Analysis and test procedures for NOR flash memory defects

Mohammad Gh. Mohammad\textsuperscript{a,*}, Kewal K. Saluja\textsuperscript{b}

\textsuperscript{a}Computer Engineering Department, Kuwait University, P. O. Box 5969, Safat 13060, Kuwait
\textsuperscript{b}Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, United States

Received 27 June 2007; received in revised form 20 November 2007

Abstract

Widespread use of non-volatile memories, especially flash memories, in diverse applications such as in mobile computing and system-on-chip is becoming a common place. As a result, testing them for faults and reliability is drawing considerable interest of designers and researchers. One of the most predominant failure modes for which these memories must be tested is called disturb faults. In this paper, we first analyze different defects that are responsible for disturb faults using a 2-dimension device simulator. We determine the impact of various defects on cell performance and develop a methodology based on channel erase technique to detect these defects. Our tests are efficient and can be converted to march tests prevalently used to test memories. We also propose a very low cost design-for-testability approach that can be used to apply the test technique developed in this paper.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Flash memories are being deployed in almost all computing devices and systems ranging from mobile devices such as cell phones and personal assistants (PDAs) and systems such as laptops and system-on-chip (SoC). It is expected that the market for flash memories of various types, such as NOR based, NAND based, embedded; in portable markets will exceed 20 billion dollars by the year 2010 [1]. In order for the flash memory market to grow as it has been for the past decade, the prices for flash memory should continue to decrease at the same rate, if not faster, to maintain its competitiveness in the memory market. One of the main contributors to the cost of flash memory is its test cost. Hence, methods to reduce the test costs are likely to be key factors in the continued growth of flash memories.

Since the introduction of flash memories substantial research has been directed towards the development of efficient tests for flash memories. The research that is reported in [2–4] has focused on modeling and test of the various failure modes that can occur in flash memories. In the first paper that dealt with modeling faults in flash memories, the authors proposed a logical fault model to model various disturb faults and developed efficient tests for their detection [2]. The authors of [3,4,6] expanded the faults that need to be considered when testing flash memories by including some traditional faults that are present in other type of memories as well as disturb faults that are described by the IEEE standard definitions and characterization of floating gate semiconductor arrays [5]. Further, in [6], built-in self-test for flash memories was proposed to reduce the cost of testing. In all these papers, special attention was paid to develop test algorithms that were efficient in detecting as well as diagnosing all known faults applicable to flash memories.

In this paper, we analyze various defects at different locations in a memory cell that are responsible for disturb faults and study their impact on cell performance using a 2D device simulator named Atlas [7]. After analyzing the behavior of the defective cells, we determine fault excitation conditions that allow fast and reliable identification of faulty cells. Using these excitation conditions, efficient tests for testing NOR type flash memories are developed.
Further, we present a design-for-testability (DFT) approach that can be adapted in a cost-effective manner for current flash memories which will allow prompt and efficient identification of various defects using the test algorithms proposed in this paper. Thus, the contributions of this paper can be stated as follows: (1) identify manufacturing defects that impact cell performance, functionality and reliability, (2) develop appropriate test conditions to excite such defects, (3) develop efficient test algorithms to detect such defects, and (4) propose a DFT approach that can be used in conjunction with the proposed algorithms.

The paper is organized as follows. Section 2 reviews the previously developed models for flash memory faults. Section 3 discusses the experimental setup used in this study for defect injection and their impact. Findings of simulation studies are provided in Section 4. Test algorithms and their detection capabilities are studied in Section 5. In Section 6, we propose a DFT approach that allows the implementation of the proposed test algorithms. Section 7 concludes the paper.

2. Flash memory faults

Semiconductor memories, including all forms of non-volatile memories (NVM) such as flash memories, suffer from defects that could occur during the manufacturing process. The characteristics of the defects, whether large or small, can alter the correct behavior of the memory cells. However, the manifestation and detection of these defects depend on their physical characteristics. For example, when a defect is large and it results in a short/open between a particular line and power supply ($V_{cc}$) or ground (GND), then it can manifest itself as a stuck-at (SAF) or stuck-open fault (SOF) [8]. Alternatively, if the short is between two lines and neither of the lines is $V_{cc}$ or GND, then such a short may result in a coupling fault (CF) or address decoder fault (AF) depending on which two memory lines are shorted together. In other instances, the defects may be too small to result into shorts or opens and may manifest as resistive defects. In which case they have substantially more complex faulty behavior, such as incorrect-read-fault (IRF), or write-disturb-fault (WDF) [9–11]. Even though there exist tests to detect these defects and resulting faults, their testing can be expensive in terms of test time and test cost unless careful attention is paid to derivation of such tests.

In NVMs, particularly in flash memory literature, the faults are classified into two major categories. The first category consists of those faults that are common between flash and all other type of semiconductor memories such as DRAMs and SRAMs. These faults include SAF, SOF, AF, and CFst (state-coupling) faults [4,6]. The second category consists of faults that are specific only to NVMs, including flash memories, and do not conform to the traditional faults known to occur in DRAMs and other types of volatile memories. These faults are known as disturb faults and can be classified as either program or read disturbs [12–15]. A disturb fault is caused by defects in the insulating layer of a core memory element or it may be induced by high/moderate electric stress conditions during the different modes of operations of memory cells. Rest of this section discusses the causes as well as the behavior of these faults and their modeling.

