1

INTRODUCTION TO NONVOLATILE MEMORY

Joe E. Brewer

COPYRIGHTED MATERIAL

1.1 INTRODUCTION

In this introductory chapter the ABCs of nonvolatile memory are reviewed. The purpose of this elementary discussion is to provide the perspective necessary for understanding the much more detailed chapters that follow. The emphasis is on communication of an overview, rather than on specifics of implementation. Simple memory concepts and terminology are presented, the parameters and features unique to nonvolatile memory (NVM for short) are examined, generic Flash memory variants are described, and finally the treatment of NVM in the International Technology Roadmap for Semiconductors (ITRS) is described.

Semiconductor memory is an essential part of modern information processors, and like all silicon technology it has been more or less growing in density and performance in accordance with Moore's law. Semiconductor NVM technology is a major subset of solid-state memory. Nonvolatility, of course, means that the contents of the memory are retained when power is removed. This book provides an in-depth description of semiconductor-based nonvolatile memory including basic physics, design, manufacture, reliability, and application. Flash memory is emphasized because for a long period of time it has been the dominate form of NVM both in terms of production volume and magnitude of sales dollars. Flash, however, is not the only alternative, and this book also attempts to describe some of the many NVM technologies that have some hope of achieving success in the marketplace.

Nonvolatile Memory Technologies with Emphasis on Flash. Edited by J. E. Brewer and M. Gill
Copyright © 2008 the Institute of Electrical and Electronics Engineers, Inc.

1.2 ELEMENTARY MEMORY CONCEPTS

All information processing can be viewed as consisting of the sequential actions of sensing, interpreting/processing, and acting. These actions cannot be accomplished without somehow remembering the item of interest at least long enough to allow the operations to take place, and most likely much longer to allow convenient use of the raw data and/or the end results.

The length of time that the memory can retain the data is the property called *retention*, and the *unpowered retention* time parameter is the measure of *nonvolatility*. A volatile memory will typically have a worst-case retention time of less than a second. A nonvolatile memory is usually specified as meeting a worst-case unpowered retention time of 10 years, but this parameter can vary from days to years depending on the specific memory technology and application.

Integrated circuit nonvolatile memories are frequently classified in terms of the degree of functional flexibility available for modification of stored contents. Table 1.1 summarizes the categories currently in frequent use [1]. This class of memory was evolved from ultraviolet (UV) erasable read-only memory (ROM) devices, and thus the category labels contain "ROM" as a somewhat awkward reminder of that heritage.

Flash memory [2] is an EEPROM where the entire chip or a subarray within the chip may be erased at one time. There are many variants of Flash, but present-day production is dominated by two types: NAND Flash, which is oriented toward data-block storage applications, and common ground NOR Flash, which is suited for code and word addressable data storage.

In general, information processing requires memory, but it is not at all clear that any constraints are placed on the structure or location of the storage relative to the processing elements. That is, the memory may be a separate entity and entirely different technology than the logic, or it may be that the logic is embedded in the memory and be a technology compatible with the logic, or any combination thereof.

At the heart of every memory is some measurable attribute that can assume at least two relatively stable states. Many common memory devices are *charge based* where charge can be injected into or removed from a critical region of a device, and

TABLE 1.1. Nonvolatile Memory Functional Capability Classifications

Acronym	Definition	Description
ROM	Read-only memory	Memory contents defined during manufacture and not modifiable.
EPROM	Erasable programmable ROM	Memory is erased by exposure to UV light and programmed electrically.
EEPROM	Electrically erasable programmable ROM	Memory can be both erased and programmed electrically. The use of "EE" implies block erasure rather than byte erasable.
E^2PROM	Electrically erasable programmable ROM	Memory can be both erased and programmed electrically as for EEPROM, but the use of "E^2" implies byte alterability rather than block erasable.

the presence or absence of the charge can be sensed. The process of setting the charge level is called *writing*, and the process of sensing the charge level is called *reading*. Alternatively, the write operation may be referred to as the *store* operation, and the read operation may be called the *recall* operation.

Dynamic random-access memory (DRAM), a volatile technology, uses charge stored on a capacitor to represent information. Charge levels greater than a certain threshold amount can represent a logic ONE, and charge levels less than another threshold amount can represent a logic ZERO. The two critical levels are chosen to assure unambiguous interpretation of a ZERO or ONE in the presence of normal noise levels. (Here the higher charge level has been called a ONE, but it is arbitrary which level is defined to be the ONE or ZERO.)

Leakage currents and various disturb effects limit the length of time that the capacitor can hold charge, and thus limits “powered” retention to short periods. The word “dynamic” in the name “DRAM” indicates this lack of ability to hold data continuously even while the circuit is connected to power. Each time the data is read, it must be rewritten in order to assure retention, and regular data refresh operations must be performed when the cell is idle. Worst-case retention time (i.e., the shortest retention time for any cell within the chip) is typically specified as about 60 ms. DRAM is a volatile memory in terms of “unpowered” retention because the charge is not maintained when the circuit power supply is turned off.

Flash memory makes use of charge stored on a floating gate to accomplish nonvolatile data storage. Figure 1.1 provides a cartoon cross-section sketch of a floating-gate transistor and its circuit symbol representation. The floating-gate electrode usually consists of a polysilicon layer formed within the gate insulator of a field-effect transistor between the normal gate electrode (the control gate) and the channel. The amount of charge on the floating gate will determine whether the transistor will conduct when a fixed set of “read” bias conditions are applied. The fact that the floating gate is completely surrounded by insulators allows it to retain charge for a long period of time independent of whether the circuit power supply voltage is present. The act of reading the data can be performed without loss of the information.

Figure 1.2 compares an imaginary idealized transistor with no charge layer in the gate insulator with a transistor that has a charge per unit area, Q, at distance d from the silicon channel surface. The impact of the charge on the threshold voltage depends on the amount of charge per unit area, its distance from the silicon surface, and the permittivity of the insulator between the charge and the silicon. In a Flash device the floating gate provides a convenient site for the charge, but other means may serve the same purpose. For example, in silicon oxide nitride oxide silicon (SONOS) transistors the charge will reside in traps within the nitride layer.

