Microelectronics Reliability 88–90 (2018) 61–66

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

Influence of temperature of storage, write and read operations on multiple level cells NAND flash memories

Julien Coutet[a,b,*], François Marc[b], Flavien Dozolme[a], Romain Guétard[a], Aurélien Janvresse[a], Pierre Lebossé[b], Antonin Pastre[b], Jean-Claude Clement[c]

[a] Thales Communications & Security SAS, Labège, France
[b] University of Bordeaux, Laboratory IMS, Talence, France
[c] Thales Research & Technology, Palaiseau, France

ARTICLE INFO

Keywords:
NAND flash memory
Reliability
Accelerated ageing
Data retention
MLC

ABSTRACT

This paper presents an analysis of the reliability of 20 nm technology NAND Flash memory components based on Multiple Level Cells (MLC). The focus of the study is to assess the influence of temperature during programming, storage and reading operations. In order to reach this goal, several memories were programmed once at many temperatures ranging from −40 °C to 85 °C, then they have been stored powered off in one case and have been activated in reading in the other case, under different thermal stresses.

1. Introduction

With the increasing amount of data-driven applications, the non-volatile Flash memories are growing in popularity in applications such as memory sticks, embedded memories, Solid-State Drives… Flash memories are now succeeding to EPROM (Erasable Programmable Read-Only Memory). EPROM and Flash-PROM can both be cleared easily by users.

Following the attempt of using COTS (commercial off-the-shelf) in military and space projects, it is important to characterize the reliability of DSM (Deep Sub-Micron) technology. The subject of the study is a 64Gb Multiple-Level Cells (MLC) NAND Flash memory that belongs to the 20 nm technology family.

In this paper, the main topic is the study of the influence of temperature during storage, writing and reading operations on MLC NAND flash memories. As of today, the two test flows described below have not reached their end but preliminary data extracts are already of interest.

2. Technology

The evolution from EPROM to EEPROM (Electrically-Erasable Programmable Read-Only Memory) was marked by the implementation of an electrical erasing process to replace the existing and constraining UV process. Both technologies are based on floating gate MOS transistors where the floating gate is located between the gate and the channel. Floating gate MOS transistors have a threshold voltage that is driven by the charge trapped in the floating gate [1]: the injection of negative charges in the floating gate increases the transistor's threshold voltage. Then, to read the cell, drain current is measured under a constant gate voltage which gives access to the threshold voltage of the cell and thus to the information stored.

In the usual technologies, a single cell stores a single bit: its threshold voltage is either above or below a reference value, in which cases the cell is respectively either said to be “programmed” (bit ‘0’) or “cleared” (bit ‘1’). These cells are called Single Level Cells (SLC).

However, a MLC can contain several bits stored as different levels of charge. As can be observed on Fig. 1, the increase of bits per cell leads to a reduction of the noise margin between two neighbor states. That is why the study of the reliability of this kind of memory is of paramount importance.

These MLC NAND flash memories are complex to read. In fact, to increase the reliability of their products, the founder has introduced the Read-Retry method [2]. This method follows the charge loss from floating gate during retention by changing the reference voltage used during reading (see Fig. 2). Without this, retention time would be shortened too much because of little noise margin between adjacent states in cell memory.

It is known from exchanges with the memory manufacturer that there are 8 Read-Retry methods that can be implemented during read operations. The first 4 are based on reference voltage shifts and the other 4 add adaptive measurement periods.

* Corresponding author at: Thales Communications & Security SAS, Labège, France.
E-mail address: julien.coutet@thalesgroup.com (J. Coutet).

https://doi.org/10.1016/j.microrel.2018.06.088
Received 4 June 2018; Received in revised form 20 June 2018; Accepted 27 June 2018

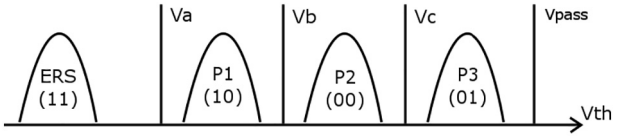

Fig. 1. Distribution of MLC states by Vth.

