Non-Volatile Memory:
The principles, the technologies, and their significance to the smart card integrated circuit

David Sowards

The success of smart card ICs is completely dependent on the memory technologies used in their implementation. The design and manufacture of non-volatile memory is a specialized field of semiconductor technology. Each design requires careful balancing of the physics and construction involved with producing the memory cell and the transistors that switch the various surrounding voltages used for read, write, and erase. The designs are invariably different from previous generations. As a result, the technology has developed a mystique and jargon of its own. The terms used are generally imprecise, and used incorrectly. For instance, "what is a Flash memory?" will yield answers from engineers that will confuse rather than help. The subject is immense and highly technical. This paper attempts to introduce the subject and define the terms for engineers within the smart card industry.

Mr. Sowards has extensive experience in many types of non-volatile memory design, spanning a variety of cell technologies. He is the Director of Engineering of Emosyn, a fabless semiconductor company specializing in smart card ICs. Emosyn is also developing a unique approach for processing FeRAM non-volatile memory technology, which is destined to go into the next generation of wireless, smart card, and other ICs that use non-volatile memory.

Introduction:

Traditional microprocessor smart card ICs have used masked ROM and EEPROM for code and data storage. EEPROM has been the predominant choice for data storage. Program code has traditionally been stored in masked ROM. The use of masked ROM for code storage has significant cost and time-to-market disadvantages associated with software changes, especially given that “high-end” smart cards are extremely software intensive.

With the evolution of different memory cell structures in the bulk memory market, there is a wider range of proven memory technologies becoming available. The possibility of replacing masked ROM with the normal EEPROM technologies has been prohibitive because of cost. Traditional EEPROM can occupy 6 times the area of silicon for the same memory size as masked ROM. Some of the new cell designs can now challenge this accepted wisdom, and offer new possibilities for use in smart card applications.

The technology of choice is a Flash technology that combines the low cost advantages of masked ROM, and the re-programmability of EEPROM. With ever-decreasing die sizes and cost of components, EEPROM has not been able to keep pace with the technology advances. From a cost point of view, the replacement of EEPROM is a logical progression. Because of its low cost, Flash technology can also replace the traditional masked ROM, with highly attractive logistical advantages for software modifications. This paper will describe the present technologies and discuss the use of Flash technology to implement the function of EEPROM.

Emosyn chose a patented and proprietary Silicon Storage Technology Inc. (SST) CMOS SuperFlash EEPROM technology for its smart card ICs. It is a split gate, floating poly memory cell with the SST field enhancing tunneling injector. [1] This cell provides many advantages for designing and manufacturing embedded Flash EEPROMs when compared to a stacked gate or two-transistor approach. These advantages translate into significant flexibility, cost, and reliability benefits for the user.

The memory array basic backgrounder

Non-volatile memory has been required for many types of systems including those with a micro-controller or microprocessor. The minimum requirement is a memory that can retain its data while the power is off. The first proposal of a floating gate to be used as a re-programmable non-volatile storage element was in 1967. A second proposed solution was an MNOS (Metal Nitride Oxide Semiconductor) device also in 1967. [2] This was not a floating gate device as the charge is actually stored in the nitride insulator and not on a floating gate. The first functional floating gate device was the FAMOS (Floating gate Avalanche injection MOS) [3]. Since then, the most common types have been n-channel stacked gate or split gate EPROM devices and two transistor EEPROM devices.

Since the mid-1980's, there have been a plethora of single transistor Flash EEPROM devices. The reference to "Flash" as a technology is not very informative. There are a great number of vastly different technologies, all of which are referred to as Flash technologies. The origin of the term is not related to the underlying technology, but to its operation in a memory array. Quite simply, Flash came from the block or sector erase, which was done in a single operation, or *in a Flash*! There are several Flash architectures, such as NAND, NOR, DINOR, etc. They use a multitude of different techniques to achieve their non-volatile function. A few of these are FN (Fowler-Nordheim) erase and FN programming, FN erase and CHE (Channel Hot Electron) program, FN erase and SSI (Source Side Injection) program, just to name a few. All have stacked gate cells or split gate cells. A description of these various technologies as well as the traditional two transistor EEPROM are well documented and can be found in the literature. [4-7, 15]

