Embedded Flash Testing: Overview and Perspectives

Olivier Ginez, Jean-Michel Daga, Patrick Girard, Christian Landrault, Serge Pravossoudovitch, Arnaud Virazel

Embedded Flash Testing: Overview and Perspectives

O. Ginez [1, 2] J.-M. Daga [2] P. Girard [1] C. Landrault [1] S. Pravossoudovitch [1] A. Virazel [1]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier – Université de Montpellier II / CNRS
161, rue Ada – 34392 Montpellier Cedex 5, France
Email: <lastname>@lirmm.fr URL: http://www.lirmm.fr/~w3mic

[2] ATMEL Rousset – Libraries and Design Tools Department – Embedded Non-Volatile Memory Group
13106 Rousset Cedex, France
Email: jean-michel.daga@rfo.atmel.com URL: http://www.atmel.com

Corresponding author e-mail:
olivier.ginez@lirmm.fr

***Abstract*— The evolution of System-on-Chip (SoC) designs involves the development of non-volatile memory technologies like Flash. Embedded flash (eFlash) memories are based on the floating-gate transistor concept and can be subject to complex hard defects creating functional faults. In this paper, we present a complete analysis of a particular failure mechanism, referred as disturb phenomenon. Moreover, we analyze the efficiency of a particular test sequence to detect this disturb phenomenon. Finally we conclude on the interest to develop new test infrastructure well adapted to the eFlash environment.**

***Keywords*—**Flash memory, 2T-FLOTOX core-cell, Fowler-Nordheim Tunneling effect, Disturb, Checkerboard, March algorithms, Coupling effect.

I. INTRODUCTION

Different types of memory can be embedded in a SoC as SRAM, DRAM, EEPROM and Flash. The increased use of portable electronic devices such as mobile phones and digital camera produces a high demand for flash memories. Flash memory is a non-volatile memory that allows programming and erasing memory data electronically [1]. The mainstream operation is based on the floating-gate concept on which charges can be stored and removed. Its low-power consumption and high density make it popular for portable devices.

The functional scheme of an embedded Flash (eFlash) memory is presented in Figure 1. As all other memories the eFlash is composed of a core-cell array (CORE), bit-line and word-line decoders and sense amplifiers. In addition, eFlash memories need particular building blocks to perform dedicated functions: a Charge Pump device for the High Voltage Generation (HVG) allowing the write and erase operations a Sense Reference Voltage used during the read operation, and data latches (DLATCH) allowing full page programming in parallel.

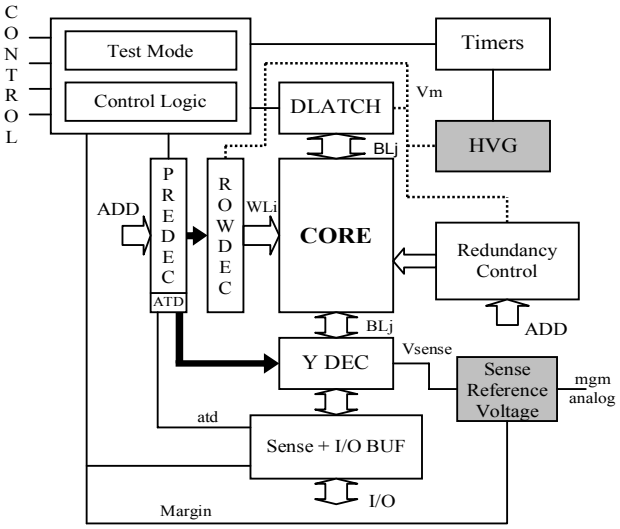

Figure 1: Functional scheme of eFlash memories

The eFlash core-cell is based on the floating-gate (FG) concept. There are two typical mechanisms to transfer electric charges from and into the FG: hot carrier injection and the Fowler-Nordheim (FN) tunneling effect. The FN tunneling effect is used for charge injection or removal in 2T FLOTOX core-cells for example. For eFlash memories, FN tunneling offers several advantages such as low power operation and very good endurance. In addition a very low current programming mechanism offers good opportunities to decrease testing time by implementing massive parallel programming modes. As a result, FN tunneling is extensively used for both Erase and Write operations in embedded flash memories. Three different operations can be performed on an eFlash: Erase, Write and Read. The Erase operation consists in injecting charges in the FG with a specific high voltage combination. To inject charges in the FG, the high voltage must be applied on the Vref node of the sense transistor while its drain must be maintained at ground (Figure 2). During the Erase operation, the core-cell