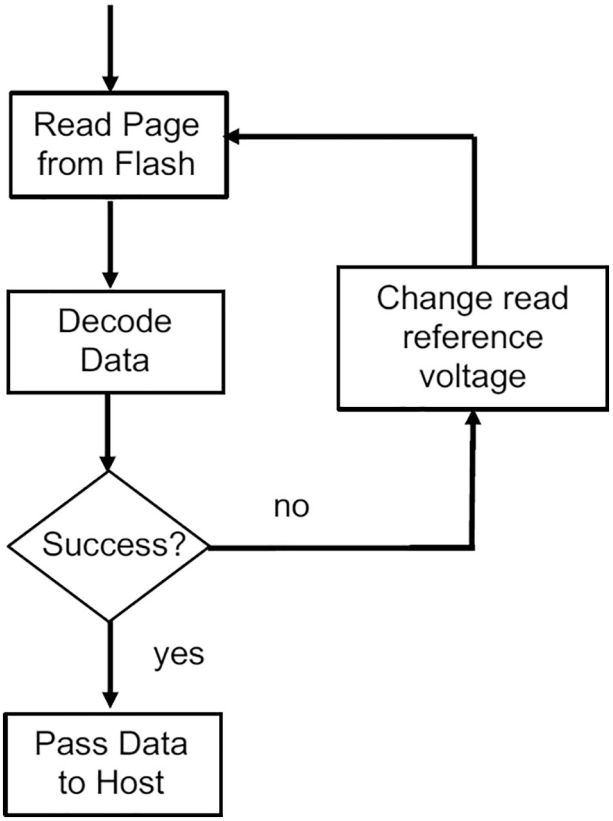

Fig. 2. Read Retry algorithm.

Practically, the user must store both the data and the associated ECC (Error Corrector Code). For each reading operation, data is read and compared to its ECC. If there is no match, the Read-Retry method to be used gets incremented through the use of an internal register and the operation starts over until a match occurs or all 8 methods have failed.

In this paper, a page (a cluster of 8 k-bytes) is then defined as "fail" if none of the 8 Read-Retry methods allows reading the content with less than 40 wrong bits (compared to initial content) per 1117 bytes. For reference, this limit is directly linked with the minimum required ECC recommended by the manufacturer to ensure that data is stored properly over the life of the NAND Flash device.

3. Static ageing

In some military applications, data must still be readable years after getting written. That fact is the basis for the retention tests detailed hereafter.

3.1. Initial writing

Write operation was performed at four temperatures. All the cells of a device were written at the same temperature. That is why several batches were created following the Table 1.

To understand the influence of data content and to be able to make a correlation with the physical aspects of the failure cases, each memory has been divided in twelve memory areas. Each of these areas contains a different data pattern (see Table 2). In static ageing each ZM (Memory Area) contains 50 blocks. Each block contains 256 pages. Indeed, cells with more charge are more prone to leakage during data retention, which are accelerated by thermal stresses. This leakage usually leads to information loss.

Table 1
Test table for the 51 memories under evaluation.

Temperature during writing	Temperature of ageing							
	25 °C		85 °C		110 °C		125 °C	
−40 °C	3	*1*	3	*4*	0	*4*	3	*0*
−10 °C	3	*0*	3	*0*	0	*0*	3	*0*
25 °C	3	*1*	3	*4*	0	*4*	3	*0*
85 °C	3	*2*	3	*4*	0	*4*	3	*0*

Quantity of devices in Static ageing.
Quantity of devices in Dynamic ageing.

Table 2
Content of data patterns written in each memory areas.

Name	Pattern
ZM1	Every cell at '11'
ZM2	Every cell at '01'
ZM3	Every cell at '10'
ZM4	Every cell at '00'
ZM5	25% of one block at '00', others remains at '11'
ZM6	50% of one block at '00', others remains at '11'
ZM7	75% of one block at '00', others remains at '11'
ZM8	Checked pattern 0xAA55
ZM9	Checked pattern 0x55AA
ZM10	Checked pattern 0x00FF
ZM11	Some '00's, '01's, '10's surrounded by '11's
ZM12	Random sequence with a length of 1 page (8 k-bytes)
ZM12'	Random sequence with a length of 8 bytes

3.2. Static ageing and readouts

Devices are then aged in ovens at three different temperatures, powered off (see Table 1). At pre-defined steps, DUTs (Device under test) readouts are performed: memory content is read and compared to initial state. These readouts are done under ambient temperature.

