Technical Paper
November 2001

INTRODUCTION

The following paper describes the patented and proprietary Silicon Storage Technology, Inc. (SST) CMOS SuperFlash EEPROM technology and the SST field enhancing tunneling injector split-gate memory cell. The SuperFlash technology and memory cell have a number of important advantages for designing and manufacturing flash EEPROMs, or embedding SuperFlash memory in logic devices, when compared with the thin oxide stacked gate or two transistor approaches. These advantages translate into significant cost time-to-market and reliability benefits for the user.

The SST SuperFlash technology typically uses a simpler process with fewer masking layers, compared to other flash EEPROM approaches. The fewer masking steps significantly reduces the cost of manufacturing a wafer. Reliability is improved by reducing the latent defect density, i.e., fewer layers are exposed to possible defect causing mechanisms. Reliability is also improved through the use of a thick tunnel oxide. Thin tunnel oxides in competing flash devices are more prone to failures from trapped charges.

The SST split-gate memory cell is comparable in size to the single transistor stacked gate cell (for a given level of technology), yet provides the performance and reliability benefits of the traditional two transistor byte alterable E^2PROM cell. By design, the SST split-gate memory cell eliminates the stacked gate issue of “overerase”, by isolating each memory cell from the bit line. “Erase disturb” cannot occur because all bytes are simultaneously erased in the same sector and each sector is completely isolated from every other sector during any high voltage operation.

FIELD ENHANCING TUNNELING INJECTOR EEPROM CELL

The field enhancing tunneling injector EEPROM cell is a single transistor split-gate memory cell using poly-to-poly Fowler-Nordheim tunneling for erasing and source side channel hot electron injection for programming. Poly-to-poly tunneling is from a field enhancing tunneling injector formed on the floating gate using industry standard oxidation and dry etching techniques. Source side channel hot electron injection is very efficient, thus allowing the use of a small on-chip charge pump from a single low voltage power supply. Cells are normally erased prior to programming.

The split-gate memory cell size is comparable to traditional stacked gate memory cells using the same process technology. This is possible because

1. the tunneling injector cell does not need the extra spacing to isolate the higher voltages and currents required for programming the stacked gate array, and
2. floating gate extensions are not needed to achieve the required stacked gate coupling ratios.

Additionally, the simplicity of the structure eliminates many of the peripheral logic functions needed to control erasing of the stacked gate device. The tunneling injector cell can be formed using standard CMOS process.

Memory arrays may use either random access or sequential access peripheral architectures.

CELL STRUCTURE

Cell Cross Sections and Layout

A top view and a cross-sectional view along the word line are presented in Figures 1 and 2 (note drawings are not to scale).

