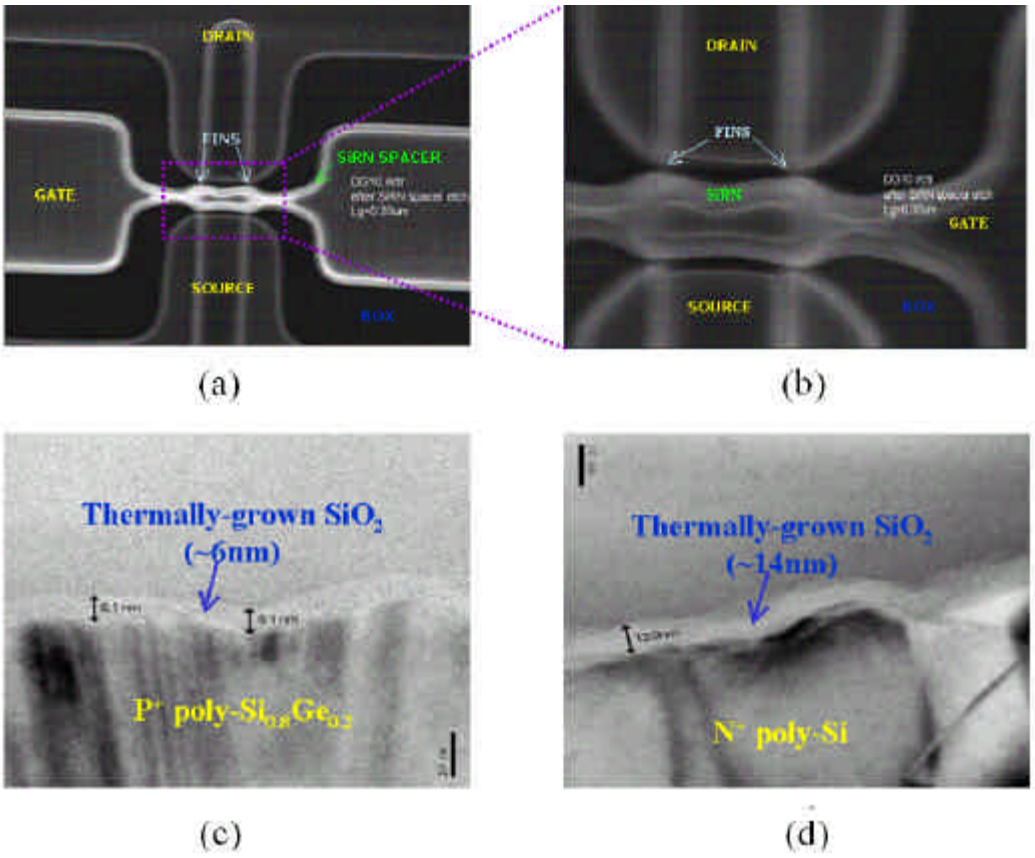


Figure 5.3: (a-b) SEM view of a p-channel GSS FinFET NVM cell after silicon-rich nitride (SiRN) spacer definition. The cross-sectional TEM images in (c-d) show the thickness of the control oxide grown along the sidewalls of the in-situ doped (c) P^+ poly-$Si_{0.8}Ge_{0.2}$ and (d) N^+ poly-Si gate films (respectively).

Incremental rapid-thermal annealing (RTA) at 900°C was used to activate and diffuse the implanted source/drain (S/D) dopants, until gate-induced drain leakage (GIDL) current at moderate drain-to-source voltage was seen, to achieve a slightly gate-underlapped S/D structure. (This is necessary to ensure optimum programming and read operation of GSS structures [5.11][5.13]). A total annealing time of ~52 seconds was required for the p-channel devices (**Figure 5.4**). Finally, a sintering step was performed (90m at 450°C in forming gas) to improve Si/SiO_2 interface properties.

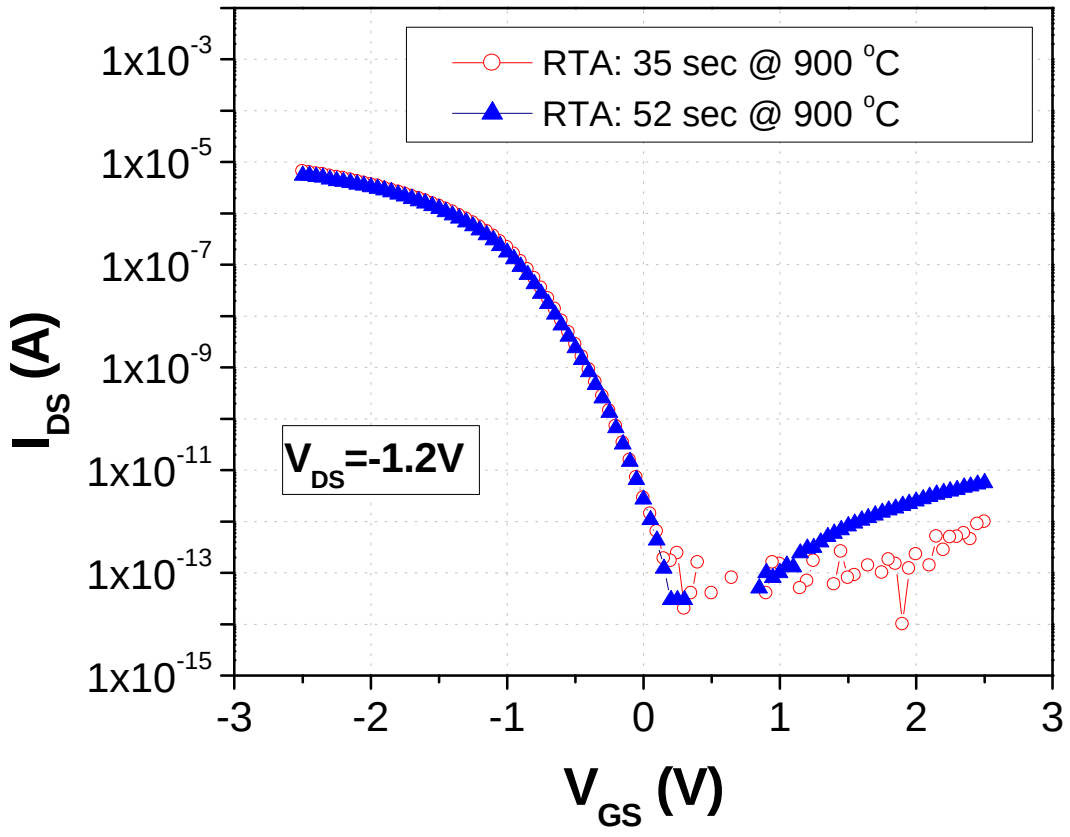


Figure 5.4: Effect of RTA conditions on the I_{DS}-V_{GS} characteristics of the p-channel GSS FinFET NVM cell design shown in Figure 5.1(a). (L_g=0.5µm, W=0.1µm, V_{DS}=-1.2V).

5.3 Device Operation and Characterization

Each bit of the p-channel GSS FinFET NVM cell can be independently programmed and read (via the BBHE and current-sensing methods, respectively) in a manner similar to that reported in [5.14] for the dual-bit, single-gate bandgap-engineered (BE) SONOS NVM cell. For instance, to program Bit 2, a negative voltage (i.e., V_D=-6V) is applied to the drain and a positive voltage (i.e., V_G=6V) is applied to the gate (**Figure 5.5**). This allows (1) band-to-band tunneling of electrons in the drain, which subsequently (2) tunnel towards the charge-trapping site close to the drain junction due to the large transverse field that is applied there.

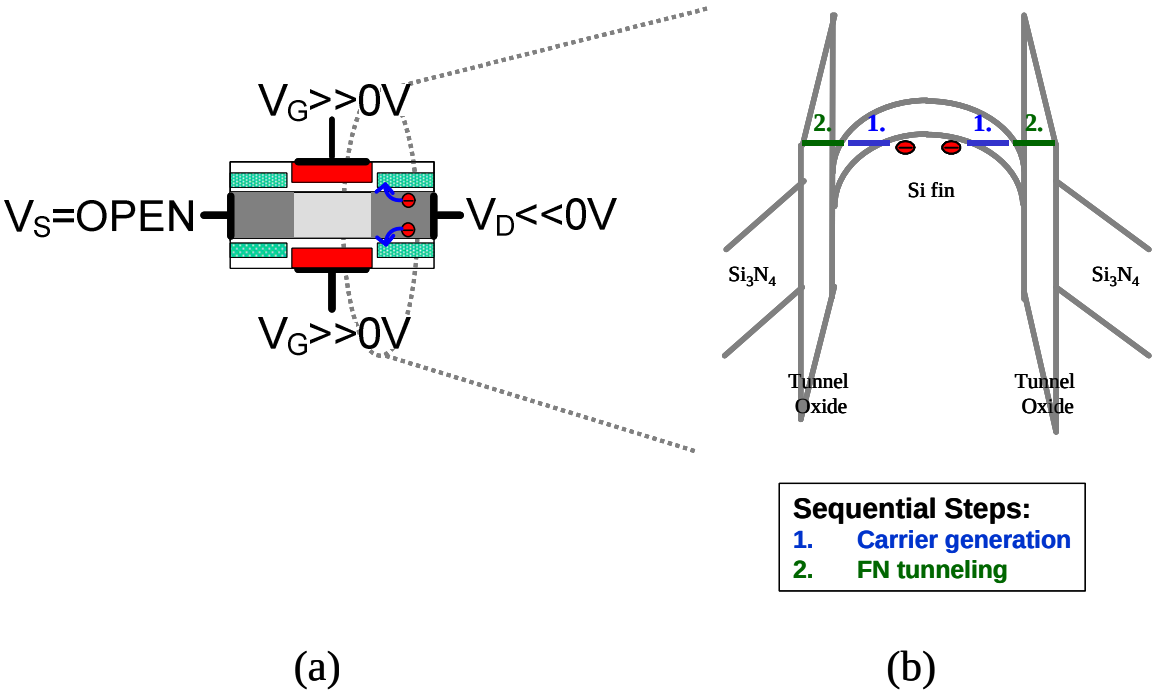
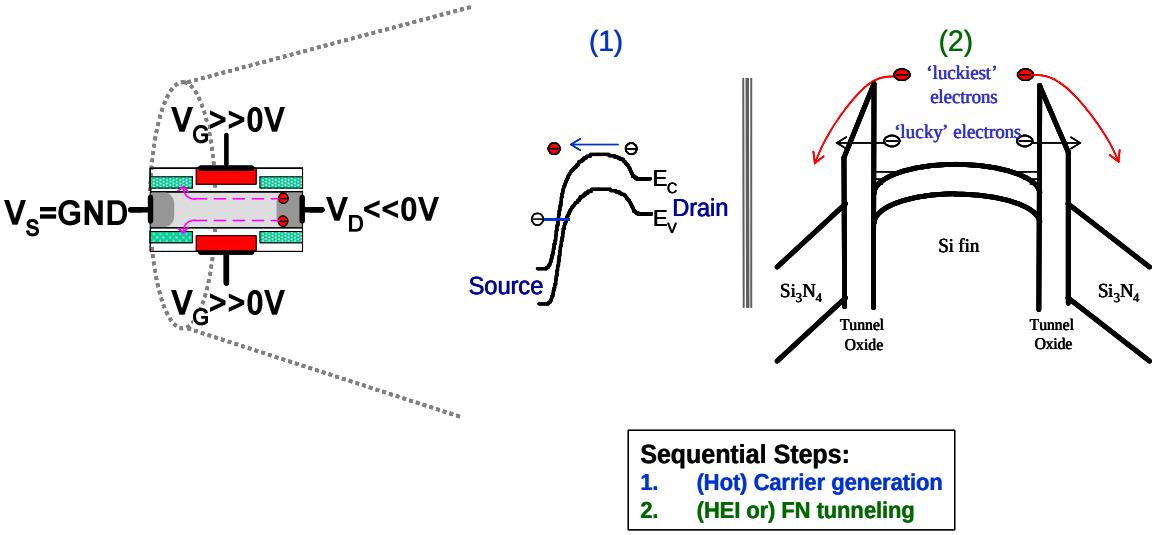


Figure 5.5: The band-to-band Hot Electron Injection (BBHE) programming mechanism [5.12],[5.14]: (a) Bias conditions. (b) Energy band diagrams (used to describe the mechanism).

To determine the (un)charged state of the cell, the transistor on-state current can be used to determine the charge-storage state of the bit near to the source. For this conventional reverse-read method, however, the source junction must be located *underneath* the charge-storage site to maximize the resulting V_T separation (ΔV_T) with charge storage at the source-side bit, *i.e.* a gate-underlapped S/D structure is required [5.11],[5.13].

Even though the BBHE programming method is more efficient (in terms of power consumption) than the HEI programming mechanism [5.12] for an n-channel GSS NVM device, it suffers from a serious programming disturbance (i.e., the injection of electrons onto on the complementary bit), especially when large programming voltages are used. As shown in the energy band diagrams in **Figure 5.6**, some electrons (either electrons that make up both inversion channels or additional electrons that are generated via band-to-band tunneling) will *not* be injected onto Bit2 and will drift towards the source electrode. A portion of these electrons eventually gets trapped (either via HEI or FN tunneling) within the charge-trapping site next to the Source electrode (thereby programming that bit). To mitigate (the effects of) this disturbance, the source electrode could be kept electrically floating when the BBHE is used to program the bit next to the Drain electrode; however, this disturbance is still present even with these settings [5.14].



(a) (b)

Figure 5.6: Programming disturbance of the BBHE programming method: (a) Bias conditions. (b) Energy band diagrams used to describe the injection of electrons in the complementary bit (next to the Source electrode).

Figure 5.7(a) shows measured I_{DS}-V_{GS} characteristics (in forward and reverse modes of operation) for a p-channel GSS FinFET NVM cell (with p^+ PolySiGe gates) before and after Bit2 (next to the drain electrode) was programmed via BBHE. A shift in the reverse-read V_T (due to charge storage on Bit2) is clearly seen; however, there is also an increase in forward-read V_T, due to either enhanced drain-induced barrier lowering (DIBL) and/or unintentional programming of the complementary bit (as previously discussed). Despite this disturbance though, the V_T window in this cell can be improved significantly by increasing $|V_{DS}|$ during the read operation (**Figure 5.7(b)**), as was noted in [5.14].

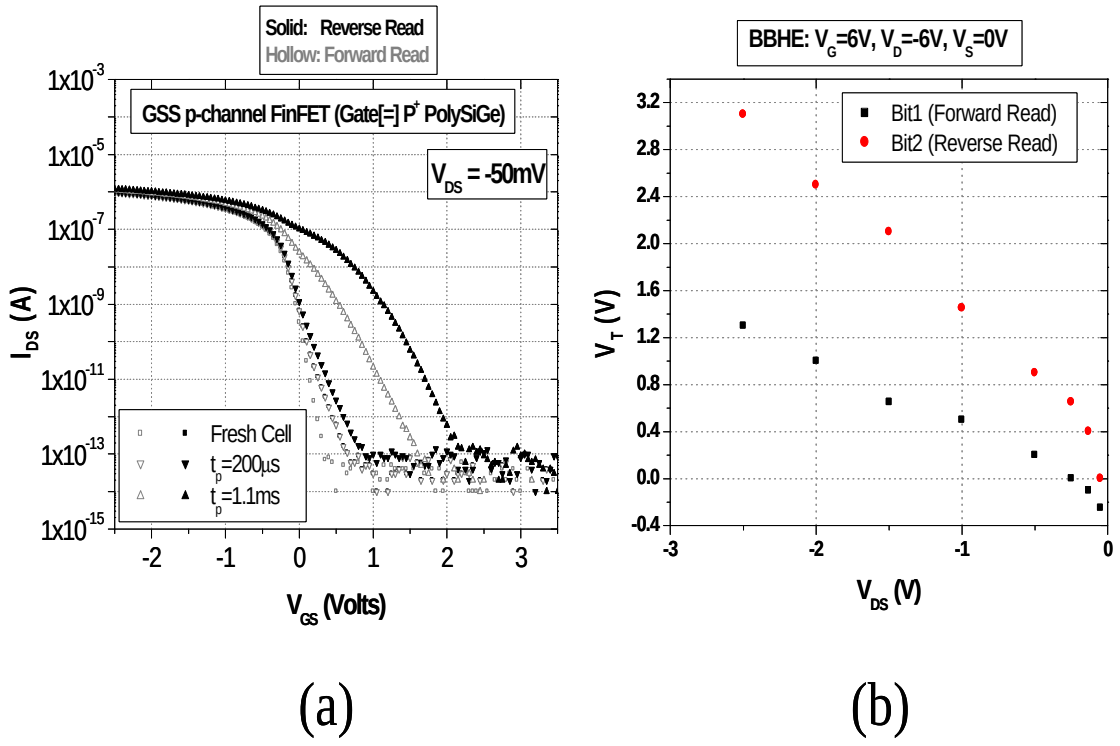


Figure 5.7: Measured I_{DS}-V_{GS} curves of a p-channel GSS FinFET NVM cell (Figure 5.1a) before and after various BBHE (V_{GS}=6V, V_{DS}=-6V) programming pulses (t_p) were applied to the structure (Lg=0.5µm, W=100nm, V_{DS}=-50mV). (b) V_T of both bits as a function of V_{DS}.

Figure 5.8 shows the change in V_T *vs.* Bit2 BBHE programming time for the GSS FinFET under study. For long programming times, significant programming disturbance is seen on the complementary bit in both cases. Thus, during BBHE programming it is preferable to float the source (although this programming disturbance has also been observed previously with these settings in a BE-SONOS NVM cell [5.14]). However (as already mentioned), the V_T window can be enhanced significantly by increasing $|V_{DS}|$ during the read operation (as demonstrated by the results shown in the figure).

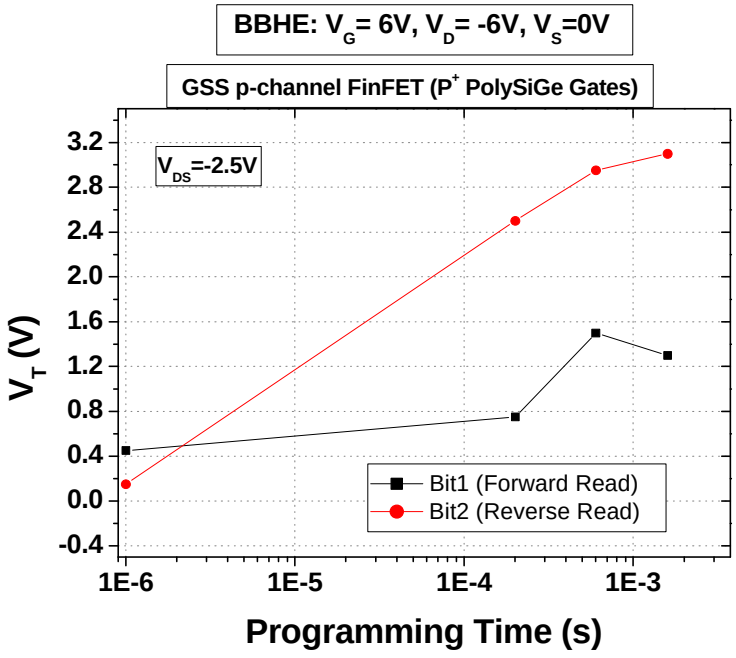


Figure 5.8: Measured BBHE programming characteristics (V_{DS}=-2.5V) of a p-channel GSS FinFET NVM cell with p+ PolySi gates. (Lg=0.5µm, W=100nm).