During readouts, the quantity of wrong bits of each page is recorded only for the most successful Read-Retry method defined by being the first with a quantity of wrong bits below the maximum quantity of bits recoverable by ECC.

3.3. Preliminary results of static accelerated ageing

After 6000 h of accelerated ageing, some drifts can be observed. As of today, failures occur only for groups written at −40 °C and stored at 125 °C.

As failures are not expected before a long time for some test groups, the amount of Read-Retry operations necessary to obtain a successful reading of the initial content has been added to the list of monitored parameters.

Based on this, results clearly show a stronger degradation for parts written at low temperatures in comparison with the ones written at higher temperatures. It can also be noted that the higher temperature during the storage operation, the more Read-Retry operations are needed (see Fig. 3).

However, since voltage step values between two consecutive Read Retry modes and the changes from Read Retry methods 4 and 5 are unknown, no precise calculations could be undertaken.

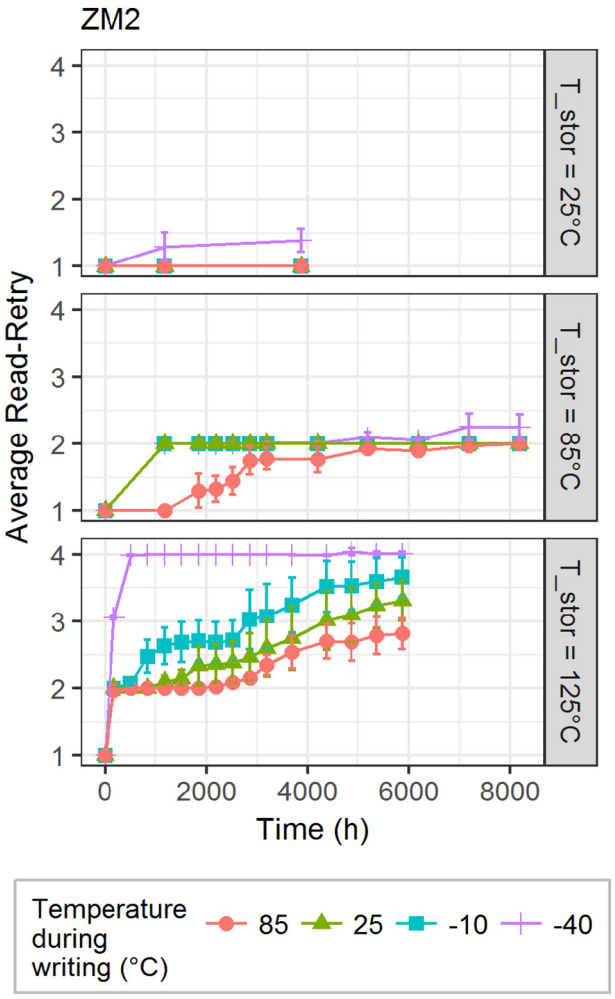

Fig. 3. Evolution of Read Retry quantities over time for three storage temperatures, ZM2, static ageing.