0-7803-9727-4/06/$20.00 ©2006 IEEE

is 'on' and allows the node BLj to be pulled-down at the Vss potential fixed at ground. It is important to notice that the Erase operation is performed simultaneously on all the cells of the same page and not cell-by-cell. At the end of the Erase operation, charges trapped in the FG have changed the VT of the sense transistor to a high VT referred as VTH in Figure 3. From a functional point of view VTH corresponds to logic '1'.

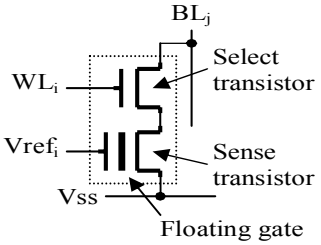

Figure 2: 2T-FLOTOX core-cell

The Write operation consists in removing electrons from the FG by putting the Vref node at ground and maintaining BLj at a high voltage. With this operation, charges of the FG are removed and so the sense transistor has a low VT, referred as VTL in Figure 3, which corresponds to logic '0'. We call VT window the difference between VTH and VTL. Note that the write operation is performed with bit granularity. The memory cell current is sensed using a dedicated sense amplifier circuit. If the sense transistor has a low VT (VTL), it delivers a current and the sense amplifier provides a logic '0' on its output. On the other hand, with the same Vref value, if the sense transistor has a high VT (VTH) there is no current through the bit-line and the sense amplifier gives a logic '1'.

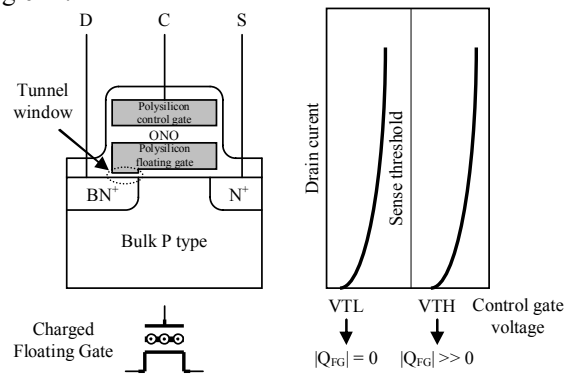

Figure 3: Illustration of the floating gate concept

The major characteristic of eFlash memories is the slow programming operation. This is mainly a technology related issue, due to the small amount of current involved during cell programming. FN tunneling current is in the 10^{-12}-10^{-11}A ranges. Figure 4 shows the typical cell programming characteristics with respect to the programming time when using FN tunneling. The threshold voltage difference between the erased and written states (VT window) must be large enough to authorize fast differentiation between the two programming states during read operation.

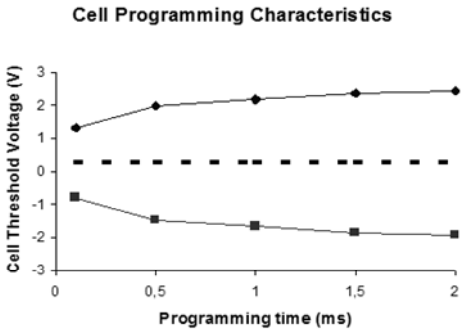

Figure 4: Threshold voltage depending on programming time

Program window narrowing across the time due to leakage mechanisms is a major reliability limitation of eFlash memories. In addition, bit programming efficiency decreases after several program cycles, resulting in a program window closure. Margins should be taken when programming to ensure an acceptable program window after the specified retention time.

II. FAULT MODELING

The previous section has introduced two main aspects that we have to consider in the eFlash testing environment, namely eFlash technology specificities and the slow programming time. The eFlash technology is important in the fault modeling process whereas the slow programming time has to be considered for test sequences or algorithms development.