Figure 5.9 shows the change in V_T *vs.* BBHI erase time. As shown, both bits of the p-channel GSS FinFET cell can be erased via the BBHI method by applying a large negative voltage to the gates (-6V) and a large positive voltage to the drain (6V). These settings allow (1) band-to-band tunneling of electrons and holes, and the holes subsequently (2) tunnel towards both charge-trapping sites (and mainly to the site closest to the drain junction) due to the large transverse field that is applied to the structure. These results demonstrate that FN tunneling can be used to either program or erase this structure, which makes it feasible for use with a NAND-type array architecture.

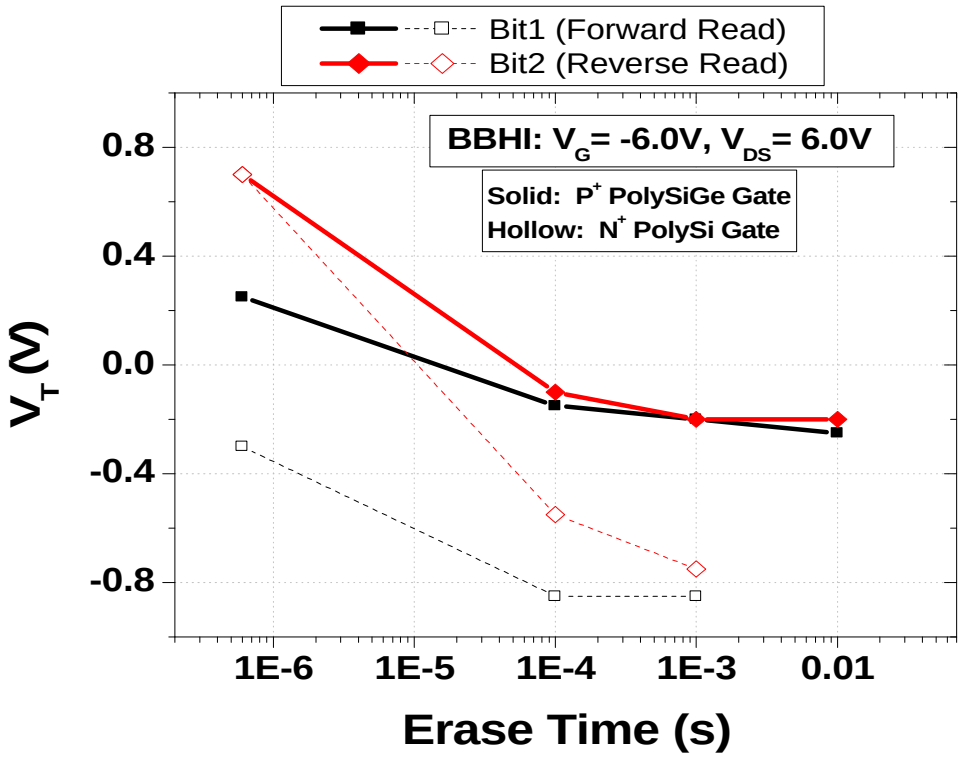


Figure 5.9: Band-to-band Tunneling-induced Hole Injection (BBHI) Erase characteristics of a p-channel GSS FinFET NVM cell.

5.4 Summary

P-channel FinFET NVM devices with gate-sidewall charge storage are demonstrated for the first time. These dual-bit cell designs are in principle more scalable and thus offer improved packing density as compared with a more conventional SONOS design. The BBHE mechanism can be used to program each bit of these structures (though partial programming of the complementary bit occurs with this method, especially when large programming voltages are used). The BBHI method can be used to erase both bits of these structures. The use of these (FN tunneling) methods should enable implementation of this cell design within a NAND-type array architecture.

5.5 References

[5.1] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor, and T.-J. King, "FinFET SONOS Flash Memory for Embedded Applications", *IEDM Technical Digest*, p. 609 (2003).

[5.2] M. Specht, R. Kommling, F. Hofmann, V. Klandzievski, L. Dreeskornfeld, W. Weber, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rosner, M. Stadele, R.J. Luyken, H. Reisinger, A. Graham, E. Hartmann, and L. Risch, "Novel Dual Bit Tri-Gate Charge Trapping Memory Devices", *IEEE Electron Device Letters*, Vol. 25, p. 810 (2004).

[5.3] C.W. Oh, S. D. Suk, Y.K. Lee, S.K. Sung, J.-D. Choe, S.-Y. Lee, D.U. Choi, K.H. Yeo, M.S. Kim, S.-M. Kim, M. Li, S.H. Kim, E.-J. Yoon, D.-W. Kim, D. Park, K. Kim, and B.-I. Ryu, "Damascene Gate FinFET SONOS Memory Implemented on Bulk Silicon Wafer", *IEDM Technical Digest*, p. 893 (2004).

[5.4] C. Friederich, M. Specht, T.Lutz, F. Hofmann, L. Dreeskornfeld, W. Weber, J. Kretz, T. Melde, W. Rosner, E. Landgraf, J. Hartwich, M. Stadele, L. Risch, and D. Richter, " Multi-level p^+ tri-gate SONOS NAND string arrays," *International Electron Devices Meeting Technical Digest*, p. 1-4 (2006).

[5.5] N. Lindert, L. Chang, Y.-K. Choi, E.H. Anderson, W.-C. Lee, T.-J. King, J. Bokor, and C. Hu, "Sub-60-nm quasi-planar FinFETs fabricated using a simplified process," *IEEE Electron Device Letters*, Vol. 22, No. 10, pp. 487-489 (2001).

[5.6] C. Hu, *VLSI Technology Symp. Dig*. 4 (2004).

[5.7] J.-D. Lee, S.-H. Hur, and J.-D. Choi, “Effects of Floating-gate Interference on NAND Flash Memory Cell Operation”, *IEEE Electron Device Letters*, p. 264 (2002).

[5.8] K. Kim, *IEDM Tech. Dig*., Paper 13.5 (2005).

[5.9] T. Mikolajic, M. Specht, N. Nagel, T. Mueller, S. Riedel, F. Beug, T. Melde, and K.-H. Kusters, “The Future of Charge Trapping Memories,” *International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pp. 112-115 (2007).

[5.10] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, “NROM: A novel localized trapping, 2-bit nonvolatile memory cell,” *IEEE Electron Device Letters*, Vol. 21, No. 11, pp. 543-545 (2000).

[5.11] M. Fukuda, T. Nakanishi, and Y. Nara, “New nonvolatile memory with charge-trapping sidewall,” *IEEE Electron Device Letters*, Vol. 24, No. 7, pp. 490-492 (2003).

[5.12] T. Ohnakado, K. Mitsunaga, M. Nunoshita, H. Onoda, K. Sakakibara, N. Tsuji, N. Ajika, M. Hatanaka, and H. Miyoshi, “Novel Electron Injection Method Using Band-to-Band Tunneling Induced Hot Electron (BBHE) for Flash Memory with a P-channel Cell,” *International Electron Devices Meeting Technical Digest*, pp. 279-282 (1995).

[5.13] A. Padilla, K. Shin, T.-J. King Liu, J. W. Hyun, I. Yoo, and Y. Park, “Dual-Bit Gate-Sidewall Storage FinFET NVM and New Method of Charge Detection,” *IEEE Electron Device Letters*, Vol. 28, No. 6, pp. 502-505 (2007).

[5.14] H.-T. Lue, S.-Y. Wang, E.-K. Lai, M.-T. Wu, L.-W. Yang, K.-C. Chen, J. Ku, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "A Novel P-Channel NAND-Type Flash Memory with 2-bit/cell Operation and High Programming Throughput (>20 MB/sec)," *International Electron Devices Meeting Technical Digest*, pp. 331-334 (2005).

Chapter 6: Design of 4bit Double-Gated NVM Cells

6.1 Background

In this chapter, the design of NVM cells that are able to store 4 bits is discussed. As already mentioned in previous chapters, enhancement in NVM density has traditionally been achieved through the use of multi-level cell (MLC) or multi-bit cell (MBC) designs, coupled with device scaling. In the MLC design, various levels of charge storage are utilized to attain distinct threshold voltage (V_T) levels and thus distinct binary states within every cell of the array (**Figures 6.1(a)**, **6.1(b)**). In this approach, each V_T level corresponds to *one* binary state; consequently, the required number of distinct V_T levels increases substantially as the number of bits stored within every cell increases (for example, 16 distinct V_T levels are required to store 4 bits on every cell), hence the implementation of a 4-bit-per-cell MLC design is very difficult. On the other hand, the MBC design (**Figures 6.1(c)**, **6.1(d)**) utilizes the cell's symmetry (with respect to the Source and Drain electrodes) to attain multiple bits of storage in every cell. This symmetry allows treatment of each Bit-Line (BL) within a (NOR-type) memory array layout as either a Source or Drain electrode in order to access or modify the data stored at either side ("A" or "B") of the cell. However, since the conventional single-gate NVM

cell (shown in **figure 6.1(c)**) contains only one axis of symmetry (shown by the dashed line), then only 2 bits can be stored on this cell with this approach.

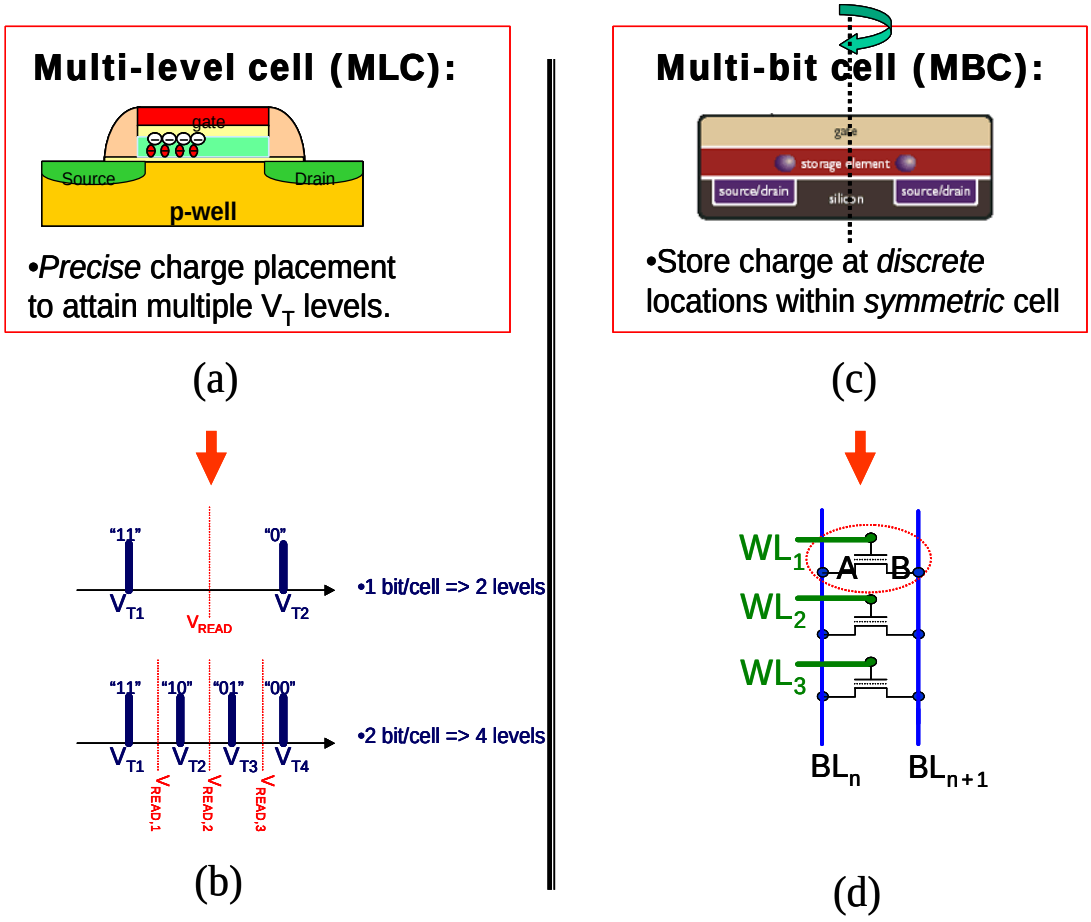


Figure 6.1: In the MLC design (a), different V_T levels, one for each binary state (b), are defined to store multiple bits in every cell. The MBC design (c) utilizes the cell's symmetry to access the information stored in either side ("A" or "B") of the cell (d).

Alternatively, both (MLC, MBC) charge storage schemes can be simultaneously utilized to increase the number of bits stored within every cell. Saifun adopted this approach when it introduced its 4-bit NROM™ NVM cell (**figure 6.2(b)**) [6.4]. In a similar approach as that used for its 2bit NROM™ cell (**figure 6.2(a)**) [6.3], this 4-bit

structure also utilizes localized charge trapping (within the nitride layer of an oxide-nitride-oxide, or "*ONO*", gate-stack) and its symmetry (with respect to the Source and Drain electrodes) to store 2 bits on each side of the structure. The structure utilizes the MLC algorithm, which requires 4 distinct V_T levels (**figure 6.1(b)**), to store 2 bits on *each* side. Thus, since each side of the structure is used to store 2 bits, then the entire cell can store 4 bits within it.

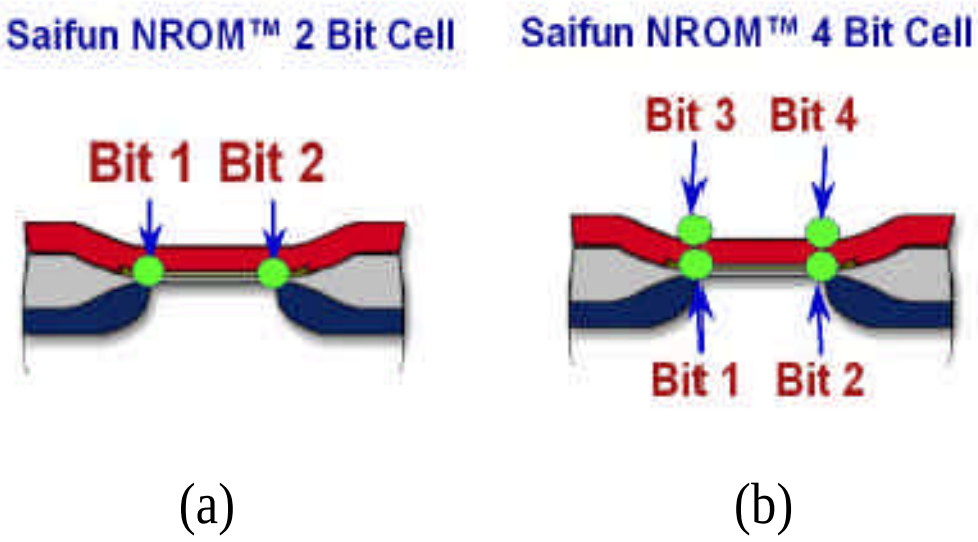


Figure 6.2: 2D Schematic view of Saifun's (a) 2-bit, and (b) 4-bit NROM™ NVM cells [6.4].

A serious challenge with this approach deals with reliability. For each charge-storage node, 4 V_T levels must be placed within a very narrow V_T window, and these 4 distinct V_T levels must have enough separation from each other to properly identify each binary state. This is difficult to achieve in practice since the Complementary Bit Disturb (CBD) issue becomes more pronounced with this cell design since more charge must be stored (to achieve 4 distinct V_T levels) on the complementary charge-storage node. Additionally, since more charge must be stored at each charge-storage node, the

charge retention of the cell is detrimentally affected since a larger amount of charge stored on each node enhances the transverse electric field around each node (even when no voltages are applied to the structure) and thus the probability of charge leakage. This is indeed a serious reliability concern since this cell utilizes the MLC algorithm, which requires very strict control of the various V_T levels for optimum operation. For these reasons, this cell design has very limited scalability and, consequently, is not the optimum 4-bit cell design.

6.2 Motivations

An improved 4-bit cell design involves the use of a double-gate (DG), thin-body field-effect transistor (FET) structure as a NVM cell (**Figure 6.3**). The enhanced symmetry of this structure (with respect to the Source (S) and Drain (D) electrodes *and* both gate electrodes, which have the same work function) permits the operation of this structure as either a 4-bit MBC design (with 4 distinct charge storage sites) or a 4-bit MLC design (with 2 distinct charge storage sites, each containing 4 different V_T levels). In the former case, the enhanced symmetry of the DG-FET structure is utilized to independently operate each charge storage site via the conventional methods (i.e. selective read, program, and erase of each bit via the reverse-read, hot electron injection, and hot hole injection methods, respectively) normally utilized with a Virtual Ground NOR-type architecture [6.2][6.2]. In the latter case, the symmetry of the DG-FET structure with respect to both gate electrodes is utilized to store 4 different V_T levels (and thus 2 bits) within each charge storage site (which is embedded within each gate electrode). The independent operation of each gate electrode thus allows the

independent operation of each charge-trapping site via the conventional methods (i.e. selective read, program, and erase of each charge-trapping site via the current-sensing, FN tunneling electron injection, and FN tunneling hole injection methods, respectively) normally utilized within an NAND-type array architecture [6.2][6.2].

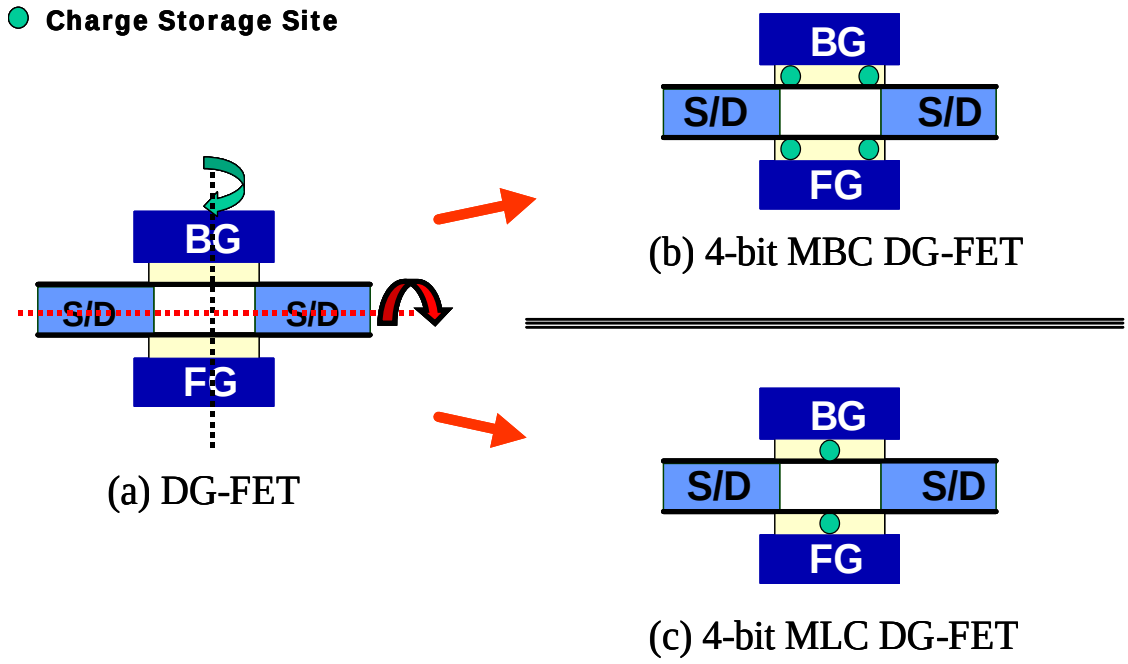


Figure 6.3: 4-bit DG-FET NVM Cell Designs: (a) The enhanced symmetry of a DG-FET structure permits its operation as either (b) a MBC DG-FET design (with 4 charge storage sites) or (c) a MLC DG-FET design (with 2 charge storage sites).