2.1. 1T cell structure and defects

The core memory element in most of today’s flash memories is the 1T floating gate transistor (FG). The structure of the 1T transistor is similar to the traditional MOS transistor, except for the addition of a floating polysilicon gate that is completely insulated by dielectric from all other conducting terminals. The floating gate is insulated from the substrate by a layer of high quality oxide, called tunnel oxide, whereas it is insulated from the top by a oxide-nitride-oxide (ONO) layer, an interpoly insulator, as shown in Fig. 1 [12]. The value stored in the memory cell is represented by the charge on the floating gate. When there is no charge present on the floating gate, the cell is referred to as erased or containing a logic “1” value. On the other hand, when the cell has a negative charge accumulated on it, it is referred to as being programmed or having a logic “0” value.

There are two commonly used methods to transfer charge onto the floating gate, namely by channel hot electron injection (CHEI) or by Fowler-Nordhiem (FN) tunneling. On the other hand, charge extraction from the floating gate can only be performed by FN tunneling. In NOR array organization, the CHEI mechanism is used to program a flash cell and FN tunneling is used to extract the charge from the floating gate [16]. When accumulating or extracting charge of the floating gate, higher than normal voltages are applied on the various terminals of the memory cell to create high electric field across the tunnel oxide to enable the transfer of charge. These high fields that are used to program and erase the cell pose reliability concerns as well. When flash cells are arranged in a memory array, such as NOR organization (shown in Fig. 2)
or NAND organization, many of the unselected cells experience the same high electric fields as the selected cell(s). For example, when addressing a cell \((i,j)\), where \(i\) is the row address and \(j\) is the column address of a cell, all cells on row \(i\) will experience the voltage applied on that row. Similarly, all cells in column \(j\) (whether selected or not for programming or erase operation) will experience the voltage on that column. Designers are aware of these high fields and they realize that these high fields could result in disturb behavior. As a result, the designers normally consider these effects when designing the memory array to avoid unwanted consequences [16] of such disturbs. How-ever some of the issues still remain. For example, the quality of the tunnel oxide must be flawless and manufacturing process must be nearly perfect to ensure the reliability of the memory cell and its correct operation.

During memory device characterization, numerous studies are carried out to find optimal design for desired voltages, array organization, as well as cell structure to meet the targeted design specifications. However, due to the nature of the manufacturing process, contaminants and other anomalies are unavoidable in the core memory cells. When these contaminants or defects are located in the insulating layers of the 1T cell, they pose a reliability concern and amplify the unwanted behavior of disturb. It has been previously argued that any single defect located in the ONO layer cannot result into faulty behavior of the cell [2]. This is due to the fact that the ONO layer is a composite layer and a single defect can not jeopardize the layer’s insulating characteristics. However, defects located in the tunnel oxide could result in faulty behavior [2]. In this paper we investigate both type of defects (i.e. tunnel oxide and ONO layer defects) and their behaviors and the results of our study are discussed in detail in Section 4.

2.2. Fault models for disturbs

In recent years, the interest in flash memory test has been increasing due to (1) special failure modes of flash, (2) its impact on the reliability, and (3) its impact on the cost of the memory. The IEEE standard definitions and characterization of floating gate semiconductor arrays includes nearly all disturbs for all possible memory array organizations and cell structures [5]. In a set of recent papers [2–4,17] the most common disturb behaviors were modeled as logical faults. The following is a short description of four most common program disturbs as well as their fault models for NOR type flash memories. Other disturbs which are not very common or which are applicable to other memory array organizations can be found in [5,17]. The description below uses the notation \(S_i; S_j; F/R\) which is commonly used in representing static coupling faults [18]. In this notation: \(S_i \in \{0,1,r0,r1\}\), is the sensitizing operation sequence (SOS) consisting of read and write operations on the state of the aggressor cell, \(S_j\) is a SOS for the state of the victim (before fault excitation) cell, “\(F\)” \(\in \{0,1\}\) is the state of the faulty cell (after excitation), and \(R \in \{0,1,\ldots\}\) is the output of the read operation. The value “–” in field \(R\) is used in case of write operation.

**Word-line erase disturb (WED)** exists when a cell being programmed (selected cell) causes another unprogrammed cell (unselected cell), sharing the same word line, to be erased. This fault is modeled as \(1w0:0/1/–\) fault.

**Word-line program disturb (WPD)** occurs when a cell being programmed (selected cell) causes another unprogrammed cell, sharing the same word line, to be programmed and is modeled as \(1w0:1/0/–\).

**Bit-line erase disturb (BED)** which is modeled as \(1w0:0/1/–\) fault takes place when a cell being programmed (selected cell) causes another unprogrammed cell, sharing the same bit line, to be erased.

**Bit-line program disturb (BPD)** arises when a cell being programmed (selected cell) causes another unprogrammed cell (unselected cell), sharing the same bit line, to be programmed \(1w0:1/0/–\).