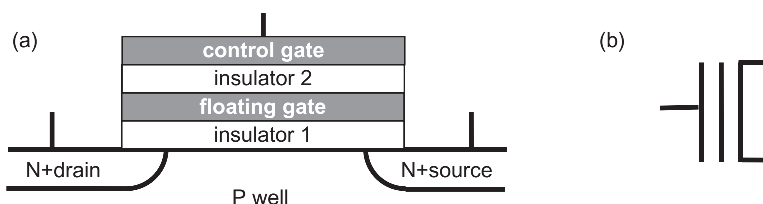

Figure 1.1. Floating-gate transistor: (*a*) elements of the transistor structure and (*b*) circuit symbol.

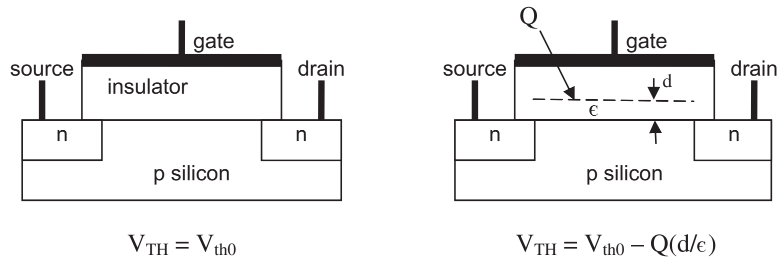

Figure 1.2. V_{TH} shift due to charge in gate insulator.

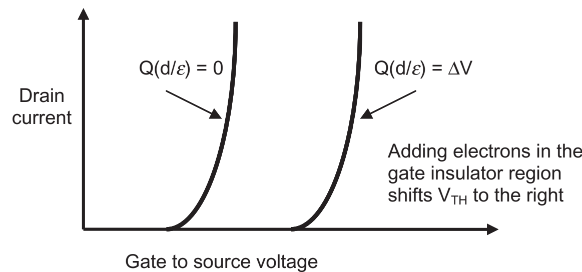

Figure 1.3. Shift of current–voltage characteristics because of inserted charge.

The threshold voltage, of course, impacts how the source-to-drain current of the transistor will change as a function of change in the gate-to-source voltage. Figure 1.3 shows how the characteristic curves can be made to shift as a function of the stored charge. As electrons are added to the charge within the gate insulator region, the curve will move in a positive direction.

For some memory technologies the process of reading destroys the data. This is referred to as *destructive readout* (DRO). For other technologies, Flash, for example, readout can be accomplished without significantly disturbing the data. This is referred to as *nondestructive readout* (NDRO). DRO memory has the disadvantage of requiring that every read operation must be followed by a write operation to restore the data.

Over time some forms of memory organization have been so firmly established that most engineers immediately assume those structures and the parameters that characterize those structures as being the norm. Probably the most pervasive assumption is that the information to be stored and recalled is in a digital binary format.

There are several schemes where a single transistor may be used to store more than just one bit. One approach is to store charge in physically separated parts of the device that can be sensed separately. Currently, the most common example of this concept is the NROM cell discussed later in this chapter. Another approach is to interpret the amount of charge stored in one physical location in the device as a representation of a multibit binary number. In this case the sensing process must reliably distinguish between different quantities of stored charge and the readout process must generate the corresponding binary number.

Consider the “one physical location” approach. A 1-bit-per-cell arrangement is a robust form of storage that allows relaxed margins and comparatively simple

sensing circuitry. The read current needs only to be unambiguously above or below a preset value in order to establish whether a ONE or a ZERO was stored. For a 2-bit-per-cell memory, the recall process must reliably distinguish between four preset levels of charge, and the readout circuitry must translate the detected level into a 2-bit digital format. The storage process and the protection of the cell from disturb conditions must accurately set and maintain those four levels under all operating and nonoperating storage conditions. Considering that the usual requirement is for nonvolatile data retention for 10 years, assuring stability of charge levels and circuit characteristics is quite a challenge. Of course, the complexity rapidly increases as a cell is required to store larger numbers of bits. For example, a 4-bit-per-cell memory must reliably cope with 16 levels and still meet all specifications.

While the heart of a semiconductor memory is the cell, the surrounding circuitry is the mechanism that makes it usable. For economic reasons, cells are packed as close together in a rectangular planar array as available integrated circuit technology and noise management concerns will allow. This X, Y array arrangement contains the cells and conductive lines that allow access to each individual cell.

The lines that run in the X direction (rows) are called *wordlines*, and they are used to select a row of cells during the write or read operations. Wordlines tie to the control gates of the cells in a row. The lines that run in the Y direction (columns) are called *bitlines*. As shown in Figure 1.4, bitline connections for the NOR array architecture are tied to drain terminals of devices in a column. One end of a bitline connects to power and then goes through the array to sensing and writing circuitry. Thus the wordlines activate a specific group of cells in a row, and the bitlines for each intersecting column connect those cells to read and write circuits.

In the example of a NOR architecture, the cells in a column are connected in parallel where all the drain terminals tie to a bitline and all the source terminals tie to a common source line (ground). In this configuration a positive read mode voltage on one wordline while all other wordlines are at zero volts will result in a bitline current that is a function only of the selected row of cells.

This, of course, assumes that a zero control gate-to-source voltage actually turns off the unselected rows. For the NOR organization it is important that the process of initializing the array, called *erase*, not proceed to the point of overerasing transistors to the extent that the threshold voltage becomes negative and the transistors change from operation in the enhancement mode to operation in the depletion mode.

The NAND array architecture, shown in Figure 1.5, achieves higher packing density. Here the bitlines are formed using series-connected strings of cells that do not require contact holes. A string is typically 8, 16, or 32 cells long. If other strings

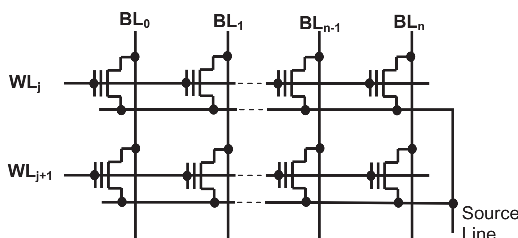

Figure 1.4. NOR array architecture.