The standard list of fault models coming from the literature is not necessary realistic because most of the faults have been derived from CMOS memories such as SRAM memories. Embedded Flash failure mechanisms, such as program disturb, endurance and retention limitations, are directly linked to the floating gate technology. These mechanisms are related to the intrinsic and aggressive analog characteristic of the programming operation. A lot of work has therefore to be done in the field of eFlash fault modeling. One of our previous works has shown that embedded flash memory can be subject to complex failure mechanisms [2]. This study allows predicting more efficiently the eFlash faulty behavior involved by a set of complex defects. The interest of such a study is to complete the actual functional fault listing reported in the semiconductor memories literature. To illustrate the previous statement, we develop one typical example of specific fault on eFlash memories.

II.1. Analysis of the disturb phenomenon

The disturb failure appears when a Write, Erase or Read operation on a targeted cell affects the state of its neighbourhoods. Most of the time, the disturbances are due to the presence of a high voltage on the core-cell nodes. Until now, all disturb mechanisms analysis have been done on NAND or NOR-based eFlash architecture with ETOX

core-cells but never on 2T FLOTOX structures. The reason is the presence of the select transistor placed serially with the sense transistor that must normally inhibit the high voltage from the bit-line when a cell is unselected for a program operation. In this section, we will show that disturb phenomenon may also appear in NOR-based eFlash architectures with 2T FLOTOX core-cells and in presence of an oxide tunnel thickness variation.

Thanks to its structural specificities, the 2T FLOTOX eFlash memory may be affected by only one disturb mechanism. This disturb is due to the bit-line coupling between the targeted cell and the victim sharing the same word-line (Figure 5). The aggressive over-scaling of eFlash technology enables two adjacent bit-lines (at a layout point of view) to create a non-negligible coupling capacitance (C1). This coupling creates a capacitive divider bridge with the equivalent bit-line capacitance (C2). Due to this coupling effect, the high voltage applied on the BLj node for a write operation involves an undesired increase of the node BLj-1 potential.

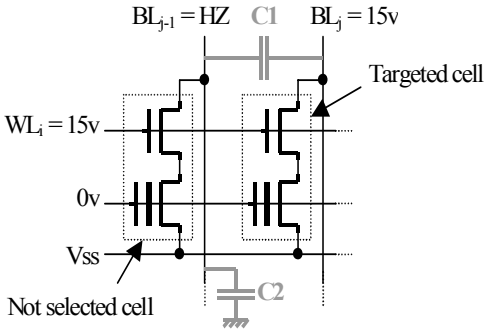

Figure 5: Disturb failure due to bit-line coupling

With advanced technology rules (<0.15µm), the ratio between the coupling capacitance C1 and the bit line capacitance C2 increases drastically as C1 is in the same order of magnitude as C2. The resulting voltage on the unselected bit line increases to reach a value that can cause a cell disturb. As the select gate of the victim cell is on, the coupled bit-line voltage value is directly applied on the drain node of the sense transistor. The presence of this voltage creates an electric field between the floating-gate and the drain diffusion (BN+) of the sense transistor. To have a well understanding of this phenomenon, we must take into account the Fowler-Nordheim current equation (1), which realizes the programming operation in a 2T FLOTOX core-cell:

$$I_{FN} = A * \alpha * E^2_{ox} * \exp(-\beta/E_{ox}) \quad (1)$$

with:

A = Tunnel window area
α = Fowler Nordheim constant
β = Fowler Nordheim constant
E_{ox} = Oxide electric field

Moreover, the threshold voltage value of a FLOTOX core-cell depends on the charge quantity stored in its floating-gate. This quantity is given by the integration of the Fowler-Nordheim tunneling equation (1) during the write time Tp:

$$VT_{cell} = K*VT_{FG} - (Q_{FG} / C_c) \quad (2)$$

with:

$Q_{FG} = Q_{FG0} + \int Tp\ I_{FN}*dt$
K = Coupling factor
VT_{FG} = Floating-gate voltage threshold
C_c = ONO (Oxide Nitride Oxide) capacitance