Figure 6.4 shows 2D schematic cross-sections of two 4-bit NVM cells that derive from a DG FET structure. As shown, these cells are comprised of an n-channel DG FET structure, modified to include the charge-trapping layers (*e.g.* made of silicon-nitride, poly-Si, or any other material that is able to store charge) embedded either underneath each gate electrode (**figure 6.4(a)**) or within the sidewalls of each gate

electrode (**figure 6.4(b)**). The former is the conventional SONOS-like NVM cell structure. The latter structure does not require a thick gate oxide layer and thus provides the added benefit of a much thinner EOT, which makes the structure even more scalable to very short gate lengths; however, it requires a rather thick control oxide layer on the sidewalls of each gate electrode to properly isolate the gates from the charge-trapping layers.

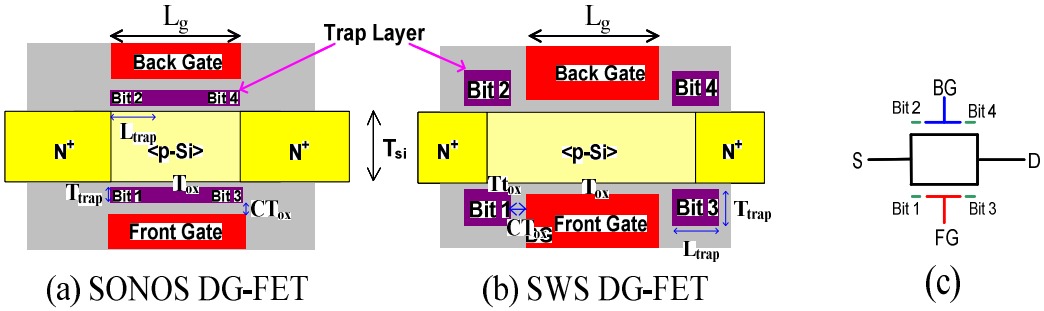


Figure 6.4: Schematic cross-sections of (a) the conventional SONOS, and (b) the Sidewall Storage (SWS) 4-bit DG-FET NVM cell designs. (c) Symbolic representation for either structure.

In practice, both DG-FET structures shown in **Figure 6.4** can be implemented with either the 3-dimensional (3D) SOI Multiple Independent Gate Field Effect Transistor (MIGFET) design [6.20], or the planar Back-Gated (BG) thin-body SOI MOSFET (BG-FET) design [6.5], as demonstrated previously with the 4-bit *Double SONOS Memory* (DSM) NVM cell design in [6.6][6.7] (**Figure 6.5**). The DSM cell, proposed by C. W. Oh *et al.*, consists of a BG-FET structure that has been modified to include an oxide-nitride-oxide ("ONO") gate stack next to each gate electrode. Due to the complex process flow of this structure, the EOT of the gate stack next to the BG

electrode is much thicker than that underneath the Front Gate (FG) electrode; consequently, the symmetry of the structure (with respect to both gate electrodes) is lost, and this results in different I_{DS}-V_{GS} characteristics for the front and back channels. This asymmetry introduces complexity in the peripheral circuitry (since different V_T levels must be utilized to distinguish the information stored at either gate). Furthermore, even though appropriate device functionality (i.e., selective read, program, and erase of each bit via the reverse-read, hot electron injection, and hot hole injection methods, respectively) was demonstrated with this structure (when operated either in MBC mode or MLC mode), significant coupling was observed between both gates (which introduced a significant gate-bias dependency on the I_{DS}-V_{GS} characteristics of both channels). In addition, this DSM cell cannot be aggressively scaled since the EOT of the BG is much thicker than that of its FG; also, any additional scaling on the EOT of either gate dielectric stack enhances the coupling between both gates, which is *not* desired in this case. Furthermore, the process flow used to fabricate this structure is quite complex (e.g., it is difficult to self-align both gates without introducing significant parasitic resistive and capacitive elements) and deviates significantly from traditional CMOS process flows.

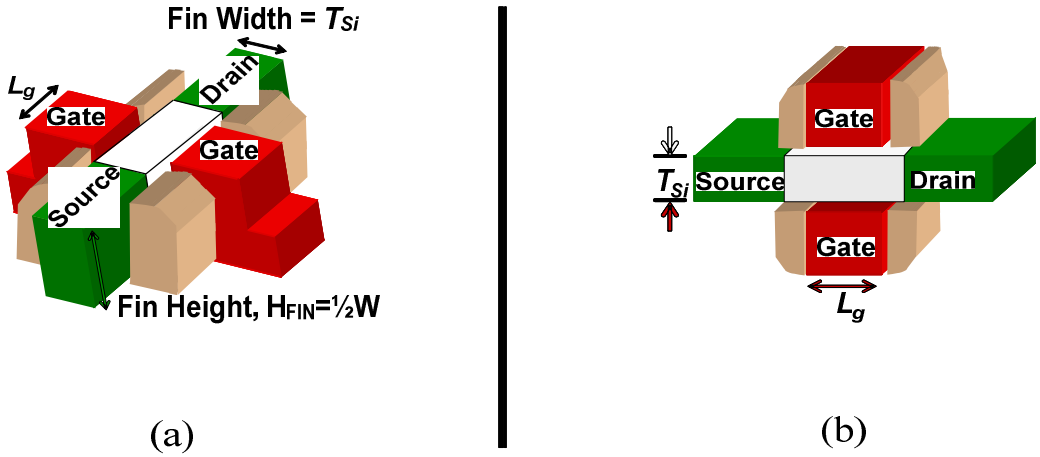


Figure 6.5: Independent Gate DG-FET designs: (a) 3D Multiple Independent Gate FET (MIGFET) design [6.20]. (b) Planar Back-Gated FET (BG-FET) design [6.5].

As an alternative, both DG-FETs shown in **Figure 6.4** can also be attained with a MIGFET structure (**Figure 6.5(a)**). The use of a DG thin-body transistor, such as the MIGFET structure, as a 4-bit NVM cell provides additional benefits including a simpler process flow, symmetric I_{DS}-V_{GS} curves at both channels (since both gate electrodes can be designed to have the same EOT), and enhanced scalability (particularly for the novel structure of **Figure 6.4(b)**), since a DG-FET can achieve good electrostatic integrity without the need of a thin gate-dielectric stack [6.9] and achieve the ideal subthreshold swing (60mV/decade at room temperature); therefore, this structure is the most scalable [6.10]. Nonetheless, the use of the MIGFET structure as a NVM cell requires additional layout area to place *both* gate electrodes on the side of the cell's active area, and for the isolation of adjacent word-lines connecting to both gates. As a result, the use of the MIGFET structure as a 4-bit NVM cell is *not* the desired choice (even though the NVM

cell is more scalable with this design) since its use will not ultimately reduce the size per bit of the unit cell (when this structure is placed within an array architecture).

In this chapter, design considerations regarding the use of either 4-bit DG-FET structure (shown in **Figure 6.4**) within both a NOR- and NAND-type array architecture are investigated. 2-dimensional (2D) device simulations are performed on both 4-bit DG-FET structures to both investigate their scalability and to demonstrate the possibility to distinguish the state of each bit via the conventional charge detection method. In addition, the operation of these structures within both array architectures is discussed (with particular emphasis on the selective read operation). Practical implementation of these structures within both architectures is also discussed in detail.

6.3 4-bit DG Cell Designs and Operation

6.3.1 Operating Principles

The operation of these novel 4-bit DG-FET cell designs is based on two concepts: 1) biaxial symmetry, and 2) independent-gate operation. As shown in **Figure 6.6(a)**, the DG-FET structure is symmetric with respect to the Source and Drain electrodes, *and* both gates. Thus, bit 3 (bit 2) becomes bit 1 (bit 4) when the source (S) and drain (D) electrodes are interchanged; similarly, bit 1 (bit 3) becomes bit 2 (bit 4) upon interchange of the Front (FG) and Back Gates (BG). The resulting symmetry of these structures allows for the treatment of each bit-line (word-line) in an array layout as either a source or drain (FG or BG) electrode, depending on the location of the bit that needs to be read, programmed or erased. Independent biasing of each gate, specifically the FG and the

BG, as seen in **Figure 6.6(b)**, can effectively modulate the threshold voltage (V_T) and sub-threshold swing (S) of a MIGFET [6.14]. This feature can thus be used to "mask" the bit information stored at the opposing gate while bit information stored at the selected gate is determined. For instance, the bit information stored at the BG can be "masked" while reading the bit information stored in the FG by placing the channel next to the BG in accumulation mode and the channel next to the FG in inversion mode (e.g., by applying $V_{BG}<0$, $V_{FG}>0$ for an n-channel FET).

As already mentioned, the biaxial symmetry of these 4-bit cell designs (with respect to the source (S) and drain (D) electrodes and both gate electrodes, which have the same work function) allows for each bit to be independently read, programmed, and erased in a manner similar to that reported in a manner similar to that reported in [6.3][6.11] for a single-gate SONOS or sidewall-charge-storage NVM cell, respectively.

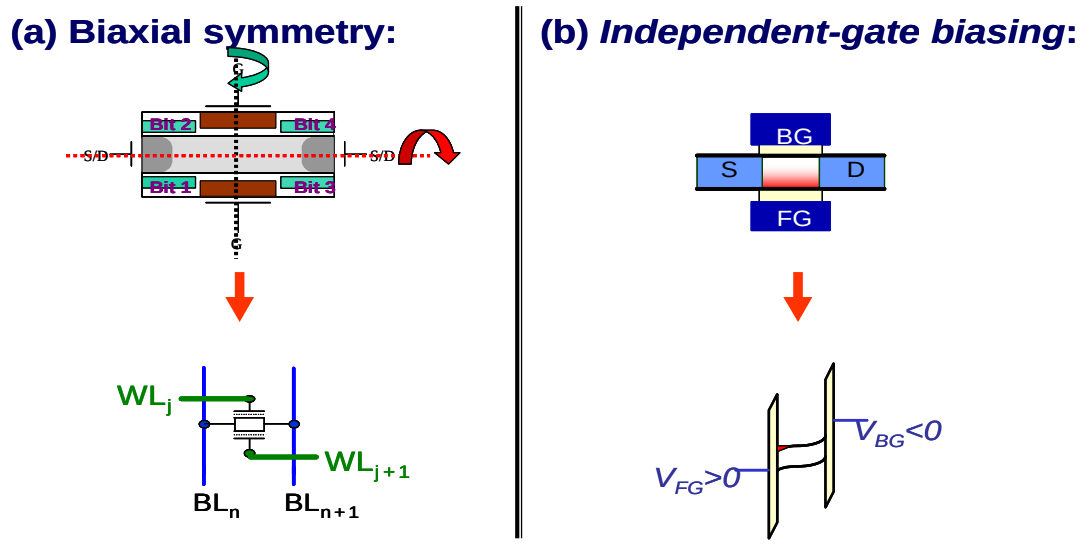


Figure 6.6: The operation of a 4-bit DG-FET NVM cell is based on two concepts: (a) Biaxial symmetry, which allows independent access of each charge storage site, and (b) Independent-gate biasing, which allows the masking of information stored at the unselected gate (e.g. the BG) as necessary.

The state of each bit can be detected in a conventional manner by sensing the transistor current in the on state, since the threshold voltage (V_T) will be affected by the presence of charge stored in the bit located next to the source electrode. By biasing the unselected gate to a negative voltage, the bits next to it can be effectively masked while the bits next to the selected gate are read. For example, to read the bit close to the source and beside the front gate (*i.e.* Bit 1), the back gate is biased to a negative voltage (V_{NR} ~ -1.5V) while the front gate is biased to a positive voltage (V_R ~ 0.5V) with a moderate drain-to-source bias (*e.g.* V_{DS} ~ 1.5V) applied (**Figure 6.7**). The charge state of Bit 1 is then determined from the transistor current: if electrons are stored in Bit 1, the threshold voltage will be high so that the read current will be low; if no electrons are stored in Bit 1, the read current will be high. The negative bias voltage applied to the unselected gate (*e.g.* the back gate in the present example) causes the channel next to it to be accumulated (p-type) so that it does not contribute any source-drain current [6.14]; therefore, the charge state of the bits next to the unselected gate have minimal influence on the read current, *i.e.* these bits are effectively masked. The bit close to the drain and beside the selected gate (*e.g.* Bit 3 in the present example) is effectively masked by applying a moderate drain-to-source bias so that the drain depletion region extends beyond the influence of this bit, allowing the source-side bit to be "read through" the drain-side bit [6.3].

Identification of charge storage at bit 1:

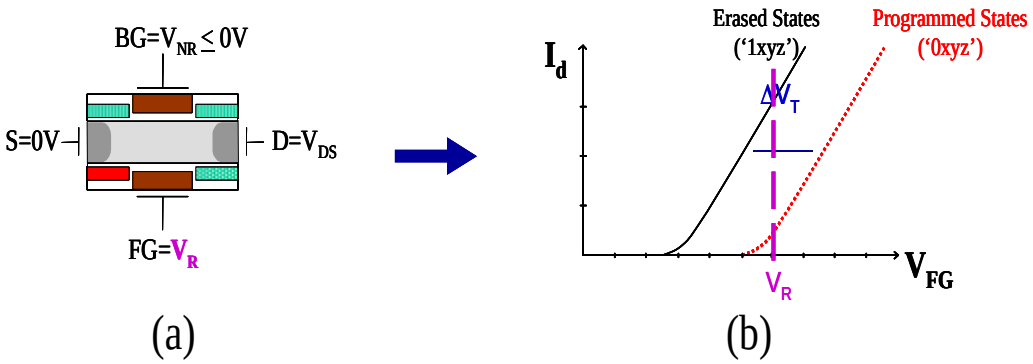


Figure 6.7: Identification of charge storage at Bit 1 of a 4-bit DG_FET structure by the conventional method. (a) Bias conditions. (b) Resulting I_{DS}-V_{FG} (constant V_{DS}) curves, describing the separation between the programmed ('0xyz') and erased ('1xyz') Bit1 states.

Each bit can be independently programmed via hot electron injection (HEI) by biasing the selected gate and drain electrodes (nearest to the bit) to a moderately high voltage (V_{PGM} and V_{DS} respectively, where $V_{PGM} \geq V_{DS}$ to re-direct the generated hot electrons towards the desired charge-trapping site) while the source electrode is grounded (**Figure 6.8(a)**). The unselected gate electrode should be grounded or biased to a slightly negative voltage ('V_{acc}') to prevent unintentional programming of the unselected bits stored within it.

Each bit of these 4-bit DG-FET structures can also be independently erased via the band-to-band tunneling induced Hot Hole Injection (HHI) method (**Figure 6.8(b)**). To erase a bit, the appropriate drain electrode (nearest to the bit) is biased to a high positive voltage, V_{DS}, and the selected gate electrode (nearest to the bit) is biased to a

high negative voltage V_{ERS_G} (where $|V_{ERS\_G}| \geq |V_{DS}|$ to re-direct the generated hot holes towards the desired charge-trapping site) while the source electrode is grounded and the unselected gate is either grounded or biased to a small positive voltage. Alternatively, the two bits next to the drain electrode can be simultaneously erased via HHI by biasing both gates to the same high negative voltage, V_{ERS_G} (even when the substrate is floating [6.12]). Alternatively, all bits can be simultaneously erased via the Fowler-Nordheim (FN) Tunneling method, by biasing both gate electrodes to a large negative voltage while both the Source & Drain electrodes are grounded (even when the substrate is floating [6.13]).

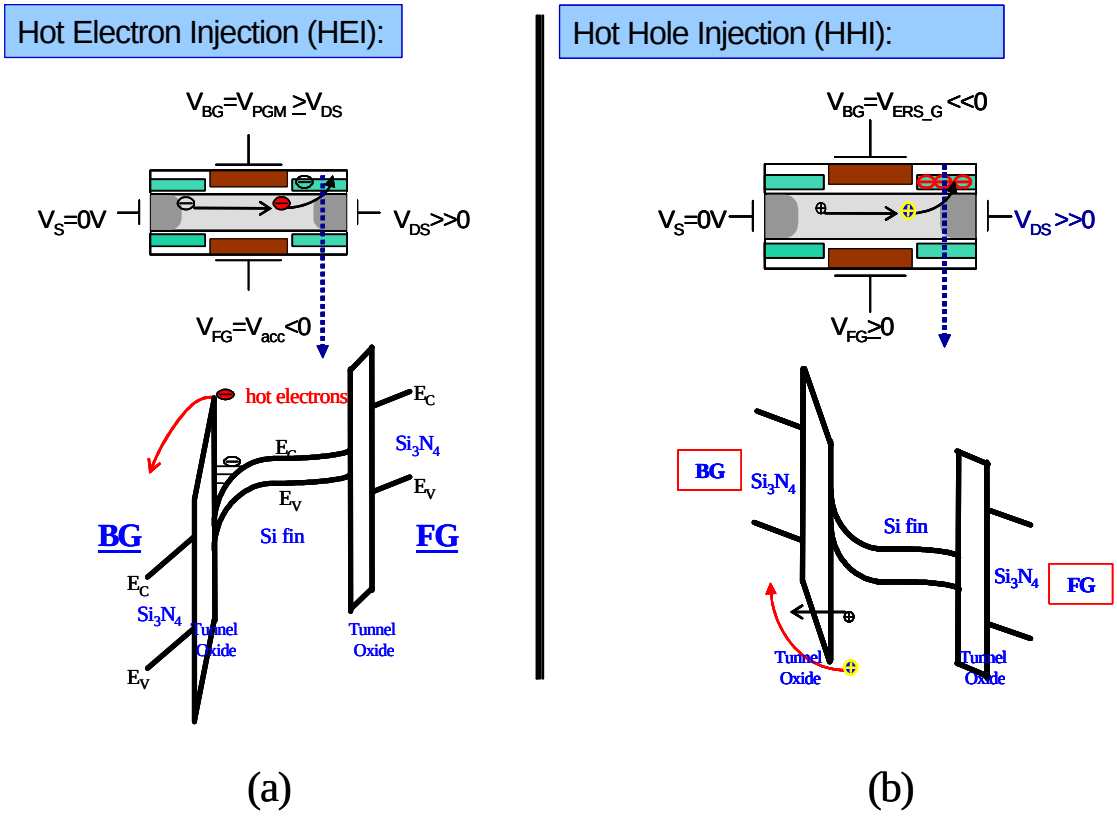


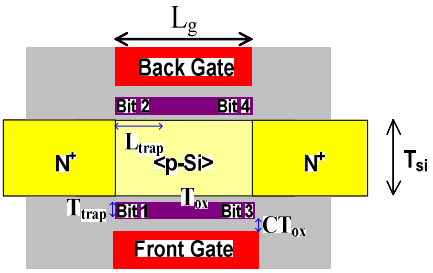
Figure 6.8: Each bit of the 4-bit DG-FET structure can be selectively programmed and erased via the conventional (a) Hot Electron Injection, and (b) Hot Hole Injection methods, respectively.