Each of these technologies has advantages and disadvantages, which has led to their use to varying degrees in different applications. EPROM has traditionally been a low cost component. It is small, but requires removal from the system and UV light exposure to erase. Another issue is the need for special programmers that can supply the high currents and voltages that are required. The reliability has also been low with most parts only specify a cycling endurance of 100 write/erase cycles. The two transistor EEPROM has offered many of the things EPROM could not deliver, such as in system re-programmability with low or modest currents, and higher endurance of 100,000 typical write/erase cycles. However, with two transistors, EEPROM has a higher cost per bit, thereby precluding widespread use. This led to a large industry push for a low cost, in system re-programmable, highly reliable technology, which could combine the beneficial aspects of both EPROMs and EEPROMs. Consequently it has led to the development of many forms of single transistor and split-gate transistor Flash technologies. Some of these technologies have made great progress with these goals in mind. However, many have done so at the expense of performance, or by requiring extremely difficult manufacturing processes. Performance refers to the ability of the device to retain its data, continue its normal operation, and be able to repeat the re-programming of the device over its lifetime. These key measures are known in the industry as data retention and cycling endurance. Additionally, the effects of "disturb" mechanisms, which occur when read write operations on one memory cell have adverse effects on its neighbors, must be minimized, and should be insignificant in normal operation. Until now, it has been difficult to solve all of these problems with a single Flash technology. One reason is that most of these technologies use a high voltage on a common bit line. The other cells on that bit line are subjected to this high voltage repeatedly.

As mentioned above, the SST SuperFlash EEPROM technology has several advantages that give it the best combined characteristics of previous EPROMs and EEPROMs. We believe it offers the best performance available from the various Flash technologies, especially for embedded applications. Three main features make this technology advantageous. Any one of which is useful in and of itself, but it is all three which make such a winning combination.

The first advantage is the use of a split gate. This allows a single control gate to control the floating gate, and act as a select transistor that can shut off the channel directly. This eliminates the "over erase" issue commonly associated with single transistor stacked gate Flash approaches.

The second advantage is the use of field enhanced tunneling through a thick oxide for erase. This has several benefits. It allows the manufacturing process to have much more margin to make a uniform and reliable oxide. The fabrication line does not have to reproduce an ultra-thin oxide to exacting standards, and is therefore very repeatable, which increases the reliability. With this thick oxide, the data retention characteristics are greatly improved as it presents a high-energy barrier to electrons that could otherwise "leak" off from the floating gate. Additionally the "wear-out" of this type of thick oxide is a much more predictable trap up type mechanism. [8] Modern processing techniques, combined with the well understood nature of the trap up wear-out, have extended the cycling capability of SST CMOS SuperFlash EEPROM technology to surpass that of its traditional thin oxide counterparts. The thick oxide is much less susceptible to tiny variations and defects that can lead to unpredictable reliability.

The third advantage is the use of source side injection (SSI) for the programming of the device. [9] This allows the very efficient generation of hot electrons in the channel while also having very high, greater than 99%, efficiency in collecting them onto the floating gate. This eliminates the high currents associated with EPROM, like drain side channel hot electron (CHE) injection methods. SSI programming improves the reliability and allows for re-programmability using on-chip charge pumps. This, in combination with the thick oxide for erase, eliminates the need for ultra-thin oxides in the process.

Detailed discussion of NVM technologies

Base cell technologies

In this section, several of the predominant technologies will be reviewed. Quite often the terminology of programming and erasing can mean different things for different technologies and even among different vendors of similar technologies. In this paper programming will generally refer to the accumulation of electrons on a floating gate or making the NVM transistor a high threshold or non-conducting device. Erasing will refer to removing electrons and/or supplying holes to a floating gate. This will leave a neutral or even positive charge on a floating gate making it a low threshold or conducting device.

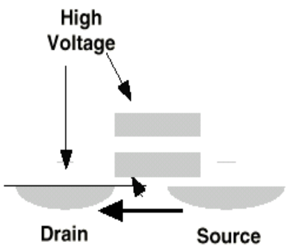
This section will not discuss technologies such as ROMs, Nitride memories (MNOS) [2, 10], Non-Volatile RAM (NOVRAM) [11], any analog [12] or multi-level cell (MLC), nor exotic structures which are not in wide spread industrial use.

EPROMs

EPROMs have developed in various forms, and were one of the earliest forms of re-programmable non-volatile memories. [3, 13, 14] Figure 1 is the most common type of EPROM, shown during programming.

EPROM erase is performed by exposure to UV light for an extended period. This leaves the floating gate in a neutral state. The device is processed such that this state has a low threshold. It is conductive when a voltage is applied to its control gate during a read operation. The resulting current in the device is then detected.

Figure 1 shows a high voltage applied to the drain and the control gate at the same time in order to program the cell. The desire is to generate hot electrons in the channel, which can overcome the oxide barrier height, and then be collected by the floating gate. To do this, a compromise is arrived at where about the same voltage is applied to the drain and the gate at the same time. The best generation of hot electrons is in the deep saturation region with a high drain voltage and a lower gate voltage, but of course enough to turn the device on. The best collection of electrons is with a very high gate voltage and a low drain voltage so that electrons which are injected over the oxide barrier are strongly attracted to the gate. Because of this, the compromise solution causes several disadvantages, such as very high programming currents and low reliability. Most EPROMs require miliamps to program and only guarantee 100 write/erase cycles.