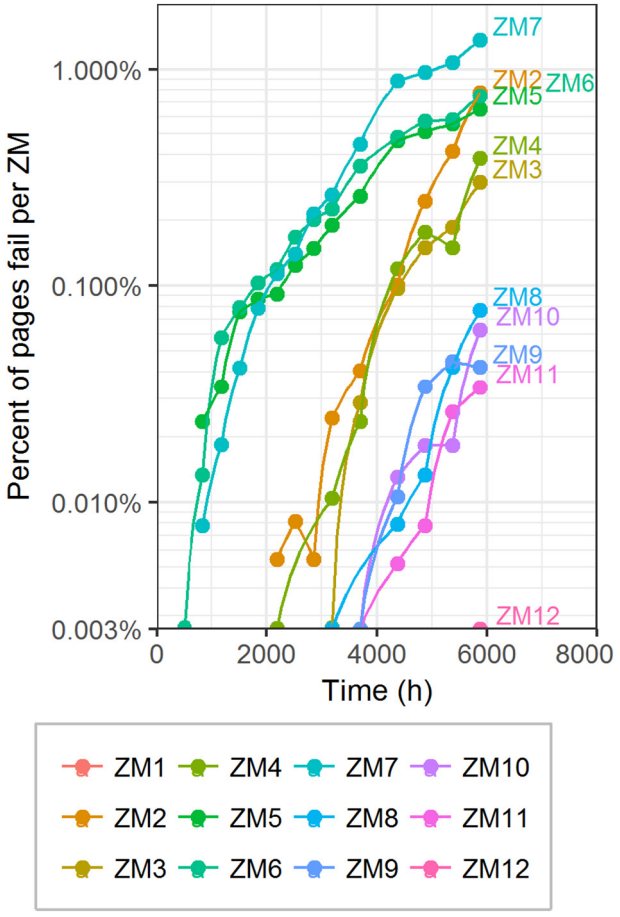

Fig. 4. Evolution of the percent of failed pages for each ZM, DUTs written at −40 °C and stored at 125 °C, static ageing.

3.4. Influence of patterns written

The number of failed pages in the most critical configuration (written at −40 °C and stored at 125 °C) is different for each ZM. This clearly indicates the data itself has an influence on retention time. According to Fig. 4, there are five groups of ZM sorted by percentage of failed pages.

The most critical subset is made up of ZM5, ZM6 and ZM7. These ZMs correspond to 25%, 50% or 75% of consecutive '00's in their first pages of each block and '11's for the rest.

The second most critical is composed of ZM2, ZM3 and ZM4 which correspond to every cell at either '01', '10' or '00'. Based on failure rates, this group can be sorted from high to low as follows: ZM2 > ZM4 > ZM3 (> ZM1).

The third group consists of ZM8, ZM9, ZM10 and ZM11. It has been naturally expected that the amount of pages failed for ZM8 or ZM9 would be between those of ZM2 and ZM3.

The area containing random data (ZM12) shows fewer errors compared to other ZMs.

As expected, ZM1 which only contains '11's still shows no errors.

Regarding ZM5 to ZM7, the study of their distributions of failed pages shows some interesting results. These ZMs contain only '11's or '00's, and it appears the ageing process only reveals significant amounts of fails for pages written with '00's (see Fig. 5).

What's more interesting is the fact that for each of these memory areas (ZM5, 6 or 7), the eighth page before the switch from '11's to '00's (respectively pages 56, 120 and 184) displays the highest failure rate.

However, ZM4, which only contains '00's, does not show any failure rate peak on page 248. The global amount of failed pages is also lower than for ZM5 through 7.

For a NAND Flash memory, and according to datasheet, pages must be programmed sequentially within a block in increasing page address order, corresponding to order of cells along the columns of transistors. A descrambling explanation of that phenomenon is so unlikely. At this time, there is no clear explanation of this phenomenon.

Furthermore, the initial implementation of several patterns of data in memories significantly helps in forming many hypotheses concerning the distribution of states (P1 to P3) and boundary threshold levels.

Firstly, in static ageing, it is possible to sort patterns ZM1 to ZM4 by their mean number of failed pages. Analysis of static results leads to an assumption of the organization of states related to their voltage threshold as follows: ZM1 < ZM3 < ZM4 < ZM2. In others words, ERS state of memory cell corresponds to '11', the P1 state is most likely to correspond to '10', P2 to '11' and P3 to '01' (see Fig. 1 for a more visual representation). In this logical organization, only one bit changes between two neighbor states, which makes ECC correction more efficient than for incremental increases. If designers had chosen a natural increasing logical organization ('11' < '10' < '01' < '00'), a voltage threshold shift from P2 '01' to P1 '10' would correspond to two bits changing simultaneously instead of one.

4. Dynamic ageing

In this case, memories are aged in a powered on state and read in-situ periodically.