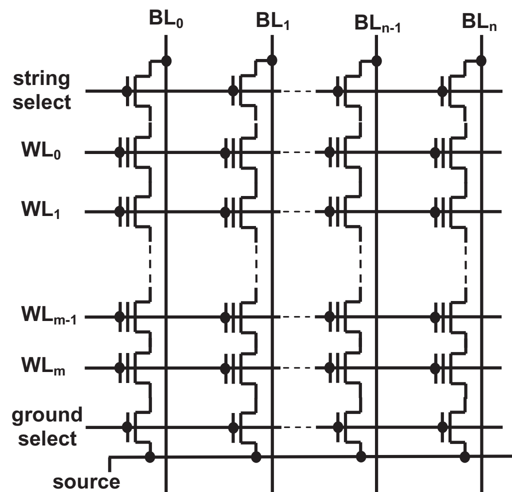

Figure 1.5. NAND array architecture.

were to be tied to the same bitlines, they would be connected in parallel between the bitline and the common source line in a manner similar to individual cells in the NOR architecture.

Reading and writing a NAND device is more complicated than dealing with a NOR device. The wordlines are used to select the transistor of interest in the string. To access data from a string, the reading process requires that all nonselected transistors be turned on while only the selected transistor is allowed to influence the current flow through the string. In contrast to the NOR architecture, it is not objectionable to allow the transistors to be shifted into depletion mode. There are, however, problems if transistors are shifted too far in the enhancement direction. It is important that the distribution of threshold voltages for the programmed state be limited to a specified design range in order to assure proper circuit operation.

There are a number of basic principles of operation that apply to both NOR and NAND organized devices. For a single read operation the individual bits that form a word are made to appear in an input/output register. For a single write operation the word present in an input/output register is used to determine the data inserted into the cells for that word. Reading or writing processes must be designed such that unselected cells are not disturbed while the selected cells are operated on. The design of the array must contend with basic circuit design issues associated with driving heavily loaded transmission lines as well as assuring proper operation of each individual cell.

In order to select a given row and column an integrated memory device is usually provided with a binary address word from external circuitry. The address word is routed to address decoder circuitry that is tightly tied to the sides of the array, and designed to drive the word and bitlines. For reasons of management of loading and data grouping considerations, large memory chips are usually partitioned into many arrays.

A modern integrated memory device incorporates control circuitry that accepts relatively simple commands as inputs and generates timed sequences of signals to accomplish writing, reading, and various other modes of operation. Also, the writing

operations may require voltages that differ from the readily available normal logic supply voltages. In order to simplify the use of the device these voltages are typically generated on-chip. This feature also has the advantage of allowing the semiconductor manufacturer to assure that proper voltages and pulse sequences are applied to the memory cells during all operating conditions without having to depend on decisions made by the end user.

The data is arranged in binary *words* of some prescribed length and structure, and the words may be handled in groups usually called *blocks* or *sectors*. Figure 1.6 summarizes in a black box format the functional interfaces of present-day semiconductor memories. This diagram can be applied to a memory chip, a memory module, or a complete memory system.

While the diagram hides the internal complexity of the memory, the performance parameters for all memories can be understood in the context of this simple block diagram. Memory inputs consist of data, address, and control signals. Memory outputs consist of data and status signals.

The major dynamic performance parameter of a memory is the time between stable address input and stable output data. This latency is the key parameter that determines information processor performance. In a memory that is completely word oriented this delay is called the *access time*. By definition, in a *random-access memory* (RAM) the access times have the same value independent of the address. For memories that are block oriented, the memory performance is characterized by the latency parameter and a data flow parameter. That is, the memory has a delay (latency) before the first word or byte appears at the data out signal terminal, and then successive words or bytes can be clocked out at some byte/second rate.

For writing, the important dynamic characteristic is the time between stable address and input data and the time when the chip can accept a new set of input data. A status output signal (e.g., ready/busy signal) is usually provided to signal when a new operation can be initiated.

Data may be organized within the memory system in various ways. It has been a frequent practice to use chips that have a one-bit-wide output, and use separate chips to provide the bits that make up a word. This is convenient for reliability purposes where spare bits (chips) can be added to form a coded word such that all single-bit errors can be detected and corrected. In this manner single-chip failures can be tolerated, and system reliability is enhanced. For large bit capacity memory chips this approach can make it difficult to efficiently implement small memory systems. Chips 4 bits, 8 bits, or 16 bits wide are more appropriate in such cases.

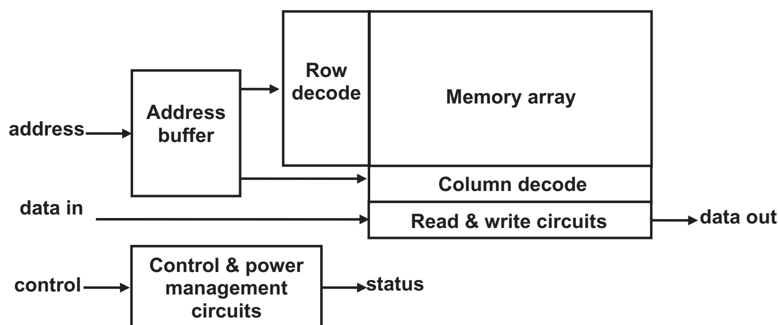

Figure 1.6. Memory functional block diagram.

For block- or sector-oriented chips it is common to serially input and output data in 8-bit-wide bytes. Internal to a chip the bytes are assembled into words and assigned to sectors. Reading and writing are accomplished by loading command and address information and then initiating an automatic sequence that accomplishes the commanded functions and moves and formats the data. Typically these on-chip automatic control sequence signal generators are implemented as state machines.

Historically, the design of information processors has been an art form where engineers made use of whatever logic and memory technologies were currently available and married the technologies to achieve viable systems. The Von Neumann model for a stored program computer was an abstraction that helped engineers to visualize specific forms of computing elements and begin to build practical systems. An accompanying concept, the memory hierarchy, provided a way to work around the limitations of whatever technology was currently available and achieve improved performance–cost trade-offs.