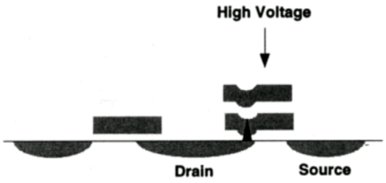


Source - Drain Electron Current ≈ 1 mA
Injection Oxide ≈ 100Å
Method: Drain-Side CHE Injection

Figure 1 STACKED GATE CELL

EEPROM

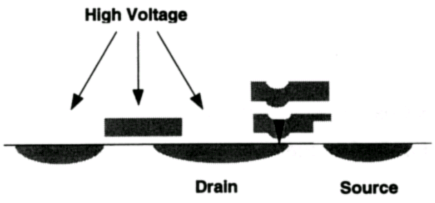
The most common EEPROM structure is a non-volatile memory with a basic floating gate transistor and a MOS select transistor. [15] The floating gate has a very thin tunnel oxide between it and a drain diffusion that lies under part of it. This oxide is extremely thin, generally less than 100 Å thick. The poly 2 above the floating gate is usually referred to as the control gate. The poly of the MOS transistor is usually referred to as the select gate and the entire transistor as the select transistor. Raising the select gate voltage and turning on the select transistor performs a read. At the same time, the control gate is taken to higher voltage or a reference voltage. Then current, or the lack of it, is detected as flowing from a metal bit line through the drain of the select transistor and floating gate transistor into the source.



Electric Field ≈ 11 MV/cm
Tunnel Oxide ≈ 85Å
Method: Drain FN Tunneling

Figure 2 2 Transistor Thin Oxide Cell

Figure 2 shows the EEPROM cell in a programming operation. A high voltage is applied to the control gate. A low voltage is applied to the source, and quite often, a low voltage is applied to the drain through the select transistor. This causes a high electric field of about 11MV/cm to be applied across the tunnel oxide. This will induce Fowler-Nordheim tunneling which will move electrons from the drain to the floating gate. The negatively charged floating gate will cause the cell to be non-conductive during a read operation.



Electric Field ≈ 11 MV/cm
Tunnel Oxide ≈ 85Å
Method: Drain-Poly FN Tunneling

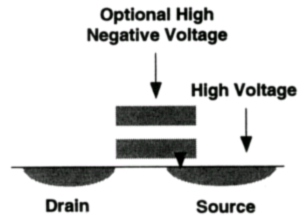
Figure 3 2 Transistor Thin Oxide Cell

In Figure 3, the EEPROM cell is in an erase operation. There is a high voltage applied to the bit line and the select gate, which will turn on and pass the high bit line voltage to the drain of the floating gate device. The source is held higher or sometimes floating. The control gate is driven to a low voltage. This causes a high electric field of about 11MV/cm to be applied across the tunnel oxide in the reverse direction. This will remove electrons from the floating gate and leave it neutral or even positively charged. This will make the device a low threshold or conducting device during a read operation.

With such high electric fields applied to the ultra-thin tunnel oxide, the quality of this oxide must be extremely high. Most major NVM vendors have made great progress at manufacturing these oxides with consistent dimensions, uniformity, and low defects. This is essential for low failures and thus high reliability. When any thinning, non-uniformity, or defect occurs, the failure is very unpredictable and difficult to catch during infant mortality screens. This leads to an increase of random failures during the "useful life" of reliability. See Figure 9, the classical "bathtub" curve, as described in a later section.

Flash

As mentioned above, "Flash" is a catch all term that refers to many different technologies. The most common types are easiest to understand by adding to the basic EPROM structure. As shown in Figure 4, the basic structure looks very similar to that of an EPROM. The read operation is essentially the same. Figure 1 shows an EPROM cell during programming, which is essentially the same as a Flash cell for this case. Again a programmed cell has a negatively charged floating gate and is non-conducting when read. Figure 4 shows the cell during erasing. Here a low or even negative voltage is applied to the control gate, a high voltage is applied to the source and the drain is relatively high or floating. With a thin channel oxide of about 100 Å, this produces a high electric field of about 10MV/cm to be applied across the gate oxide on the source side. This high field causes Fowler-Nordheim tunneling through the gate oxide, which removes electrons from the floating gate. This leaves a neutral or more often a positively charged floating gate. This cell is then conducting during a read operation. Unlike EPROMs which can only be UV erased to neutral, this positive charging of the floating gate can cause undesirable issues. When a device is deselected, while the word line is discharged to ground, the positive charge on the floating gate can still keep the channel of the device on or conducting. Many "deselected" cells in this state, on the same bit line, can discharge that bit line and the "selected" cell will be detected as an "on" cell instead of an "off" or non-conducting cell. This is referred to in the literature as an "over erase" issue. [16] Additionally, if EPROM-like programming is used then all the disadvantages of high current, high hot electron generation, and low collection efficiency remain. This usually results in compromises to achieve any comparable level of reliability. Many variations of this basic form of Flash have been attempted. In most cases, solving one issue degrades some other performance of the device.



