Reliability Issues of Flash Memory Cells

SEIICHI ARITOME, RIICHIRO SHIROTA, GERTJAN HEMINK, TETSUO ENDOH, AND FUJI0 MASUOKA

Invited Paper

The reliability issues of Flash electrically erasable programmable read-only memory (Flash EEPROM) are reviewed in this paper. The reduction of the memory cell size and improvement in the reliability have been realized by several breakthroughs in the device technology; in particular, the reliability of the ETOX and NAND structure EEPROM will be discussed in detail. Flash EEPROM is expected to be a very promising device for a large nonvolatile memory market. One of the most promising applications is the replacement of the conventional magnetic hard disk by nonvolatile memories.

I. INTRODUCTION

Recent progress in computers requires further efforts in developing higher density and higher reliability nonvolatile semiconductor memories. A breakthrough in the field of nonvolatile memories was the invention of the Flash EEPROM [1]. The Flash EEPROM has many advantages in comparison with other nonvolatile memories. Therefore, the Flash EEPROM explosively accelerated the development of higher density EEPROM’s. This paper describes the technical trends and reliability issues of Flash EEPROM’s.

In the following section, the trends in Flash EEPROM development will be reviewed. After that, the reliability of several types of Flash memories will be discussed. And finally, the reliability of thin oxide and interpoly dielectrics are discussed.

II. HISTORY OF THE FLASH EEPROM

Table 1 shows the history of Flash EEPROM development. The first modern Flash EEPROM was proposed at the 1984 IEDM by Masuoka et al. [1]. Fig. 1 shows the cell structure of this first Flash EEPROM. The Flash EEPROM is realized in a three-layer polysilicon technology. The first polysilicon is used as the erase gate, the second polysilicon as the floating gate, and the third polysilicon is used as the control gate. Programming can be performed by hot electron injection from the channel, similar to that of the ultraviolet erasure-type EPROM (UV-EPROM). The erasure is carried out by extracting the electrons from the floating gate to the erase gate for all bits at the same time. This EEPROM structure was called a Flash EEPROM because the complete memory array could be erased very quickly. At present, the name Flash EEPROM is used for all EEPROM’s in which all or a large number of cells, called a block or a page, are erased at the same time.

The Flash EEPROM has several advantages in comparison with UV-EPROM. The erasing time is less than 1 s, while the erase time for an UV-EPROM is about 10 min. Moreover, the UV-EPROM needs an expensive package because of the UV transparent quartz window, and the UV-EPROM has to be taken out of the system for the erase. Thus, some companies have stopped developing UV-EPROM in favor of the development of the Flash EEPROM.

For erasure, the conventional EEPROM can achieve selective erase of each byte, while the Flash EEPROM erases all bits at the same time. This is a disadvantage of the Flash EEPROM in comparison with the conventional EEPROM. However, the cell size of the Flash EEPROM is less than that of the conventional EEPROM, and is even comparable with EPROM and DRAM. Therefore, the cost per bit is greatly reduced. Now, the market of the Flash...
Table 1  History of Flash Memory