Other disturbs that occur under low electric field stress, such as read disturb, are caused by the same physical defects. Therefore, by identifying defects that cause program disturb faults, we can also detect read disturb faults. In addition to the common disturb faults discussed above, flash and many other NVMs, suffer from various different kinds of failures that are induced by cycling [14,19–22]. Examples of these failures which are considered some of the most important degradations induced by cycling are erratic bits, ultra-fast bits and tail bits. These degradations are believed to be caused by traps in the tunnel oxide which may result in faulty behaviors or performance degradation. Many studies were performed to analyze the cause as well as the behavior of erratic bits and the other failure modes. Appropriate measures to reduce, if not eliminate, such degradations/failures [23–25] have also been developed. Even though such faulty behaviors are important for flash memory reliability, it is
beyond the scope of this paper. In this paper we restrict the discussion to the test of virgin cells.

3. Simulation setup

In order to investigate the various defects in 1T cell, 2D device simulation tools, namely Athena and Atlas, which are part of the Virtual Wafer Fab (VWF) software were used [7, 26]. Using Athena, we constructed a fault-free 1 μm × 1 μm 1T structure with a tunnel oxide of A105 and combined ONO stack thickness of A400. The cell is designed to be programmed by the CHEI using a 1 μs pulse and erased using FN tunneling from the source side. The erase operation takes 10 ms. The bias conditions required to accomplish these operations and the resulting threshold voltages ($V_t$) are shown in Table 1. In this table all values are given in volts (V) and CG, D, S, and B represent the voltages applied at control gate, drain, source, and base terminals respectively. The program/erase $I$–$V$ characteristics of the developed 1T structure are shown in Fig. 3. The erase operation described in Table 1 is the most common approach used for erase. In this approach, known as negative-gate erase (NGE), erase voltages is applied to the control gate as well as to the source terminal of the memory cell. This technique for erase operation is superior than source side erase technique which uses high voltage on the source terminal while grounding the control gate of the cell. The NGE technique results in lower power consumption as well as better cell reliability [16, 27].

Table 1 Program/erase biases and threshold voltages

<table>
<thead>
<tr>
<th>Operation</th>
<th>CG</th>
<th>D</th>
<th>S</th>
<th>B</th>
<th>$V_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>10</td>
<td>6</td>
<td>GND</td>
<td>GND</td>
<td>7.95</td>
</tr>
<tr>
<td>Erase</td>
<td>−8</td>
<td>Floating</td>
<td>7</td>
<td>GND</td>
<td>1.11</td>
</tr>
<tr>
<td>Read</td>
<td>3.3</td>
<td>0.5</td>
<td>GND</td>
<td>GND</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. $I$–$V$ characteristics.

3.1. Stress time calculation

It is discussed in Section 2 that high electric fields in any array organization result in stresses that can lead to device failure. Further, duration and amplitude of these stresses in flash memories are based on the number of cells in each row/column of the array organization. The number of cells in each row or column define the worst case gate or drain stress, respectively, that a cell can undergo. The duration of the worst case stress can be calculated by multiplying the program time with the number of cells in a row/column (i.e. $T_{stress} = T_P \times (N - 1)$, where $T_P$ and $N$ are program time and number of cells in a row/column respectively for an $N \times N$ array organization) [28]. Actual stress time varies from 0.1 ms to 2 ms depending on the program time and the array organization used. In our experiments, we assume a memory array organized as an 128 × 128 grid. This implies that the duration of the worst case gate/drain stress for such an array is 127 μs for a program time of 1 μs. Instead of simulating all stress durations, we chose the following three stress durations: 127 μs, 635 μs, and 10 ms, for the reasons that follow. The first choice represents a typical stress time for the chosen array organization, whereas the remaining two choices are applicable to memories with somewhat different organizations (e.g. 512 × 512 array) and different program time (e.g. 5 μs) combinations. Further, during stress simulation studies, all appropriate physical models that can effect the study of stress conditions, were turned ON to simulate realistic disturb behavior. The models that were activated during the study include FN and Band-to-Band tunneling, Selberherr’s Impact ionization, and lucky hot electron injection [7].

3.2. Defective cell implementation

Defective cells were constructed using the same processing steps as the fault-free cell with the exception of the introduction of a defect in a particular region. The locations of the defects were limited to the various oxide layers in the structure, namely the tunnel oxide or the oxide layers in the ONO stack. For the study of impact of a defect in the tunnel oxide, five different defect locations were identified and for each such defect location a defective device was created. Thus each defective device had one injected defect in the structure, and these are shown in Fig. 4. As is evident from the figure, the defects in the tunnel oxide are located in the oxide in the area above the diffusion region of the 1T structure (i.e. drain/source overlap) or in the channel region. The ONO stack defects were limited to the bottom or top oxide layers.

One way to characterize the defects is by specifying the effective oxide thickness at the defect site. For example, the fault-free value of the tunnel oxide thickness was A106 and a defect will results in an oxide thickness smaller than this value. For example, a A100 defect results into effective oxide thickness of A100 instead of the design value of A106. The remaining characteristics of the defect size are

Please cite this article in press as: Mohammad MGH, Saluja KK, Analysis and test procedures for NOR flash memory defects, Microelectron Reliab (2008), doi:10.1016/j.microrel.2008.01.004
kept constant in the “x” (length) and “z” (along the width of transistor) directions. A total of 11 defective devices were created for simulation as shown in Table 2. In this table, ONO_T and ONO_B refers to defects in the top and bottom oxide layers of the ONO stack.

Two studies were carried out on defective cells. The first study was to determine the impact of the shape of the defects on the cell performance and to find the important characteristics/size of killer defects. Then, killer defects were injected in different regions of the insulating layers of the 1T cell and their impact on the cell operation during normal use of the memory as well as during stress were analyzed. The findings of these studies are discussed next.