Electric Field ≈ 10 MV/cm
Tunnel Oxide ≈ 100Å
Method: Source-Poly FN Tunneling

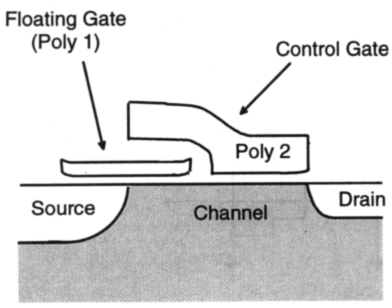
Stacked Gate Cell

Figure 4

SST CMOS SuperFlash EEPROM Cell

As mentioned above, the SST CMOS SuperFlash EEPROM technology has several advantages that can be exploited in a smart card memory design.

Figure 5 shows a cross-sectional view of the cell along the bit line, it is a very good representation of the actual cross-section as shown in the SEM picture Figure 6. The SEM shows a metal line used to connect all the drains below this line together, and is referred to as the bit line. Two cells share each N+ diffusion drain, which are "mirrored" with respect to each other along the bit line. In the orthogonal direction, a common polysilicon line is used, with or without silicide, to connect the control gates together. This is referred to as the word line, and one word line is referred to as a row of cells, or simply a row. An N+ region is used for the source of each cell. Such a region is shared in common between two rows. A pair of "even" and "odd" rows which share a common source form a page. A design using this technology can erase all cells in a page at the same time. Then any programming, or write operation, is done for all cells within the page, even if the data is changed in only one cell. A cell implant is used to control the intrinsic threshold of the floating gate, as well as the "punch through" voltage of the storage transistor. The select gate is separated from the channel by a 400 Å oxide. The floating gate is separated from the channel and source diffusion by a 150 Å gate oxide. The floating gate is separated from the control gate by a 400 Å oxide on the sidewall and a 2000 Å oxide vertically between the gates. A tunneling injector is formed on the floating polysilicon.



Cross-section along the Bit Line

Figure 5

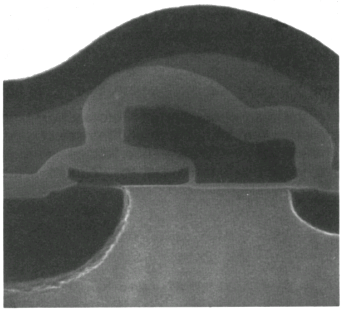
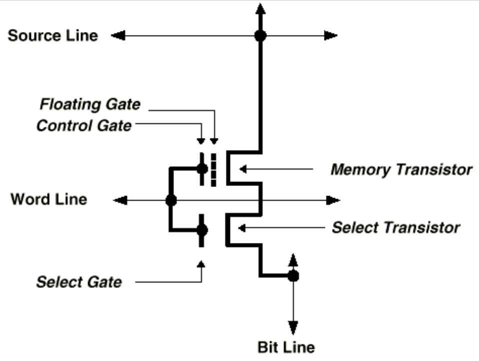


Figure 6 Cross-section SEM Picture

Figure 7 shows a schematic representation of the physical cell shown in Figures 5 and 6. This is only an equivalent circuit. It shows two transistors, a memory transistor and a select transistor. This representation clearly shows that the channel, and hence the current through it, is controlled by the combination of the select gate and the memory gate transistors in a series like configuration. In reality, the control gate and select gate are one and the same in a "split' gate configuration. Depending on the charge stored on the floating gate, the memory transistor is either in a high or low threshold state.



EQUIVALENT MEMORY CELL STRUCTURE FOR Q1

Figure 7

Figure 8 shows a schematic representation of an array with simplified symbol, Q1, for each memory cell of Figure 7. This logical configuration shows the word lines, bit lines, and common source lines.