In the equation 1, we see that the oxide electric field Eox takes an important part on the Fowler-Nordheim current generation and we know that this electric field directly depends on the voltage between the drain and the floating gate node. With the help of equations 1&2, we can find a relation between the threshold voltage variation of an erased cell (ΔVTH) under a disturb voltage and its exposition duration. Thus, we know that a disturb voltage can occur on the bit-line node of an unselected core-cell due to a coupling effect between two bit-line nets. From a theoretic point of view, the core-cell is designed to avoid a VT changing under some considerations, for example a minimal electric field is required to shift the threshold voltage of the cell. Indeed, a stand-alone disturb voltage does not impact so much the core-cell voltage threshold but in presence of a defect in the tunnel oxide thickness [3] we can observe a large variation of this voltage. This is the reason why in this section we add to the disturb voltage a possible variation of the tunnel oxide thickness, Tox. We consider that the tunnel oxide thickness can vary from 60Å to 80Å. This Tox variation is referred as Df9 Figure 6 below:

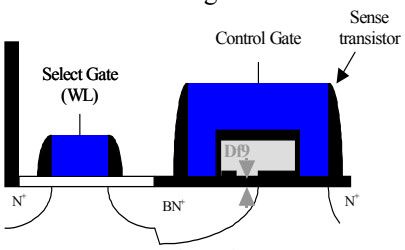

Figure 6: Cross section of a FLOTOX core-cell

II.2. Evaluation of the disturb impact

We have seen in the equation 2 that the programming time has to be considered to fix the final threshold voltage value of the FLOTOX core-cell. This is the reason why with the help of a theoretical model established previously, we have analyzed the impact of Df9 with a disturb voltage on the bit-line node of an unselected cell for two different programming time, Tp= 2ms and 5ms, with 2ms the nominal condition. Note that, the previous equations 1&2 describe a continuous phenomenon, and we have to digitize the previous expressions to provide numerical results. Tables 1 and 2 summarize the results obtained in this analysis. In these tables, the first column gives the Tox variation and the following ones give the threshold voltage variation from a nominal value (ΔVTH) for different disturb voltages. Note that, the threshold voltage nominal value of an erased core-cell is about +2.6V.

With the help of the Tables 1 and 2, we see the main impact of a disturb bit-line voltage when the tunnel oxide thickness is less than its nominal value. Indeed, even if we have an important Tox variation (ΔTox), we observe an impact on

the VTH of an erased cell when the disturb voltage is equal or higher than 7V. In such a case, the VT value of the erased cell can shift from a logic '1' (VTH) to a logic '0' (VTL) when the VT variation is close to 2.5V.

Table 1: ΔVTH of a disturbed FLOTOX core-cell during 2ms

		Disturb Voltage (volts)					
		5v	6v	7v	8v	9v	10v
Tunnel Oxide Thickness (Å)	60	5m	208m	1.01	1.96	2.92	3.87
	62	2m	104m	0.78	1.72	2.67	3.63
	64	1m	50m	0.57	1.48	2.44	3.39
	66	0	21m	0.38	1.24	2.2	3.14
	68	0	9m	0.23	1	1.95	2.9
	70	0	4m	0.12	0.78	1.71	2.65
	72	0	2m	62m	0.57	1.47	2.42
	74	0	1m	30m	0.39	1.23	2.18
	76	0	0	14m	0.25	1	1.93
	78	0	0	7m	0.14m	0.78	1.7
	80	0	0	3m	76m	0.57	1.45

Table 2: ΔVTH of a disturbed FLOTOX core-cell during 5ms

		Disturb Voltage (volts)					
		5v	6v	7v	8v	9v	10v
Tunnel Oxide Thickness (Å)	60	11m	362m	1.25	2.21	3.16	4.12
	62	4m	208m	1.03	1.97	2.93	3.88
	64	1m	105m	0.8	1.74	2.73	3.65
	66	0	48m	0.6	1.51	2.46	3.41
	68	0	21m	0.4	1.28	2.23	3.18
	70	0	9m	0.24	1.05	1.99	2.95
	72	0	3m	135m	0.83	1.76	2.71
	74	0	1m	70m	0.62	1.53	2.48
	76	0	0	33m	0.43	1.3	2.25
	78	0	0	16m	277m	1.07	2.01
	80	0	0	7m	164m	852m	1.77