Figure 1.7 shows the 2005 typical hierarchy for a personal computer. The notion of a hierarchy is that a central processing unit (CPU) within a processor is serviced by a series of memory technologies where the memory closest to the CPU, the cache, provides very fast access to instructions and data. The information stored in cache is an image of a portion of the data stored in the next highest level of memory, that is, the information immediately needed by the program code being executed. The much larger, slower, less expensive main, or primary, memory stores the bulk of the programs and data to be processed at the present time. This relationship also occurs between the primary and secondary memory. The primary memory retrieves program and code from the secondary store and writes information to be retained in the form of files back to the secondary store. The cache memory is word oriented (addressable to the word level). The primary memory is both word and block oriented. It can be addressed at the word level and then roll out a page (block) of data in a burst mode. The secondary memory is nonvolatile and block oriented.

Variants of the diagram in Figure 1.7 appear in most texts on computer architecture. Note that nonvolatile semiconductor memory does not explicitly appear in Figure 1.7. For personal computers the nonvolatile magnetic disk is currently a much lower cost option for secondary memory as long as the use environment does not impose high mechanical stresses. NVM semiconductor drives can and do perform the same functions as the magnetic secondary store where requirements other than cost allow them to compete. For personal computers the semiconductor option is firmly established as relatively small manually transportable storage that can be accessed via a universal serial bus (USB) port, and for physically small mobile

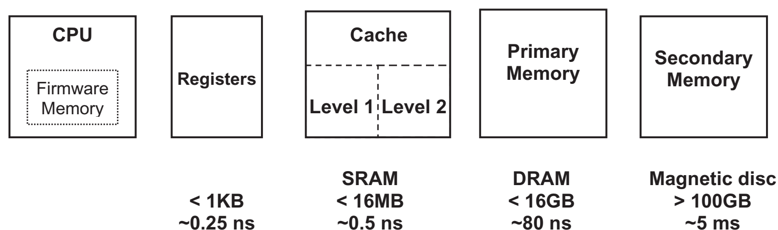

Figure 1.7. Memory hierarchy showing approximate latency and byte capacity bounds.

devices such as digital cameras the semiconductor option is dominate. As the integrated circuit technology progresses to densities where multiprocessors per chip dominate the market, it is likely that a mix of memory technologies (including nonvolatile memory) will be included on the devices, and most of the chip area will be devoted to memory circuitry.

Information processing systems dedicated to specific functions take different approaches to the mix of logic and memory. Engineering of processing systems is an ad hoc art focused on the specific needs of the product in question. Design for items intended to be produced in volume at low cost may attempt to meet requirements using a single processor chip that includes embedded solid-state memory of various kinds. High-performance systems may use multiple processors and a finely tuned mix of supporting memory in order to meet processing throughput requirements. Because the capabilities and options for both logic and memory grow rapidly, the system designer is always faced with a moving baseline. The optimum design of today most probably will not be optimum tomorrow.

1.3 UNIQUE ASPECTS OF NONVOLATILE MEMORY

Nonvolatile memory devices, like all semiconductor memory technology, have electrical alternating current (ac) and direct current (dc) characteristics that can be specified and measured, thermal and mechanical ratings, and reliability characteristics. However, the requirement for nonvolatile retention of data is unique to NVM. Each NVM technology makes use of some physical attribute to achieve nonvolatility, and the particular mechanism employed brings with it a number of technology-specific characteristics.

Device data sheets need to specify retention, and either the data sheet or supporting application information should explain the basis for retention claims. Stresses associated with the processes of writing and reading an NVM device may degrade retention capability and/or other properties of a device. This stress is generally related to the number of writes and/or reads and is summarized in an "endurance" specification stated in terms of the number of write and/or reads that can be tolerated. These matters may be different in nature for different NVM technologies. The particular features of Flash are examined here to illustrate the major issues involved.

1.3.1 Storage

Perhaps the easiest way to understand the nature of the floating-gate memory device is to consider the energy levels involved. Figure 1.8 shows the band structure for a simple floating-gate device where the silicon substrate is shown on the left. The N-type control gate is on the right, and an N-type polysilicon floating gate is in the middle, sandwiched between two silicon dioxide layers. The floating gate, embedded within insulators, is isolated from the exit or entry of charge by the high-energy barrier between the conduction band in the polysilicon and the conduction bands in the top and bottom SiO_2 layers. These barriers, much greater than the thermal energy, provide nonvolatile retention of the charge. In order to change the amount of charge stored on the floating gate it is necessary to change the potential of the floating gate relative to the potential on the opposite side of either SiO_2 layer until

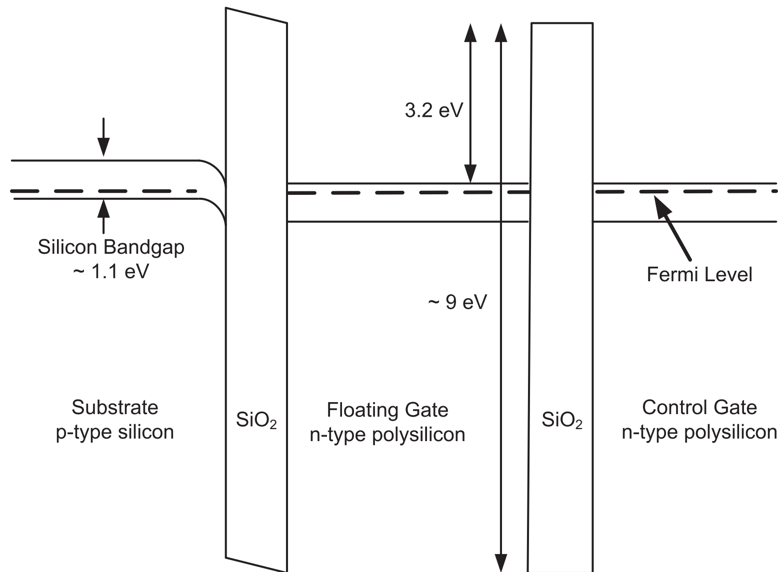

Figure 1.8. Energy band diagram for a typical floating-gate structure.

some conduction mechanism is invoked that can overcome or tunnel through the barrier.

Different strategies may be selected to overcome the energy barriers. Two conduction mechanisms in common use are channel hot-electron (CHE) injection and Fowler–Nordheim (FN) tunneling. Hot carrier injection may be used to add electrons to a floating gate (i.e., programming). FN tunneling may be used to remove or add electrons to a floating gate (i.e., erase or program). Present-day NOR devices typically use CHE to program floating gates and FN to erase floating gates. NAND devices employ FN for both program and erase.