<table>
<thead>
<tr>
<th>Year</th>
<th>Feature</th>
<th>Authors</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>EAROM</td>
<td>D. C. Guterman et al. [37]</td>
<td>TI</td>
</tr>
<tr>
<td>1984</td>
<td>Flash memory</td>
<td>F. Masuoka et al. [1]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1985</td>
<td>Flash memory (256 kb)</td>
<td>F. Masuoka et al. [2, 3]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1985</td>
<td>Source-erase type Flash</td>
<td>S. Mukherjee et al. [4]</td>
<td>EXCEL</td>
</tr>
<tr>
<td>1987</td>
<td>Drain-erase type Flash (128 kb)</td>
<td>G. Samachisa et al. [5, 6]</td>
<td>Seeg, UCB</td>
</tr>
<tr>
<td>1987</td>
<td>NAND structure EEPROM</td>
<td>F. Masuoka et al. [7]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1987</td>
<td>Source-erase type Flash</td>
<td>H. Kume et al. [8]</td>
<td>Hitachi</td>
</tr>
<tr>
<td>1988</td>
<td>ETOX-type Flash (256 kb)</td>
<td>V. N. Kynett et al [9, 10]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1988</td>
<td>ETOX-type Flash</td>
<td>S. Tam et al. [12]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1988</td>
<td>ETOX-type Flash, reliability</td>
<td>G. Verma et al. [13]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1988</td>
<td>NAND EEPROM</td>
<td>M. Momodomi et al. [14]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1988</td>
<td>Contactless Flash</td>
<td>M. Gill et al. [16]</td>
<td>TI</td>
</tr>
<tr>
<td>1989</td>
<td>Contactless Flash (256 kb)</td>
<td>S. D’Arrigo et al. [17]</td>
<td>TI</td>
</tr>
<tr>
<td>1989</td>
<td>Gate-negative erase</td>
<td>S. Haddad et al. [18]</td>
<td>AMD</td>
</tr>
<tr>
<td>1989</td>
<td>NAND EEPROM (4 Mb)</td>
<td>M. Momodomi et al. [19]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1989</td>
<td>ETOX-type Flash (1 Mb)</td>
<td>V. N. Kynett et al. [20]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1989</td>
<td>Contactless Flash</td>
<td>M. Gill et al. [22]</td>
<td>TI</td>
</tr>
<tr>
<td>1989</td>
<td>Punchthrough erase</td>
<td>T. Endoh et al. [23]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1990</td>
<td>Well-erase</td>
<td>S. Aritome et al. [24]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1990</td>
<td>Contactless Flash, ACEE</td>
<td>B. Riemenschneider et al. [26]</td>
<td>TI</td>
</tr>
<tr>
<td>1990</td>
<td>NAND EEPROM, well erase</td>
<td>R. Kirisawa et al. [27]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1990</td>
<td>FACE cell</td>
<td>B. J. Woo et al. [28]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1990</td>
<td>Gate-negative erase</td>
<td>N. Aikawa et al. [29]</td>
<td>Mitsubishi</td>
</tr>
<tr>
<td>1990</td>
<td>Contactless Flash</td>
<td>M. Gill et al. [30]</td>
<td>TI</td>
</tr>
<tr>
<td>1990</td>
<td>Bipolarity Write/Erase</td>
<td>S. Aritome et al. [31]</td>
<td>Toshiba</td>
</tr>
<tr>
<td>1991</td>
<td>PB-FACE cell</td>
<td>B. J. Woo et al. [32]</td>
<td>INTEL</td>
</tr>
<tr>
<td>1991</td>
<td>Burst-pulse erase</td>
<td>N. Kodama et al. [33]</td>
<td>NEC</td>
</tr>
<tr>
<td>1991</td>
<td>Sector-erase</td>
<td>H. Kume et al. [34]</td>
<td>Hitachi</td>
</tr>
<tr>
<td>1991</td>
<td>Flash cell, scaling</td>
<td>K. Yoshikawa et al. [35]</td>
<td>Toshiba</td>
</tr>
</tbody>
</table>

EEPROM is as large as that of the conventional EEPROM, and will grow up.

Because of these advantages, the Flash EEPROM has been actively developed in many LSI companies, as shown in Table 1.

In 1985, Mukherjee et al. proposed a source-erase type Flash memory cell [4], called the ETOX (EPROM with tunnel oxide) cell. The structure of this cell is the same...
as that of the UV-EPROM. The cell is programmed to a high threshold state by means of channel hot electron injection, with the control gate and the drain connected to a high voltage. Erasing the cell to a low threshold state is performed by Fowler-Nordheim tunneling of electrons from the floating gate to the source diffusion layer by grounding the control gate and applying a high voltage to the source diffusion. Erasure can be achieved with less than 15 V on the source diffusion layer. Since 1985, the performance of this ETOX flash memory cell has been improved [Fig. 2(b)] by several companies; as a result, 8 Mb flash memories can be fabricated at this moment.

At the 1987 ISSCC, Samachisa et al. [5] proposed a drain-erase type Flash EEPROM cell with two levels of poly-silicon, as shown in Fig. 2(c). The cell can be thought of as two transistors in series. One is a floating-gate memory transistor, similar to an EPROM cell. The other one is a simple enhancement transistor controlled by the control gate. This series enhancement transistor is used as a select transistor. The cell is programmed by hot electron injection and erased by electron tunneling between the floating gate and drain.

In 1987, a NAND structured cell was proposed by Masuoka et al. [7]. This structure reduces the cell size without scaling of the device dimensions. The NAND structure cell arranges a number of bits in series, as shown in Fig. 3. The current EPROM cell has one contact area per 2 b. However, for a NAND structure cell, only one contact hole is required per two NAND structure cells. As a result, the NAND cell can realize a smaller cell area per bit than the current EPROM. Several improvements of the NAND structure cell make it possible to operate the cell with a single 5 V power supply and to use sector erase [11], [14], [19], [24], [25], [27].