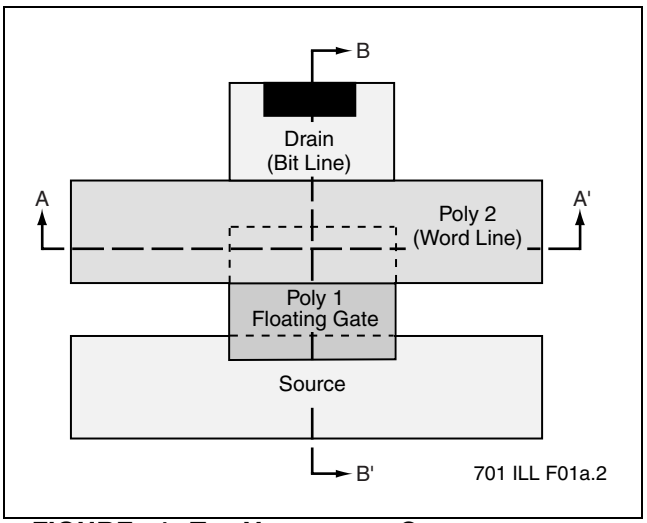

FIGURE 1: TOP VIEW OF THE CELL

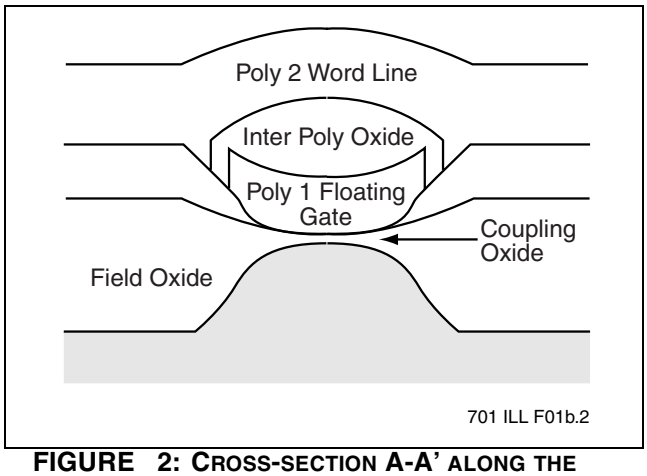

FIGURE 2: CROSS-SECTION A-A' ALONG THE WORD LINE

A cross-sectional view along the bit line and a SEM cross-section are presented in Figures 3 and 4. Polysilicon or polysilicon with silicide is used to connect control gates along the word line (row). Metal is used to connect the drain of each memory cell along the bit line (column). A common source is used for each sector, i.e., each pair of bits sharing a common source along a row pair (even plus odd row). A single word line is referred to as a row; the combination of the even and odd rows form a sector, which is erased as an entity. Programming may be either byte-by-byte individually or for all bytes within the same sector simultaneously.

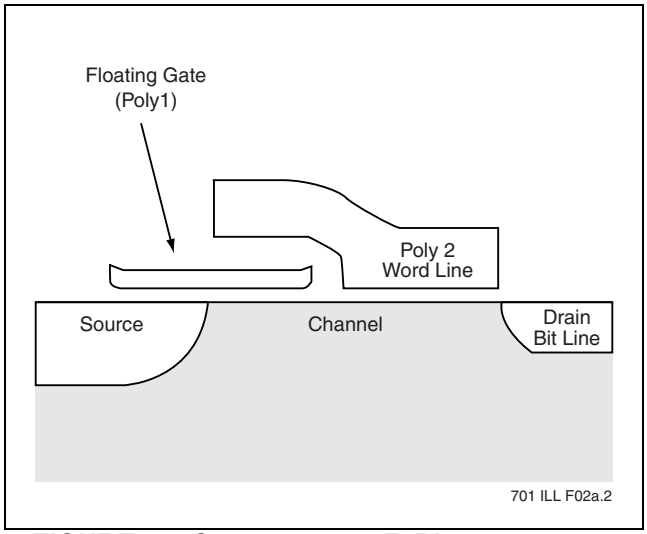

FIGURE 3: CROSS-SECTION B-B' ALONG THE BIT LINE

The drain region consists of an n+ S/D diffusion, which is aligned with the edge of the Poly 2 control gate. The source region consists of an n+ S/D diffusion, which overlaps the floating poly. A cell implant beneath the floating gate is used to control the intrinsic cell threshold (V_T) and the punch through voltage.

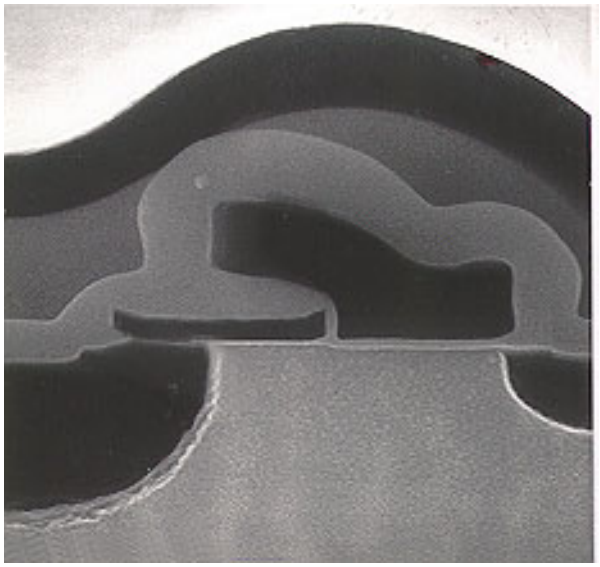

FIGURE 4: CROSS-SECTION SEM PICTURE

The word line is separated from the channel by a thin oxide. The floating gate is separated from the channel and source diffusion by a thermally grown thin gate oxide. The floating gate is separated from the word line by a thin oxide on the sidewall and a thick oxide vertically between the gates. The tunneling injector on the floating gate is formed by oxidation of the polysilicon, similar to the formation of the field oxide "bird's beak" on single crystal silicon, followed by a reactive ion etching of polysilicon. A silicide can be formed on the control gate to reduce the poly word line resistance.