To invoke CHE injection a lateral channel electric field on the order of 10^5 V/cm is required to accelerate electrons to energy levels above the barrier height. Some of those electrons will be scattered by the lattice and be directed toward the floating gate. To actually reach the floating gate, the scattered electrons must retain sufficient energy to surmount the silicon to insulator barrier and cross the insulator to the floating gate. CHE is an inefficient method in that less than about 0.001% of the channel current will be directed to the floating gate.

Fowler–Nordheim tunneling of cold electrons will occur when a field on the order of 8 to 10 MV/cm is established across the insulator next to the floating gate. The process is slower than CHE injection, but it is better controlled and more efficient.

The floating-gate memory transistor can be viewed as a capacitively coupled network as illustrated in Figure 1.9. A pulse applied to the control gate (or any other terminal) will be capacitively coupled to the floating gate, and the resulting potential on the floating gate relative to the other terminals can cause the movement of electrons to or from the floating gate.

Using the terminology of Figure 1.9, the total capacitance on the floating gate, C_t, is the sum $C_g + C_d + C_b + C_s$. If the network is driven from the control gate, the coupling ratio, k_g, would be C_g/C_t. If the network is driven from the drain, the

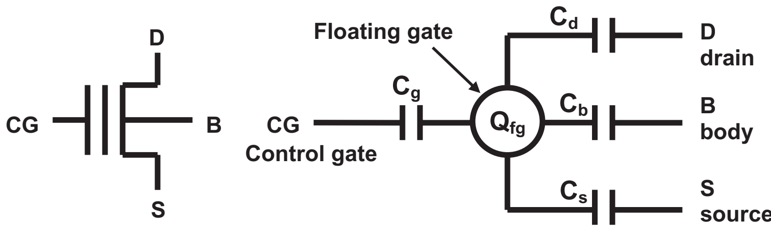

Figure 1.9. Capacitor model for a floating-gate transistor.

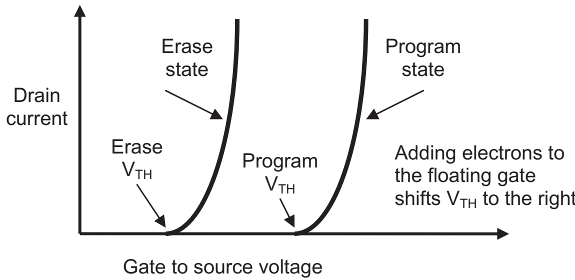

Figure 1.10. Erase state and program state current–voltage characteristics.

coupling ratio, k_d, would be C_d/C_t. In a similar fashion the coupling coefficients k_b and k_s would be C_b/C_t and C_s/C_t. The voltage on the floating gate resulting from pulses applied to the four terminals would be

$$V_{fg} = k_g V_{CG} + k_d V_D + k_s V_S + k_b V_B$$

The usual process of writing (storing) data into a Flash memory requires two operations. First, all of the cells in a common tub (i.e., a sector) are "erased" to initialize the state of the cells. Erasing refers to the removal of all charge from the floating gates. By convention this is usually taken to mean that all of the cells have been cleared to a ONE state. Second, the cells within the tub that are selected by a row address are "programmed" to ZERO or left at ONE in accord with the input data signals. The programming operation may continue over row addresses until all of the data sites in the sector have been programmed.

Erasing moves the threshold voltage in a negative direction while writing moves the threshold voltage in a positive direction. Figure 1.10 shows the relative relationship of the erase and program states in terms of the resulting current–voltage characteristics. Both the magnitude of shift and the statistical distribution of threshold voltages after an erase or program is a major design issue.

The erase procedure is complicated by several concerns. First, erasing (removal of electrons) shifts the threshold voltage negatively from a positive value toward a value nearer to zero. If continued too far, the threshold voltage can go through zero and become negative. This is referred to as *overerase*, and the transistor changes from enhancement mode to depletion mode. In NOR arrays depletion must be avoided in order for the array to work properly.

Second, the erase process is sensitive to the initial state of the transistor. Erase voltages are applied to transistors located in the same tub simultaneously. If some

of those transistors are in a programmed state while others are in the erased state, the resulting final threshold voltages will have quite a spread. It is likely that for some devices overerase will occur. To manage this situation, it is common practice to program all the transistors before applying an erase voltage.

Third, caution should be exercised to avoid overstressing the insulator between the floating gate and the channel. A transistor in the programmed state has a static charge on the floating gate. A fast transient erase pulse to the device will be coupled through the capacitive divider circuit of Figure 1.9 and add directly to the potential across the insulator. This can result in very high electric fields, and repeated application of this stress can rapidly degrade the insulator. Slowing the leading edge of the pulse will allow the floating gate to begin to discharge as the applied voltage increases and avoid the peak stress condition. Because modern Flash chips generate erase pulses on-chip, this risk is easily managed by the chip designer.

1.3.2 Storage Mechanisms

Each nonvolatile memory technology exploits some physical or chemical approach to capture and retains a representation of information independent of the presence of a power source. Many technology options are charge based. Flash memory, as pointed out in the previous examples in this chapter, makes use of charge held on a conductive floating gate surrounded by insulators as its storage mechanism. Variants of the floating-gate approach such as nanocrystal memory operate similarly. Another common technology, SONOS (silicon oxide nitride oxide silicon) uses charge retained in nitride traps as its storage mechanism. Single and few electron transistor approaches confine charge on small islands.

There are several non-charge-based approaches being actively pursued. Ferroelectric memories make use of the switchable polarization of certain materials to store information and detect the resultant change in capacitance. Solid-state magnetic memory approaches currently being explored in integrated circuit form are based on either the giant magneto-resistance effect or magnetic tunnel junctions. In both cases readout is accomplished by detecting a resistance change. Phase change memories based on chalcogenide materials are reversibly switched between a low-resistance crystalline state and an amorphous high-resistance state through the application of heat.

This incomplete list of mechanisms is intended to convey a notion of the broad scope of approaches. It has been noted by many workers in the field that whenever it is observed that a material exhibits two or more relatively stable switchable states, that material has been considered a candidate for forming a memory technology.

1.3.3 Retention

The elapsed time between data storage and the first erroneous readout of the data is the *retention time*. Each nonvolatile storage technology employs a particular storage mechanism, and properties associated with that mechanism and its implementation format will determine the retention characteristics of the device. For Flash the storage mechanism is to represent data by quantities of charge held on a floating gate.