Concerning the programming time, its impact on the VT value variation is less important compared to the association of a disturb voltage with Df9. Thanks to this analysis we can easily imagine the functional model related to these two failure mechanisms. In another hand we can also imagine that Df9 could be a hard defect, either a short between the BN+ diffusion (channel) and the floating-gate (FG) or an open between this two layers. In the case of an open, if there is no tunnel window between the FG and the channel, the write and the erase operations will be inhibited and thus insensitive to disturb voltage. In this case the cell remains close to its virgin state with a standard threshold voltage. With a short between the FG and the channel, there is a very good bit programming but there is no data retention because the floating-gate is not isolated from all others nodes. The two previous cases are extreme ones because in almost all practical cases only a little variation of Tox on a defective bit may occur. In the case of -ΔTox, we have a bad retention of charges by the FG but a good VT window. It means that the duration of the stored information is not maximal. Moreover, a little negative ΔTox variation can increase the stress of the oxide and so, affect the reliability of the core-cell. When the tunnel window thickness variation is positive, the electric field is smaller and the charges are less injected or depleted in the floating-gate. The VT window is affected.

III. TEST ALGORITHMS AND FAULT COVERAGE

From a test point of view due to the intrinsic very low speed of the programming operation, eFlash-testing strategy is very different from other types of memory. March like algorithms are not suitable, and a very limited number of patterns can be used to test the memory in order to keep the testing cost acceptable. Assuming page programming with 2ms to erase and 2ms to write a page, it takes 4ms to program a whole page. Based on 256 bytes per page architecture, 512 pages are necessary to build a 1Mb memory, resulting in more than 2 seconds to write one pattern to the 1Mb eFlash using the page mode. Consequently, programming a set of basic patterns such as '00', 'FF', checkerboard and inverse checkerboard using a page mode programming will result in a testing time close to 10s per die. Decreasing the testing time per die is technology dependent, as parallel access to full or large portions of the array is mandatory to speed up the programming of test patterns. Programming a huge number of cells in parallel is only possible if a very low current programming mode such as FN tunneling is used. Executing dedicated test modes, one time programming of large sectors to '00' or 'FF' is possible in a few milliseconds. A checkerboard pattern can also be programmed in a few ms using a partial programming mode. In any case, even if only one testing pattern is programmed using the user mode (page programming), testing time of medium to large eFlash memories will be in the range of seconds, to compare with milliseconds in ROM testing. A 5N testing sequence is typically used to test eFlash memories. The fault coverage analysis of this 5N sequence has been performed and we reach a 100% coverage rate on the test of Stuck-At, Stuck-Open, Addressing and Transition faults. The particular programming mode used during 5N test sequence is not able to detect more complex fault. Indeed, the writing of a large amount of cells in parallel does not allow testing particular coupling faults. For example, all combinations of idempotent coupling fault (CFid) from the CMOS memories testing literature are not detected by the 5N test sequence. From an algorithmic point view, the 5N testing flow is composed as follows:

(1N)	Wppm CKI + Read CKI
(2N)	Wpage CKB + Read CKB
(3N)	Wpage Diag0 + Read Diag0
(4N)	CE + Read '1'
(5N)	CW + Read '0'

Where CKB (CKI) stands for checkerboard (inverse checkerboard), and Diag0 is a diagonal pattern of '0'. Wppm is a test mode allowing high speed programming of the CKB (CKI). CE and CW are specific test modes allowing one time programming of the full array to 'FF' and '00' respectively. Wpage is the page mode programming which corresponds to the user mode.

Now, if we consider the 1Mb eFlash memory architecture cited above (256 bytes on 512 pages), this 5N sequence will take close to 5sec to test the whole memory. We can easily imagine the test cost of such memory compared to the equivalent SRAM or ROM memory architecture.
Such test sequence will have an impact on the test quality of 2T FLOTOX eFlash memory. For example if we consider the previously detailed failure mechanism involving an oxide thickness variation and a bit-line coupling, such test strategies present a weakness. From a functional point of view, this failure mechanism has a faulty behavior that can be modeled as an idempotent coupling fault and more precisely a CFid (↓,↓). In the 5N test sequence, we know that the patterns CKB and CKI have the ability to detect CFid (↓,↓). But these faults occur only between two adjacent core-cells sharing the same word line signal. In this figure, the cells are located according to the i and j axes:

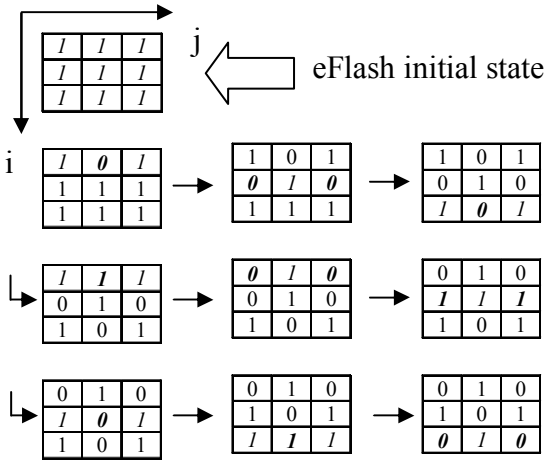

Figure 7: CKI and CKB patterns on a 3*3 eFlash array

In Figure 7, the pattern written between two steps is represented in italic whereas the bits changing are highlighted in bold. For example if we look the first page of the memory, the CFid (↓,↓), in which cell (0,1) is the aggressor and cell (0,0) or (0,2) are the victim cells, is never tested. If we want to detect all possible combinations for this kind of fault we have to use a March like algorithm that is the most often used in a CMOS memory context [4]. eFlash memories are word oriented, and in order to detect intra-word Coupling Fault, a 'Bit Oriented Memory' March test could be converted to a 'Word Oriented Memory' March test. This could be done by replacing the bit wide operations (`r0', `r1', `w0' and `w1') by operations reading and writing a data background of n bits [5]. In the case of CFid detection, we have to define the number of data background (NB_{DB}) to apply to the memory. In [5] the formula is given as following:

$$NB_{DB}=3+3*\log_2(B) \qquad (3)$$

where B is the number of bits of a word
Now, we consider an eFlash memory composed of 64K words of 32 bits. The memory has 1024 pages of 64 words composed of 32 bits. To detect all possible CFid in such a memory, we have to calculate the number of data backgrounds that we need using expression (3). To simplify the context, we consider that a word corresponds to an entire page, either an equivalent word of 2048 bits. In applying the expression 3, we find that 36 different data backgrounds (3+3*11) are needed to detect CFid. We have seen previously that an erase followed by a write takes 4ms in an eFlash memory. Thus the test of all possible CFid in our memory will take 1024*36*4ms, either 147sec. From this estimation, it is evident that March like algorithms are not a cost effective solution. A realistic list of faults suitable for this type of memory is first mandatory. According to this list of faults, 5N test algorithms can be improved to achieve acceptable fault coverage, while minimizing as much as possible its impact on the test time. In addition, the test infrastructure should be adapted to deal with the eFlash based SoCs.

IV. TEST INFRASTRUCTURE

From a memory design point of view, dedicated circuitry must be added to manage parallel programming modes used to speed up the test operation. Considering now the eFlash access for test, the following conditions must be respected:

- The embedded flash should be independently accessible (controllable) and observable to allow concurrent testing of the eFlash and the rest of the circuit in the single insertion approach.
- The number of pins necessary to test the memory should be minimized as much as possible in order to test the maximum of dies in parallel
- Enough flexibility to provide debug capabilities must be maintained.

According to these requirements, the use of a Serial Test Interface appears as a good compromise [6]. The number of dedicated pins can be decreased to a small number using a STI, and there are no limitations to the test patterns that can be applied, providing a lot of flexibility for debugs purpose. Typical STI limitations in terms of speed limitations and impact on testing time are limited with this technology, due to the intrinsic long programming delay. However, BIST strategies could be considered as a way to minimize testing cost. Even if they are a fraction of the total testing time, delay overheads due to the STI could be removed. In addition, at-speed read operation on high-speed eFlash [7] could become mandatory, and is only possible with BIST. A comparison was performed on several products embedding high-speed flash memories (25ns random read access time) ranging from 32KB to 1024KB. Due to the intrinsic low speed of the programming operation, the delay overhead between the STI approach and the BIST is mainly due to the read sequence, and increases with the memory capacity. Assuming at-speed read operation with BIST, and a 16MHz maximum frequency for the STI (mainly limited by hardware) the delay overhead was 9% for the 32KB memory capacity, and increased up to 35% with the 1024KB configuration. As a result, BIST appeared as a good opportunity to save testing delay, especially on large memories.