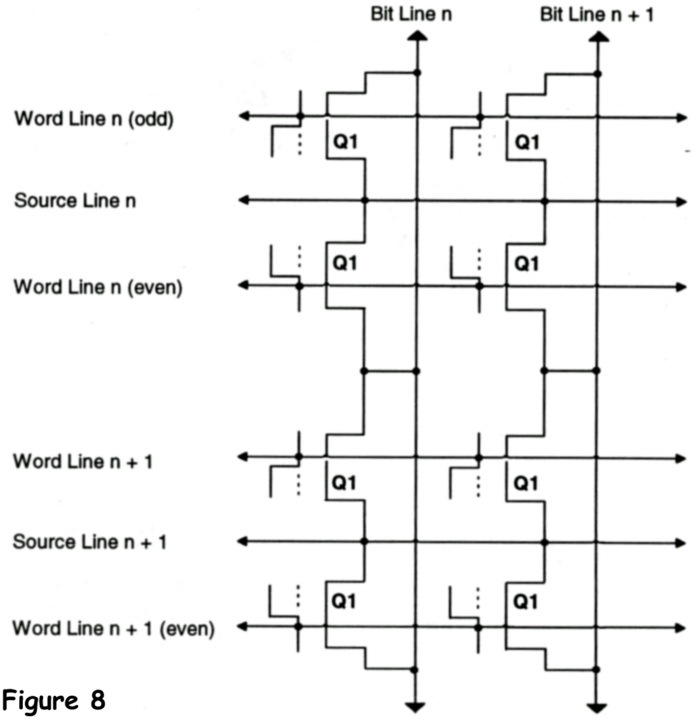


Figure 8

During a read operation, the word line is driven to a reference voltage. If the selected cell is in the "high threshold" state, then the memory transistor will not conduct. This lack of current, flowing through Q1 along the selected bit line, will be detected and output as a logic “0”. If the selected cell is in the low or "negative threshold”, then the memory transistor will conduct. This current, flowing through Q1 along the selected bit line, will be detected and output as a logic “1”.

Description of Operations

Table 1 shows the normal operating conditions for the memory cell terminals during the erase, program, and read operations. These are nominal conditions for a generic 1-micron process. Vdd is the power supply voltage. Vss is ground. Vt is the cell threshold. Vref is the reference voltage used to access the memory cell during the read cycle. An on-chip charge pump generates the high voltages on the word line during erase and on the source line during programming.

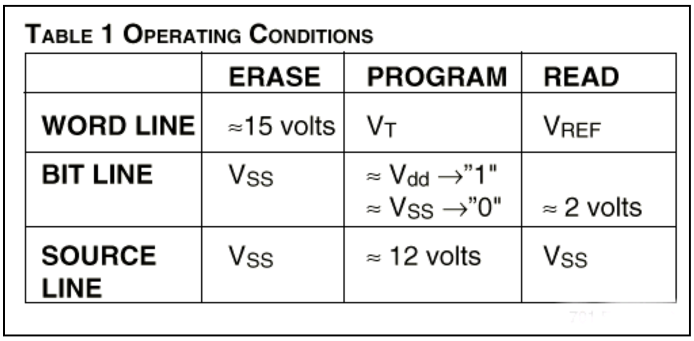


Table 1 Operating Conditions

	ERASE	PROGRAM	READ
WORD LINE	≈15 volts	V_T	V_{REF}
BIT LINE	V_{SS}	≈ V_{dd} →"1" ≈ V_{SS} →"0"	≈ 2 volts
SOURCE LINE	V_{SS}	≈ 12 volts	V_{SS}

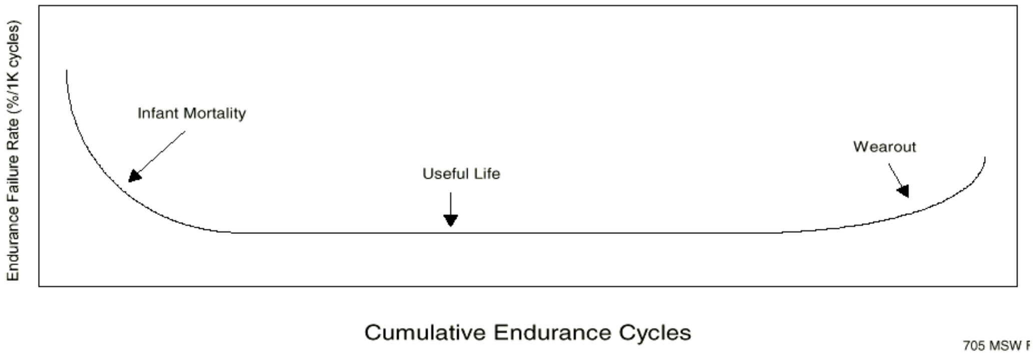
The cell erases using Fowler-Nordheim tunneling from floating gate to control gate. The floating gate poly oxidation process provides a uniform field enhanced tunneling injector along the edges of the floating gate. This repeatable manufacturing process provides consistent oxide integrity that minimizes endurance-induced degradation, i.e., charge trapping or oxide rupture. During erasing, the source and drain are grounded and the word line is raised to about 15 volts. The conditions for erasing are in Table 1. The low coupling ratio between the control gate and the floating gate provides a significant voltage across the interpoly oxide between poly 1 and poly 2. A local high electric field is generated primarily along the edge of the tunneling injector. Charge transfer occurs and is eventually limited by the accumulation of positive charge on the floating gate. The positive charge raises the floating gate voltage until there is insufficient voltage across the interpoly oxide to sustain Fowler-Nordheim tunneling. The removal of negative charge leaves a net positive charge on the floating gate. The positive charge on the floating gate decreases the memory cell’s threshold voltage, such that it will conduct during a read, giving a value of “1”.

