

An Optimized Test During Burn-In for Automotive SoC

Original

An Optimized Test During Burn-In for Automotive SoC / Appello, D., Bernardi, P., Bugeja, C., Cantoro, R., Pollaccia, G., Restifo, M., Ernesto, S., Venini, F.. - In: IEEE DESIGN & TEST. - ISSN 2168-2356. - STAMPA. - (2018), pp. 46-53. [10.1109/MDAT.2018.2799807]

Availability:

This version is available at: 11583/2698573 since: 2018-05-18T15:31:51Z

Publisher:

IEEE

Published

DOI:10.1109/MDAT.2018.2799807

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

An Optimized Test During Burn-In for Automotive SoC

**Davide Appello¹, Paolo Bernardi², Conrad Bugeja¹, Riccardo Cantoro²,
Giorgio Pollaccia¹, Marco Restifo², Ernesto Sanchez², Federico Venini²**

¹ *STMicroelectronics, Italy*

² *Politecnico di Torino, Italy*

Abstract

The cost of Burn-In is a major concern for the testing of Automotive Systems-on-Chip (SoCs). This paper proposes an optimized Test-During-Burn-In (TDBI) flow that takes advantage of the parallel execution of several types of stress procedures in which many components are carefully interleaved. The proposed methodology permits to significantly reduce the BI time and enables production monitoring by providing detailed test data-logging capabilities helping the debug of potential yield issues largely caused by the ageing of Burn-In tester consumable parts. The paper describes an experimental scenario about TDBI of an automotive SoC manufactured by STMicroelectronics.

Keywords: SoC, Automotive, Test-During-Burn-In, Burn-In Equipment Issues

1. Introduction

The purpose of the Burn-In (BI) process [1] is to give rise to infant mortalities (early life latent defects) that naturally affect populations of electronic devices. The Burn-In approach uses high temperature to accelerate the rate at which the latent defects appear [1]. A special board called *Burn-In-Board* (BIB) hosts hundreds of chips to perform the process in parallel. A BI tester is composed of a climatic chamber, that accommodates a set of BIBs, and an Automatic-Test-Equipment (ATE), which performs the test procedures on all the chips in parallel.

Burn-In has long testing times and high costs, which make it a bottleneck in the IC manufacturing process. Engineers have struggled to reduce time and cost over the past several decades [1].

A typical System-on-Chip (SoC) usually integrates at least one microprocessor core, I/O, advanced peripherals, RAM, and Flash. The Burn-In of SoCs for safety-critical applications comprises several types of stress procedures depending on the modules integrated into the SoC. Stress on microprocessor targets the ageing of all logic components of a CPU based system; ageing on RAM is accelerated by a specific amount of high-voltage read and write operations, while erase and verify operations are used for Flash memories. Generally, internal stress procedures induce junction level stress [2] by making the circuit toggling and thus raising the internal temperature. In this way, the internal stress complements the external acceleration factor obtained by applying high temperatures. Burn-In testers may provide also another factor of stress, which is related to the voltage supplied during the electrical stress application.

Typically, a large timespan is spent to stress embedded Flash memories during the Burn-In of a SoC; this is done by erasing them many times to meet strict quality standards (*Flash erase cycling*). The Flash erase duration is not known a priori, and it changes from one erase to another due to the erratic erase effect discussed in [3][4]. Moreover, Flash memories present a strong temperature dependency concerning the program/erase speed [5]. For example, in a scenario with a 4MB 90nm technology Flash Memory, the erase duration may last from 25 to 45 seconds according to environmental conditions.

An additional purpose of BI is the logic gate stress [6][7], where logic stimuli stress the SoC gates. Logic gate stress uses scan chains, Built-In Self-Test (BIST) engines or functional programs, depending on the type of Design-for-Test infrastructure available on-chip.

Hereinafter, we refer to the *Test-During-Burn-In* (TDBI) process as the BI process where the ATE can drive and monitor the execution of test procedures. The primary concern for BI is

a reduction of the Flash erase time required to achieve a satisfactory stress. BI time reduction requires an optimization of both stress procedures and tester electrical capabilities.

Additional criticalities of the BI process can apparently cause yield issues affecting the overall test time. The harsh environment of the climatic chamber heavily wears the BIB and the board's sockets, thus possibly provoking misbehaviours. This means that sometimes the chip may not communicate properly with the ATE. Disconnections between a chip and the ATE are quite common and lead to the identification of a set of "suspect" devices that are affected either by a real failure or, most likely, by showing a false fail behaviour. Suspect devices need to be completely or partially re-processed or "recycled". There is no single and well-identified reason that causes disconnection problems; industry accepts several hypotheses including: variation of the nominal impedance of the pins of the device and the connectors of the socket, weaknesses in the communication protocol, and integrity issues between clock and communication signals. The mitigation of the occurrence of these disconnections and their correct management are crucial to optimize the throughput of a BI facility.