Each technology can be expected to have some natural processes where the data representation changes with time. Flash has some intrinsic charge decay char-

acteristics that define the ultimate retention potential of the approach. Natural decay tracking points to very large retention on the order of thousands or millions of years. Natural decay is so slow as to not be a factor in determining practical retention specifications. At the present time a typical unpowered retention specification is 10 years.

Defects associated with materials, details of device geometry, or aspects of circuit design can impact retention. Each of these three potential problem areas can result in the addition or removal of charge to/from the floating gate. Gate insulator or interpoly insulator defects are typical causes of retention degradation. Phenomena associated with ionic contamination or traps can also be contributing factors. Management of these issues is a function of the general state of the integrated circuit reliability arts and of the specific practices and equipment of given manufacturers.

1.3.4 Endurance

Achievement of nonvolatility depends on exploitation of some natural phenomena peculiar to a given technology approach. In most nonvolatile technologies the normal processes employed to write and/or read cells will result in stresses that eventually degrade the properties of the memory or disturb the contents of the memory. *Endurance* is the term used to describe the ability of a device to withstand these stresses, and it is quantified as a minimum number of erase–write cycles or write–read cycles that the chip can be expected to survive. For quite a number of years the industry has used 100,000 cycles as the minimum competitive endurance requirement.

Knowledge of the reliability physics of the memory technology is essential in order to develop meaningful endurance specifications. The endurance capabilities of a device are a function of both the intrinsic properties of the technology and the quality control of the production line. The impact of cycling stress on retention is a key aspect of endurance, and window closure or shift is a concern. The *window* for a Flash memory is the difference between the erased state threshold voltage and the programmed state threshold voltage. This is the difference that must be reliably detected by a readout sense amplifier.

Both retention and endurance pose difficult test and verification quandaries. Economics and the limits of the human life span make direct observation of retention periods of decades impractical. The process of testing endurance by cycling each part implies that the testing would consume much of the useful life time of the product.

1.4 FLASH MEMORY AND FLASH CELL VARIATIONS

Flash has been used as the example technology throughout the discussions above where a simple stacked gate transistor was used to illustrate several points. In practice the design of Flash cells has several flavors each of which has advantages and disadvantages. Figure 1.11 groups some of the existing Flash varieties by addressing method and by program and erase technique (see Figure 5 in Bez et al. [2]). As mentioned earlier, the common ground NOR used for both code and data storage, and the NAND used for mass storage, presently dominate the market.

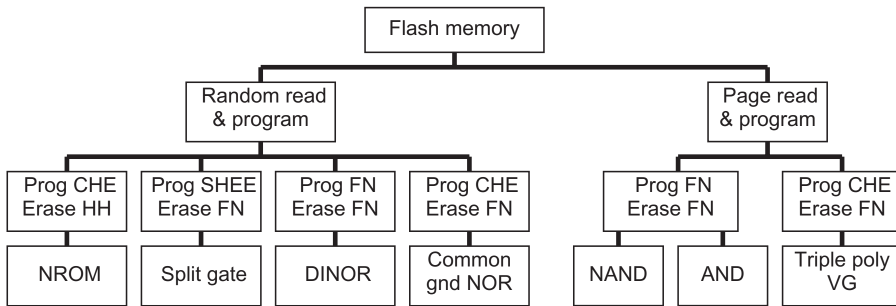

Figure 1.11. Flash cell architecture family tree suggested by [2].

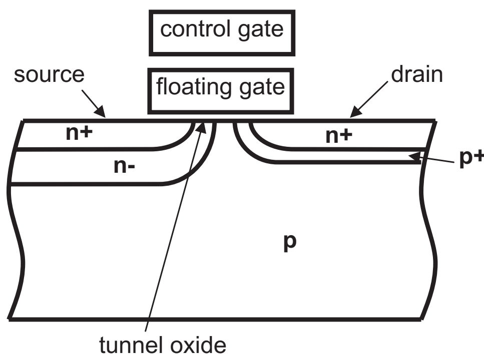

Figure 1.12. Source-erase stacked gate Flash cell [7].

The first modern Flash EEPROM was proposed by Masuoka et al. of Toshiba [3] at the 1984 IEEE International Electron Devices Meeting. The term *Flash* was coined to indicate that the complete memory array is erased quickly. This label is generally applied to all EEPROM devices where a block or page of data is erased at the same time. As indicated by the Figure 1.11 family tree, a large (and growing) number of variations of Flash have emerged that differ in the combination of cell structures, erase and program schemes, and array circuits. (These circuits are described later in this book.) The history of Flash [4] has been characterized by constant and aggressive introduction of modifications that seek to improve density, cost, performance, and reliability.

The Toshiba Flash cell [3, 5, 6] accomplished erase by having a special erase gate. Electrons were removed from the polysilicon floating gate by field emission directly to the polysilicon erase gate in response to a positive voltage pulse. A source-erase stacked gate Flash was proposed in 1985 by Mukherjee et al. [7]. As shown in Figure 1.12, this device looked like a conventional stacked gate UV-EEPROM with two modifications. The source junction was graded to support high voltages, and the oxide under the floating gate was thinned to allow FN tunneling.

It was not until 1988 when the reliability of the emerging devices was proven that volume production really began [8], and that structure, named ETOX (Intel trademark) for EPROM tunnel oxide [9], became an industry standard Flash cell. The ETOX device had a tunnel oxide thickness of about 10 nm, a graded source junction, and an abrupt drain junction. Erase was accomplished by FN tunneling of electrons from the floating gate to the source diffusion using a typical voltage of

12.5 V. Programming was done by CHE. The ETOX type of cell has been employed dominantly for NOR arrays. Historically, NAND arrays have employed a similar transistor structure, but the simple daisy chain interconnection of individual transistors to form a string allows the sharing of interconnection overhead over 8, 16, or 32 cells resulting in a much smaller area per bit.

Split-gate cells are distinctly different in structure because the floating-gate transistor incorporates a series nonmemory device that provides a select and over-erase mitigation function. As illustrated in Figure 1.13 a portion of the channel is controlled by the floating gate while a second portion is under direct control of the control gate. The nonmemory transistor assures that the overall transistor structure remains in the enhancement mode independent of the status of the memory transistor portion. The isolation afforded by the nonmemory structure is also beneficial in reducing disturb conditions.