Cell Array Schematic

Figure 5 is an equivalent memory cell, showing how the split-gate cell provides the logical equivalent of a select transistor and a memory transistor. The voltage applied to each terminal during normal operations is listed in Table 1. The split-gate behaves as a series combination of a select transistor and a memory transistor. The memory transistor is either in high or low negative threshold state depending on the amount of stored electric charge on the floating gate.

During the Read operation, a reference voltage (V_{REF}) is applied to the control gate and the select gate via the word line. The reference voltage will "turn on" the select gate portion of the channel. If the floating gate is programmed (high threshold state), the memory transistor portion of the channel will not conduct. If the floating gate is erased (low or negative threshold state), this memory cell will conduct. The conducting state is output as a logic "1", the nonconducting state is a logic "0."

The cell schematic is presented in Figure 6, showing the logical organization of the memory array. This illustration represents a section of a typical cross-point memory array, arranged as 8 memory cells in 2 columns (bit lines), 2 source lines, and 4 word lines (rows). Note that the word line is split into an even and odd row, which isolates the source line in the illustration from all other source lines on the array.

Table 1 gives the conditions for the memory cell terminals during the Erase, Program, and Read operations. V_{DD} is the power supply voltage. V_{SS} is ground. V_{REF} is the reference voltage used to access the memory cell during the Read cycle. The high voltages on the word line during erase and the source line during programming are generated by an on-chip charge pump.

TABLE 1: OPERATING CONDITIONS

	Erase	Program	Read
Word Line (Control Gate)	High Voltage	V_T	V_{REF}
Bit Line (Drain)	V_{SS}	$\approx V_{DD}$ r"1" $\approx V_{SS}$ r"0"	~1V
Source Line	V_{SS}	High Voltage	V_{SS}