4. Simulation results and discussion

The study targets three different objectives. The first objective is to identify the size of the defects that can adversely impact the reliability of the cell either during normal operation, or under stress conditions. The second objective is to study the effectiveness of various stress tests on detecting defects. The third objective is to develop a test that will allow the detection of all defects being investigated in this paper in the most efficient manner.

For each simulation study, the defect is first injected in the specified region (e.g. drain overlap, source overlap, etc.), and then an operation is performed on the cell (e.g. gate stress). After that, the \( I-V \) characteristics are measured and the threshold voltage is extracted. The process is repeated for all combinations of different operations and defect locations.

4.1. Defect characteristics and cell performance

The first set of experiments was carried out to analyze the impact of the size of defects on the performance of the flash memory cell. In these experiments, defects of different shape/size but with the identical total volume were simulated. Instead of doing the experiments for all defect locations, we chose inserting defects on the drain overlap region to carry out these experiments. The outcomes are shown in Table 3. In this table, the first column represents the operation performed and the remaining columns represent the resulting threshold voltages of three different defects (labeled by their effective tunnel oxide thickness) under various modes of operations. Further, in the table, the rows labeled gate/drain stress_E represent cells that underwent a 127 \( \mu \)s of gate/drain stress when they were assumed to be initially erased cells; whereas the rows labeled gate/drain stress_P correspond to cells that underwent stress assuming that they were initially programmed cells. Also, the erase operation in this study was performed using the channel erase operation (discussed later in Section 4.3). We will argue, from defect detection point of view, that Channel erase is superior than the common NGE approach, but it may require some DFT overhead.

It is apparent from the table that even though the total volume of each defect is identical, the oxide thickness of the defect plays a major role in determining the performance (in this case threshold voltage) of the cell. In the case of \( \text{A}^{100} \) and \( \text{A}^{80} \) defects, the cells were depleted when they were erased, depicting the well known over-erase phenomenon [16,21,27]. Furthermore, the \( \text{A}^{49} \) defect was the only defect that was identified as a detectable defect using the drain stress_E experiment, whereas all other cells did not show any significant shift in the threshold voltage before and after the stress test. In particular, the \( \text{A}^{49} \) defect shifts the threshold voltage of the cell from a programmed \( V_t \) of

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective tunnel oxide of various defects</td>
</tr>
<tr>
<td>Defect location</td>
</tr>
<tr>
<td>Drain overlap</td>
</tr>
<tr>
<td>Source overlap</td>
</tr>
<tr>
<td>Channel</td>
</tr>
<tr>
<td>ONO_B</td>
</tr>
<tr>
<td>ONO_T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drain defect size characterization</td>
</tr>
<tr>
<td>Operation</td>
</tr>
<tr>
<td>Defect</td>
</tr>
<tr>
<td>Program</td>
</tr>
<tr>
<td>Erase</td>
</tr>
<tr>
<td>Gate stress_E</td>
</tr>
<tr>
<td>Drain stress_E</td>
</tr>
<tr>
<td>Gate stress_P</td>
</tr>
<tr>
<td>Drain stress_P</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Mohammad MGh, Saluja KK, Analysis and test procedures for NOR flash memory defects, Microelectron Reliab (2008), doi:10.1016/j.microrel.2008.01.004
6.4 to 3.7, i.e. in excess of a 2 V shift, whereas all other defects cause only a marginal and insignificant shift in the threshold voltage. Even though the 2 V threshold shift in a A49 does not result in value flip in the memory cell (hence not detected using normal read operation), other techniques, such as margin reads [28], can detect such defects. The main conclusion of this set of experiments is that a defect which reduces the effective tunnel oxide thickness to A80 or less represents a killer defect and proper test must be conducted for their detection.

4.2. Impact of tunnel oxide defects on cell performance

The second set of experiments was carried out on defective cells with effective tunnel oxide thickness at the defect site to be A80, and the defect was located in one of the various tunnel oxide regions as specified in Section 3. For all these cases we used the negative-gate erase algorithm (discussed in Section 3).

The simulation studies were carried out on tunnel oxide defects in source/drain overlaps and in the channel regions. The cells were simulated to see the impact of each defect on the cell program/erase characteristics as well as their behavior under stress conditions. Table 4 summarizes the findings.

It is evident from this table that only the source overlap defect can potentially be identified in these studies as explained below. For this defect, the erase operation resulted in a depleted cell (cell with negative threshold voltage), causing in faulty cell behavior, where as for all other defects the threshold shift is only marginal. Thus in the case of source overlap defect, when reading a cell in the same column as the depleted cell, the read data will be corrupted if the addressed cell has a logic “0” value. The stress time for all defective cells was increased to five time (5 × 127 μs) the worst case stress duration in order to see if an undetected defect would become detectable. However, the defects did not show different behavior under such stress conditions (only a minor shift in $V_t$ for drain stress experiment). Even though the drain overlap and channel defects do not seem to impact the performance of the memory cell, they will pose reliability concerns. Further, these undetected defects most likely will be excited by cycling and will reduce the useful lifetime of the memory module and will results in an in-field failure, one of the major concerns about the success of flash memories.