6.3.2 Read Simulation Setup

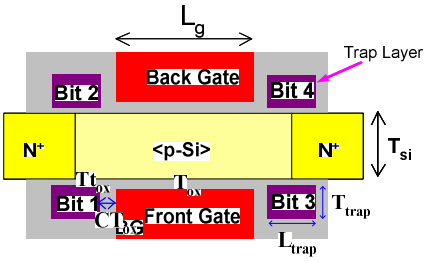
To demonstrate the feasibility of the conventional read method, and the enhanced scalability of the novel Side-Wall Storage DG-FET (SWS DG-FET) structure shown in **Figure 6.4(b)**, 2-D device simulations were performed on both structures shown in **Figure 6.4** using the Taurus [6.15] device simulator. For optimum performance, both simulated structures utilize a very lightly doped (1×10^{13} cm^{-3}) p-type silicon body (to provide high carrier mobilities and to minimize statistical dopant fluctuation effects) with thickness T_{si} chosen to be less than ~0.45*L_{eff} (for instance, T_{si}=20nm for L_g~40nm, or L_{eff}~50nm) to suppress SCE [6.16][6.17]so that leakage from unselected cells will not be an issue for a NOR array architecture. Also (in both structures), the gate electrodes are each n+ poly-Si (work function = 4.17eV), the charge-trapping material was chosen to be either undoped poly-Si or Si_3N_4, and the areal density of charge at the bottom interface (nearest to the channel) of the charge-trapping film was set to a value of $5x10^{12}$ cm^{-2}, comparable to that used by other investigators [6.18]. In the novel (SWS DG-FET) structure, the edges of the source and drain junctions must fall between the charge-trapping sites in order to maximize the V_T shift due to charge storage [6.11], so that a gate-underlapped source/drain structure is required (*i.e.* the effective channel length L_{eff} is greater than L_g) with this structure. Thus, L_{eff} was selected to be ~50nm for this structure, and ~80nm for the SONOS DG-FET structure (since the latter is not as scalable as the former [6.11]). Furthermore, for optimum programming (or erasing) of the SWS DG-FET device, the thickness of the tunnel oxide (T_{tox}) between each charge-trapping region and the channel was selected to be thin enough (~3nm) to allow for efficient drift of electrons or holes from the channel into the charge-trapping region (or vice versa) at

low operating voltages (≤5V). Additionally, the gate-oxide thickness (T_{ox}) must be thicker than T_{tox} to allow for efficient programming and was therefore selected to be 6nm; also, the control oxide film (that separates each gate from each charge-trapping region) has thickness CT_{ox}=6nm, and each charge-trapping region has thickness T_{trap}=15 nm and length L_{trap}=15 nm. Table 6.1 provides a list of the optimized parameters used in both structures. Appendix B contains sample simulation input files used in these simulations.

Parameter	a) SONOS DG-FET	b) SWS DG-FET
Gate Oxide, T_{ox}	3 nm	6 nm
Tunnel Oxide, T_{tox}	n/a	3 nm
Control Oxide, CT_{ox}	5 nm	6 nm
Body Thickness, T_{si}	40 nm	20 nm
Gate length, L_g	$2*T_{si}$=80 nm	$2*T_{si}$=40 nm
Trap-layer length, L_{trap}	25 nm	15 nm
S/D Doping (n^+ region)	2e20 #/cm^2	2e20 #/cm^2
$Q_{ox,max}$	-5e12 #/cm^2	-5e12 #/cm^2
Φ_G (n+ PolySi)	4.17 eV	4.17 eV
Body (constant) doping	1e13 #/cm^3	1e13 #/cm^3
Applied voltage to non-selected gate	-2.5 V	-1.5 V
Applied V_{DS} voltage	1.5 V	1.5 V



(a) SONOS DG-FET



(b) SWS DG-FET

Table 6.1: Parameter settings used in Taurus simulations (for optimum performance).

6.3.3 Read Simulation Results

Figure 6.9(a) shows the simulated drain current *vs.* selected gate voltage (I_{DS}-V_{FG}) curves for the optimized SWS DG-FET structure (L_g=40nm, T_{si}=L_g/2=20nm) with the unselected gate (V_{BG}) biased at -1.5V and drain-to-source bias V_{DS} = 1.5V. **Figure 6.9(b)** shows a similar plot for the optimized SONOS DG-FET structure (L_g=80nm, T_{si}= L_g/2=40nm). In both plots, the curves are clearly seen to be clustered into 2 groups ('0xyz', '1xyz'), separated by a shift in threshold voltage (ΔV_T) according to charge stored at the selected bit near to the source (*i.e.* Bit 1).

These plots also show the enhanced spread in V_T obtained with the SONOS DG-FET (even though this cell has larger dimensions than its counterpart) since this structure contains the charge-trapping sites underneath the gate electrode (which makes it more susceptible to SCE). Nonetheless, the V_T separation observed in both structures is large enough to allow easy detection of charge stored at the selected bit (*e.g.* by applying V_{FG} = 0.5V and sensing the transistor current). Due to the symmetry of these structures, the state of each bit can be determined by a sequence of 4 read operations. For example, to read the bit next to the source electrode and beside the back gate (*i.e.* Bit 2), the back gate is biased to a positive voltage (V_{BG} ~ 0.5V) while the front gate is biased to a negative voltage (V_{FG} ~ -1.5V) with a moderate drain-to-source bias applied (V_{DS} ~ 1.5V). To read the complementary bits (*i.e.* Bits 3 and 4), the roles of the source and drain electrodes are interchanged (*i.e.* the transistor is operated in reverse mode).

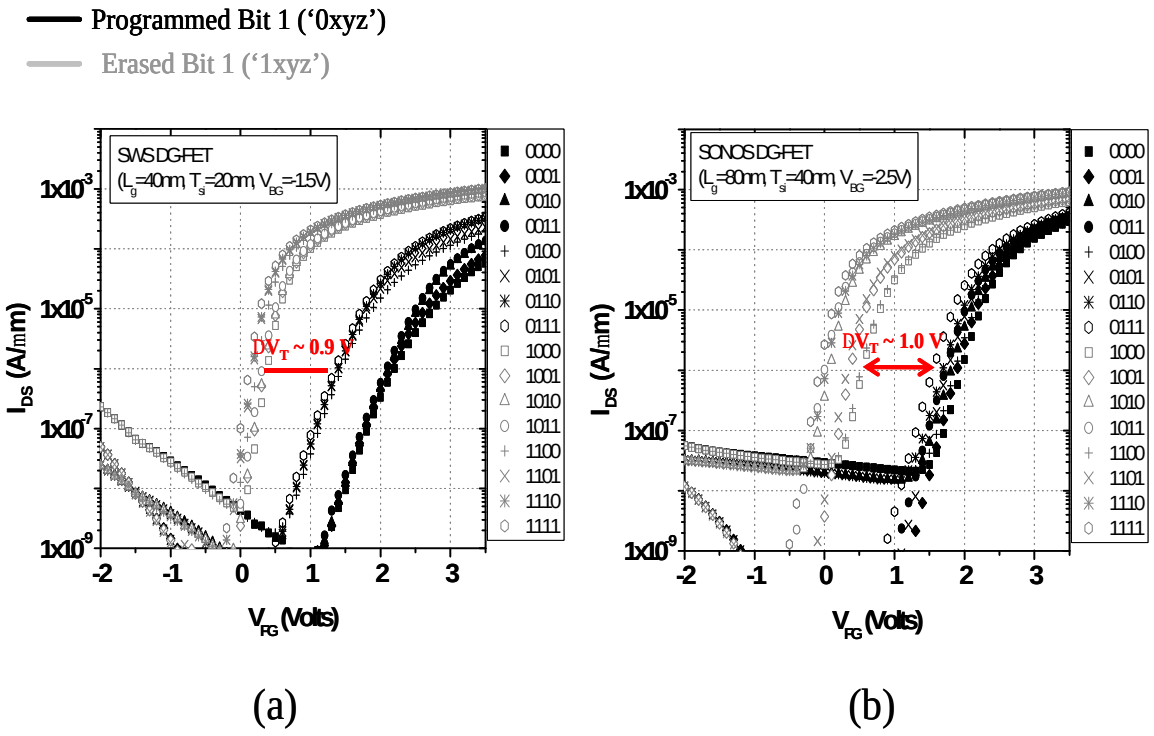
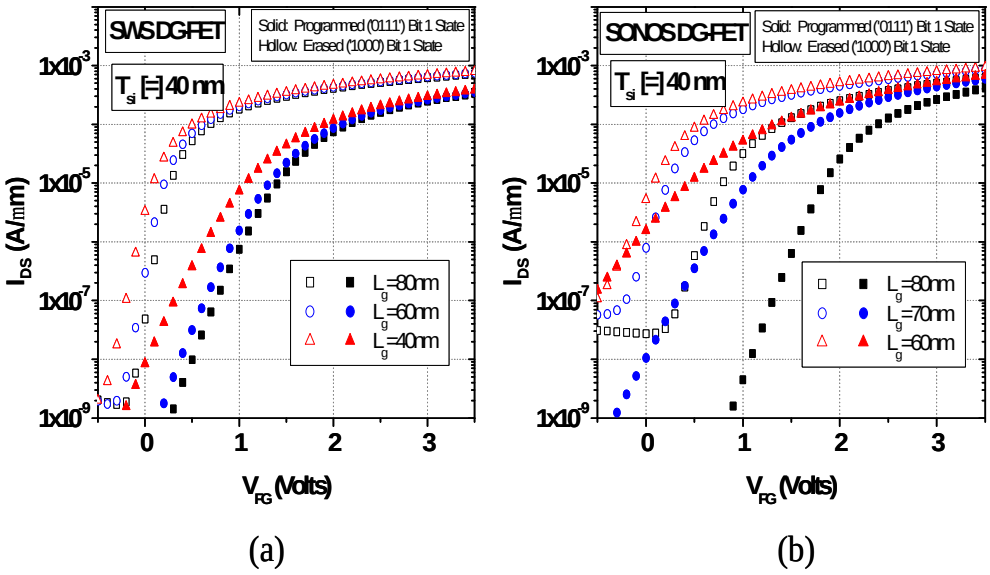


Figure 6.9: Simulated drain current (I_{DS}) vs. front-gate voltage (V_{FG}) characteristics for each of 16 possible charge-storage combinations (states ranging from 0000 to 1111) of (a) a 4-bit SWS DG-FET structure (with T_{si}=20nm, L_g=40nm), and (b) a 4-bit SONOS DG-FET structure (with T_{si}=40nm, L_g=80nm). There are two distinct groupings of curves (corresponding to states '0xyz' or '1xyz') separated due to a shift in threshold voltage (ΔV_T) resulting from charge stored in Bit 1. Thus, the conventional method of transistor on-state current sensing can be used to determine state of Bit 1, regardless of the state of the other bits.

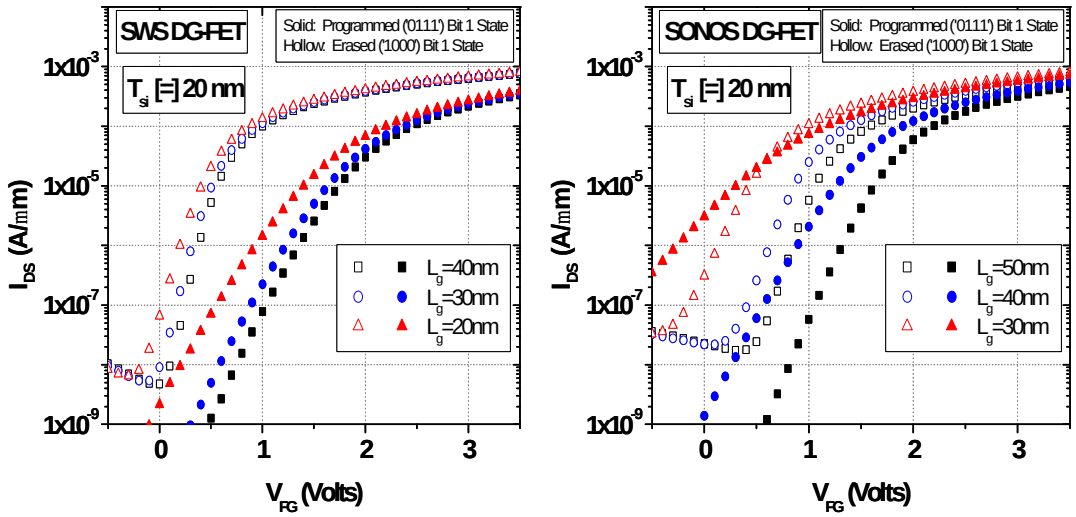
To assess the scalability of both DG-FET structures, additional simulations were performed on structures that had a specific silicon body thickness (T_{si}) but different gate lengths (L_g). **Figure 6.10(a)** shows the simulated (I_{DS}-V_{FG}) curves of a SWS DG-FET structure with T_{si}=40nm (V_{BG}=-1.5V, V_{DS}=1.5V) for the 2 binary states with *smallest* V_T separation ("DV_T") due to charge stored at Bit 1 (*i.e.* binary states '*1000*' and '*0111*'), for

various L_g. **Figure 6.10(b)** shows a similar plot for the SONOS DG-FET structure (T_{si}=40nm, V_{BG}=-2.5V, V_{DS}=1.5V). As expected for both structures, scaling of the cell's L_g degrades its sub-threshold swing (S), *especially* for the programmed Bit 1 state. This swing degradation is nonetheless more significant for the SONOS DG-FET since this structure is more susceptible to SCE (due to both its thicker EOT and the fact that the charge-trapping sites are placed underneath the gate electrode [6.11]). Additional simulations performed on structures that have a thinner a thinner silicon body thickness (**Figure 6.11**) support these observations.



(a) (b)

Figure 6.10: Simulated I_{DS} vs. V_{FG} characteristics for the 2 binary Bit 1 states with smallest V_T separation (ΔV_T) of a (a) SWS DG-FET, and (b) SONOS DG-FET structure (for various gate lengths (L_g) and for a specific silicon body thickness (T_{si}=40nm)). As shown, scaling of the cell's L_g degrades the cell's sub-threshold swing (particularly, for the programmed state of the SONOS DG-FET) and thus reduces the ΔV_T between the programmed and erased states. This effect limits the scalability of 4-bit SONOS DG-FETs.



(a) (b)

Figure 6.11: Simulated I_{DS} vs. V_{FG} characteristics for the 2 binary Bit 1 states with smallest V_T separation (ΔV_T) of a (a) SWS DG-FET, and (b) SONOS DG-FET structure (for various gate lengths (L_g) and for a specific silicon body thickness (T_{si}=20nm)).

Figure 6.12 contains a plot of the resulting DV_T between these two binary states as a function of both L_g and T_{si} for both structures. The observed swing degradation makes the distinction between the two Bit 1 states more difficult (since DV_T gets smaller because of it) and thus imposes a serious limitation on the scalability of the 4-bit SONOS DG-FET. Note however, that the swing degradation is not as significant in the 4-bit SWS DG-FET (since this cell does *not* contain the charge-storage sites underneath the gate electrodes); consequently, this structure is more scalable.

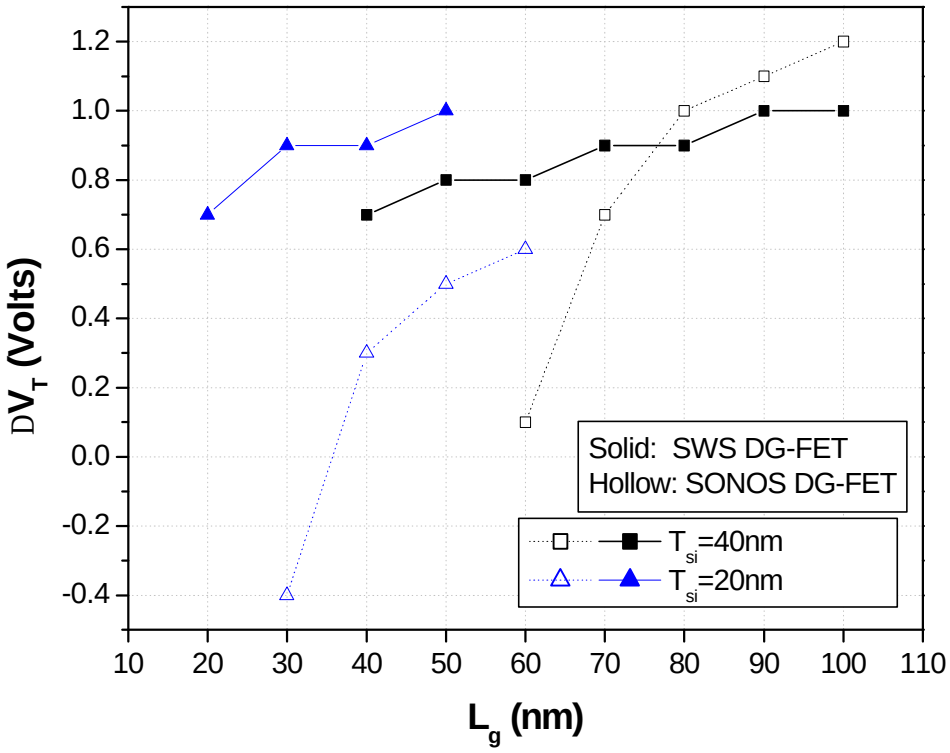
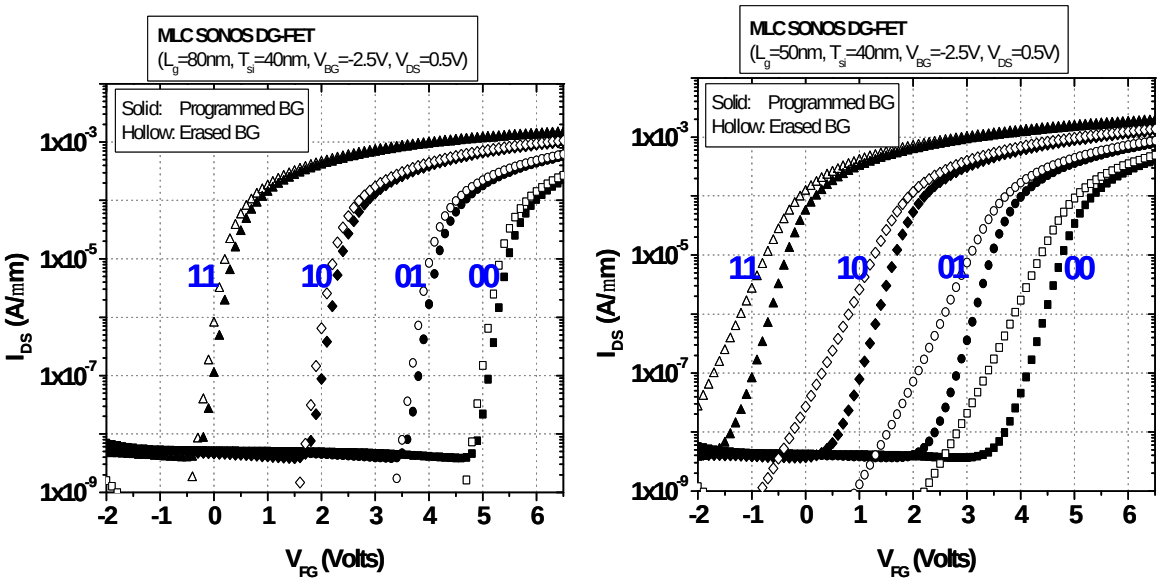


Figure 6.12: Simulated threshold voltage shift (ΔV_T) due to charge storage at Bit 1 as a function of the cell's gate length (L_g). As can be seen, the 4-bit SWS DG-FET structure is more scalable than the 4-bit SONOS DG-FET.