The cell is programmed using high efficiency source side channel hot electron injection. The conditions for programming are in Table1. During programming, a voltage approximately equaling the threshold Vt of the select transistor is placed on the control gate via the word line. This is sufficient to turn on the channel under the select portion of the control gate. The drain is held to Vss if the cell is to be programmed. If the drain is at Vdd, then programming is inhibited. The drain voltage is transferred across the select channel because of the voltage on the control gate. The source is at 12 volts. The source to drain voltage differential generates channel hot electrons. The source voltage is capacitively coupled to the floating gate. The field between the floating gate and the channel sweeps those channel hot electrons that cross the Si-SiO2 barrier height of 3.2 eV to the floating gate. This operation is highly efficient, approaching 100%.

The programming effect is self-limiting as negative charge accumulates on the floating gate. The programming source-drain current is very low. This enables the source voltage to be generated by a charge pump internal to the die. The programming is fast because of the high efficiency of the source side injection. The addition of negative charge to the floating gate neutralizes the positive charge generated during erasing. This causes the cell to be non-conducting when a reference voltage is applied during a read, giving a value of “0”. Additionally, the native cell floating gate threshold is adjusted so that its intrinsic (i.e. UV erased,) value is positive. This means that if charge is lost for a reason which is non-destructive, then that cell would read as a “0”.

Reliability considerations

In general terms, reliability is the ability of a device to continue its operation for an extended period of time. Figure 9 is a representation of the classical bathtub curve for reliability. The three main regions are “Infant Mortality”, “Useful Life”, and “Wear-out”. Infant mortality usually occurs right after manufacturing and normally is caused by latent defects or irregularities. With our technology, these are eliminated by extensive prescreening tests developed specifically for this purpose. Useful life is characterized by a very low or non-existent failure rate over the expected life of the product. For more information on industry standards see [IEEE Std 1005-1991]. The “Wear-out” is usually the point where a larger number of devices begin to fail, due to phenomena that are inherent to the device operation, such as defects, etc. For most products this should be well understood and predictable in nature given the technology, design, and manufacture of the product.



Oxide integrity

All oxides are subject to time-dependent dielectric breakdown, (TDDB). For a given oxide and electric field, the oxide will eventually breakdown. [17] The lower the electric field and the less time the field is applied, the longer the time to breakdown. For oxides used in normal low voltage ranges, this time is essentially infinite. However, in non-volatile memories that use high voltages, (and rely on moving charge through the insulator,) the time of oxide exposure to high electric fields can contribute to the intrinsic device reliability.

The SST memory cell uses a 4MV/cm electric field during erasing. This value is significantly lower than the 10 MV/cm used by other stacked gate Flash technologies, or the 11 MV/cm used by the thin oxide EEPROM and NAND Flash technologies. The SST cell is exposed to the lower electric field for less time during erase than other technologies. Since the oxide TDDB rate is an exponential function of the field strength, the SST memory cell intrinsically has a much lower failure rate than a thin oxide cell for oxide breakdown. Indeed the major failure mechanism for thin oxide devices is breakdown. [18]

Data Retention

The field enhancing tunneling injector cell uses relatively thick oxides, compared with other EEPROM or Flash approaches; therefore, intrinsic data retention is robust. The quality of thin oxides have been greatly researched and improved over the years. However the ability to maintain this uniformity die to die, wafer to wafer, lot to lot, and day to day is enormously difficult. The thicker SST oxides minimize initial and latent oxide defects, thus improving yield, oxide integrity, and lowering manufacturing cost. The lower voltages used for erase and programming combined with the relatively thicker oxides reduce the endurance related extrinsic data retention failure rate.

Endurance

Since the SST field enhancing tunneling injector cell uses a relatively thick oxide for the Fowler-Nordheim tunneling oxide, the primary endurance limitation is due to charge trapping in the interpoly oxide. Since both erasing by tunneling and the source side channel hot electron programming utilize relatively weaker electric fields across the poly 1 insulating oxides, the oxide rupture failure rate is low.

Trapping occurs mainly in a 20 Å shallow region adjacent to the tunneling injector. Within this distance, direct tunneling de-trapping occurs in the quiescent times between erase/program cycles. In practice, this means the endurance of the device in real world applications will be greater than the endurance demonstrated in a test environment, where the device is being erase/program cycled at the maximum possible frequency. And indeed, when a first cycling failure is detected in laboratory experiments the individual failure usually recovers within a short time after the end of the fast cycling operations.

As the neutral threshold is already detected as a “0”, the margin for cycling is in favor of the programming. This means that a cycling failure usually occurs first for an erasure of a cell. Again, the field enhancing injector thick oxide will virtually eliminate failures from oxide ruptures. As oxide ruptures are a catastrophic failure, they are much less predictable and dominated by small imperfections, irregularities, and defects. The main failure mechanism for thick oxide is trap up. [19] As such the normal lifetime behavior and the end of life behavior is much more stable and predictable.