The above study suggests that the stress experiments may not be the most efficient way of detecting tunnel oxide defects. In particular, we notice that the stress tests do not result in any noticeable shift in the $I-V$ characteristics for any of the defects. It is apparent that starting with a programmed/erased cell, the threshold voltage of the cell does not change. However, it is clear that source overlap defect does impact cell performance (cell becomes depleted after erase). This suggests that the erase operation could be the key to efficient detection of tunnel oxide defects. In the next section we describe a method that can be utilized to detect these defects.

4.3. Channel erase and tunnel oxide defect detection

In the previous section, it was shown that the drain-overlap and channel defects could not be detected by any of the stress conditions as well as by erase and program conditions, thus they may remain undetected by any of the tests. The source overlap defect on the other hand could be detected because it resulted in a depleted cell when the cell underwent an erase operation. Investigating further for the reason for this behavior, we found that during erase operation, the overlap area undergoes high electric field stress due to the biases applied to the gate/source terminals of the memory cell. Therefore, we felt that in order to excite and detect defects in any of the tunnel oxide regions, appropriate electric field stress must be present in every region that needs to be investigated and tested. One possible approach that offers the opportunity to stress all regions of the cell is the channel erase operation discussed below.

Unlike the commonly used NGE erase operation, which utilizes negative-gate–positive-source bias condition, channel erase concept results in a uniform electric field stress in all regions of the tunnel oxide. The source side erase technique restricts the high electric field region to only the source overlap area. The channel erase approach is accomplished by biases applied either to the control gate only, or by using gate and substrate biases (i.e. in triple well technology). Two typical bias voltages for the two channel erase approaches are shown in Table 5. Further, for the cell structure created in our study, a channel erase approach would require 70 ms erase time, which is substantially longer time compared to the NGE erase approach which requires only 10 ms.

In order to analyze the effectiveness of channel erase technique in identifying tunnel oxide defects, we ran the

---

**Table 4**

<table>
<thead>
<tr>
<th>Operation</th>
<th>Threshold voltage (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fault-free</td>
</tr>
<tr>
<td>Program</td>
<td>7.95</td>
</tr>
<tr>
<td>Erase</td>
<td>1.11</td>
</tr>
<tr>
<td>Gate stress $E_{gc}$</td>
<td>1.11</td>
</tr>
<tr>
<td>Drain stress $E_{gc}$</td>
<td>1.11</td>
</tr>
<tr>
<td>Gate stress $E_{ds}$</td>
<td>7.95</td>
</tr>
<tr>
<td>Drain stress $E_{ds}$</td>
<td>7.95</td>
</tr>
</tbody>
</table>

**Table 5**

<table>
<thead>
<tr>
<th>Method</th>
<th>CG</th>
<th>D</th>
<th>S</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method$_1$</td>
<td>-14</td>
<td>Float</td>
<td>Float</td>
<td>6</td>
</tr>
<tr>
<td>Method$_2$</td>
<td>-20</td>
<td>Float</td>
<td>Float</td>
<td>GND</td>
</tr>
</tbody>
</table>

---
same simulation studies that were performed previously, but this time using the channel erase approach. We first used the voltages specified in Table 5 in the row labeled Method 1. In each case, as before, the required operation was performed and the I–V characteristics of the device were measured and the threshold voltage was extracted. The results of our study are shown in Table 6. It is evident that every defect in this case results in a depleted cell. We also performed a similar study by utilizing the voltages listed in Table 5 for method 2 channel erase. We found the results were similar to those shown in Table 6. These observations suggest that the channel erase technique is far more effective and superior in detecting all tunnel oxide defects.

4.4. ONO defects and impact of cell performance

Next we expanded our investigation to study defects in the ONO layer. We used the same approach as before and faulty cells with ONO defects were constructed and simulated. Two types of defects were simulated. First, a defect is created in the bottom oxide layer of the ONO layer (see Fig. 4d). The fault-free value of this layer was approximately A96 and in the presence of a defect it has an effective thickness of A67. Second, a defect in the top oxide layer of the ONO layer (see Fig. 4d) was also created and simulated. The fault-free thickness in this case was A100 and that of the defective cell was A71. The results of the study of these defects are compared to the fault-free case as shown in Table 7. It is apparent that defects in the ONO oxide layers do not impact the performance of the flash memory cell and hence can be ignored. This finding supports what was previously argued and suggested in [2,17] using logical reasoning only. Yet another reason for these defects not to result into faulty behavior is the fact that the voltage drop in the ONO layer is small compared to that in the tunnel oxide. The high coupling factor between the control and the floating gates makes the electrical stress on the ONO layer small compared to that across the tunnel oxide resulting in higher reliability of the ONO layer. Thus, in either case, the conclusion is that the ONO layers are more reliable than oxide layer and less prone to disturb behaviors and hence faults in ONO layer can be ignore. Furthermore, the gate coupling ratio in our simulation was approximately 0.5 which is considered small compared to that used in current flash memory. However, even with this conservative coupling ratio, no ONO defect showed impact on cell performance. We believe that with high gate coupling ratio, the impact of such defects will be even less and hence the conclusion of ignoring ONO defects (in this context) is still valid.

5. Test algorithms and cell reliability

Based on the results of our study described in Section 4, we summarize the important findings about various defects as follows. These findings will be used to develop efficient tests for various defects in the IT cell based flash memories.