As an alternative, the SONOS DG-FET structure can also utilize the MLC algorithm to store 4 bits in a manner similar to that reported in [6.7] for the 4-bit DSM Cell. As noted previously, this algorithm requires the use of 4 distinct V_T levels (to store 2 bits) at each gate *and* the ability to mask the information stored at the unselected gate (by biasing its channel into accumulation) as necessary. The use of the MLC algorithm with this cell allows treatment of each charge-trapping layer as a *single* charge storage site; consequently, the FN tunneling mechanism can be used to either program or

erase each charge-trapping site as necessary (by biasing each gate accordingly), and this makes this structure applicable to NAND-type architectures. **Figure 6.13(a)** shows the simulated I_{DS}-V_{FG} curves of an optimized MLC SONOS DG-FET structure (L_g=80nm, T_{si}==40nm, with V_{BG}=-2.5V, V_{DS}=0.5V), with the BG either completely erased or programmed to the highest V_T level. As shown, significant V_T separation is achieved between the 4 states; however, a slight shift in V_T is noticed at all V_T levels when the BG is programmed (due to the strong coupling between both gates), and this behavior was also noticed on the 4-bit DSM cell reported in [6.7]. The observed V_T shift gets worse when the cell is not properly scaled to address SCE, as shown in **Figure 6.13(b)**. Even though this V_T shift could be reduced through application of a larger negative voltage at the unselected gate (to properly screen the information stored there) [6.19], this is *not* desired since this setting induces a potential read disturbance (i.e., charge leakage from either gate due to application of large transverse fields), which must be avoided.



(a) (b)

Figure 6.13: Simulated I_{DS} vs. V_{FG} characteristics (V_{BG}=-2.5V, V_{DS}=0.5V) of a MLC SONOS DE_FET cell with T_{si}=40nm, and (a) L_g=80nm, (b) L_g=50nm. In both cases, the 4 distinct states of the FG are shown with the BG either completely erased or programmed to the highest V_T level. As shown, the (un)charged state of the BG affects the V_T level of all states of the FG (and vice versa), especially for the latter structure, which is not properly scaled to address SCE.

To assess the scalability of the MLC SONOS DG-FET structure, additional simulations were performed on structures with various L_g and for a specific T_{si}. **Figure 6.14** shows the simulated I_{DS}-V_{FG} characteristics for the lowest V_T state (i.e., binary state '11') of a MLC SONOS DG-FET cell with various L_g (V_{BG}=-2.5V, V_{DS}=0.5V), with the BG bit either completely erased or programmed to the highest V_T level. As previously shown with the MBC SONOS DG-FET design, scaling of this cell's L_g degrades its sub-

threshold swing, especially when the BG bit is completely erased. This swing degradation severely affects the observed V_T shift due to the programmed state of the unselected gate, which is of course not desired. In addition, the enhanced gate-to-body coupling attained in DG FET structures with thinner T_{si} [6.19] further limits the scalability of MLC SONOS DG-FET cells precisely for this reason (i.e., the enhanced gate-to-gate coupling observed on DG-FET structures with thinner T_{si} makes it more difficult to screen the information stored at the unselected gate, as shown in **Figure 6.14(b)**). Consequently, this structure has limited scalability, even when operated in MLC mode.

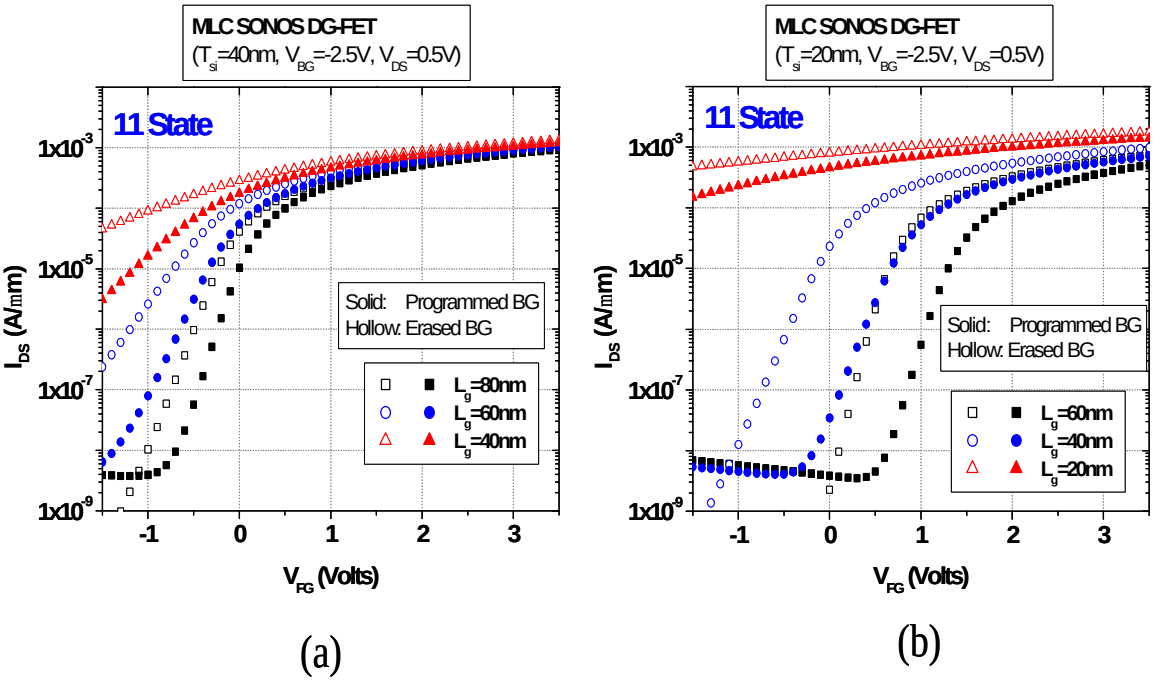


Figure 6.14: Simulated I_{DS} vs. V_{FG} characteristics (V_{BG}=-2.5V, V_{DS}=0.5V) for the erased state of a MLC SONOS DE_FET cell, with the BG either fully erased or programmed to the highest V_T level, for various L_g. (a) T_{si}=40nm. (b) T_{si}=20nm.

6.4 Memory Architectures

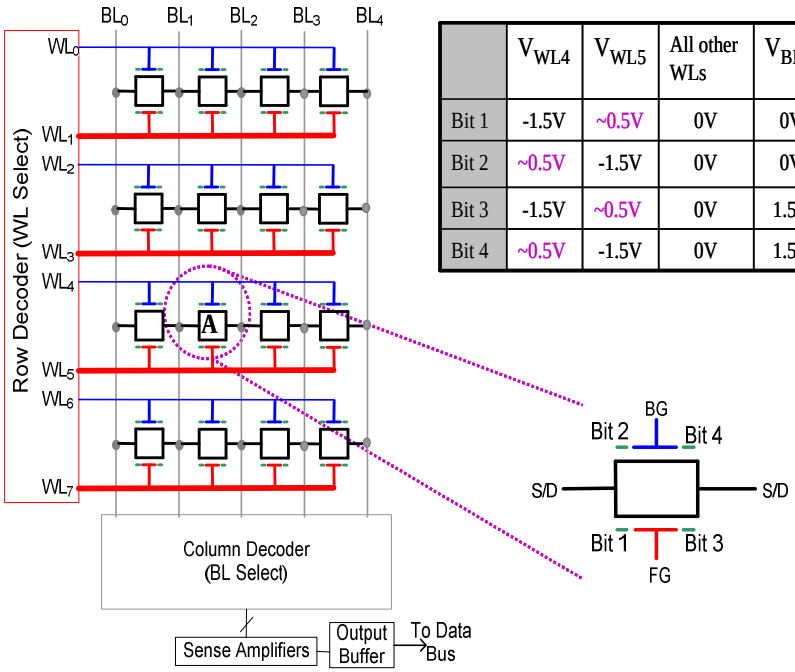
In this section, the use of either (SONOS or SWS) 4-bit DG-FET structure (shown in **Figure 6.4**) within both a NOR-type and NAND-type architecture is discussed. First, the selective operation of each bit within a 4-bit DG-FET cell (within a NOR or NAND array architecture) is discussed, with particular emphasis on the selective read operation. In practice, these structures can be implemented with either the BG-FET design [6.5], as already demonstrated previously with the DSM NVM cell design in [6.6][6.7], or the SOI Multiple Independent Gate Field Effect Transistor (MIGFET) design [6.20] (**Figure 6.5**). Consequently, the size of each unit cell are estimated and compared (based on these FET designs) within these architectures by using the design rules listed in **Table 6.2** below. In some of the drawings that follow (in the next sub-sections), the word-lines connecting to either the Front Gate (FG) or the Back Gate (BG) of each unit cell are drawn with a different color for clarity purposes only.

Table 6.2) Adopted Design Rules

Parameter	Size
Minimum Word-Line Width (WLW):	F
Minimum Gate-Length Width (L_g):	F
Minimum WL-to-WL spacing:	F
Minimum Bit-Line Width (BL):	F
Minimum BL-to-L_g spacing (or minimum gate-sidewall spacer width)	F/2
Minimum *Active* Area thickness (includes T_{si})	F

6.4.1 'Virtual Ground in SOI' NOR-type Architecture

Figure 6.15 shows a circuit diagram of DG-FET NVM cells arranged in a (Virtual Ground) NOR array configuration, along with the word-line (WL) and bit-line (BL) bias conditions required to read each bit within Cell A. In this architecture, cells are arranged in a two-dimensional (2D) array, where the front- or back-gates of all cells in the same row are connected to the same WL, and all the source (or drain) electrodes of all cells in the same column are connected to the same BL. In this architecture (as noted previously), each bit of either 4-bit DG-FET NVM cell can be *selectively* programmed or erased via the HEI, HHI methods, respectively (as demonstrated previously for the 4-bit DSM NVM cell reported in [6.6][6.7]).



	V_{WL4}	V_{WL5}	All other WLs	V_{BL0}	V_{BL1}	V_{BL2}	V_{BL3}	V_{BL4}
Bit 1	-1.5V	~0.5V	0V	0V	0V	1.5V	1.5V	1.5V
Bit 2	~0.5V	-1.5V	0V	0V	0V	1.5V	1.5V	1.5V
Bit 3	-1.5V	~0.5V	0V	1.5V	1.5V	0V	0V	0V
Bit 4	~0.5V	-1.5V	0V	1.5V	1.5V	0V	0V	0V

Figure 6.15: Schematic circuit diagram of 4-bit DG-FET NVM cells arranged in a NOR array architecture. Bias voltages used to read each bit of Cell 'A' are indicated in the Table.

To read Bit 1 of Cell A (**Figure 6.15**), its front gate electrode is biased to a positive voltage (e.g. V_{WL5} ~ 0.5V) with a moderate drain-to-source voltage applied (*e.g.* V_{BL2}-V_{BL1} = 1.5V) to mask Bit 3; its back gate electrode is biased to a negative voltage (*e.g.* V_{WL4} = -1.5V) to mask Bits 2 and 4. If no charge is stored at Bit 1, then the front channel of the NVM cell will be turned on so that significant current (>100μA/μm channel width) will flow in bit-lines BL_2 and BL_1. If charge is stored at Bit 1, V_T will be shifted in the positive direction (**ref. Figure 6.9**) so that only leakage current will flow (~1nA per μm channel width) in bit-lines BL_2 and BL_1. Thus, the state of Bit 1 in Cell A can be distinguished by sensing the current flowing in either one of the bit-lines BL_2 and BL_1. All of the other word lines are biased at 0V (or a negative voltage) to ensure that non-selected cells are turned off so that they contribute negligible bit-line current. In order to prevent the non-selected cells along the same word lines (WL_4 and WL_5) from contributing any bit-line current, the bit lines must be biased such that V_{DS} = 0V for each of these non-selected cells.

Note from **Figure 6.9** that V_T for the erased state may be close to 0V – or even slightly negative – if n+ poly-Si is used as the gate electrode material, consistent with [6.21]. In this case, negative word-line biasing is necessary to ensure that the unselected cells are turned off. A metallic gate material with a larger (near-midgap) work function can be used instead to achieve positive V_T for the erased state, to avoid the need for negative word-line biasing.

Figure 6.16 shows a basic layout (not to scale) of an embodiment of the NOR-type array architecture shown (as a circuit diagram) in **Figure 6.15**. As shown, this architecture uses SOI technology, and the unit cell in this architecture is the MIGFET

structure (modified accordingly as a 4-bit SONOS DG-FET, as an example). In this layout, the word lines (WLs) are n^+ Poly Si stripes, and the source or drain (S/D) bit-lines (which consist of n^+-Si stripes) are shared between adjacent cells to reduce space, as done with the *Virtual Ground* NOR-type NROM™ architecture [6.3]. The required placement of WLs (connecting to the same cell) on the side (instead of on top) of the active area of each cell in the same row, and the required isolation between adjacent WLs significantly increases the size of each unit cell. As a result, the size of the unit cell with this FET design is slightly larger (~$12F^2$, or ~$3F^2$ per bit) than that of the 2-bit NROM™ cell (~$2.5F^2$ per bit) [6.3], which indicates that the use of the MIGFET structure is *not* the optimum cell design in terms of memory density (since it does not reduce the size per bit).

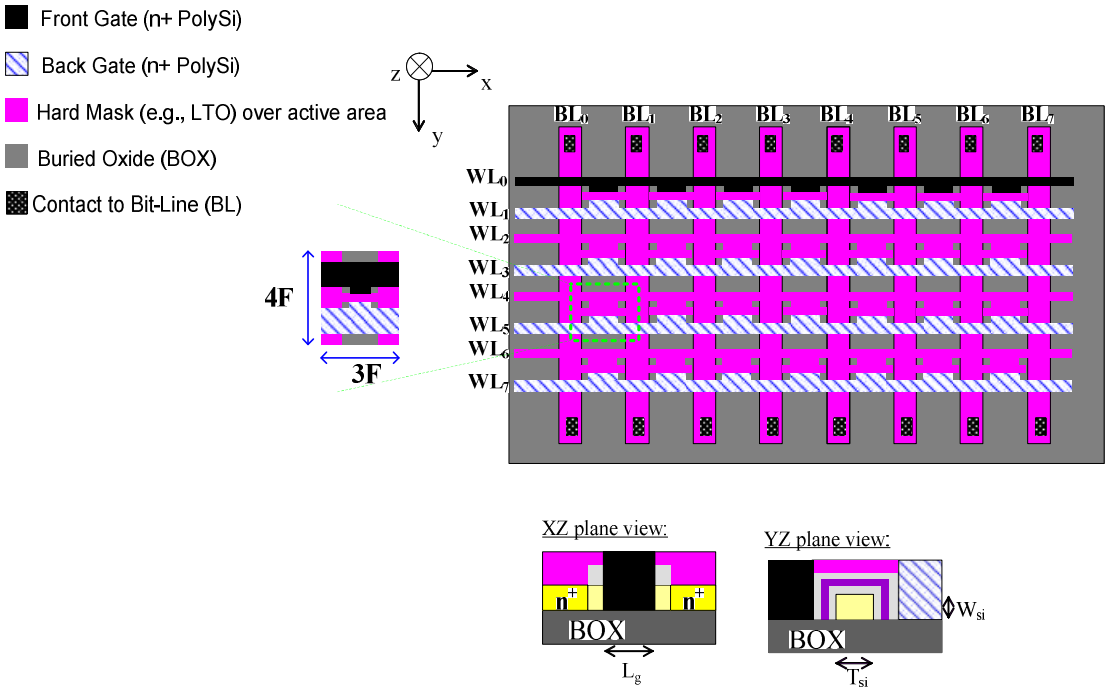


Figure 6.16: NOR-type array architecture layout (based on SOI technology) that utilizes the MIGFET structure as its unit cell. The placement of both word-lines on the side (instead of on top) of the active area and their required isolation significantly increases the size of its unit cell.

As an alternative, both 4-bit DG-FET structures (shown in **Figure 6.4**) can also be implemented with the planar BG-FET design. In this case, buried diffusion (n^+-Si) stripes can be used to define the WLs connecting to the back gates of each cell in the same row [6.6]; consequently, the word lines connecting to each cell in the same row can be placed on top and bottom the active region, thereby resulting in a more compact unit cell (~$6F^2$, or ~$1.5F^2$ per bit), as shown in **Figure 6.17(b)**. Clearly, the use of the more scalable 4-bit SWS DG-FET device (**Figure 6.4(b)**), implemented with the more compact planar BG-FET design (**Figure 6.5b**), is the optimum choice for use with this architecture.

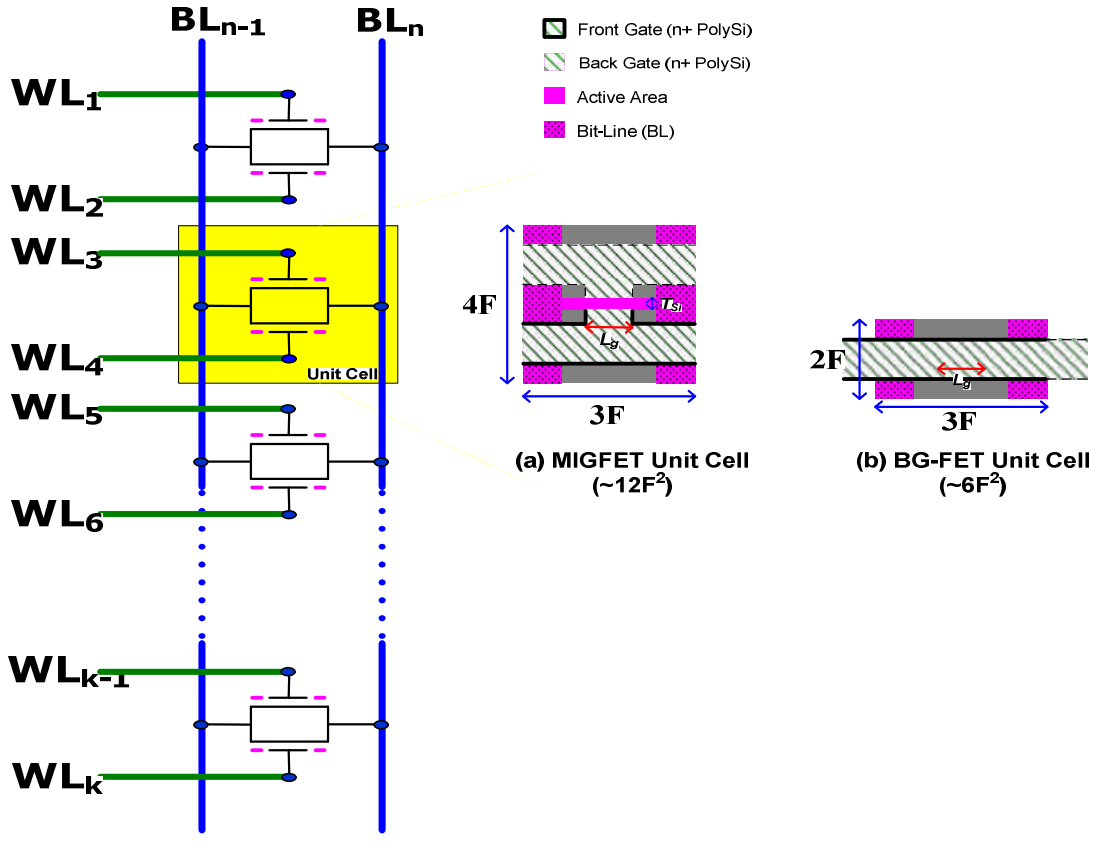
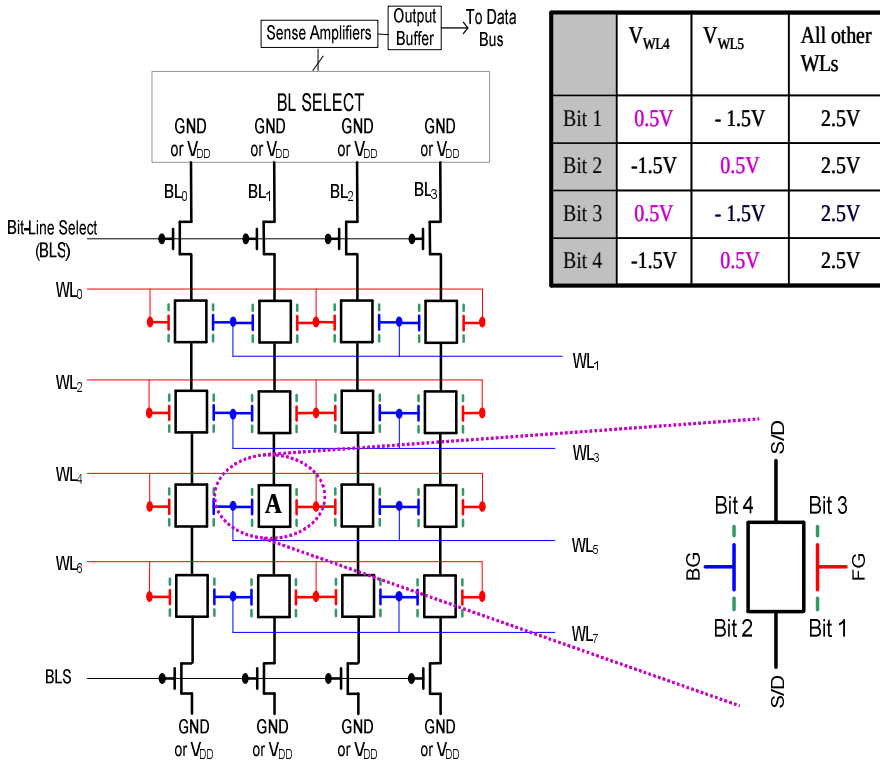


Figure 6.17: Unit cell layout of (a) a MIGFET cell design, and (b) a planar BG-FET cell design within a NOR-type array architecture. The use of a planar BG-FET as a 4-bit DG-FET cell significantly reduces the size of each unit cell (~$6F^2$) with this architecture [6.6].