Disturbs

A major concern of re-programmable non-volatile memories is that of “disturb” phenomena. This is where a different location in the memory array, other than the one being selected, is altered by the operation on the selected cell or cells. This usually occurs during a high voltage operation such as an erase or program. The SST cell has several advantages to reduce the possibilities for a disturb:

a) There is no high voltage placed on the bit line during any high voltage operations as with conventional stacked gate or two transistor approaches. In addition, the split gate cell isolates each memory storage node from all other nodes along the bit line. Thus, even a small disturb from a spurious Vdd spike or other intermediate voltage on the bit line is just not possible. While significantly reducing the disturb mechanisms, the split gate cell also eliminates the "over erase” issue, which is commonly associated with other single transistor stacked gate technologies.

b) The device uses a page erase, whereby, all bytes in the page are erased simultaneously. The array is organized into pairs of even and odd rows. Each row pair shares a common source line and each row pair has the word line at the same voltage potential during erasing. As they all see the same high voltage at the same time, there cannot be any disturb within a page. Furthermore, each page is physically isolated from each other so all other word lines do not receive the erasing high voltage, and no page to page disturb is possible.

c) The device uses a unique source line for each page, unlike most other approaches, which have the source line common to the entire array. This limits the exposure to disturb conditions to only the cells within a page during the time that page is being programmed. This is the reason that all cells (both up and down) that share a source line are included in the same page. As such, all are erased together and when one side of a source line is programmed, then the other side will be programmed in an immediately following operation as well. The circuit design architecture ensures that these cells are within the same page. The two types of possible program disturbs are a reverse tunnel disturb and a punch through disturb.

d) A reverse tunnel disturb can occur for unselected erased cells within the page sharing a common source line, but on the other row of the selected page to be programmed when its word line is grounded. The source voltage is capacitively coupled to the floating gate of the unselected erased cell. If there is a defect in the oxide between the control gate, Fowler-Nordheim tunneling may occur. This could program the unselected erased cell. Proper design, processing, and screening assures the reverse tunnel voltage is significantly higher than any applied voltage. Including a reverse tunnel voltage screen in the 100% testing operations eliminates defects. Forward tunneling is defined as occurring when electrons are transferred from poly 1 to poly 2, thereby erasing the cell. Reverse tunneling is defined as occurring when electrons are transferred from poly 2 to poly 1.

e) Punch through disturb can occur for unselected erased cells within the page sharing a common source line and bit line with the selected page to be programmed. An inhibited word line is grounded to prevent normal channel hot electron injection. If there is a defect that reduces the effective channel length, then it can allow punch through along the unselected select gate channel, and hot electrons could be available to program the inhibited erased cell. Proper design, processing, and screening assures the punch through voltage is significantly higher than any applied voltage. Including a punch through voltage screen in the 100% testing operations eliminates defects.

This base technology can support many thousands of writes to one side of the source line before a significant disturb is detected on the other. As mentioned above, only one write is allowed to one side of the source line without programming the other. Additionally only the adjacent cells need to be examined for disturb after a program operation. This greatly reduces the testing for disturbs, and with proper screening these types of disturbs are virtually eliminated.

Emosyn's use of the SST SuperFlash technology

Emosyn has developed a technique, where a small block size of only 8 bytes can be altered using this technology. The result is dramatic. It allows the direct replacement of EEPROM, and the numerous benefits of cost and performance are passed on to the user. The small block size of 8 bytes, instead of 64 or 128, allows the cycling to be increased almost an order of magnitude. For these array sizes, it is expected to exceed 100,000 cycles, which is standard for most EEPROMs. Data retention for this technology is specified as 100 years, which is more than sufficient for most smart card applications.

Parts are shipped as cleared or all “1” in the memory field. As mentioned previously the most predominant type of failure will be that of an erased cell appearing to be programmed, or simply the change of a “1” to a “0”. This is addressed by application note "Software considerations for smart card memory integrity ", which describes that any data which must have high reliability, in terms of data retention, should be written as “0” where ever possible.

Security is also enhanced by our approach. As this is a floating gate technology, then the accumulation or depletion of electrons from the floating gate is not readily apparent through optical or other visually enhanced means. Ascertaining program execution, by examining the memory cell contents in an attempt to breach security, is made much more difficult, if not impossible, even using the most advanced techniques. Since code storage is not fixed during wafer processing, the ability to easily change its physical position becomes possible. This introduces a new technique to further confuse the hacker, and improve security. This is unlike most masked ROM arrays, where the data can be either directly visibly apparent, or detectable with cross sectioning or other physical deprocessing. This means the data stored in OTP is more secure than traditional masked ROM. It also eliminates the very unlikely event that someone can get the mask or data used to generate the masked ROM code.