**Defect excitation:** It was apparent that stress tests are not very effective in defect excitation. It was shown that the erase operation is a more effective way to excite tunnel oxide defects.

**Fault detection:** Channel erase technique was shown to be superior in detecting all defects when compared to NGE erase method which was only capable of detecting one defect type out of the three simulated tunnel oxide defect types.

**ONO defects:** After simulating ONO defects, it was found that no single defect in the ONO layer will result in faulty behavior. Hence, all test algorithms that have special test patterns for ONO defects can be simplified by removing those patterns.

**Depleted cell behavior:** All defects in the tunnel oxide layer were shown to have depleted threshold voltages when using channel erase technique. Therefore, a test for depleted cell, rather than erased/programmed cell, as previously suggested in [2–4], is likely to a more efficient method for detecting such defects.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Threshold voltage (V)</th>
<th>Fault-free</th>
<th>Source</th>
<th>Drain</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
<tr>
<td>Erase</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Gate stressE</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Drain stressE</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Gate stressP</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
<tr>
<td>Drain stressP</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
</tbody>
</table>

Table 7
ONO defects simulation

Operation | Threshold voltage (V) | Fault-free | Bottom oxide | Top oxide
---|-----------------------|------------|--------------|------------
Program    | 7.95                  | 8.54       | 8.54         |
Erase      | 1.10                  | 1.17       | 1.14         |
Gate stressE| 1.10                | 1.17       | 1.14         |
Drain stressE| 1.10               | 1.17       | 1.14         |
Gate stressP| 7.95                 | 8.54       | 8.54         |
Drain stressP| 7.95                | 8.54       | 8.54         |

Table 6
Channel erase experiments

<table>
<thead>
<tr>
<th>Operation</th>
<th>Threshold voltage (V)</th>
<th>Fault-free</th>
<th>Source</th>
<th>Drain</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
<tr>
<td>Erase</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Gate stressE</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Drain stressE</td>
<td>1.10</td>
<td>–1.47</td>
<td>–1.97</td>
<td>–2.37</td>
<td></td>
</tr>
<tr>
<td>Gate stressP</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
<tr>
<td>Drain stressP</td>
<td>7.95</td>
<td>7.90</td>
<td>7.12</td>
<td>7.97</td>
<td></td>
</tr>
</tbody>
</table>

Please cite this article in press as: Mohammad MGh, Saluja KK, Analysis and test procedures for NOR flash memory defects, Microelectron Reliab (2008), doi:10.1016/j.microrel.2008.01.004
assumed to be fault-free as it was previously argued in [2]. We further conclude that to develop a test to detect tunnel oxide defects, the following conditions must be met:

**Programmed initial state:** All cells to be tested must be programmed (i.e. set to logic “0” state).

**Channel erase fault excitation:** Programmed cells to be tested must be erased using channel erase technique.

Fig. 5 gives a new test procedure called Flash-CE test, which can be used to detect all defects in the tunnel oxide layer of 1T cells organized in a NOR array. Since most disturb faults are assumed to be caused by defects in the tunnel oxide, we can claim that this algorithm can detect all disturb faults. In Fig. 5, $n$ and $m$ represent the number of rows and columns in the memory array, respectively.

The working of the algorithm is as follows. Step 1 initializes the array to a programmed state in order to prepare it for the channel erase operation. Step 2 erases all cells in the array using channel erase approach, hence exciting all tunnel oxide defects. The third step programs each cell in the first row ($i = 0$) of the array and reads each element of that row, expecting a value of “0”. In case there is any depleted cell in any column (defective cell), the read operation will fail and the value will be read as “1”. This is so because of the excessive depletion of the defective cell and as a result the column containing the defective cell will read a logic 1 value. After this step, the only cells that remain to be tested are those in the first row. Steps four, five and six initialize, excite, and detect these remaining faults in a similar manner.

In recent years “March tests” have gained popularity and have been used in many test algorithms for testing flash memories [3,4,17]. This is due to their simplicity, their fault detection capability, and their ease of implementation in built-in self-test (BIST) environment. We have developed an efficient (minimum length) March algorithm, called March-CE, which can detect all tunnel oxide defects and it is shown in Table 8. The algorithm March-CE is written using conventional march algorithm syntax [8]. Symbols “||” and “*” represent the addressing sequences that proceed in increasing or descending address order respectively. The $R_x$ and $W_x$ represent read or write operation for the value “x” on the addressed cell. All operations that are between brackets are done on the same cell before proceeding to the next cell. For example, the march element $(R1,W0,R0)$ performs a read operation while expecting a value of “1” during read operation, followed by write “0”, and then followed by a read operation while expecting a value of “0” from the addressed cell (say cell $i$) before going to the next cell (cell $i + 1$). In this algorithm, the term $E_{Ch}$ represents a “w1” (Erase) on the whole array since a selective “w1” in flash memories is not permissible. Further, the subscript “Ch” in the erase operation signifies the fact that the erase operation uses channel erase instead of the conventional source side or negative-gate erase operation. Further, the symbols SAF, SOF, TF, AF, CF$_{st}$, CF$_{in}$, and CF$_{id}$ refer to stuck-at, stuck-open, transition, address, state-coupling, state-inversion, and idempotent coupling faults respectively [8].