6.4.2 NAND-type Architecture

Figure 6.18 illustrates a circuit diagram of a NAND-type array architecture that utilizes either DG-FET NVM cell (shown in **Figure 6.4**) as its unit cell, along with the word-line (WL) and bit-line (BL) bias conditions required to read each bit within cell A (when the cell is operated in MBC design mode). In this architecture, cells belonging to the same bit-line are connected in series between two bit-line select (BLS) transistors, which must each be able to pass a high voltage (V_{DD}) as well as a low voltage (GND) to allow for forward-read as well as reverse-read operation of each cell.



	V_{WL4}	V_{WL5}	All other WLs	BLS	V_{BL1} (top)	V_{BL1} (bottom)
Bit 1	0.5V	- 1.5V	2.5V	V_{DD}	V_{DD}	GND
Bit 2	-1.5V	0.5V	2.5V	V_{DD}	V_{DD}	GND
Bit 3	0.5V	- 1.5V	2.5V	V_{DD}	GND	V_{DD}
Bit 4	-1.5V	0.5V	2.5V	V_{DD}	GND	V_{DD}

Figure 6.18: Schematic circuit diagram of 4-bit DG-FET NVM cells arranged in a NAND array architecture. Bias voltages used to read each bit of Cell 'A' are indicated in the Table.

Each bit of each 4-bit DG-FET NVM cell could be selectively programmed via the Band-to-Band Tunneling Induced Hot Electron (BBHE) injection method. Though this method has only been demonstrated to date on p-channel SONOS-type NVM cells [6.22][6.23], its use has already been demonstrated with a NAND array architecture composed of dual-bit p-type Band-gap Engineered SONOS ('BE-SONOS') NVM cells [6.23]. To read a specific cell in the array, each of the other cells that share the same bit-line are turned on strongly (by applying moderate word-line biases) so that these simply serve as pass transistors. Current flow through the bit-line is then determined by the state of the selected bit in the selected cell. Note that cells sharing the same word lines in a NAND array can be read simultaneously, in contrast to cells sharing the same word lines in a NOR array which must be read sequentially.

As noted previously, the SONOS DG-FET cell design can *also* utilize the MLC algorithm to store 4 bits, and this makes this structure directly applicable to a NAND-type architecture (without the need of forward- or reverse-read operation to determine the state of the cell). In this mode of operation, the conventional methods can be used to selectively read (via the on-state current sensing method), program (via the FN tunneling method), and erase (via the FN tunneling method) the information stored at each gate [6.2]. For instance, each charge-trapping layer can be *selectively* programmed (to the desired V_T level) via the FN tunneling mechanism by biasing the selected gate to a large positive voltage, and the unselected gate of the cell to a large negative voltage (in order to inject electrons via FN tunneling onto the selected gate).

Figure 6.19 shows a column of the NAND-type array architecture of **Figure 6.18**, along with the basic layout (not to scale) of the unit cell which corresponds to a 4-bit DG-

FET NVM cell that is implemented with the planar BG-FET design. As shown, the size of the unit cell within this architecture is very small (~$4F^2$, or ~$1F^2$ per bit), which indicates that this architecture is indeed the optimum array architecture design in terms of memory density (since it significantly reduces the size per bit). In this case, the use of the MLC SONOS DG-FET structure (implemented with the more compact planar BG-FET design) is the optimum choice for use with this architecture (since this cell utilizes the conventional methods of operation normally used with a NAND-type architecture).

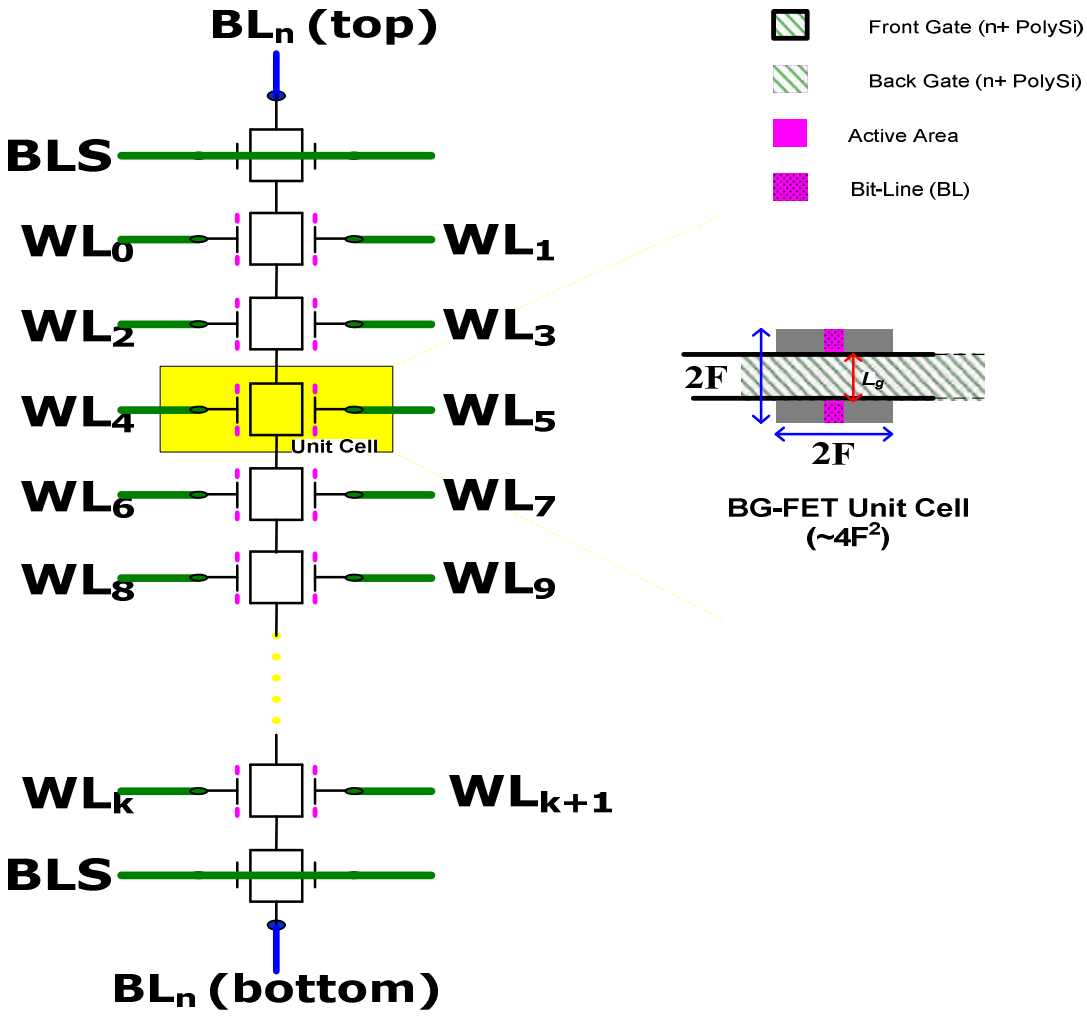


Figure 6.19: Unit cell layout of a BG-FET cell design within a NAND-type Array Architecture.

6.5 Conclusions

In this chapter, two 4-bit NVM cell designs (utilizing a double-gate FET structure) are investigated via device simulation. To store charge, one (SONOS DG-FET) cell design utilizes a charge-trapping site underneath each gate electrode, whereas the other (SWS DG-FET) cell design utilizes gate-sidewall charge-trapping sites. For both cell designs, the possibility to distinguish the state of each bit is verified via 2-dimensional (2-D) device simulations. Read operation of each cell design within NOR and NAND memory array architectures is also discussed. Since the SONOS DG-FET cell design is very susceptible to Short-Channel Effects (SCE), it is less scalable than the SWS DG-FET design. Nonetheless, the SONOS DG-FET design can utilize either the Multi-Bit Cell (MBC) or the Multi-Level Cell (MLC) algorithm to store 4 bits, and this feature makes it directly applicable to both NOR- and NAND-type array architectures. For both architectures, the optimum practical implementation of these structures involves the use of the planar BG-FET design, since its use most effectively reduces the size per bit of each unit cell within these architectures. Both cell designs are promising to enhance storage density in future non-volatile memories.

6.6 References

[6.1] M. L. French, Chun-Yu Chen, H. Sathianathan, and M. H. White, "Design and scaling of a SONOS multi-dielectric device for nonvolatile memory applications," *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, Part A, Vol. 17, No. 3, pp. 390-397 (1994).

[6.2] R. Liu, *et al.*, "Memory Technologies for 45nm and Beyond", *2006 International Electron Device Meeting Short Course*, (2006).

[6.3] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NROM: A Novel Localized Trapping, 2-bit Nonvolatile Memory Cell," *IEEE Electron Device Letters*, Vol. 21, No. 11, pp. 543-545 (2000).

[6.4] B. Eitan, G. Cohen, A. Shappir, E. Lusky, A. Givant, M. Janai, I. Bloom, Y. Polansky, O. Dadashev, A. Lavan, R. Sahar, and E. Maayan, "4-bit per Cell NROM Reliability", *IEDM Technical Digest*, p. 539-542 (2005).

[6.5] H.-S. P. Wong, K. K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," *International Electron Devices Meeting Technical Digest*, pp. 427 (1997).

[6.6] C. W. Oh, S. H. Kim, N. Y. Kim, Y. L. Choi, K. H. Lee, B. S. Kim, N. M. Cho, S. B. Kim, D.-W. Kim, D. Park, and B.-I. Ryu, "A 4-Bit Double SONOS Memory (DSM) with 4 Storage Nodes per Cell for Ultimate Multi-Bit Operation," *Symposium on VLSI Technology Digest of Technical Papers*, p. 50 (2006).

[6.7] C. W. Oh, N. Y. Kim, S. H. Kim, Y. L. Choi, S. I. Hong, H. J. Bae, J. B. Kim, K. S. Lee, Y. S. Lee, N. M. Cho, D.-W. Kim, D. Park, and B.-I. Ryu, "4-Bit Double

SONOS Memories (DSMs) Using Single-Level and Multi-Level Cell Schemes," *International Electron Devices Meeting Technical Digest*, pp. 967 (2006).

[6.8] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How far can Si go?," *International Electron Devices Meeting Technical Digest*, p. 553-556 (1992).

[6.9] B. Yu, Y.-J. Tung, S. Tang, E. Hui, T.-J. King, and C. Hu, "Ultra-thin-body silicon-on-insulator MOSFETs for terabit-scale integration," *International Semiconductor Device Research Symposium*, pp. 623-626 (1997).

[6.10] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, Vol. 85, No. 4, pp. 486-504 (1997).

[6.11] M. Fukuda, T. Nakanishi, and Y. Nara, "New Nonvolatile Memory with Charge-Trapping Sidewall", *IEEE Electron Device Letters*, Vol. 24, No. 7, p. 490-492 (2003).

[6.12] M. Specht, R. Kommling, F. Hofmann, V. Klandzievski, L. Dreeskornfeld, W. Weber, J. Kretz, E. Landgraf, T. Schulz, J. Hartwich, W. Rosner, M. Stadele, R.J. Luyken, H. Reisinger, A. Graham, E. Hartmann, and L. Rish, "Novel Dual Bit Tri-Gate Charge Trapping Memory Devices", *IEEE Elecron Device Letters*, Vol. 25, No. 12, pp. 810-812 (2004).

[6.13] F. Hofmann, M. Specht, U. Dorda, R. Kommling, L. Dreeskornfeld, J. Kretz, M. Stadele, W. Rosner, and L. Rish, "NVM based on FinFET device structures", *Solid-State Electronics*, Vol. 49, pp. 1799-1804 (2005).

[6.14] L. Mathew, Y. Du, A. V-Y Thean, M. Sadd, A. Vandooren, C. Parker, T. Stephens, R. Mora, R. Rai, M. Zavala, D. Sing, S. Kalpat, J. Hughes, R. Shimer, S. Jallepalli, G. Workman, W. Zhang, J.G. Fossum, B.E. White, B.-Y. Nguyen, and J. Mogab, "CMOS Vertical Multiple Independent Gate Field Effect Transistor (MIGFET)," *Proceedings of the IEEE SOI Conference*, pp. 187-189 (2004).

[6.15] Synopsys "Taurus Process & Device User Manual" 2003. http://www.synopsys.com.

[6.16] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "*Scaling Theory for Double-Gate SOI MOSFET's*," *IEEE Transactions on Electron Devices*, Vol. 40, No. 12, pp. 2326-2329 (1993).

[6.17] P. Xuan, M. She, B. Harteneck, A. Liddle, J. Bokor, and T.-J. King, "FinFET SONOS flash memory for embedded applications," *International Electron Devices Meeting Technical Digest*, pp. 609-612 (1999).

[6.18] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan, "Characterization of Channel Hot Electron Injection by the Subthreshold Slope of NROM^{TM} device," *IEEE Electron Device Letters*, Vol. 22, No. 11, pp. 556-558 (2001).

[6.19] C.-H. Lin, X. Xi, J. He, L. Chang, R. Q. Williams, M. B. Ketchen, W. E. Haensch, M. Dunga, S. Balasubramanian, A. M. Niknejad, M. Chan, and C. Hu, "Compact Modeling of FinFETs Featuring Independent-Gate Operation Mode," *IEEE VLSI-TSA International Symposium on VLSI Technology Proceedings of Technical Papers*, pp. 120 (2005).

[6.20] L. Mathew, Y. Du, A. V.-Y. Thean, M. Sadd, A. Vandooren, C. Parker, T. Stephens, R. Mora, R. Rai, M. Zavala, D. Sing, S. Kalpat, J. Hughes, R. Shimer, S.

Jallepalli, G. Workman, W. Zhang, J.G. Fossum, B. E. White, B.-Y. Nguyen, and J. Mogab, “CMOS Vertical Multiple Independent Gate Field Effect Transistor (MIGFET),” *Proc. IEEE International SOI Conference*, pp. 187-189 (2004).

[6.21] L. Chang, S. Tang, T.-J. King, J. Bokor, and C. Hu, "Gate length scaling and threshold voltage control of double-gate MOSFETs,” *International Electron Devices Meeting Technical Digest*, pp. 719-722 (2000).

[6.22] T. Ohnakado, K. Mitsunaga, M. Nunoshita, H. Onoda, K. Sakakibara, N. Tsuji, N. Ajika, M. Hatanaka, and H. Miyoshi, “Novel Electron Injection Method Using Band-to-Band Tunneling Induced Hot Electron (BBHE) for Flash Memory with a P-channel Cell,” *International Electron Devices Meeting Technical Digest*, pp. 279-282 (1995).

[6.23] H.-T. Lue, S.-Y. Wang, E.-K. Lai, M.-T. Wu, L.-W. Yang, K.-C. Chen, J. Ku, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “A Novel P-Channel NAND-Type Flash Memory with 2-bit/cell Operation and High Programming Throughput (>20 MB/sec),” *International Electron Devices Meeting Technical Digest*, pp. 331-334 (2005)

Chapter 7: Conclusions

7.1 Summary

Semiconductor flash memory has played a major role in the vast evolution observed in portable electronics in the last two decades. This evolution has been made possible (in part) due to the continued scaling (and thus the enhancement in memory density) of conventional flash memory technologies, and to the compatibility of these technologies with mainstream CMOS technology. The proliferation of portable electronic devices has now spawned demand for ultra-high-density non-volatile semiconductor memory (NVM); consequently, research in this area is necessary (to meet this demand).

Historically, enhancement in flash memory density has been achieved with the use of multi-bit (or multi-level) charge storage algorithms, coupled with device scaling. Further scaling of conventional (floating-gate, SONOS) flash memory devices (beyond the 45nm technology node) is nonetheless a major challenge since these structures are highly susceptible to short-channel effects (SCE) due to their thick gate-stack equivalent oxide thickness (EOT). The tunnel oxide thickness in a floating-gate memory device cannot be too thin (below ~8nm); otherwise stress-induced leakage current (SILC) can cause charge to leak away from the floating gate, resulting in memory volatility. As an

alternative, a SONOS memory device has better scalability than a floating-gate memory device because charge is stored in discrete traps within a non-conductive nitride layer, thereby allowing for more aggressive scaling of its tunnel oxide. Furthermore, SONOS NVM cells avoid the floating-gate coupling interference issue (between adjacent cells) and thus are preferred for future high-density flash memory technologies. Still, a SONOS memory device has a much thicker EOT (~10nm) than a logic device (~2nm) and hence its electrostatic integrity (*i.e.* scalability) will be worse.

7.2 Contributions

In this dissertation, some possible alternatives for scaling flash memory have been proposed and demonstrated. The general approach involves the use of either the use of alternative multi-bit, charge-trapping NVM cell structures that are more scalable, or the use of a novel charge detection method that is less sensitive to disturbance from the complementary bit(s), or both. This section summarizes some of key contributions arising from this work.

The use of a novel charge detection method was demonstrated in a dual-bit, n-channel SONOS FinFET NVM cell. This charge detection method utilizes a change in the cell's off-state (more specifically, GIDL) current to detect charge stored on the bit next to the Drain electrode. This detection method is less sensitive to charge stored in the complementary bit and therefore (its use alone) enhances the scalability of this structure.

A prototype silicon-on-insulator (SOI), dual-bit n-channel FinFET-based NVM cell design with two separate gate-sidewall charge storage sites was presented for the first time. This Gate-Sidewall-Storage (GSS) FinFET NVM cell is in principle more scalable than the conventional SONOS FinFET NVM cell since its charge storage sites are physically separated, and the former utilizes a thinner gate-stack effective oxide thickness (EOT). The proposed structure can utilize either the conventional and/or the new charge detection method (introduced before) to identify the charge storage state of each bit in the cell. The new read method is compatible with a gate-overlapped source/drain structure which offers improved on-state conductance, in contrast to the conventional read method. The dual-bit FinFET cell design can be used to achieve very high NVM storage density because of its high scalability and compatibility with standard CMOS process technology.