One of the great benefits of using the SST technology is that the traditional ROM for most program code can be replaced with SST technology configured as OTP, all on the same die. This allows the user to change and adapt the program code without going through a costly and time-consuming ROM mask change procedure. It is also useful for adding additional applications to existing cards. Thus, a superior technology can be implemented at a much smaller die size, lower cost, and lower power than other approaches that cannot implement both the OTP and EEPROM in the same technology. The result is a fundamental improvement in flexibility to introduce new or modified software, and this logistical advantage is without a cost or performance penalty.

Conclusion

Traditional ROM and EEPROM solutions for microprocessor smart cards have inhibited widespread use. The large effort taken to turn a design and its software into new ICs with masked ROM causes very long development times and is inherently inflexible. Additionally, the large size of the EEPROM cells has increased the cost of the silicon, which is a crucial issue for the larger memory sizes that the market demands. The bulk memory market has been developing memory cells using various techniques to increase the capability, reduce the die size, and improve the reliability of the device. These new generations of chips do not fit comfortably within the NVM categories commonly applied within the smart card industry. They are neither conventional EEPROM, Flash, nor EPROM, but take certain characteristics from all three.

SST has developed a novel Flash cell, which is an optimum combination of these characteristics. Emosyn has built on the SST technology with a carefully tailored design that maximizes performance. This design is the basis for a new generation of smart card devices. It allows the same non-volatile memory cell to be used for both code and data storage. This development brings “One Time Programmable” capability into the smart card arena at no additional cost. The accepted NVM solutions in smart cards, which have been unchanged for many years, are now being challenged. Combining code and data storage into a single low-cost technology is revolutionary for this marketplace.

By selecting and optimally designing the memory technology that fits the unique requirements of the smart card industry, Emosyn will stimulate new markets and increase the use of the smart card. To fully capitalize on this revolution, one must

understand the significance of the new technologies, and this document has attempted to convey that understanding.

The technology that Emosyn has chosen for its products allows an extremely flexible and secure device. It also enables a continued migration to lower voltage, lower power, and more secure smart card devices. In the future, we anticipate that low-cost FeRAM will become a viable alternative to charge storage NV memories, and a different set of resources will be added to the memory designer's toolkit. Emosyn will continue to identify the best memory technology for the application, bring those technologies to market, and enable superior products.

References

All figures are taken from SST Technical Papers, which include "Endurance Testing of EEPROMs", "Technical Comparison of Floating Gate Reprogrammable Nonvolatile Memories", and "SuperFlash EEPROM Technology"

[1] Silicon Storage Technology Inc., "FLASH MEMORY 1998 DATA BOOK", pp (6-1) – (6-42)

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

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

[4] S. Tam, S. Sachdev, M. Chi, G. Verma, L. Ziller, G. Tsau, S. Lai, and B. Dham, "A high density CMOS 1-T electrically-erasable non-volatile (Flash) memory technology," Symp. VLSI Tech., pp. 31-32, 1988.

[5] M. Momodomi et al, "An experimental 4-Mbit CMOS EEPROM with a NAND structure cell," IEEE J. Solid-State Circuits, vol. SC-24, no. 10, pp. 1238-1243, 1989.

[6] S. Haddad et al., "An investigation of erase-mode dependent hole trapping in Flash EEPROM memory cell," IEEE Elect. Dev. Lett., pp. 514-516, 1990.

[7] H. Onoda et al., "A novel cell structure suitable for a 3V operation, sector erase Flash memory," IEEE IEDM Tech. Dig., pp. 599-602, 1992.

[8] N. Mielke, A. Fazio, and H.C. Liou, "Comparison of Flotox and textured-poly EEPROMs," Proceedings of the 1987 International Reliability Physics Symposium, pp. 85-92, 1987.

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

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

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

[12] Van Tran, H., et. al., "A 2.5V 256-Level Non-Volatile Analog Storage Device Using EEPROM Technology", ISSCC Digest of Technical Papers, pp270-271, Feb., 1996.

[13] B. Rossler, "Electrically erasable and reprogrammable read-only memory using the n-channel SIMOS one-transistor cell," IEEE Trans. Elect. Dev., vol. ED-24, pp. 606-610, May 1977.

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

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

[16] J. Kupec et al., "Triple level polysilicon EEPROM with single transistor per bit," IEEE IEDM Tech. Dig., p602, 1980

[17] I. C. Chen, S. Holland, and C. Hu, " A quantitative physical model for time-dependent breakdown in SiO2," Proceedings of the 1985 International Reliability Symposium, pp. 24-31, 1985.

[18] S. K. Lai, “Oxide and interface issues in non-volatile memory,” presented at the Santa Clara Valley Section of the IEEE Electron Devices Society 1991 Symposium: Advances in Semiconductor Technologies, pp. 1-19, 1991.

[19] M. Liang and C. Hu, “Electron trapping in very thin thermal silicon dioxides,” Technical Digest of the 1981 IEEE IEDM, pp. 396-399, 1981.