However, keeping a good level of flexibility and diagnosis capability is mandatory, and a mixed solution STI/BIST could be considered as an ideal solution.

Increasing test parallelism is of primary importance to decrease testing cost. A typical solution is to have multiple wafer probing steps using a highly parallel memory tester for the eFlash test, and a mixed-signal tester to test the logic and analog parts of the circuit. This strategy has the advantage to minimize the time spent on the more expensive tester. It has also some drawbacks: it increases the investments in dual set of probe cards and interface hardware, and it requires a balancing of the number of each tester type, based on the relative time spent to test the eFlash and the rest of the circuit. As the eFlash testing time depends on the memory capacity, it may be difficult to balance the test equipment on large families of products [8]. Multi site testers allowing concurrent testing of the embedded flash, MCU and logic are now available [9]. To take full benefit from the tester capabilities, design for test techniques must evolve accordingly. The eFlash should be independently accessible and observable through a serial test interface (STI) to minimize test pins as much as possible, and allow concurrent test of the flash and the rest of the circuit. As a result, the overall chip testing time reduces to the time needed to test the eFlash memory. The cost effectiveness of such strategy will depend on several factors such as: extra charge of multi purpose testers versus low-end memory testers, eFlash memory density, complexity and testing time of the logic and analog parts of the chip.

V. CONCLUSION

Flash based SoCs are more and more popular, mainly because eFlash adds a lot of flexibility to the system. Unfortunately there is a price to pay. Among the contributors to the flash additional cost such as increased process complexity, testing cost must be carefully evaluated. This is mainly due to the intrinsic characteristics of the FG device, resulting in a very low speed programming operation. As a result, the eFlash testing problematic is the following: how to guarantee acceptable fault coverage with a limited number of programming operations? This requires a very good understanding of eFlash specific defects, resulting in a realistic list of faults, to be considered for test pattern generation. These aspects have been illustrated in the paper by a detailed description of one specific failure mechanism related to disturb. The fault coverage based on a cost effective 5N testing flow has been provided, and compared to a March approach in terms of cost. In addition, the dedicated test infrastructure associated with flash based SoCs has been discussed. Several options such as STI or BIST have been compared, taking into account the flexibility for debug purpose, and the testing time. Finally, different options regarding the tool set optimization considering the overall SOC test have been reported.

REFERENCES

[1] P. Pavan, R. Bez, P. Olivo and E. Zanoni, "Flash Memory Cells – An Overview", Proc. of the IEEE, vol. 85, no. 8, pp. 1248-1271, 1997.

[2] O. Ginez, J.-M. Daga, M. Combe, P. Girard, C. Landrault, S. Pravossoudovitch and A. Virazel, "An Overview of Failure Mechanisms in Embedded Flash Memories", Proc. of IEEE VLSI Test Symposium., Berkeley, 2006, pp. 108-113.

[3] M.G. Mohammad, K.K. Saluja, "Simulating Program Disturb Faults in Flash Memories Using SPICE Compatible Electrical Model", IEEE Transactions on Electron Devices, vol. 50, no. 11, pp. 2286–2291, 2003.

[4] A.J. van de Goor, "Testing Semiconductor Memories, Theory and Practice", COMTEX Publishing, Gouda, The Netherlands, 1998.

[5] A.J. van de Goor, I.B.S. Tlili, "March Tests for Word-Oriented Memories," Proc of DATE, Paris, 1998, pp. 501-509.

[6] JM Daga, “Test and repair of embedded flash memories”, Proc of ITC, Panel P7.1, Baltimore, 2002, p.1219.

[7] JM Daga, “Embedded EEPROM speed optimization using system power supply resources”, Proc. of PATMOS, Santorini, 2004, pp. 381-391.

[8] Roger Barth, “Selective optimization of test for embedded flash memory”, Proc of ITC, Panel P7.4, Baltimore, 2002, p.1222.

[9] J. Agin, “Overcoming Test Challenges Presented by Embedded Flash Memory”, Proc. of the International Manufacturing Technology Symposium, 2003, pp. 197-200.