March-CE algorithm is not efficient in detecting other types of faults, such as SAF and SOF. Alternate versions of march test, based on March-CE, are shown in Table 8 with their detection capabilities as calculated by RAMSES fault simulator [29]. Note that the detection capabilities are computed by the simulator assuming a 1-bit wide rather than a word-oriented memory.

Table 8

<table>
<thead>
<tr>
<th>March algorithms details</th>
<th>Detection capabilities</th>
<th>SAF (%)</th>
<th>SOF (%)</th>
<th>TF (%)</th>
<th>AF (%)</th>
<th>CF$_{st}$ (%)</th>
<th>CF$_{in}$ (%)</th>
<th>CF$_{id}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>March-CE</td>
<td>$(</td>
<td></td>
<td>w0; E_{Ch};</td>
<td></td>
<td>(w0,0,r0); E_{Ch};</td>
<td></td>
<td>(w0,r0))$</td>
<td>50</td>
</tr>
<tr>
<td>March-CER</td>
<td>$(</td>
<td></td>
<td>w0; E_{Ch};</td>
<td></td>
<td>(r1,0,w0); E_{Ch};</td>
<td></td>
<td>(w0,r0))$</td>
<td>100</td>
</tr>
<tr>
<td>March-CERR</td>
<td>$(</td>
<td></td>
<td>w0; E_{Ch};</td>
<td></td>
<td>(r1,0,w0); E_{Ch};</td>
<td></td>
<td>(r1,w0,r0))$</td>
<td>100</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Mohammad MGh, Saluja KK, Analysis and test procedures for NOR flash memory defects, Microelectronics Reliab (2008), doi:10.1016/j.microrel.2008.01.004
5.2. Reliability aspect of flash memory

The test algorithms developed and given in this paper are used to test virgin memory units in an efficient manner. The issue of flash memory reliability (cell performance after cycling), on the other hand, requires different methods to insure the proper operation of a flash memory array. For example electron/hole trapping/detrapping that occurs after cycling can result in erratic behavior of flash memory cells which could lead to over-erased cells or cells with poor program/erase characteristics. These issues are normally circumvented by the use of post erase repair (PER) method or program verify algorithms which guarantee proper threshold voltages for program and erased cells in the memory array [30,31]. These methods and algorithms are used to insure proper operation of the cells while being used in the field.

Our approach, on the other hand, assumes that such algorithms are disabled when performing tests on virgin units, which commonly is the case. Therefore, our goal is to identify defects in the oxide layer in an efficient manner without considering the impact of cycling on the cell behavior. Further, we assume that the program and erase operations are designed to reduce the probability of shifts in erase operation using methods discussed in [23–25,32]. Hence, by incorporating both methods, namely our method of testing and other methods such as PER for repair, which are complementary to each other, a reliable production of flash memories is possible. The test approach can be used as a first tier screen for cells with gross defects whereas the other algorithmic approach can be used to improve the reliability of the units while in field.

6. Design-for-testability technique

Unlike conventional memories which use only one voltage (supply voltage) throughout the memory module, flash and other non-volatile memories use many different voltages for program, read, and erase operations. Generally the memory modules have high voltage switches in various units in the memory module to route program, erase, or read voltages as these values are often higher (and different) than the supply voltage. The various algorithms developed in the previous section require that the erase operation be performed using the channel erase technique. In order to support this mode of operation, modification to the design of various switches that connect the memory cell to the word, bit, and source line voltages is needed. This modification will help supply the necessary voltages during various modes of operations. For example, to be able to program, read, and perform channel erase operation on every cell, we must be able to apply 10 V, 3.3 V, and −20 V to the control gate of each cell (gate of the cell is connected to the word line), hence these signals are supplied to the word line through various switches. Various voltages that need to applied to the terminals of the 1T cells of a memory consisting of NOR array organization are shown in Fig. 6. This figure shows the biases applied to different terminals of memory cells during program, read, and erase operations (values of signals given from top are for program, read, and erase voltages respectively). The figure also contains biases for memory cells provided the method used for erase operation is negative-gate, source, or channel erase. Notice that the substrate of the memory cell is assumed to be grounded irrespective of the operations, and therefore substrate bias is not shown.

The following subsections discuss the modification(s) to the memory organization to support channel erase method provided the original memory is designed to utilize negative-gate erase (NGE) method. Changes similar to the changes discussed in the next subsections are possible for designs that utilize source or some other erase technique.

6.1. DFT approach for memory

In order to support the channel erase approach, as proposed above, during test of flash memories, the design of the supporting circuitry for the memory cells needs to be altered/enhanced from its original design. Biases that are different than those used under normal mode of operations need to be available during test. For example, for the NGE approach discussed previously, −8 V, 7 V, and float are applied to the gate, source, and drain, respectively, of all cells in the sector to be erased. However, during channel erase (i.e. test), the required signals for gate, source, and drain terminals are −20 V, float, and float, respectively. The high voltage needed during channel erase can be generated by an on-chip charge pump, or supplied externally by the tester and applied through the $V_{pp}$ pin that is normally available in most flash memory designs. The former approach is more expensive in terms of chip area than the latter one since it requires an additional charge pump or a larger charge pump than that needed for normal operation.
In any case, additional control logic, control signals, and switches may be needed and these are discussed below.