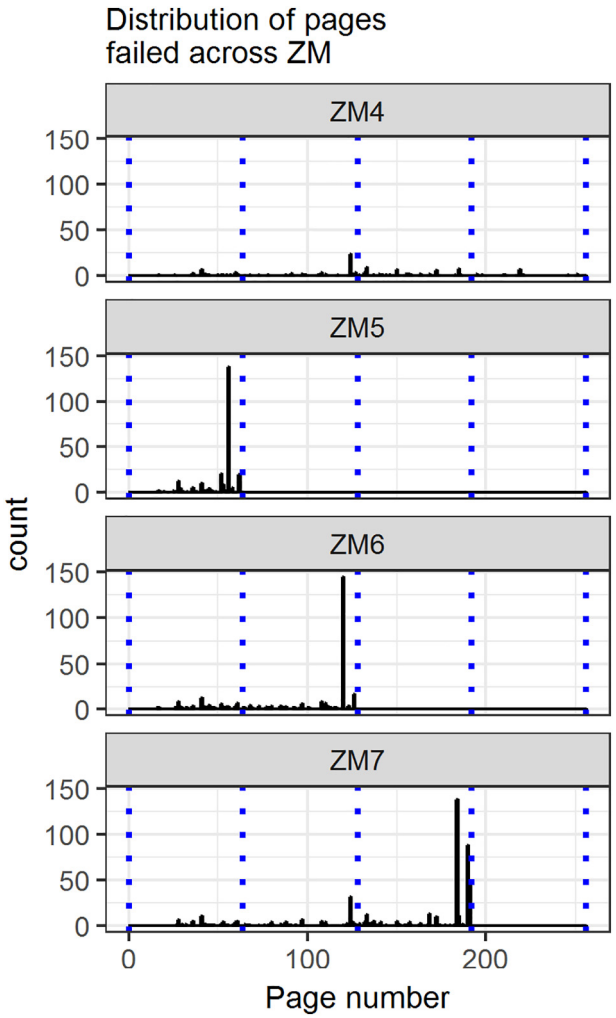

Fig. 5. Distribution of failed pages across 4 ZMs after 5873 h, pieces written at −40 °C and stored at 125 °C, static ageing.

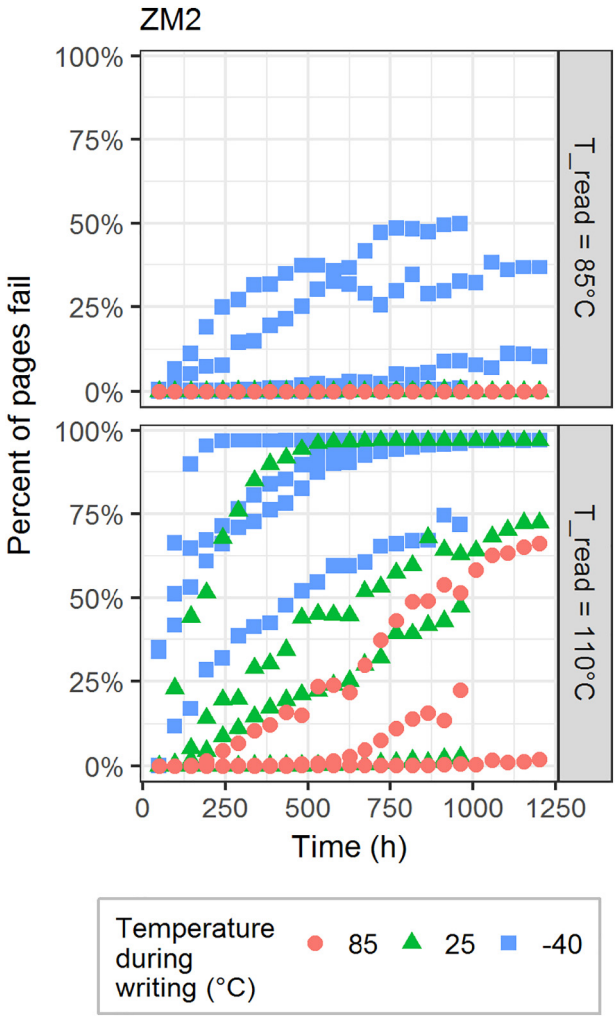

Fig. 6. Evolution of the percent of failed pages over time and by DUT, for two out of three temperatures of reading and activation, ZM2, read in continuous, dynamic ageing.

4.1. Initial writing

Used methodology is slightly simpler than the one for "static storage" since DUTs were programmed at only three different temperatures and the number of patterns was brought down to only 5: ZM1 through ZM4 plus ZM12. In the dynamic ageing, each ZM contains 300 blocks read continuously (memory cells read every 45 min) and 300 other blocks read once a day.