T1.2 701

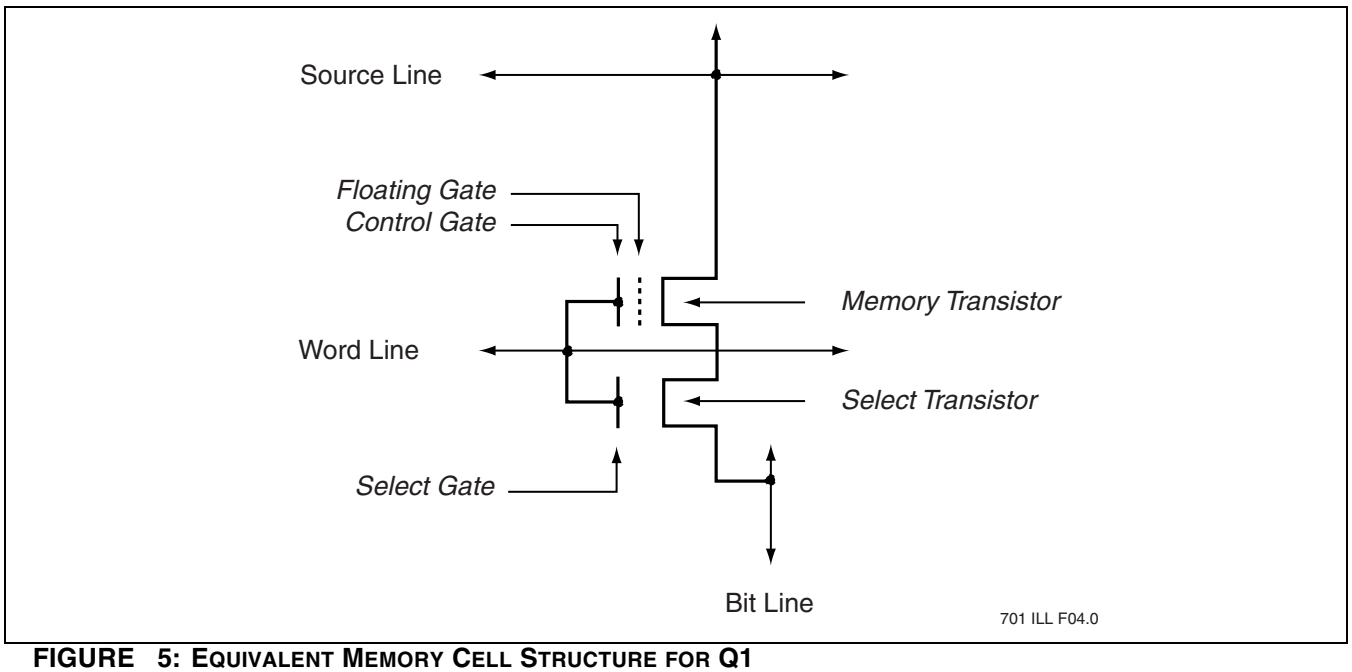

FIGURE 5: EQUIVALENT MEMORY CELL STRUCTURE FOR Q1

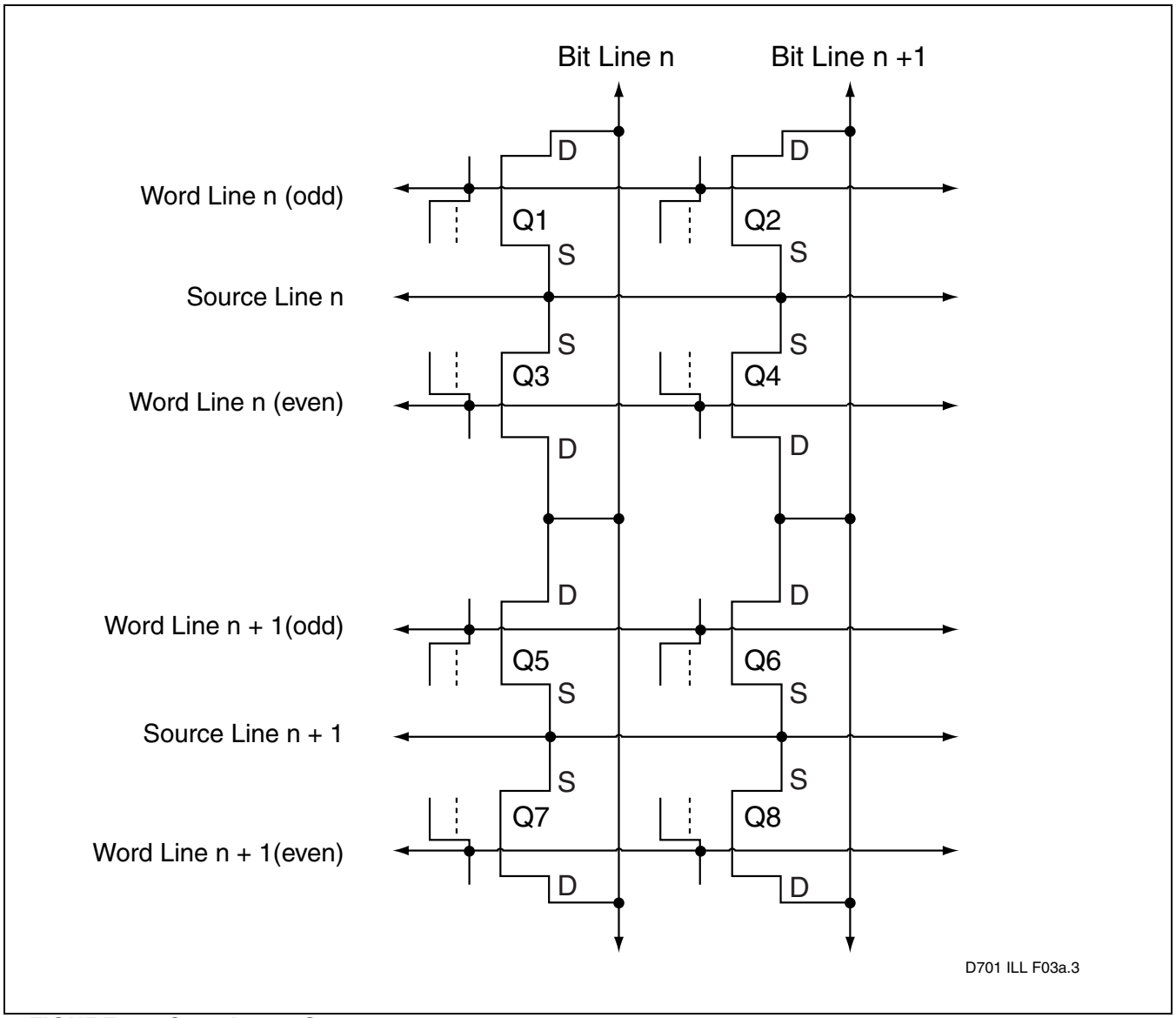

FIGURE 6: CELL ARRAY SCHEMATIC

CHARGE TRANSFER MECHANISMS

Erasing

The cell erases using Fowler-Nordheim tunneling from the floating gate to the control gate. The floating gate poly oxidation process provides a 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 a high voltage. The conditions for erasing are in Table 1, reference Figures 3 and 4 for identification of terminals. The low coupling ratio between the control gate and the floating gate provides a significant Δ V across the interpoly oxide. A local high electric field is generated primarily along the edge of the tunneling injector. Charge transfer is very rapid and is eventually limited by the accumulation of positive charge on the floating gate. This positive charge raises the floating gate voltage until there is insufficient Δ V 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 the memory cell will conduct when the reference voltage is applied to the memory cell during a Read cycle. The reference voltage is sufficient to turn on both the select transistor and the erased memory transistor in the addressed memory cell.

Erasing is performed using fixed erase pulses generated by an internal timer.

Erase Disturb

The enhanced field tunneling injector devices are internally organized by 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. Thus, all bytes are simultaneously erased along the common word lines. No other word lines receive the erasing high voltage. Therefore, erase disturb is not possible. The column leakage phenomenon caused by "overerase" in stacked gate cells is not possible, because the split-gate provides an integral select gate to isolate each memory cell from the bit line.

Programming

The cell programs using high efficiency source-side channel hot electron injection. The conditions for programming are in Table 1, reference Figures 3 and 4 for identification of terminals. The intrinsic (i.e., UV erased) floating gate threshold is positive; thus, the memory cell is essentially non-conducting, with the word line at the reference voltage during a Read cycle.

During programming, a voltage higher than the threshold V_T 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 at $\approx V_{SS}$, if the cell is to be programmed. If the drain is at V_{DD}, programming is inhibited. The drain voltage is transferred across the select channel because of the voltage on the control gate. The source is at $\approx$ 12 volts. The source to drain voltage differential (i.e., 12 volts - $\approx V_{SS}$) generates channel hot electrons. The source voltage is capacitively coupled to the floating gate. The field between the floating gate and the channel very efficiently ($\approx$ 100%) sweeps to the floating gate those channel hot electrons that cross the Si-SiO_2 barrier height of $\approx$ 3.2 eV.