III. RELIABILITY OF THE ETOX FLASH EEPROM

In this section, the cell operation and the reliability of the ETOX-type flash cell will be discussed. First, the operation of the cell will be discussed. After that, reliability-related issues, such as disturb, endurance, and retention, will be discussed.

A. Cell Operation

Programming of the cell is done in a similar way as in traditional EPROM's, by channel hot electron injection (CHE), as shown in Fig. 4. The cell is programmed to a high threshold state "0," as shown in Fig. 5. A shallow drain junction and an optimized channel profile are used to enhance CHE injection. Electrical erasure is accomplished by Fowler-Nordheim tunneling of electrons from the floating gate to the source diffusion, as shown in Fig. 6. The cell is erased to a low threshold state. A graded source junction [8] permits the application of high voltages.

B. Disturb Mechanisms

The two principal memory cell disturb mechanisms that can occur during programming of an array are called gate-disturb (DC program) and drain-disturb (program disturb)
[13]. These mechanisms can occur in memory cells sharing a common word line (WL) or a common bit line (BL) while one of the cells is programmed. The effect of these disturbs on the different cells of the memory array is shown in Fig. 7.

Gate-disturb occurs in unprogrammed or erased cells which are connected to the same word line as the cell that is being programmed. These cells have a low cell threshold voltage. During the programming operation, the common word line is connected to a high voltage. The electric field across the tunnel oxide becomes high, and may cause tunneling of electrons to the floating gate from the substrate. The threshold voltage of the cell will increase, and in severe cases the cell is programmed unintentionally.

Drain-disturb occurs in programmed cells, which are on the same bit line as the cell that is being programmed. These cells will experience a high electric field between the floating gate and the drain. This may cause electrons to tunnel from the floating gate to drain, and lead to a reduced cell threshold voltage.

An important design consideration is a proper selection of the programming voltages to minimize these disturb. The write/erase cycling of the cell also affects these disturbs. Verma et al. [13] showed that the drain-disturb characteristics of the flash memory devices are excellent with no measurable change until several thousand cycles. However, write/erase cycling has an influence on the gate-disturb behavior. Fig. 8 shows the gate-disturb time as a function of cycling. Before cycling, the disturb time is about 100 s. After 100 cycles, this margin is decreased about two-three orders of magnitude. This degradation of the gate-disturb is caused by hole trapping during erasing, as will be shown in the following section [13].

C. Overerasing

As shown in Fig. 5, the threshold voltage of a cell in the programmed state is higher than 6.5 V. And the threshold voltage of a cell in the erased state is limited from 0.5 to 3.5 V. Electrical erase is not self-limiting. Therefore, electrical erase can leave the floating gate positively charged, thus turning the memory transistor into a depletion-mode transistor; this is called overerasing. The bit-line leakage current caused by this depletion-mode transistor may cause problems during the reading of the cells.

In order to solve this overerasing problem, several methods were proposed [21], [9], [20]. A series enhancement transistor can be used to prevent the leakage current, as has been shown in the previous section. Another similar method is to use the sideways select-gate type cell. Naruke et al. proposed a Flash EEPROM cell which consists of a stacked-gate MOSFET with a sidewall select-gate on the source side (SISOS cell) [20], as shown in Fig. 9. The cell has a select-gate which prevents undesirable leakage current due to overerasing. This cell has some additional advantages. The cell is programmed by channel hot electron injection at the source side, and is erased by Fowler–Nordheim tunneling of electrons from the floating gate to the drain. The programming by source side injection makes it possible to optimize the drain junction for erase by Fowler–Nordheim tunneling without having an influence on the generation of hot electrons.

The verified-erase method [20] is another means used to prevent overerase. An erase step is carried out by raising the source junction of all cells in the memory array to the erase voltage with all control gates grounded. Subsequently, a read operation is performed with a voltage of 3.2 V applied to the control gate. This 3.2 V is the upper limit for the threshold voltage of a cell in the erased state. If some
bits require more time to reach the erased state, erasing is performed again. The erase verify sequence is repeated until all cells in the array have a threshold voltage less than or equal to the 3.2 V maximum. It is shown in [20] that this verified erase method effectively suppresses overerasing.

A self-convergence erasing scheme, which is proposed by Yamada et al. [36], can also be used to prevent overerasing. This method uses avalanche hot carrier injection after erasure by Fowler–Nordheim tunneling. The threshold voltages converge to a certain "steady state," as shown in Fig. 10. The steady state is caused by a balance between avalanche hot electron injection and avalanche hot hole injection into the floating gate. A tight distribution of the threshold voltages and stable erasure without overerase can be accomplished in this way.