A prototype SOI, dual-bit p-channel GSS FinFET NVM cell design was also presented for the first time. Each bit of this structure can be programmed via band-to-band tunneling induced hot electron injection (BBHE) and erased via band-to-band tunneling induced hole injection (BBHI), which make it suitable for NAND-type array architectures. The conventional reverse-read method can be used to identify the state of each bit in the cell. This dual-bit FinFET cell design can be used to achieve very high storage density because of its superior scalability and compatibility with standard CMOS process technology.

Finally, two different NVM cell designs, that utilize a double-gate (DG) FET structure (with either 2 or 4 physically separate charge-storage sites) to store 4 bits of information, were also presented. These cell designs are comprised of an n-channel

double-gate field-effect transistor (DG-FET), modified to include the charge-trapping layers (*e.g.* made of silicon-nitride, poly-Si, or any other material that is able to store charge) embedded either underneath each gate electrode (i.e., the SONOS DG-FET structure) or within the sidewalls of each gate electrode (i.e., the GSS DG-FET structure). A benefit of gate-sidewall charge storage is that the EOT of the gate dielectric (in-between the gate and the channel) can be thinner, which makes this structure even more scalable to very short gate lengths. The symmetry of these structures allows for the independent access or modification of each bit via the conventional read, program, and erase methods. By biasing each gate independently, the information stored at the unselected gate can be decoupled, so that the information at the selected gate can be selectively accessed or modified. The scalability of both cell designs is investigated via numerical device simulation. As expected, the GSS DG-FET structure is more scalable due to its thinner EOT. Read operation of these cells within NOR- and NAND-type array architectures is also discussed. Both cell designs are promising to enhance storage density in future non-volatile memories.

7.3 Suggestions for future work

Practical implementation of the new charge detection method (that is proposed in this dissertation) requires a larger (current or voltage) signal for easy detection. In other words, the magnitude of the GIDL current that was observed on the tested FinFET structures (in the order of tens of nano-amperes) is not large enough for easy detection and requires amplification. GIDL current can be boosted by enhancing the transverse electric field near to the drain electrode, or by enhancing band-to-band tunneling within that region. This can be achieved in practice (for example) by applying a larger gate-to-drain (V_{GD}) bias voltage, by storing more electrons within the charge-trapping layers, by using a gate material with a larger work function (i.e., p+ poly-Si), by using a body with a lower energy band-gap (i.e., germanium instead of silicon), or by enhancing the S/D doping concentration underneath the charge-trapping layers. Each of these different approaches requires further investigation.

In addition, the use of the new charge detection method on the conventional planar, Single-Gated (SG) NVM cells needs to be investigated as well. Clearly, this read method may also be utilized with either the conventional SG SONOS NVM cell (**Figure 2.1b**), or either SG (Gate-Sidewall Storage) structure shown in **Figure 7.1**. An assessment on the potential enhancement scalability of these structures (due to the use of this new read method) needs to be performed in detail.

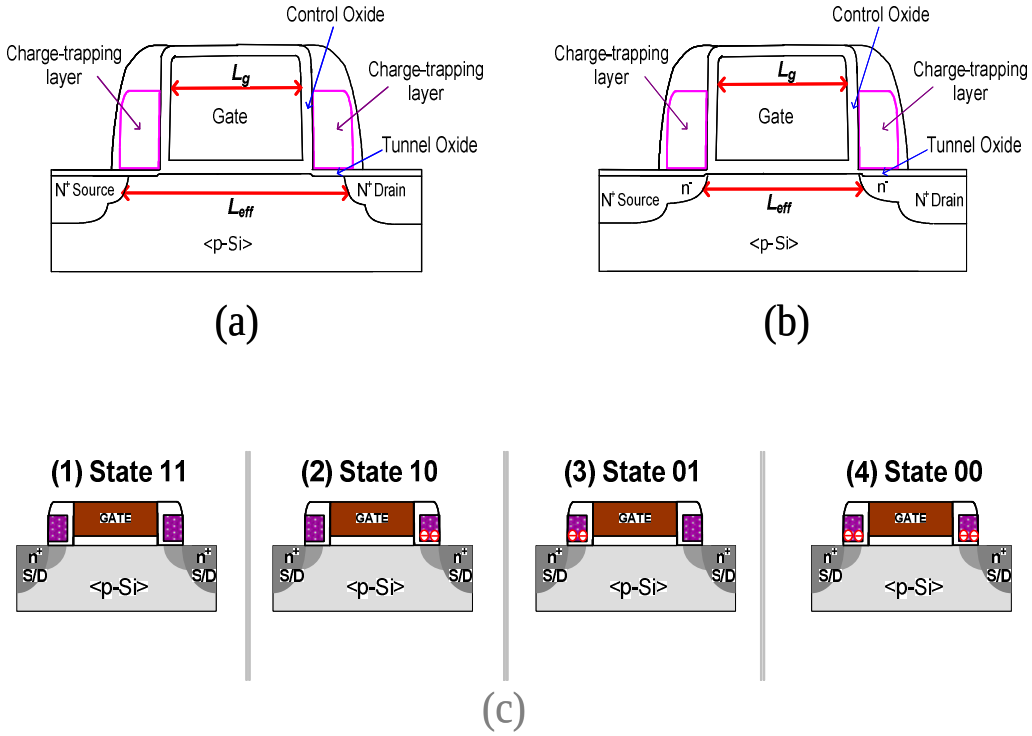


Figure 7.1: 2D schematic cross-sections of (a) a gate-underlapped ($L_{eff} > L_g$), and (b) a gate-overlapped ($L_{eff} \sim L_g$) NVM cell structure with charge storage sites located at the gate sidewalls. (c) Definition of the 4 charge-storage states of this dual-bit cell.

Finally, a new (NOR- or NAND-type) array architecture that utilizes this new read method (either independently, or in conjunction with the conventional read method) needs to be properly defined *and* its functionality needs to be verified accordingly. Implementation of this read method with the proposed 'Virtual Ground in SOI' array architecture (covered in Section 3.6.1) will be a serious challenge since it will be difficult to suppress the (off-state) leakage current contribution from all cells connected to the same Bit-Lines (even when these cells are connected to grounded or negatively-biased word lines to suppress this leakage). Consequently, a thorough investigation in this topic is also necessary.

Appendix A:

Fabrication Process (GSS FinFET NVM cell)

	PROCESS	CONDITIONS	EQUIPMENT	COMMENTS
	Initial wafers	6" (100) p-type 100nm thick SOI with 400 nm BOX		Unibond SOI
0.0	**LABELING:**			
0.1	Label		Diamond pencil	
0.2	DI Rinse	3-cycle DI Water Rinse	Sink6	
0.3	Thickness Measure	Silicon	NanoDuv	R:4
0.4	HF Cleaning	10:1 HF, 5 min.	Sink6	
1.0	**BODY THINNING:**			
1.1	Piranha Cleaning	120 ^{o}C, 10 min.	Sink6	
1.2	Wet Oxidation	Wet O_2, 850 ^{o}C, 100 min.	Tystar2	R: SWETOXB
1.3	Wet Etching	10:1 HF, 10 min.	Sink6	
1.4	Thickness Measure	Silicon	NanoDuv	R:4
1.5	Piranha Cleaning	120 ^{o}C, 10 min.	Sink6	
1.6	Dry Oxidation	Dry O_2, 900 ^{o}C, 4 min.	Tylan6	R: SGATEOX
1.7	Piranha Cleaning	120 ^{o}C, 10 min.	Sink6	
2.0	**ALIGNMENT KEY FORMATION:**			
2.1	PR Coat	Coating(HMDS/Shipley UV 210 .9um/Soft bake 130C 60s)	svgcoat6	(#1/#2/#1)
2.2	Align Key Photo	Exposure (Reticle: 45024001**A066**; LayerID: PM)	asml	
2.3	PR Development	Develop(PEB 130C, 60s/LDD-26W, 45s/No Hard bake)	svgdev6	(#1/#1/#9)
2.4	Thickness Measure	Silicon	NanoDuv	R:4
2.5	Align Key D/E	13mT/200Ws/40Wb/100CF4/70sec (S/E with dm-3)	lam5 ('5003')	Si & Box trench ~120nm
2.6	PR Ashing	3.75T/400W/200C/40% O2/ 1min30sec	matrix ('std')	
2.7	Post Cleaning	Piranha, 120C, 10min	sink8	

	PROCESS	CONDITIONS	EQUIPMENT	COMMENTS
3.0	**FIN FORMATION (spacer lithography)**			
3.1	Pre-cleaning	Piranha, 120C, 10min	sink6	dm-3
3.2	α-Si deposition	SiGeHTD.019 (150nm target) Nucleation: 300mT/200SiH4/530C/1sec Si CVD: 300mT/200SiH4/530C/120min	Tystar19	
3.2	PR Coating	Coating(HMDS/Shipley UV 210 .9um/Soft bake 130C 60s)	svgcoat6	(#1/#2/#1)
3.4	Spacer Island Photo	Exposure LayerID: 'Nwell' (#2) Control Mode:'M"; E=25 mJ/cm^2; Offset=0.07	ASML	Reticle: 'SpacerFinFET'
3.5	PR Development	Develop(PEB 130C, 60s/LDD-26W, 45s/No Hard bake)	svgdev6	(#1/#1/#9)
3.6	PR Hard Bake	Recipe: "U"	uvbake	Recipe: "U"
3.7	α-Si Dry Etch	BT:13mT/200Ws/40Wb/100CF_4/10 s ME:15mT/300Ws/150Wb/50Cl_2/150 HBr/ EPD (delay=5s, Norm=10s, Norm-value=5000, trigger 30%) //max t (step5):30sec OE:15mT/250Ws/120Wb/5O_2/200 HBr/15s	Lam5 expected EPD: 26s/25s (back/front)	Recipe: "5003"
3.8	PR Ashing	3.75T/400W/200C/40% O2/ 1min30sec	Matrix	Standard recipe
3.9	Post Cleaning	Piranha, 120C, 10min	sink8	
3.10	Step height	Standard recipe	asiq	Verify step height
3.11	Inspection	Si surface, profile	leo	Verify CD, alignment.
3.12	Piranha cleaning	Piranha, 120C, 10min	sink6	
3.13	PSG CVD	11SDLTOA, 100sec (450A target)	tystar11	
3.14	PSG Spacer *and* α-Si Dry Etch	ME: 13mT/200Ws/40Wb/100CF_4/EPD + 6sec OE:15mT/250Ws/120Wb/200HBr/ 5O_2/60s	Lam5	
3.15	Inspection	CD, profile	leo	
4.0	**Active Area Definition**			
4.1	PR Coating	Coating(HMDS/Shipley UV 210 .9um/Soft bake 130C 60s)	svgcoat6	(#1/#2/#1)
4.2	S/D Pad Photo	Exposure LayerID: 'PWELL' (#1) Control Mode:'W"; E=25 mJ/cm2;	ASML	Reticle: 'SpacerFinFET'

	PROCESS	CONDITIONS	EQUIPMENT	COMMENTS
4.3	PR Development	Develop(PEB 130C, 60s/LDD-26W, 45s/No Hard bake)	svgdev6	(#1/#1/#9)
4.4	PR Hard Bake	Recipe: “U”	uvbake	Recipe: “U”
4.5	Hard Mask Dry etch	Standard Oxide recipe :10sec	Centura_mxp	Recipe: “mxp_variable_ox”
4.6	Polymer removal	100:1 HF 10sec	sink7	
4.7	Active Area Dry Etch	BT: 13mT/200Ws/40Wb/100CF_4/7sec ME:15mT/300Ws/150Wb/50Cl_2/150 HBr/ 10sec OE:none	Lam5	SOI Dry Etch //EPD: Si
4.8	PR Ashing	3.75T/400W/200C/40% O2/ 1min30sec	matrix	Standard recipe
4.9	Post Cleaning	Piranha, 120C, 10min	sink8	
4.10	Inspection	CD, profile	leo	
4.11	T_{BOX} Measurement	SiO_2 Thickness	nanoduv	
4.12	Step height	Standard recipe	asiq	Verify T_{SOI}
5.0	**GATE-STACK FORMATION**			
5.1	Pre-cleaning	Piranha, 120C, 10min	sink6	dm-3
5.2	Sacrificial oxidation	Dry oxidation: O_2, 850C, 1min; Post-N_2 anneal: 950C 20min	tystar1	//target: 3nm
5.3	SiO_2 Wet etch	25:1 HF Tank (ER~ 110 A/min) => 15 sec	sink6	//remove sac'l oxide
5.4	Dry Oxidation (gate oxide)	Dry oxidation: O_2, 850C, 00:17:40; Post-N_2 anneal: 950C 20min	Tystar1	**//target: 6.0nm**
5.5	In-situ N^+ PolySi CVD	tcvd=65min; Tcvd=615 C; P[=]375mT; PH3 /Si [=]4, FRSIH4[=]120 sccm	Tystar10	R: '10SDPLYA' //target: 100-150nm
5.6	LTO deposition	11SULTOA, 12min, //target: 150nm	tystar11	
5.7	PR Coating	Coating(HMDS/Shipley UV 210 .9um/Soft bake 130C 60s)	svgcoat6	(#1/#2/#1)
5.8	Gate Photo	Exposure LayerID: 'Poly' (#6) Control Mode:'M"; E=25 mJ/cm2; Offset=0.07	ASML	Reticle: ‘DEV_CLR_CRIT’
5.9	PR Development	Develop(PEB 130C, 60s/LDD-26W, 45s/No Hard bake)	svgdev6	(#1/#1/#9)
5.10	PR Hard Bake	Recipe: “U”	uvbake	Recipe: “U”
5.11	Inspection	CD	leo	Verify CD, alignment
5.12	Hard Mask Etch (LTO)	ME:200mT/ 700W /150Ar / 15 CF_4/ 45 CHF_3 /60sec	Centura_mxp	R: ‘MXP_OX_ETCH’
5.13	PR Ashing	3.75T/400W/200C/40% O2/ 1min30sec	matrix	Standard recipe
5.14	Polymer removal	100:1 HF 10sec	sink7	

	PROCESS	CONDITIONS	EQUIPMENT	COMMENTS
5.15	Inspection	Gate alignment	leo	Verify CD, alignment.
5.16	PolySi Dry Etch	BT:13mT/200Ws/40Wb/100CF_4/10 s ME:15mT/300Ws/150Wb/50Cl_2/150 HBr/ EPD (trigger 30%) OE:15mT/250Ws/120Wb/5O_2/200 HBr/15s	Lam5	Etch LTO, N^+ PolySi
5.17	Post Cleaning	Piranha, 120C, 10min	sink8	
5.18	T_{BOX} Measurement	SiO_2 Thickness	nanoduv	T_{BOX} Measurement
5.19	Step height	Standard recipe	asiq	Verify T_{SOI}
5.20	Inspection	Si surface, profile	leo	Verify CD, alignment.
6.0	**ONO Spacer Definition**			
6.1	Pre-cleaning	Piranha, 120C, 10min	sink6	dm-3
6.2	Sacrificial oxidation	Dry oxidation: O_2, 850C, 1min; Post-N_2 anneal: 950C 20min	tystar1	//target: 3nm
6.3	SiO_2 Wet etch	25:1 HF Tank (ER~ 110 A/min) => 20 sec	sink6	//remove sac'l oxide
6.4	Dry Oxidation (gate oxide)	Dry oxidation: O_2, 850C, 1min; Post-N_2 anneal: 950C 20min	Tystar1	**//target: 3.0nm**
6.5	SiRN CVD	t_{CVD}=01:30:00; T_{CVD}=750 C; P[=] 300 mTorr; T_{LOAD}=650 C FR_{NH3} [=]24 sccm; FR_{DCS} [=]25 sccm, FR_{N2} [=]100 sccm; Post-depn NH3 Purge[=]1sec	Tystar9	R: '9VNITA'
6.6	SiRN Spacer Dry Etch	ME:50mT/ 300W / 90Ar / 45 CF_4/ 10 CHF_3 / 10 O_2 /20 Gauss Sine / EPD #2 (~20 sec) OE:50mT/ 450W / 50Ar / 20 CH_3F/ 7 O_2 /~13sec	Centura_MxP	
6.7	Inspection	CD, profile	leo	
6.8	Step height	Standard recipe	asiq	Verify T_{SOI}

	PROCESS	CONDITIONS	EQUIPMENT	COMMENTS
7.0	**CONTACT FORMATION**			
7.1	Pre-cleaning	Piranha, 120C, 10min	sink6	
7.2	N+ S/D IIP	Phosphorus/ 5e15 cm^{-2}/ 15 KeV/ 7 degrees	Implanter	Outside vendor
7.4	PR Coating	Coating(HMDS/Shipley UV 210 .9um/Soft bake 130C 60s)	svgcoat6	(#1/#2/#1)
7.5	Contact Litho	Exposure LayerID: 'NSEL' (#4) Control Mode:'M"; E=25 mJ/cm2; Offset=0.07	ASML	Reticle: 'DEV_DRK_CRIT'
7.6	PR Development	Develop(PEB 130C, 60s/LDD-26W, 45s/No Hard bake)	svgdev6	(#1/#1/#9)
7.7	PR Hard Bake	Recipe: "U"	uvbake	Recipe: "U"
7.8	Inspection	CD	leo	Verify CD, alignment
7.9	Contact opening wet etch	1min in 5:1 BOE (Sink8), plus 1minin 100:1 aq. (5:1) BOE.		
7.10	PR Ashing	3.75T/400W/200C/40% O2/ 1min30sec	matrix	Standard recipe
7.11	Post Cleaning	Piranha, 120C, 10min	sink8, sink6.	
8.0	**FINAL (ANNEALING) STEPS**			
8.1	S/D activation	RTA, N2, 900C, 20sec	heatpulse3	
8.2	Sintering	400C, 30min, 10N2/1H2	tylan18	