The programming effect is eventually self limiting as negative charge accumulates on the floating gate. The programming source-drain current is very low; thus, the source voltage can be generated by a charge pump internal to the die. The program time is fast because of the high efficiency of source side injection. The addition of negative charge to the floating gate neutralizes the positive charge generated during erasing; thus, the cell is nonconducting when the reference voltage is applied during a Read cycle.

Programming is performed by fixed program pulses generated by an internal timer.

Program Disturb

The memory cells are arranged in a true cross point array, using a word line and bit line for address location selection; thus, unselected cells within a sector will see the programming voltages. There are two types of possible program disturbs with the field enhanced tunneling injection cell, both of which are described in the following paragraphs. Both mechanisms are preventable by proper design and processing. Defects are screenable with testing. Devices with this memory architecture do not have program disturb caused by accumulated Erase/Programming cycles because each sector is individually isolated. Each cell is only exposed to high voltage within the selected sector along the row or source line; there is no high voltage on the bit line.

Reverse Tunnel Disturb

Reverse tunnel disturb can occur for unselected erased cells within the sector sharing a common source line, but on the other row of the selected sector to be programmed; thus, the 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 and the floating gate, Fowler-Nordheim tunneling may occur. This could program the unselected erased cell. Proper design and processing assures the reverse tunnel voltage is significantly higher than any applied voltage. Defects are eliminated by including a reverse tunnel voltage screen in the 100% testing operations. Forward tunneling is defined as occurring when electrons are transferred from poly 1 (the floating gate) to poly 2 (the control gate), thereby erasing the cell. Reverse tunneling is defined as occurring when electrons are transferred from poly 2 to poly 1, thereby programming the cell.

Punch through Disturb

Within a sector, punch through disturb can occur for erased cells in the adjacent inhibited word line, that share a common source line and bit line with the cell being programmed. An inhibited word line is grounded to prevent normal channel hot electron injection. If there is a defect that reduces channel length and creates punch through along the select gate channel, there could be hot electrons available to program the inhibited erased cell. Proper design and processing assures the punch through voltage is significantly higher than any applied voltage. Defects are eliminated by including a punch through voltage screen in the 100% testing operation.