4.2. Ageing and test

Dynamic ageing is performed on a FPGA based test bench. All along the ageing process, a master FPGA periodically reads the contents of the DUTs and compares them with initial values. The amount of fails is sent to a computer and recorded whenever a new test is done.

This test bench, designed by Thales, allows continuous tracking of the degradation process. Moreover, each DUT is heated independently thanks to local heaters and a PID controller. A preliminary series of measures of the DUTs' protective diode under several temperatures of the local heater has been conducted in order to estimate the gradient of temperature across the plastic package, which is used as a correction factor in the PID controller.

Two reading periods were implemented. To allow that, the ranges of addresses swept by the test bench are 50% or 100% of cells written. The read operation of all the memory is performed once a day.

The monitored parameter is the amount of failed pages (the number of bit fail is above the maximum number bit recoverable by ECC) for each DUT and each ZM.

4.3. Preliminary results of dynamic accelerated ageing

With 1000 h of accelerated life test data currently available, limited retention time of DUT written under low temperature, can be observed, as highlighted by static accelerated ageing. Also, writing temperature has a noticeable impact on data retention time: the lower the temperature the shorter the retention time (see Fig. 6).

During this dynamic test, there is no access to the amount of Read-retry operations used in order to read the right information.

4.4. Influence of patterns written

As with static ageing, the pattern of data written in memories has a noticeable influence on retention time. In dynamic ageing, only 5 patterns were written (ZM1, ZM2, ZM3, ZM4 and random ZM12'). However the ZM12' with random content is different to the static ZM12. In fact, in static ageing, every bit of a complete page is random, but this sequence is repeated in other pages in ZM12. And in dynamic ageing, only 8 bytes are generated via random process, and this sequence is repeated in every page in ZM12'.

The ZM2, ZM3, ZM4 and ZM12' seem to follow the same trend. ZM1 still shows no error. ZM2, ZM3, ZM4 and ZM12' display quite the same

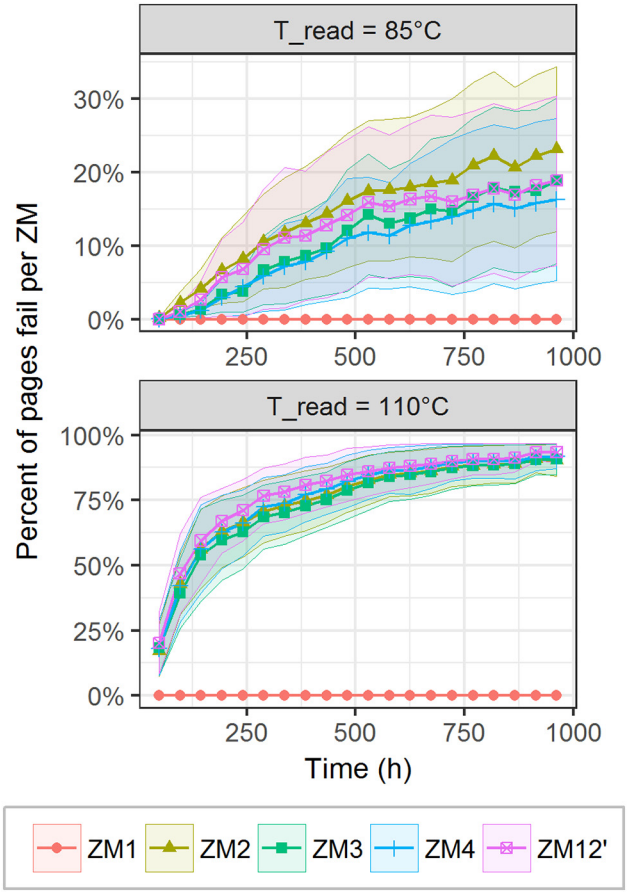

Fig. 7. Evolution of the mean percentage of failed pages for each ZM, pieces written at −40 °C and read at 85 °C or 110 °C, read in continuous, dynamic ageing. Error envelopes are the empirical unbiased standard deviation.