**D. Write/Erase Cycling Endurance**

Due to the presence of a high voltage at the gated-diode junction during erasing, holes are inevitably generated by band-to-band tunneling [38], [39], and a small amount of them are injected into the oxide after being accelerated in the depletion region. Hot-hole injection during erasing has been reported to cause variations in the erased threshold voltages of the cells in the memory array [8], and trapped holes in the oxide were shown to degrade the charge retention of memory cells [13], [18].

In order to suppress the degradation of memory cells, several solutions have been proposed. The simplest method is to reduce the applied voltage to the source. Table 2 shows the data loss of a flash memory cell after various cycling conditions. The average charge loss is small in the case of a low applied voltage to the source because the hot-hole injection due to band-to-band tunneling is suppressed. The low electric field across the tunnel oxide during erasing also results in excellent write/erase endurance characteristics. Fig. 11 shows the $V_{th}$ distribution versus write/erase cycles [20]. The plot shows that even after one million write/erase cycles, a program read margin of more than 2 V exists.

In order to realize both a high erase speed and an acceptable write/erase endurance, negative gate erasure has been proposed [18]. Haddad et al. [18] compared the gate negative erase and the source erase method. Fig. 12 shows the gate-disturb after negative gate erasing and sources erasing. No threshold voltage shift is observed for the cell erased under the gate negative erasing. This indicates that hot-hole injection is effectively suppressed by gate negative erasing. However, the gate-disturb after source erasing is significant. These data strongly support a model of hot-hole injection and hole trapping in the oxide during source erasing.

**E. Data Retention**

In floating-gate memories, the stored charge can leak away from the floating gate through the gate oxide or through the interpoly dielectrics. This leakage, caused by mobile ions, oxide defects, or other mechanisms, results
in a shift of the threshold voltage of the memory cell. The threshold voltage shift can also be caused by the detrapping of electrons or holes from oxide traps.

Different charge loss mechanisms have been described in [46], [47]. These mechanisms are briefly discussed below. The first mechanism is thermionic emission over the image force lowered potential barrier. However, this mechanism is not dominating since the barrier height for thermionic emission is considerably higher than the activation energies for the other charge loss mechanism. The second mechanism is called electron detrapping. The activation energy for this type of charge loss is about 1.4 eV. Because the intrinsic charge loss rate decreases with time and stops after a threshold voltage drop of about 0.5 V [47], this type of charge loss is associated with the detrapping of electrons that are trapped in the oxide. The third mechanism is related to defects in the oxide. This mechanism can result in both charge loss and gain, depending on the biasing condition. The activation energy for this kind of charge loss is about 0.6 eV [47]. The fourth mechanism is related to contamination. Positive ions entering the memory cell may compensate a part of the negative charge stored on the floating gate. The activation energy for charge loss by contamination is about 1.2 eV [47].

Kynett et al. showed the results of data retention tests for a 1 Mb Flash memory. Accelerated retention bake experiments done at 250°C for 168 h indicate that after 10⁶ write/erase cycles, flash memory will exhibit only a 0.7 V program Vth shift [20]. Furthermore, a retention bake of 52 h produced less than a 0.5 V Vth shift in programmed devices which have been through more than one million write/erase cycles. The program read margin is more than sufficient to guarantee reliable operation.

IV. RELIABILITY OF THE NAND STRUCTURE CELL

A. Operation of the NAND Structure Cell

In order to realize a small EEPROM memory chip, the NAND structure EEPROM cell was proposed in 1987 [7]. In this structure, the memory cells are arranged in series. Fig. 3 shows a top view and an equivalent circuit with 8 b, arranged in a NAND structured cell. By using 1.0 µm design rules, the cell size NAND structure EEPROM is 12.9 µm². This is only 44% of the cell size required by a NOR-structure full-function EEPROM cell [19], [25], and 85% of the cell size required by an ETOX cell [20].

The operation mechanism of a single-memory transistor of the NAND structure is comparable with that of the conventional EEPROM. However, the programming and reading the memory cells are more complex. Therefore, additional peripheral circuitry is required, and the reading speed is lower than that of the conventional type because a number of memory cells are connected in series.