Appendix B: Sample Simulation Code for Read Simulations (4-bit DG-FETs)

```
#############################################################
#######                                                     #
#   filename: ADG_finfet_msh.pdm                            #
# This is the Taurus file that defines the geometry, mesh,  #
# and doping profiles for the 4-bit/cell DG_FET device.     #
#                                                           #
# NOTEs: This one is used for READ simulations.             #
#                                                           #
# ap (11/19/05)                                             #
#############################################################

# Definition of constants:  (note: 1[=]FG, 2[=]BG)
Define     (Lgate=.050)
Define     (Lgate2=$Lgate)
Define     (Ltrap=.015)
Define     (Leff=$Lgate)
Define     (WF_BG=4.17)       #Work Fxn for n+PolySi [=]BG
Define     (WF_FG=4.17)       #Work Fxn for p+PolySi [=]FG


Define     (Tpoly=.0195)
Define     (Tox1=.006)
Define     (Tox2=.006)
Define  (Ttox1=.003)
Define  (Ttox2=.003)
Define  (CTox1=.006)
Define  (CTox2=.006)
Define     (Tsi=.020)
Define  (Ttrap=.0225)

Define     (pdope=1e20)
Define     (bdope=1e13)
Define  (halodope=1e13)
Define     (sddope=1e13)            #LDD doping

Define     (sdxchar=.001)
Define     (sdychar=.001)
```

```
# tags:
Define  (const1=0.015)           #0.005 constant!!
#Define  (const1=.005)      #constant used to indicate the
sep'n % Trap_Layer
                     #& the S/D electrodes.
Define  (Lsd =expr(Ltrap + CTox1 + const1) )
Define     (xmin=expr(0-Lsd-Lgate/2))
Define     (xmax=expr(Lsd+Lgate/2))
#Define    (ymin=expr(0-Tox1-Ttrap-CTox1-Tpoly-Tsi/2))
Define  ( ymin=expr(0-Tsi/2-Tox2-Tpoly)  )
#Define    (ymax=expr(Tox2+Ttrap+CTox2+Tpoly+Tsi/2))
Define  (ymax=expr(Tsi/2+Tox1+Tpoly)  )
Define     (xoffset=expr(Lgate/2 + Lsd/2.0) )

#use 'xoffset' below for complete overlap of Xj with Ltrap;
o.w., use
#expression above.
#Define    (xoffset=expr(Lgate/2 + CTox1) )

#for completely underlapped, use:
#Define    (xoffset=expr(Lgate/2 + CTox1+Ltrap+Tox1) )


Define     (halodepth=expr(Lgate/2 + Lsd/2) )

Define     (cont=0.0005)       #constant used to dictate
'thickness' of S/D
                     #electrodes.

# Enable device mode:
Taurus {device}


##############################################################
## i) Structure Generation:
##############################################################


Include(variables)


# Define the device size, list the regions, and specify
fixed mesh lines:

DefineDevice (
  minX=expr(xmin-cont), maxX=expr(xmax+cont),
```

```
  minY=expr(ymin-cont), maxY=expr(ymax+cont),

  Region (name=fin,     material=silicon),
#  Region (name=fin,     material=germanium),

#FG:
  Region (name=trapFG_L,  material=polysilicon),
  Region (name=trapFG_R,  material=polysilicon),
  Region (name=oxideFG,   material=Oxide),
  Region (name=FG,       material=polysilicon),
  Region (name=FG_contact,material=electrode),

#BG:
  Region (name=trapBG_L,  material=polysilicon),
  Region (name=trapBG_R,  material=polysilicon),
  Region (name=oxideBG,   material=Oxide),
  Region (name=BG,       material=polysilicon),
  Region (name=BG_contact,material=electrode),

#S/D:
  Region (name=source,   material=electrode),
  Region (name=drain,    material=electrode),


#'special' mesh lines:
  x=expr(xmin-cont), x=$xmin, x=expr(xmin+cont), x=expr(0-
xoffset),
  x=expr(0-Lgate/2), x=0nm, x=expr(Lgate/2), x=$xoffset,
  x=expr(xmax-cont), x=$xmax, x=expr(xmax+cont),

  y=$ymin, y=expr(ymin/2), y=expr(0-Tsi/2-Tox1), y=expr(0-
Tsi/2), y=0nm,
  y=$ymax, y=expr(Tsi/2+Tox2), y=expr(Tsi/2), y=expr(ymax/2)
)


# Define the silicon substrate region
DefineBoundary (
  region=fin,
  Polygon2D (
    Point (x=$xmin, y=expr(0-Tsi/2)), Point (x=$xmax,
y=expr(0-Tsi/2)),
    Point (x=$xmax, y=expr(Tsi/2)), Point (x= $xmin,
y=expr(Tsi/2))
  )
)
```

```
####################
## BG:
####################

# Define the back oxide region:
DefineBoundary (
  region=oxideBG,
  Polygon2D (
   Point(x=expr(xmin-cont), y=expr(ymin-cont) ),
#Point 5
   Point(x=expr(0-Lgate/2), y=expr(ymin-cont) ),
#Point 6
   Point(x=expr(0-Lgate/2),y=expr(0-Tsi/2-Tox2) ),
#Point 7
   Point (x=expr(Lgate/2), y=expr(0-Tsi/2-Tox2) ),
#Point 8
   Point(x=expr(Lgate/2),y=expr(ymin-cont) ),
#Point 9
   Point(x=expr(xmax+cont), y=expr(ymin-cont) ),
#Point 10
   Point(x=expr(xmax+cont), y=expr(0-Tsi/2)),
#Point 11
   Point(x=expr(xmin-cont),y=expr(0-Tsi/2))
#Point 12
  )
)

# Define TrapBG_L:
DefineBoundary (
  region=trapBG_L,
  Polygon2D (
   Point(x=expr(0-Lgate/2-CTox2-Ltrap), y=expr(0-Tsi/2-
Ttox2-Ttrap) ),    #Point 5b
   Point(x=expr(0-Lgate/2-CTox2), y=expr(0-Tsi/2-Ttox2-
Ttrap) ),         #Point 6b
   Point(x=expr(0-Lgate/2-CTox2), y=expr(0-Tsi/2 -Ttox2) ),
#Point 7b
   Point(x=expr(0-Lgate/2-CTox2-Ltrap), y=expr(0-Tsi/2 -
Ttox2) ),        #Point12b
  )
)

# Define TrapBG_R:
DefineBoundary (
  region=trapBG_R,
  Polygon2D (
```

```
   Point(x=expr(Lgate/2+CTox2), y=expr(0-Tsi/2-Ttox2-
Ttrap) ),           #Point 9b
   Point(x=expr(Lgate/2+CTox2+Ltrap), y=expr(0-Tsi/2-Ttox2-
Ttrap) ),     #Point 10b
   Point(x=expr(Lgate/2+CTox2+Ltrap), y=expr(0-Tsi/2-
Ttox2) ),           #Point 11b
   Point(x=expr(Lgate/2+CTox2),      y=expr(0-Tsi/2 -
Ttox2) )          #Point 8b
  )
)


# Define the BG electrode gate region:
DefineBoundary (
  region=BG,
  Polygon2D (
    Point (x=expr(0-Lgate/2), y=expr(ymin)),
#Point 6
    Point (x=expr(Lgate/2), y=expr(ymin)),
#Point 9
    Point (x=expr(Lgate/2),  y=expr(0-Tsi/2-Tox2) ),
#Point 8
    Point (x=expr(0-Lgate/2),y=expr(0-Tsi/2-Tox2) )
#Point 7
  )
)

######################
## FG:
######################

# Define the front oxide region:
DefineBoundary (
  region=oxideFG,
  Polygon2D (
   Point(x=expr(xmin-cont), y=expr(Tsi/2) ),
#Point 13
   Point(x=expr(xmax+cont), y=expr(Tsi/2) ),
#Point 14
   Point(x=expr(xmax+cont), y=expr(ymax+cont)),
#Point 15
   Point(x=expr(Lgate/2),y=expr(ymax+cont)),
#Point 16
   Point(x=expr(Lgate/2),y=expr(Tsi/2 +Tox1) ),
#Point 17
   Point(x=expr(0-Lgate/2),y=expr(Tsi/2 +Tox1) ),
#Point 18
```

```
   Point(x=expr(0-Lgate/2),y=expr(ymax+cont) ),
#Point 19
   Point(x=expr(xmin-cont), y=expr(ymax+cont) )
#Point20

  )
)

# Define TrapFG_L:
DefineBoundary (
  region=trapFG_L,
  Polygon2D (
   Point(x=expr(0-Lgate/2-CTox1-Ltrap), y=expr(Tsi/2
+Ttox1) ), #Point13b
   Point(x=expr(0-Lgate/2-CTox1), y=expr(Tsi/2 +Ttox1) ),
#Point18b
   Point(x=expr(0-Lgate/2-CTox1),
y=expr(Tsi/2+Ttox1+Ttrap) ),  #Point19b
   Point(x=expr(0-Lgate/2-CTox1-
Ltrap),y=expr(Tsi/2+Ttox1+Ttrap))#Point20b
  )
)

# Define TrapFG_R:
DefineBoundary (
  region=trapFG_R,
  Polygon2D (
   Point(x=expr(Lgate/2 +CTox1), y=expr(Tsi/2 +Ttox1) ),
#Point 17b
   Point(x=expr(Lgate/2+CTox1+Ltrap), y=expr(Tsi/2 +
Ttox1) ), #Point 14b
   Point(x=expr(Lgate/2+CTox1+Ltrap),
y=expr(Tsi/2+Ttox1+Ttrap) ),     #Point 15b
   Point(x=expr(Lgate/2 +CTox1), y=expr(Tsi/2+Ttox1+Ttrap))
#Point 16b
  )
)

# Define the FG electrode gate region:
DefineBoundary (
  region=FG,
    Polygon2D (
    Point(x=expr(0-Lgate/2),y=expr(Tsi/2+Tox1) ),
#Point 18
    Point(x=expr(0+Lgate/2),y=expr(Tsi/2+Tox1) ),
#Point 17
```

```
    Point (x=expr(Lgate/2), y=expr(ymax)),
#Point 16
    Point (x=expr(0-Lgate/2), y=$ymax)
#Point 19
  )
)

###############
# Add'l electrodes:
#############

# Define the source electrode region:
DefineBoundary (
  region=source,
  Polygon2D (
    Point (x=expr(xmin-cont), y=expr(0-Tsi/2)),
    Point (x=expr(xmin), y=expr(0-Tsi/2)),
    Point (x=expr(xmin), y=expr(Tsi/2)),
    Point (x=expr(xmin-cont), y=expr(Tsi/2))
  )
)

# Define the drain electrode region:
DefineBoundary (
  region=drain,
  Polygon2D (
    Point (x=expr(xmax), y=expr(0-Tsi/2)),
    Point (x=expr(xmax+cont), y=expr(0-Tsi/2)),
    Point (x=expr(xmax+cont), y=expr(Tsi/2)),
    Point (x=expr(xmax), y=expr(Tsi/2))
  )
)

# Define the BG_contact electrode region:
DefineBoundary (
  region=BG_contact,
  Polygon2D (
    Point (x=expr(0-Lgate/2), y=expr(ymin-cont)),
    Point (x=expr(Lgate/2),   y=expr(ymin-cont)),
    Point (x=expr(Lgate/2),   y=$ymin),
    Point (x=expr(0-Lgate/2), y=$ymin)
  )
)

# Define the FG_contact electrode region:
DefineBoundary (
  region=FG_contact,
```

```
  Polygon2D (
    Point (x=expr(0-Lgate/2), y=expr(ymax+cont)),
    Point (x=expr(Lgate/2),   y=expr(ymax+cont)),
    Point (x=expr(Lgate/2),   y=$ymax),
    Point (x=expr(0-Lgate/2), y=$ymax)
  )
)

QuantumBox ( name=channelBox,
  minX=-8nm, maxX=8nm,
  minY=expr(0-Tsi/2-Tox1), maxY=expr(Tsi/2+Tox1),
  sliceDirection=Y,
  useBoundaryNodes=true,
  ymesh(depth=10A, h1=2A, h2=.5A),
  ymesh(depth=4nm, h1=2A, h2=4A),
  ymesh(depth=10A, h1=0.5A, h2=2A),
  ElectronSchrodinger( NLadders=2, NSubbands=10),
  HoleSchrodinger( NLadders=2, NSubbands=10)
)

#################################################################
########
## ii) Doping Profile Specs:
#################################################################
########
# Substrate Doping: P-type, Uniform:
Profile (name=Ptype, region=fin, Uniform (value=$bdope))

# FG (n+PolySi) Doping: n-type Uniform:
#Profile (name=Ptype, region=FG, Uniform (value=$pdope))
Profile (name=Ntype, region=FG, Uniform (value=$pdope))

# BG (p+PolySi) Doping: p-type Uniform:
#Profile (name=Ptype, region=BG, Uniform (value=$pdope))
Profile (name=Ntype, region=BG, Uniform (value=$pdope))

# Source doping: N-type Gaussian
Profile (
  name=Ntype, region=fin, addtoexisting=true
  Gauss (
    peakValue=2e20, depthvalue=$bdope, depth=5nm,
lateralRatio=1,
    Polygon (Point (x=expr(xmin),     y=expr(0-Tsi/2)),
             Point (x=expr(xmin),     y=expr(Tsi/2)),
             Point (x=expr(0-xoffset), y=expr(Tsi/2)),
             Point (x=expr(0-xoffset), y=expr(0-Tsi/2))
            )
```

```
  )
)


# Drain doping: N-type Gaussian
Profile (
  name=Ntype, region=fin, addtoexisting=true
  Gauss (
    peakValue=2e20, depthvalue=$bdope, depth=5nm,
lateralRatio=1,
    Polygon (Point (x=expr(xmax),    y=expr(0-Tsi/2)),
            Point (x=expr(xmax),    y=expr(Tsi/2)),
            Point (x=expr(xoffset), y=expr(Tsi/2)),
            Point (x=expr(xoffset), y=expr(0-Tsi/2))
           )
  )
)

################################################################
########
## iii) Mesh Generation:
################################################################
########
# Initial coarse regrid
Regrid (gridProgram=taurus, meshSpacingX=5.0nm,
meshSpacingY=5.0nm) #2.0nm x/y

# Regrid on doping
Regrid (
  gridProgram=taurus, meshSpacing=0.5nm, region=fin,
  Criterion (name=NetDoping, delta=.5, type=asinh)
)

# Regrid in channel
Regrid (
  gridProgram=taurus, region=fin, meshspacingy=1nm,
meshspacingx=2nm,
  minX=expr(xmin), maxX=expr(xmax)
)

# Regrid in channel and Tunnel Ox Regions:
Regrid (
  gridProgram=taurus, meshspacingy=1.0nm, meshspacingx=1nm,
  minX=expr(xmin), maxX=expr(xmax), minY=expr(0-Tsi/2-Tox1-
Tox1), maxY=expr(Tsi/2+Tox2+Tox2)
)
```

```
# Zero-carrier solve at equilibrium
numerics
(
      linearsolver=ilucgs, itresid=1e-6, maxiiter=400,
      linearsolver=ilugmres, linScale=1,iterations=40,
          itresid=5.e-2, relativeerror=1.e-2, maxiiter=400,
      linearsolver=ilugmres, linScale=2,iterations=40,
          itresid=5.e-2, relativeerror=1.e-2, maxiiter=400,
      linearsolver=ilugmres, linScale=3,iterations=40,
           itresid=5.e-2, relativeerror=1.e-2, maxiiter=400
)

Symbolic (carriers=0)
Solve {}

# Save structure
Save (meshfile=ADG_finfet.tdf)

#####################################
# physics models are given below:  #
#####################################

# Select Lombardi mobility model
Physics(
  Global(
   Global(
      FermiStatisticsActive=true
      DirectTunneling(
      PostProcessing=false,
      TCMethod=Gundlach
      #, GridSpacing=5A, MaxDistance=100A,
      CBET( Active=true, Barrier(UseAffinity=true) ),
      #VBHT( Active=true, Barrier(UseAffinity=true) ),
      VBET( Active=true,
      ConductionBarrier( UseAffinity=true ),
      ValenceBarrier( UseAffinity=false, Height=4.30 )
      ))
      )
   )
  Oxide(
   Poissons(
       #ElectronQMModel (
       #active=true,
       #qmmodel=Schrodinger,
       #Schrodinger(
       #  tailActive=true, tailModel=tail2)
       #)
```

```
    )
)

Silicon(
  Poissons(
      Bandgap
     (
     BGNActive=True,
     BGNModel=JainandRoulston
     )
    # ElectronQMModel (
    #active=true,
    #qmmodel=Schrodinger,
    #Schrodinger(
    #   tailActive=true, tailModel=tail1)
    #)
    )
  ElectronContinuity(
    Recombination
    (
    SRHRecombination(
     ElectronLifeTime(
      Cdependent=True
      )
    HoleLifeTime(
      Cdependent=True
      ))
    AugerRecombination
    )
    BTBTunneling( BTBTActive=True, BTBTModelType=2 )
    Mobility(
     LowFieldMobility(
     ConModelActive=True,
     #CCScatteringModelActive=True,
     #CCSModel=DorkelLeturcqModel,
     SurfModelActive=True,
     SurfModel=LombardiSurfaceModel
    )
    #highFieldMobility=true
    HighFieldMobilityActive=True,
    HighFieldMobility (
    HighFieldModel=CaugheyThomasModel,
    )
    )
  )
  HoleContinuity(
    Mobility(
```

```
       LowFieldMobility(
       ConModelActive=True,
       #CCScatteringModelActive=True,
       #CCSModel=DorkelLeturcqModel,
       SurfModelActive=True,
       SurfModel=LombardiSurfaceModel
      )
      #highFieldMobility=true
      HighFieldMobilityActive=True,
      HighFieldMobility (
      HighFieldModel=CaugheyThomasModel,
      )
     )
   )
  )
)

#################################################
# read simulation (state '0000') starts below:  #
#################################################

Include (variables)

# Enable device mode
Taurus {device}

#load meshed  device
DefineDevice (meshfile=ADG_finfet.tdf)

#load physics models:
Include(ADG_finfet_fis.pdm)
numerics (iterations=40)

# Put zero bias on all contacts & set Vd=1.5V:
#SetBias(value=0.0)
{Contact(name=source,type=contactvoltage) }
#SetBias(value=0.05)
{Contact(name=drain,type=contactVoltage) }
Voltage( electrode=BG_contact, value=-1.5 )
Voltage( electrode=FG_contact, value=-3.0 )
Voltage( electrode=source, value=0.0 )
Voltage( electrode=drain,  value=1.5 )
```

```
#############################################################
## ADD CHARGE HERE:
#############################################################

#Add charge to FG, BG  trap layers:

      Interface
(                                       #FG_L
                qf=-05e12
                material(m0=Oxide, m1=polysilicon),
                region (r0=oxideFG, r1=trapFG_L),
                boundingbox(xmin=expr(0-Lgate/2-CTox2-
Ltrap),
                            xmax=expr(0-Lgate/2-CTox2) )
            )

      Interface (
                qf=-05.0e12                             #BG_L
                material(m0=Oxide, m1=polysilicon),
                region (r0=oxideBG, r1=trapBG_L),
                boundingbox(xmin=expr(0-Lgate/2-CTox1-
Ltrap),
                            xmax=expr(0-Lgate/2-CTox1) )
            )


      Interface
(                                       #FG_R
                qf=-05e12
                material(m0=Oxide, m1=polysilicon),
                region (r0=oxideFG, r1=trapFG_R),
                boundingbox(xmin=expr(Lgate/2+CTox2),

xmax=expr(Lgate/2+CTox2+Ltrap) )
            )

      Interface
(                                       #BG_R
                qf=-05.0e12
                material(m0=Oxide, m1=polysilicon),
                region (r0=oxideBG, r1=trapBG_R),
                boundingbox(xmin=expr(Lgate/2 + CTox1),
                            xmax=expr(Lgate/2 + CTox1 +
Ltrap) )
            )
```

```
#################################################################
Symbolic (carriers=1 electron=true)
# Initial-condition, Poisson-only solve:
Solve{ couple(iterations=40,LinearSolver=iterative)
{Poissons}}
#Solve{ couple(iterations=20,LinearSolver=iterative)
{Poissons, electroncontinuity, holecontinuity}}
Solve
{
      Ramp (
            logfile=ADG_finfet_idvg0000.data,
            # Sweep both FG, BG from -1V to 2.5V:
#           Voltage (electrode=BG_contact, startValue=-1.5,
vStep=.1, nSteps= 0 )
            Voltage (electrode=FG_contact, startValue=-3.0,
vStep=.1, nSteps= 65 )
           )
      { couple {Poissons, ElectronContinuity} }
}

#################################
#extract VT:    //remember: FG turns 'on' first.
#################################
Extract (ThresholdVoltage (gateContact=FG_contact,
drainContact=drain) )

# Save final result
Save( MeshFile=ADG_finfet_0000.tdf )
```