6.1.1. Modification in memory row decoder

Fig. 7 shows the organization of the row decoder and the word-line switch of a typical NOR memory array. This figure also shows the proposed DFT modification. The Word-Line Switch shown in the figure, in most instances is integrated within the decoder and is often not shown as a separate entity. The signals “E”, “P”, and “R” represent the control signals that are asserted during erase, program, and read operations, respectively. On the other hand, the signals “$V_{EW}$”, “$V_{PW}$”, and “$V_{RW}$” are the erase, program, and read voltages that are routed through the word line (“W”) to the addressed cells.

For the technology used in such a memory organization, normal mode erase operation requires the negative voltage on the word line to be supplied by the charge pump ($V_{EW}$ in Fig. 7). If this memory is to use channel erase method during test mode, the required voltage can be supplied externally from $V_{pp}$. But this will require changes to the switch associated with original decoder design as shown in the lower box in Fig. 7. The modifications consist of adding a selector switch that selects between the erase voltages (charge pump or $V_{pp}$ pin) and a “Test” control signal which is active only during test mode to select the appropriate erase voltage.

6.1.2. Modification in bit-line/sector switch

As for the bit line (or data line), the column decoder and bit-line switches are organized similar to the decoder and switch in the row decoder. However, the internal design of the switch is somewhat different. Fig. 8 shows the design of such a switch which connects the selected bit line to the appropriate program/read-bit-line voltages during program/read operations. The switch also has the ability to disconnect these signals (open) during erase operation. However, it can be seen that neither the bit-line switch nor the bit-line decoder require any modification to support channel erase operation.

Fig. 9 shows the internal design of the source/sector switch. This switch is responsible for grounding or applying $V_{Erase}$ (i.e. 7 V) to all cells in the selected sector during read/program or erase operation respectively. In order to allow channel erase during test mode, the switch must be modified as shown in Fig. 10. This switch has the capability to apply $V_{Erase}$ to the source line of all cells that are in the sector to be erased during normal erase operation whereas under test mode, as proposed in this paper, the source terminal can be float.

6.2. DFT cost analysis

In order to justify the proposed DFT approach, the implementation of the DFT must be economical and it
must not result in major design changes. The cost must include the number of additional transistors as well as the pins needed to implement the DFT approach. It is evident from the discussions in the previous sections that to implement our test methodology only minor changes to the row decoder and the source-line switch are needed. Further, two pins, “Test” control signal and \( V_{pp} \), are also needed to run the test. In most of today’s single supply flash memories, \( V_{pp} \) already exists, therefore the only extra pin required to implement the DFT as suggested in this paper is the “Test” control pin. Thus, the effective cost of the DFT approach in terms of pin count would be one pin.

The number of extra transistors needed to implement the proposed DFT approach can be determined by computing the number of additional transistors needed for each word line and/or sector. These modifications are discussed in the previous section and the cost can be computed as follows. Each word line would require approximately 16 additional transistors (1 high voltage “HV” switch, 1 AND, 1 NOT) whereas each sector will require an additional 14 transistors (1 HV switch, 1 AND). In the above calculation we assume that a “HV” switch requires as many as 10 transistors. Thus, for an \( N \times N \) sector size, the cost of the DFT approach can be calculated as (in terms of transistor count) follows:

\[
T_{\text{Cost}} = \frac{\text{DFT}_{\text{row}} + \text{DFT}_{\text{sector}}}{N \times 16 + 14} = \log_2 N(2N + 2) + 42N + \frac{\log_2(N^2 + 2)}{4} + 24N + 10 + N^2
\]

In the above equation, \( \text{DFT}_{\text{row}}, \text{DFT}_{\text{sector}} \) represent DFT transistors per row and sector switches respectively. The terms “Dec”, \( \text{Tot}_{\text{row}}, \text{Tot}_{\text{col}}, \text{Tot}_{\text{sector}}, \) and \( \text{No}_{\text{cell}} \) represent the total transistors for the row decoder, row, column, sector switches, and total number of memory cells in the sector, respectively. Note that this calculation assumes a column decoder that is shared between four sectors but with dedicated bit-line switches for each sector (i.e. term \( \log_2(N^2 + 2) + 24N \)) and does not include the cost of sense amplifiers. Therefore, for an \( 128 \times 128 \) sector size, the total cost for DFT is 7.6% of the totals cost of a sector. However, in modern flash memories the sector size is often larger and therefore for the more prevalent sector size of \( 512 \times 512 \), the cost overhead is even smaller (i.e. 2.7%) which makes it a very attractive and cost effective approach.

7. Conclusion

In this paper we first studied different defects that are responsible for disturb faults in 1T flash cell using a 2D device simulator. It was found that stress tests are not efficient when it comes to detecting tunnel oxide defects. Oxide-nitride-oxide layer defects were found to be benign and did not result into faulty behavior and hence they can be ignored. Efficient tests based on channel erase techniques were developed to detect tunnel oxide defects (hence disturb faults) as well as other type of faults such as SAF and AF faults. A DFT approach to implement the proposed test methodology was proposed and its cost was analyzed.

Acknowledgements

The authors would like to thank Faisal Al-Hasawi of the Public Authority of Applied Education and Training (PAAET, Kuwait) for his help in preparing this manuscript. This work was supported by Kuwait University Research Grant No. [EO 01/04].

References

[19] Satoh S, Hemink G, Hatakeyama K, Aritome S. Stress induced leakage current of tunnel oxide derived from flash memory read-