The reading method is shown in Fig. 13 and is essentially the same as that of the NAND-type MASK ROM. 0 V is applied to the gate of the selected memory cell, while 5 V is applied to the gate of the other cells. Therefore, all of the other memory transistors, except for the selected transistor, serve as transfer gates. As a result, in the case where a “0” is being written, the memory transistor is in the depletion mode and a current flows. On the other hand, current does not flow in the case where a “1” is written because the memory transistor is in the enhancement mode. The state of the cell is detected by a sense amplifier that is connected to the bit line. The difference between a “0” and a “1” stands for whether negative charge is stored in the floating gate or not. If negative charge is stored in the floating gate, the threshold voltage becomes higher and the memory transistor is in the enhancement mode. This is comparable with the conventional EEPROM.

With regard to the programming and erasing methods, two different methods have been proposed. At the first stage of the NAND EEPROM development, a nonuniform write and uniform erase technology was introduced. After that, a uniform write and uniform erase technology was proposed to improve the endurance and the retention.

In the case of nonuniform write and uniform erase technology, electrons are injected into the floating gates of all of the memory transistors at first, which shifts all
of the memory transistors into the enhancement mode. Subsequently, the negative charge is removed only from the floating gate of the memory transistor which is writing "0" data. As a result, this memory transistor turns to the depletion mode.

Fig. 15 shows an example of writing data in cell 1. 0 V is applied to the select transistor on the earth side to prevent a bit-line current during writing. The gate voltage of the selected memory cell 1 for writing is 0 V. To the other gates in the selected NAND cell block, 22 V is applied. As a result, the seven cell transistors from cell 2 to cell 8 serve as transfer gates for the bit-line voltage which is transferred to the drain of cell 1. If the bit-line voltage is 22 V, electrons will tunnel from the floating gate to the drain of cell 1. This will turn the cell in the depletion mode. On the other hand, if a voltage of about 10 V is applied to the bit line, no tunneling will occur and the cell will stay in the enhancement mode. Next, Fig. 14(b) shows the case when cell 2 is written. Similar to the case when cell 1 is written, 0 V is applied to the gate of the select transistor on the earth side in order to cut off the current. On the other hand, 0 V is applied to the gate of the previously written cell 1. 0 V is also applied to the gate of cell 2 that is selected for writing. Hence, the drain voltage of cell 2 is not transferred to cell 1. Therefore, a high voltage that is applied for writing cell 2 does not have an effect on the previously written cell 1. Using this procedure, the data are written successively in cells 1–8. The programming of the cells is carried out by Fowler–Nordheim tunneling at the drain side, and is therefore nonuniform.

The erasure is performed simultaneously for 8 b of memory transistors connected in series. 0 V is applied to the earth and data line. 17 V is applied to the gates of the select transistors and the eight memory transistors. Under this condition, electrons can be injected into the floating gates of the memory transistors, and the memory transistors will be in the enhancement mode after this step. The erase is uniform because electrons will tunnel from the whole substrate area to the floating gate, not only from the drain or source.

In order to prevent the degradation of the memory cells due to band-to-band tunneling stress during writing, a new uniform write and uniform erase technology was proposed by Kirisawa et al. [27]. Erasing and writing are both accomplished uniformly over the whole channel area instead of writing nonuniformly at the drain. A new device structure has been proposed to achieve this programming method, as shown in Fig. 15. The NAND cell array and peripheral circuitry are located in different p-well regions (p-well 1 and p-well 2), which are separated from each other. This structure has the advantage of making it possible to apply $V_{pp}$ at p-well 2 while keeping p-well 1 and the n-well at, respectively, $V_{pp}$ and $V_{cc}$. The write operation is performed by injecting electrons from p-well 2 to the floating gate uniformly through the tunnel oxide. On the other hand, the erase operation is performed by emitting electrons from the floating gate to p-well 2 uniformly.

Fig. 16 shows the equivalent circuit and the operating voltages. During erasing, 20 V is applied to both p-well 2 and the n-substrate while keeping the bit lines floating and all control gates grounded. Electrons will tunnel from the floating gate to the substrate, and the threshold voltage for all memory cells becomes negative. During writing, 20 V is applied to the selected control gate, and the bit lines are grounded; electrons tunnel from the substrate to the floating gate, resulting in a positive threshold voltage.