OTHER RELIABILITY CONSIDERATIONS

Oxide Integrity

All oxides are subject to time dependent dielectric breakdown (TDDB), i.e., for a given oxide and electric field, eventually the oxide will breakdown. The lower the electric field and the less time the field is applied, the longer the time to breakdown. For oxides used in normal TTL voltage circuits, this time is essentially infinite; however, in flash memories that use high voltages, the time of oxide exposure to high electric fields can contribute to the intrinsic device reliability.

SST's memory cell uses an electric field during erasing that is roughly half that used by stacked gate flash approaches, thin-oxide E^2PROM and NAND flash approaches. Since the oxide time-dependent breakdown rate is an exponential function of the field strength, the SST memory cell intrinsically has a much lower failure rate than stacked gate cell for oxide breakdown. Note, the SST cell is exposed to the lower electric field for significantly less time during erase, compared with stacked gate approaches.

Contact Integrity

All memory arrays contain metal to silicon contacts, typically from the metal bit line to the diffused drain of the memory cell. Stacked gate and the SST memory cells use a standard cross-point array, whereby a contact is shared by every two memory cells; thus, there are many contacts in a large memory array, e.g., a 4 Mbit chip contains over 2,000,000 contacts. Contacts must have a very low failure rate because there are so many of them. Contacts and associated metal lines are subject to failure based on the current density passing through the contact and metal line. The lower the current density, the lower the potential failure rate due to contact damage or electromigration mechanisms.

The source-side channel hot electron injection current used in programming SST cells is significantly lower than the drain-side channel hot electron injection current used in programming stacked gate cells. During programming, SST cells use less than 1-10 µA of source-drain current; this is much less than the read cell current. In contrast, a stacked gate cell requires 500-1,000 µA of source-drain current during programming; which is much higher than the read cell current. The high programming current density in stacked gate cells results in a higher probability of failure due to contact damage or electromigration. Since the programming current for the SST cell is much lower than the read current, there is no increase in the reliability failure rate due to programming induced current density failure mechanisms.

Fowler-Nordheim tunneling used for erase is intrinsically a low current operation. Therefore, both the SST and stacked gate cells are not measurably affected by current density during the Erase operation.

Data Retention

The field enhancing tunneling injector cell uses relatively thick oxides, compared with other E^2PROM or flash EEPROM cells; therefore, intrinsic data retention is robust. The thicker oxides minimize initial and latent oxide defects; thus, improving yield and oxide integrity. 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 field enhancing tunneling injector cell uses a relatively thick oxide for the Fowler-Nordheim tunneling transfer 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 an ≈ 20 Angstroms 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.

Disturbs

A major concern of reprogrammable nonvolatile memories is that of “disturb” phenomena, i.e., where a different location than the one being erased or programmed is altered. “Disturbs” can occur whenever a high voltage is applied to the gate, source, or drain of a memory cell that is not being intentionally erased or programmed. The SST cell has several design advantages to reduce the possibilities for a disturb:

1. There is no high voltage placed on the bit line, as is common for stacked gate approaches. In addition, the split-gate cell isolates each memory storage node from all other nodes along the bit line. Thus, a disturb via the bit line (connected to the drain) is not possible.
2. The device uses a Sector Erase, whereby, all bytes in the sector are erased simultaneously, i.e., see the same high voltage at the same time. Since each sector is isolated from every other sector by the word line selection circuitry, disturbs along the word line (connected to the gate) during erasing are not possible.
3. The device uses a unique source line for each sector, unlike most stacked gate devices that have the source line common to large sectors or the entire array. This limits exposure to disturb conditions to only the cells within a sector during the time that sector is being programmed. This greatly reduces the probability of a disturb and eases the detection, i.e., only the sector being programmed need be verified after any programming operation.

Life Test (Dynamic Burn-in)

The field enhancing tunneling injector cell uses standard CMOS technology in both the periphery and memory array; therefore, the life test results will be comparable to other devices built with the same process technology. As with all floating gate reprogrammable nonvolatile memories, life test results for a given technology will generally be better than other memories, e.g., SRAMs, built with the same technology because of the standard endurance and data retention infant mortality screening.

Silicon Storage Technology, Inc. • 1171 Sonora Court • Sunnyvale, CA 94086 • Telephone 408-735-9110 • Fax 408-735-9036
www.SuperFlash.com or www.sst.com