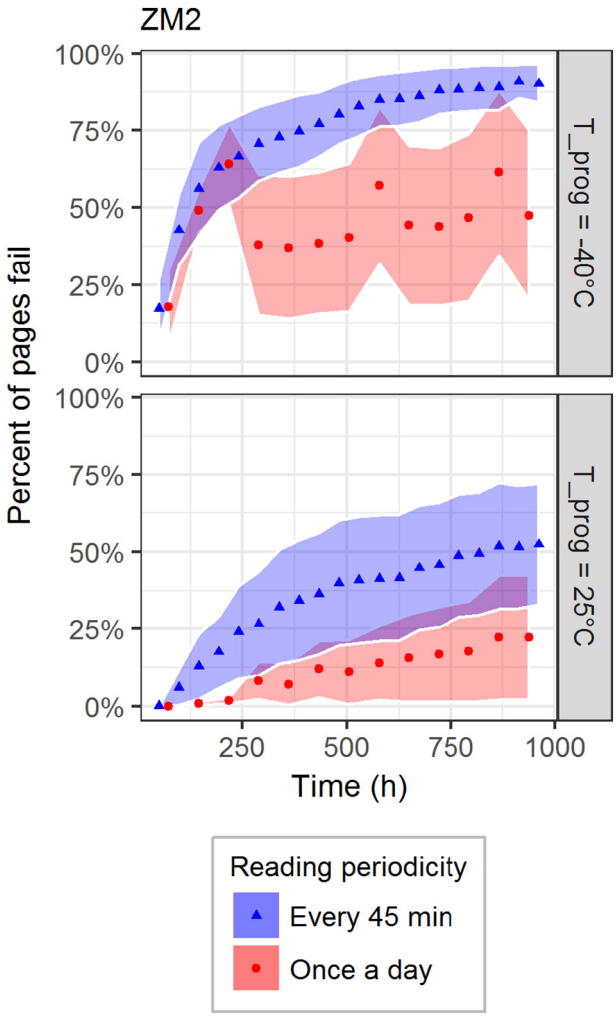

Fig. 8. Evolution of the mean percentage of failed pages for two programming temperatures depending on the reading periodicity, pieces read at 110 °C, ZM2, dynamic ageing. Error envelopes are the empirical unbiased standard deviation.

behavior (see Fig. 7).

4.5. Influence of reading periodicity

In order to evaluate the skew induced by reading disturbances, two periods of reading have been established. Half of each ZM is read in continuous, and once a day the entire ZM is read.

The behavior of less frequently read cells is roughly the same as the others, except for the number of failed pages which differs by a factor of 2 (see Fig. 8).

An increase of the number of read operations per day seems to lead to a decrease in reliability (shorter retention time). This result exacerbates suspicions of reading disturbances mechanisms in this type of flash memory.

The main observation done with dynamic test is that biasing strongly decreases data retention time. The influence of reading periodicity on reliability is important but even with only one reading per day, error rates of cells activated under high temperatures remain high after 1000 h.

5. Discussion

First results match the ones that can be found in various publications [3, 4]. Indeed, the most critical conditions in regards to the reliability of NAND Flash memories are: low temperature for write operations and high temperature for data retention [5]. Under activation and continuous reading cycles at high temperature, failures are shown to occur even earlier. The low reliability of data retention at high temperature is well known and documented in plenty of papers and is due to charge leakages [3, 4, 6].

The influence of low writing temperatures on reliability has not been considered in most studies. After correlation of observed results, a possible explanation is given below.

Most NAND Flash memories implement the Fowler-Nordheim tunneling effect [1] in order to inject charges through the floating gate [7] during write operation. During write cycles, the programming circuit controls the charge of cells to ensure a sufficient margin of voltage threshold. It is assumed that the writing management circuit probably drifts with low temperatures. Indeed, transistor parameters (threshold voltage and gain) vary with temperature which in turn induces drain current shifts.

This shift leads to a change of the reference voltage used during program and read iteration in order to inject the relevant quantity of charge in the floating gate. This finally results in a misalignment of threshold voltages and cell states (see Fig. 9). Also, when they are read at higher temperature, threshold voltages of cells written at low temperatures are very close to the expected boundaries of cell states (P1, P2 or P3) at usual temperatures. Hence, for an identical drift of their voltage threshold, they reach the boundary value a long time before the cell written at higher temperature.