B. Write and Erase Endurance

The reliability of these two write/erase methods in the NAND-structured EEPROM have been compared [24]. The write and erase endurance characteristics are shown in Fig. 17. The uniform write and uniform erase technology guarantees a wide cell threshold window of as large as 4 V, even after 1 million write/erase cycles [Fig. 17(a)]. However, the threshold window obtained by the uniform erase and nonuniform write technology begins to reduce rapidly at around $10^9$ write and erase cycles, and fails at $10^9$ write and erase cycles [Fig. 17(b)] because Fowler–Nordheim tunneling current during nonuniform writing is confined to a small region at the drain. Therefore, electron traps in the tunnel oxide are generated at a high rate near the drain area, and these electron traps impede electron injection.
and emission between the floating gate and substrate. This effect may be aggravated by hot holes that are generated by band-to-band tunneling.

In uniform write and uniform erase technology, the threshold voltage of the erased cell is dependent on the number of write/erase cycles. However, the threshold voltage of the written cell is not dependent on the number of write/erase cycles. This can be explained as follows. The oxide traps and interface traps are generated uniformly over the entire channel area because Fowler–Nordheim tunneling of electrons is performed uniformly during both the writing and erasing operation. The uniformly trapped oxide charges over the channel area affect not only the electron tunneling current through the oxide, but also the flatband voltage. The threshold voltage of the erased cell decreases slightly up to $10^3$ cycles due to hole trapping. Hole traps affect the increase of stored positive charge on the floating gate as well as the decrease of the flatband voltage. After $10^3$–$10^6$ cycles, the threshold voltage of the erased cell increases due to electron detrapping. On the other hand, the threshold voltage of the written cell remains almost constant up to $10^6$ cycles in spite of charge trapping in the oxide because the influence of stored charge on the floating gate and trapped oxide charge on both the flatband voltage and injection field cancel each other out [48].

C. Data Retention

Data retention characteristics of the memory cell programmed by the two write and erase technologies measured at 300°C after different numbers of write and erase cycles from 10 to $10^6$ were carried out (Fig. 18). In the case of uniform erase and nonuniform write technology, the stored positive charges gradually decay as baking time increases, so the threshold window decreases [Fig. 18(b)]. However, in the case of uniform write and uniform erase technology, the stored positive charges effectively increase up to 100 min baking time due to the detrapping of electrons from the gate oxide to the substrate during the retention bake [Fig. 18(a)]. This increase of effectively stored positive charges up to 100 min retention bake becomes larger with an increasing number of cycles because the number of trapped negative charges in the thin oxide increases.

The effect of detrapping electrons is equivalent to the effect of trapping holes in the gate oxide. As a result, the detrapping of the electrons suppresses the data loss of the positively charged cell. As a result, the stored positive charges which are effectively increasing at the beginning of the bake extend the data retention time of the memory cell programmed by the uniform write and uniform erase. Fig. 19 shows the data retention time after write and erase cycling. The data retention time can be extended by using uniform write and uniform erase technology, especially beyond $10^5$ write and erase cycles.

V. RELIABILITY OF THE TUNNEL OXIDE AND THE INTERPOLY DIELECTRICS

A. Stress-Induced Oxide Leakage Current

The reliability of the tunnel oxide is very important for Flash EEPROM's. These devices are programmed and erased by high-field (Fowler–Nordheim) injection of electrons in a very thin dielectric film to charge and discharge the floating gate. Unfortunately, this current through the oxide degrades the quality of the oxide and eventually leads to breakdown. The wearout of tunnel oxide films during a high-field stress has been correlated with the buildup of both positive or negative trapped charges.

Recently, it has been shown that high-field stress also induces a low-field leakage current in thin oxides [40], [41], with a thickness less than 10 nm. The mechanism of this leakage current has been attributed to the generation of localized defects or weak spots [40] and trap states near...
the injecting interface [41], [42]. These low-field leakage currents degrade the data retention of the EEPROM cell [41], [43], [31]. Unfortunately, these oxide leakage currents increase with decreasing oxide thickness [40], [43], and make it difficult to scale down the oxide thickness of the memory cell.

Aritome et al. [31] studied the influence of the waveform of the stress voltage on the degradation of the tunnel oxide. Three types of high-field dynamic stress were used to study the thin oxide leakage currents. Fig. 20 shows the applied dynamic stress waveforms. In the case of bipolarity stress, positive high voltages are applied to the gate or to the substrate and source/drain (S/D) regions. Fowler–Nordheim tunneling occurs both from the substrate to the gate and from the gate to the substrate alternately. Electron-emitted stress, emission of electrons from the gate and electron-injected stress, emission of electrons from the substrate are performed for comparison. The applied gate voltage is comparable to the floating gate voltage of a memory cell during the write operation. The substrate voltage ($V_{\text{sub}}$) is chosen in such a way that the tunnel current during emitted and injected stress is the same.