In 2000 Eitan et al. [10] introduced the NROM (Saifun trademark) concept, which established a marked change from the conventional floating-gate structure. This cell makes use of localized charge trapping at each end of an N-channel memory transistor to physically store 2 bits. Figure 1.14 shows a cross section of a cell along a wordline. In this structure the gate left junction is at 0 V, and the wordline is at 9 V. The accelerated electrons will be injected into the oxide nitride oxide (ONO) near the right junction. To read bit 1 the right junction is held at 0 V, left junction at 1.5 V, and the wordline at 3 V. This bias condition maximizes the effect of the bit 1 charge on the threshold voltage. Reversing the bias conditions for the left and right junctions enables the programming and reading of bit 2. Erase is accomplished by establishing positive voltage on the bitlines relative to the wordlines. This causes band-to-band tunneling of hot holes in the depletion layer under the ONO above the n+ junctions. Chapter 13 treats the NROM cell in detail.

The bitlines (BLs) are buried diffusions and the wordlines (WLs) are polystripes. The array does not require contacts and is very flat because there is no field

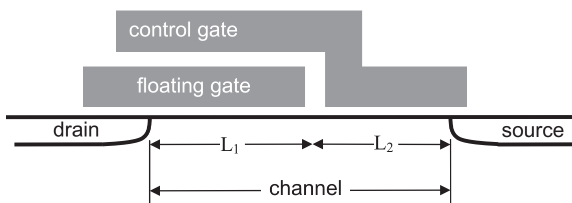

Figure 1.13. Typical split-gate cell structure cartoon.

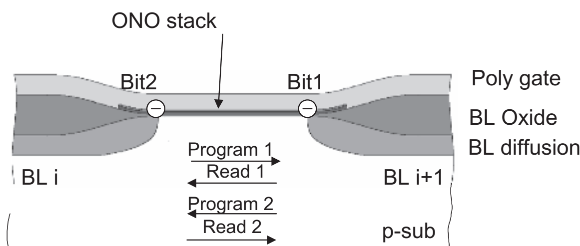

Figure 1.14. NROM cell showing localized electron storage regions.

isolation. NROM arrays are very efficient in utilization of area. A virtual ground array can be laid out using an ideal $2F$ wordline pitch and a $2.5F$ bitline pitch. This yields a $5F^2$ cell area that, since the cell contains 2 bits, is $2.5F^2$ per bit. In 2005 Eitan et al. [11] described the extension of the NROM concept to achieve 4 bits per cell. Here four threshold voltage levels are provided on each side of the cell to achieve an area per bit of $1.25F^2$.

1.5 SEMICONDUCTOR DEVICE TECHNOLOGY GENERATIONS

Integrated circuit technology has historically improved functional density at an exponential rate. This exponential trend was first pointed of by Gordon Moore in a 1965 article [12] where he observed that the number of transistors per chip was doubling every year. This trend has since slowed, but an exponential rate has been maintained. This so-called Moore's law has been the guiding rule for the International Technology Roadmap for Semiconductors (ITRS). The industry progresses by means of development of technology such that a new generation, referred to as a technology *node*, with minimum feature (F) sizes $0.7x$ of the previous generation emerges at 3-year intervals. The 2004 ITRS includes data for five nodes: 2004, 90 nm; 2007, 65 nm; 2010, 45 nm; 2013, 32 nm, and 2016, 22 nm. The most recent version of the ITRS is available at http://public.itrs.net.

Nonvolatile memory technology is treated in the ITRS in several places. The chapter developed by the Process Integration, Devices and Structures (PIDS) technology working group follows memory technologies that have matured to the point of being in production and addresses requirements for each technology to be realized in future technology nodes. Table 1.2 lists the cell size requirements given for DRAM and five NVM technologies.

Figure 1.15 provides a graphical view of this same data. The expected density advantage of Flash NAND is evident.

The ITRS plays an important part in the evolution of NVM. It is a document formulated and read by integrated circuit manufacturers worldwide. Each manufac-

TABLE 1.2. Memory Cell Sizes (μm^2) as Tabulated in the 2004 ITRS Update

Year	DRAM	Flash NAND	Flash NOR	FeRAM	SONOS	MRAM
2004	0.0650	0.0446	0.1010	0.4860	0.0660	0.4000
2005	0.0480	0.0352	0.0800	0.2030	0.0550	0.2000
2006	0.0360	0.0270	0.0610	0.2030	0.0450	0.1800
2007	0.0280	0.0190	0.0530	0.1730	0.0290	0.0900
2008	0.0190	0.0146	0.0420	0.1210	0.0250	0.0700
2009	0.0150	0.0113	0.0340	0.1210	0.0180	0.0600
2010	0.0122	0.0091	0.0270	0.0800	0.0150	0.0450
2011	0.0100	0.0072	0.0220	0.0650	0.0120	0.0360
2012	0.0077	0.0055	0.0170	0.0510	0.0100	0.0270
2013	0.0061	0.0046	0.0170	0.0390	0.0080	0.0230
2014	0.0043	0.0035	0.0110	0.0330	0.0070	0.0170
2015	0.0038	0.0028	0.0100	0.0260	0.0050	0.0140
2016	0.0025	0.0022	0.0080	0.0200	0.0040	0.0100
2017	0.0021	0.0018	0.0070	0.0180	0.0037	0.0090
2018	0.0016	0.0015	0.0050	0.0160	0.0030	0.0070

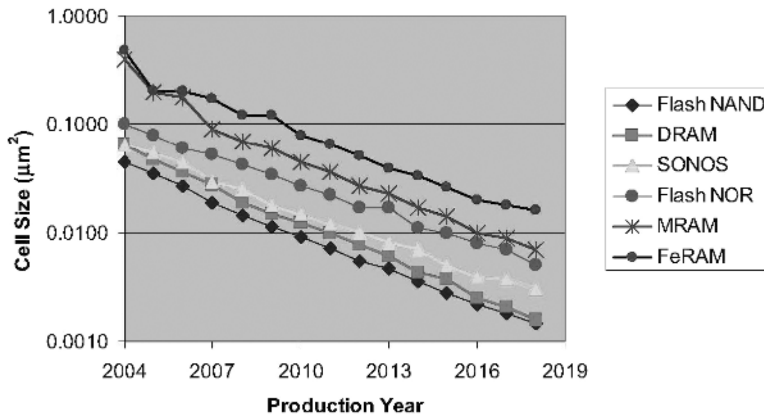

Figure 1.15. Plot of 2004 ITRS memory cell size requirements.

turer recognizes that the roadmap indicates approximately where the competition's capability will be at any given time, and each company strives to beat the roadmap goals as a matter of survival.