For this reason, some DUT whose programming had been done at −40 °C did not show any error bits just after write operation (read operation performed at −40 °C in order to confirm a correct initial state), but showed error bits starting from the first read operation at

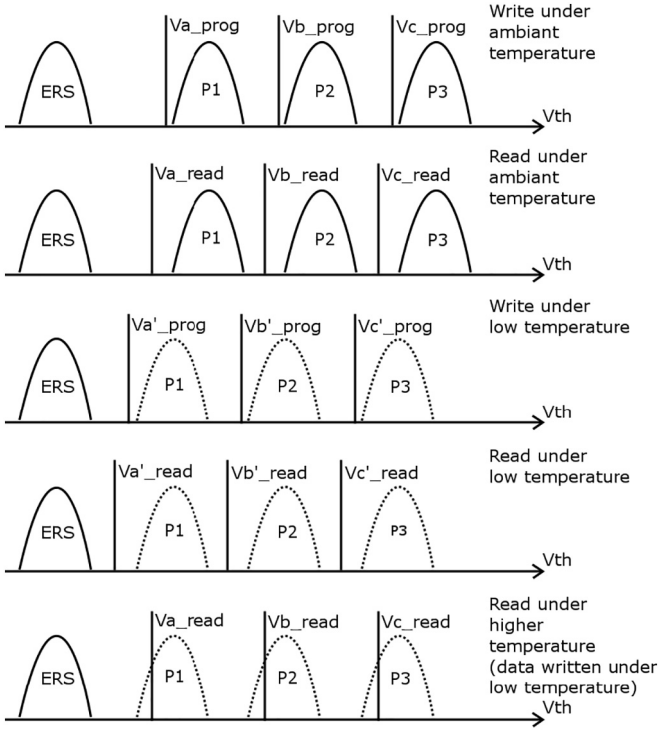

Fig. 9. Effects of reference voltage drifts during write operation at low temperatures on the distribution of threshold voltage at ambient temperature.

110 °C (dynamic ageing). It confirms the fact that the peripheral circuit's behavior is affected by temperature. It does not detect any problems during writing because of its own skew.

Hence, the charge loss required to corrupt the data is lower as the threshold voltage is closer to limits of expected state distribution.

All that said, with no knowledge of internal design, it is still complicated to accurately determine the mechanism responsible for the inappropriate writing of memory cells.

6. Conclusion

Thermal effects cannot be ignored in the estimation of the data retention time of NAND Flash memories, during both storage and writing.

Write operations at low temperatures lead to a decrease in data retention time, probably not due to a degradation of the cell but due to parametric drifts of the die embedded electronics dedicated to write operations.

Storage at high temperatures punctuated by multiple reading accesses also decreases retention time, due to a charge leakage phenomenon described in other papers.

Acknowledgements

These tests have been realized with the support of DGA ('Direction Générale de l'Armement') within the framework of the study named PISTIS (contract n°2015 91 0907).

References

[1] L. Crippa, R. Micheloni, I. Motta, M. Sangalli, Nonvolatile memories NOR vs. NAND architectures, Memories in Wireless Systems, Springer-Verlag, 2008, pp. 29–53.
[2] M. Hawes, How micron FortisFlash technology improves performance and endurance, A Micron Technical Marketing Brief, 2015.
[3] B. Govoreanu, J.V. Houdt, On the roll-off of the activation energy plot in high tempreature flash memory retention tests and its impact on the reliability assessment, IEEE Electron Device Lett. 29 (2008) 177–179 n° %12.
[4] B.D. Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumot, P. Candelier, Experimental and theoretical investigation of nonvolatile memory data-retention, IEEE Trans. Electron Devices 46 (1999) 1518–1524 n° %17.
[5] I. Olson, NAND Flash Memory Reliability in Embedded Computer Systems, Schweitzer Engineering Laboratories, 2014.
[6] K. Lee, M. Kang, Y. Hwang, H. Shin, Accurate lifetime estimation of sub-20-nm NAND flash memory, IEEE Trans. Electron Devices 63 (2) (2016) 659–667.
[7] S. Aritome, Reliability of NAND flash memory, in: I. The Institute of Electrical and Electronics Engineers (Ed.), Nand Flash Memory Technologies, First ed., John Wiley & Sons, Inc., 2016, pp. 195–272.