The $I_g-V_g$ characteristics before and after dynamic stress for a 5.6 nm oxide thickness are shown in Fig. 21. The thin oxide leakage currents subjected to three types of high-field dynamic stressing are compared. It is observed that the thin oxide leakage current induced by bipolarity stress is about one order of magnitude smaller than that induced by both the electron-emitted stress and the electron-injected stress. This result shows that the origin of the thin oxide leakage current can be suppressed by a reverse high-field stress.

The data retention characteristics of the Flash memory cell programmed by two different write/erase (W/E) technologies are compared [31]. Fig. 22 shows these two W/E technologies. First, we have the bipolarity W/E technology, which is a uniform write and erase technology. During the write operation, a high voltage ($V_{\text{wp}}$) is applied to the control gate, with the substrate and source/drain regions grounded. Electrons are injected from the substrate to the floating gate over the whole channel area of the memory cell. In the erase operation, a high voltage ($V_{\text{sub}}$) is applied to the substrate and source/drain regions, with the control gate grounded. Electrons are then emitted from the floating gate to the substrate. In this write/erase method, the high-field stress of the thin oxide corresponds to bipolarity stress. The other write/erase method is a channel hot electron (CHE) write and Fowler-Nordheim tunneling erase technology. This technology is a nonuniform write and uniform erase technology. The erase operation is the same as in the bipolarity W/E technology. However, during the write operation, high voltages are applied to the control gate and the drain. Thus, channel hot electrons are generated by the lateral electric field, and electrons are injected from the substrate to the floating gate. In this case, the high-field stress of the thin oxide is electron-emitted stress. The data retention will be different between these W/E technologies because the thin oxide leakage current is different between the bipolarity stress and the electron-emitted stress of the thin oxide. Moreover, in the conventional erasing method of Flash memories [12], a high voltage is applied to the source. However, in our experiment, a high voltage is applied to the substrate as well as the source/drain regions in order to prevent the degradation of the thin oxide due to hole injection caused by band-to-band tunneling [12].

Fig. 23 shows the write and erase endurance characteristics of both W/E technologies. No closure of the cell threshold window occurs up to $10^9$ write/erase cycles in both technologies.
Data retention characteristics are measured under various gate voltage conditions in order to accelerate the retention test. The memory cells are subjected to $10^5$ write/erase cycles. In the case of the CHE write and F-N tunneling erase technology, the stored positive charges rapidly decay as a function of time; as a result, the threshold window decreases (Fig. 24). However, in the case of the bipolarity F-N tunneling WE technology, data loss of the stored positive charge is significantly reduced. This phenomenon can be explained by the fact that the thin oxide leakage current is reduced by the bipolarity F-N tunneling stress.

Fig. 25 shows the data retention time after write and erase cycling as a function of the tunnel oxide thickness. The data retention time is defined by the time at which $V_{th}$ reaches $-1.0$ V during the gate voltage stress. In devices with a 7.5 nm tunnel oxide, the data retention time obtained with the bipolarity F-N tunneling write/erase technology is 50 times longer than that of CHE write and F-N tunneling erase technology after $10^5$ cycles. However, in devices with a 9 nm tunnel oxide thickness, the data retention time is almost the same for both technologies. For very thin (<9 nm) tunnel oxides, the bipolarity write/erase technology offers improved data retention times in comparison with the CHE write and F-N tunneling erase technology. Therefore, this technology may facilitate the down-scaling of the tunnel oxides. Reducing the tunnel oxide thickness results in lower programming voltages and in faster read operations because the read current is increased.

B. Time-Dependent Dielectric Breakdown
Under High-Frequency Stress

The time-dependent dielectric breakdown (TDDB) characteristics of the tunnel oxide will be discussed here. In particular, it will be shown that TDDB characteristics are dependent on the stress waveform and its frequency.

Fig. 26 shows that the TDDB characteristics depend on both the stress waveform and its frequency. The stress waveforms are the electron-emitted stress, the electron-injected stress, and the bipolarity stress. The time to breakdown during bipolarity stress is longer than that during both the electron-emitted stress and the electron-injected stress. Moreover, it is shown that the TDDB characteristics depend on the frequency of the stress wave. The breakdown time under unipolarity stress does not depend on frequency. However, the breakdown time of bipolarity stress strongly depends on the frequency. The bipolarity stress lifetime in the megahertz range is about 40 times higher than that at 10 Hz at the same applied voltage. It is suggested in [45] that the improved TDDB
obtained with the bipolarity stress is due to the field-assisted detrapping of holes at the Si/SiO2 interface.