Up to this point in time the industry has been successful in maintaining the frantic pace of development set out in the ITRS. Every year at every major semiconductor meeting and in private discussions among technologists there has been someone who argues that the problems on the horizon are insurmountable and scaling will end. Usually the proponent of doom has one or two specific issues that he or she feels are going to be fatal. As the minimum feature size progresses closer to atomic dimensions, no doubt the nature of density improvement will shift from simple scaling to alternative ways of advancing functional integration. The challenge for nonvolatile memory is perhaps even more difficult than for CMOS logic because of the need to manage both dielectric stresses and maintain retention characteristics.

The history of Flash is an excellent demonstration of the innovative ways that engineers have met the scaling challenge. Intel, for example, has a publication track record for ETOX that shows how the desirable characteristics of a proven cell structure can be maintained while continuously improving the utilization of array area [8, 9 13–17]. For each technology generation an improved lithographic capability has become available, but the core portion of a Flash cell (i.e., that portion that actually stores charge) is difficult to scale. Engineering attention has focused on identifying and minimizing the area devoted to nonmemory portions of the array such as contacts, isolation, and alignment tolerances. Of course, the multibit cell was a major tactic in the fight to continue to reduce area per bit.

Another portion of the ITRS also includes consideration of memory: the Emerging Research Devices (ERD) section. ERD recognizes that simple scaling will eventually reach limits, and it attempts to monitor and evaluate published data for devices in the early research stage in hopes of identifying candidate technologies capable of maintaining desired density and performance growth [18]. The ERD chapter of the ITRS is necessarily a snapshot of a rapidly changing picture that must be revised every year. New technologies are almost always described overoptimistically by initial advocates, and very few approaches will survive to the development stage. The ERD working group strives to provide a balanced assess-

ment of comparative technology potential that will be useful to the semiconductor industry for making research investment decisions.

REFERENCES

1. IEEE Std. 1005-1998, IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays, Feb. 1999.
2. R. Bez, E. Camerlenghi, A. Modelli, and A. Visconi, "Introduction to Flash Memory," *Proc IEEE*, Vol. 91, pp. 489–502, Apr. 2003.
3. F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, and S. Tanaka, "A New Flash EEPROM Cell Using Triple Polysilicon Technology," *IEEE IEDM Tech. Dig.*, pp. 464–467, 1984.
4. M. Gill, and S. Lai, "Floating Gate Flash Memories," in W. D. Brown and J. E. Brewer (Eds.), *Nonvolatile Semiconductor Memory Technology*, IEEE Press, Piscataway, NJ, 1998.
5. F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, and S. Tanaka, "A 256 K Flash EEPROM Using Triple Polysilicon Technology," *IEEE ISSCC Tech. Dig.*, pp. 168–169, 1985.
6. F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, N. Tozawa, and S. Tanaka, "A 258 K Flash EEPROM Using Triple Polysilicon Technology," *IEEE J. Solid-State Circuits*, Vol. SC-22, pp. 548–552, 1987.
7. S. Mukherjee, T. Chang, R. Pan, M. Knecht, and D. Hu, "A Single Transistor EEPROM Cell and Its Implementation in a 512 K CMOS EEPROM," *IEEE IEDM Tech. Dig.*, pp. 616–619, 1985.
8. G. Verma and N. Mielke, "Reliability Performance of ETOX Based Flash Memories," *Proc. IRPS*, pp. 158–166, 1988.
9. V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, S. Wells, and M. D. Winston, "An In-System Reprogrammable 256 K CMOS Flash Memory," *IEEE J. Solid-State Circuits*, Vol. 23, No. 5, pp. 1157–1163, Oct. 1988.
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 Lett.*, Vol. 21, No. 11, pp. 543–545, Nov. 2000.
11. 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," *IEEE IEDM Tech. Dig.*, pp. 539–542, Dec. 5, 2005.
12. G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, Vol. 38, No. 8, Apr. 19, 1965.
13. V. Kynett, M. Fandrich, J. Anderson, P. Dix, O. Jungroth, J. Kreifels, R. Lodenquai, B. Vajdic, S. Wells, M. Winston, and L. Yang, "A 90 ns One Million Erase/Program Cycle 1-Mbit Flash Memory," *IEEE J. Solid-State Circuits*, Vol. 24, No. 5, pp. 1259–1264, Oct. 1989.
14. A. Fazio, "A High Density High Performance 180 nm Generation ETOX^{TM} Flash Memory Technology," *IEEE IEDM Tech. Dig.*, pp. 267–270, Dec. 5–8, 1999.
15. S. N. Keeney, "A 130 nm Generation High Density ETOX^{TM} Flash Memory Technology," *IEEE IEDM Tech. Dig.*, pp. 2.5-1–2.5-4, Dec. 2–5, 2001.
16. G. Atwood, "Future Directions and Challenges for ETOX Flash Memory Scaling," *IEEE Trans. Devices Mater. Reliabil.*, Vol. 4, No. 3, pp. 301–305, Sept. 2004.
17. R. Koval, V. Bhachawat, C. Chang, M. Hajra, D. Kencke, Y. Kim, C. Kuo, T. Parent, M. Wei, B-J. Woo, and A. Fazio, "Flash ETOXTM Virtual Ground Architecture: A Future Scaling Direction," paper presented at the 2005 Symposium on VLSI Technology Digest of Technical Papers, June 14–16, 2005, pp. 204–205.
18. J. E. Brewer, V. V. Zhirnov, and J. A. Hutchby, "Memory Technology for the Post CMOS Era," *IEEE Circuits Devices Mag.*, Vol. 21, No. 2, pp. 13–20, Mar.–Apr. 2005.