C. Interpoly Dielectrics

ONO (oxide/nitride/oxide) interpoly dielectrics are currently used for nonvolatile memories. In the case of nonvolatile memories, the important issues are not only low defect density and long mean time to failure, but also charge retention capability. Mori et al. investigated the ONO interpoly dielectrics thickness scaling effect on charge retention characteristics in nonvolatile memories systematically [49].

For the top oxide layer, a certain thickness (3 nm) of top oxide, which can block the hole injection, is required. SiN thickness reduction leads to reduced initial rapid charge loss. However, too much reduction sometimes leads to enhanced charge loss in the long bake test. Bottom polypoxide thinning, down to 10 nm, does not result in degradation. The NO (nitridedioxide) double-layer structure shows a large initial rapid charge loss, which may be prevented by applying the ONO structure. However, if the bottom oxide layer is thin and its quality is poor, a part of the bits may suffer from rapid initial charge loss. When scaling down the ONO thickness, one must consider these key factors concerning charge retention as well as dielectric reliability.

VI. CONCLUSIONS

Reliability issues of Flash EEPROM's are reviewed. The reliability of both the source-erase type (ETOX) Flash memory and the NAND structure EEPROM have been discussed. Disturbs during programming, read/erase endurance, and charge loss of both devices are reviewed. The reliability of the tunnel oxide and the interpoly dielectric have also been addressed. The bipolarity F-N programming/erase, which is used in the NAND EEPROM, improves the charge to breakdown (Qbd) and decreases the stress-induced leakage current. Therefore, the bipolarity write/erase method is very promising for application in future Flash EEPROM devices.

VII. ACKNOWLEDGMENTS

The authors would like to thank R. Kirisawa, R. Nakayama, T. Yamaguchi, T. Hatakeyama, S. Iizuka, S. Shutoh, M. Momodomi, T. Tanaka, Y. Tanaka, H. Nakamura, H. Oodaira, S. Watanabe, and K. Ohuchi for their helpful discussion on this work.

REFERENCES

Seishi Arirome was born in Hiroshima, Japan, on August 9, 1960. He received the B.E. and M.E. degrees in electronic engineering from Hiroshima University, Japan, in 1983 and 1985, respectively. He joined the Toshiba Research and Development Center, Kawasaki, Japan, in 1985. Since then he has been engaged in the development of high-density DRAM's and EEPROM's at the Toshiba ULSI Research Center.

Mr. Arirome is a member of the Japan Society of Applied Physics.

Riichiro Shirota was born in Hyogo, Japan, on February 22, 1954. He received the B.E., M.E., and Ph.D. degrees in physics in 1977, 1979, and 1982, respectively, from Nagoya University.

He joined the Toshiba Corporation, Kawasaki, Japan, in 1982. Since then he has been engaged in research and development of high-density DRAM's and EEPROM's at the Toshiba ULSI Research Center.

Gertjan Hemink was born in The Netherlands in 1964. He received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Twente in 1988 and 1992, respectively.

At the University of Twente he worked on the development and application of nonvolatile memory cells with a buried injector structure. In July 1992 he joined the Device Research Department of the Toshiba Research and Development Center, Kawasaki, Japan. His main research interests are nonvolatile memory devices in general, hot electron modeling, and the modeling of tunnel oxide degradation.

Tetsuo Endoh was born in Tokyo, Japan, on October 30, 1962. He received the B.S. degree in physics from the University of Tokyo, Japan, in 1987.

He joined the Toshiba Research and Development Center, Kawasaki, Japan, in 1987. Since then he has been engaged in research and development of high-density EEPROM's at the Toshiba ULSI Research Center.
Fujio Masuoka was born in Takasaki, Japan, on May 9, 1943. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Tohoku University, Sendai, Japan, in 1966, 1968, and 1971, respectively. In 1971 he joined the Toshiba Research and Development Center, Toshiba Corporation, Kawasaki, Japan, where he has been working on MOS LSI memory. He is currently managing the Device Research Department of Toshiba ULSI Research Center. His research interests have been in memory cells such as EPROM, EEPROM, MROM, SRAM, and DRAM.

Dr. Masuoka received the Watanabe Award and the National Invention Award in 1977 and 1978, respectively. He is a member of the Institute of Electronics, Information, and Communication Engineers of Japan.