This paper illustrates a TDBI setup that addresses both the BI time reduction and the mitigation of the disconnection problem. We propose to parallelize the Flash erase cycling phase and the execution of functional programs that stress (and test) the digital domain of the SoC; this parallelization brings a substantial BI time saving. Moreover, we illustrate the characteristics of an improved flow aimed at collecting data concerning disconnection phenomena. Such a data-logging feature takes advantage of shadow Flash sectors and permits to move a false fail device to the good bin without recycling it.

Experimental results are gathered on a population of SoCs manufactured by STMicroelectronics.

1.1. Test-During-Burn-In flow

The TDBI process is composed of several crucial phases for quality and cost effectiveness. It is important that all the components of the device are suitably stressed and tested [8]. Therefore, the TDBI process encompasses the following steps:

- PRETEST of device liveness and connection to ATE – PARAMETER Check
- Flash cycling performs a predetermined number of Flash erases (e.g., 500)
- Dynamic Burn-In: aims to maximize the stress and it is divided into sub-phases:
 - ATPG stress and test patterns, which are applied through scan chains to toggle and test the entire digital domain in a single procedure
 - RAM stress obtained by multiple BISTs execution
 - Write/Read stress RAM memory cycles using checkerboard patterns
 - Functional stress programs, often derived from verification and test scenarios, activating the digital domain.
- Gate stress phase, where continuous memory reads aim at functionally exciting the gate-oxide interface of floating gate transistors and, in general, the Flash Memory control logic.

Flash cycling consists in erasing the whole memory multiple times and finally verifying its correct functionality. Flash cycling represents the most time-consuming phase in the whole process due to the erasing times imposed by the technology of the Flash module and the quality requirements determined by reliability standards.

Dynamic Burn-In performs a stress process on the digital domain and RAM memories of the device. Dynamic Burn-In procedures are composed of a *stress phase*, which targets latent faults, and a *test phase* that detects the spotted faults. The most relevant stress factors in this phase are circuit activity, chip surface temperature, and current consumption [2]. It is worth mentioning that applying higher supply voltages than the nominal also accelerates the dynamic Burn-In [9].

1.2. ATE access to SoC during BI

One of the major ATE problem is the large number of signals per chip to be controlled and their routing on the BIB. The goal is to decrease the overall number of communication signals, thus the number of contacting pins. The employed protocol is normally JTAG, which requires driving 3 signals and returning results on a single line. The BI tester communicates with the SoCs by driving IEEE 1149.1 and internal structures such as IEEE 1500 wrappers through the JTAG port by means of reading and writing registers and RAM locations. Even though this tester configuration looks simpler than full contacting solutions, there is a large number of devices tested in parallel, which provokes several issues in their test access, as categorized in Table I.

Table I: Macro-factors and factors affecting BI, and malfunctioning effects produced on the TDBI.

Macro-factors	Factors	Effects
Board topology	Long and varied lines reaching all DUT positions	Different voltage conditions per stimuli reaching different positions
	Power supply feedback loop closed on edge connector proximity	Topological yield loss because of RLC network
Ageing	Socket contact resistance on a specific position	Voltage drop on actual DUT pad under programming
	Distributed mechanical and thermal drift may impact the effectiveness and accuracy of the electrical connection between devices and BI tester. The Inhomogeneity is to be considered as an additional factor of noise during the run	Communication error (e.g., loss of contact temporary or permanent)
	Drift on capability of fuse element in avoiding a short to the ground in case a device fails. The fuse element might behave as a resistor differently from the expected open circuit	Not-usable position, sensitivity to current transient of voltage supply
Device sensitivity	Package options with lower supply pin count	Test instabilities impacting randomly yield

All the described factors, many of them related to various power supply issues, can cause deviations in the communication, that can be temporary or permanent and it might invalidate the whole process for the affected devices. In this harsh context, the insurgence of any suspect fail is a serious concern; a set of chips on the BIB may disconnect from the tester due to communication problems, being unable to exchange the result values and to inform the tester

about the correct completion of a given stress phase. These devices are labeled as *suspect fail* and may undergo a repetition of the process (i.e., they will be "recycled") exacerbating the cost of BI and potentially leading to over stressed devices.

2. Proposed strategies for an optimized TDBI flow

This paper illustrates two relevant aspects of the TDBI process. First, it addresses key concepts for the reduction of the BI process duration through an optimized concurrent stress/test execution schema. Secondly, it introduces a set of monitoring facilities able to collect information about TDBI criticalities that are responsible for false fail detection, potentially causing recycling.

2.1. Time saving by concurrent execution of stress procedures

When pursuing time saving, we propose to identify stress/test procedures that can be executed in parallel: interleaving various phases, which were previously executed sequentially. This permits to reducing the overall BI time significantly.

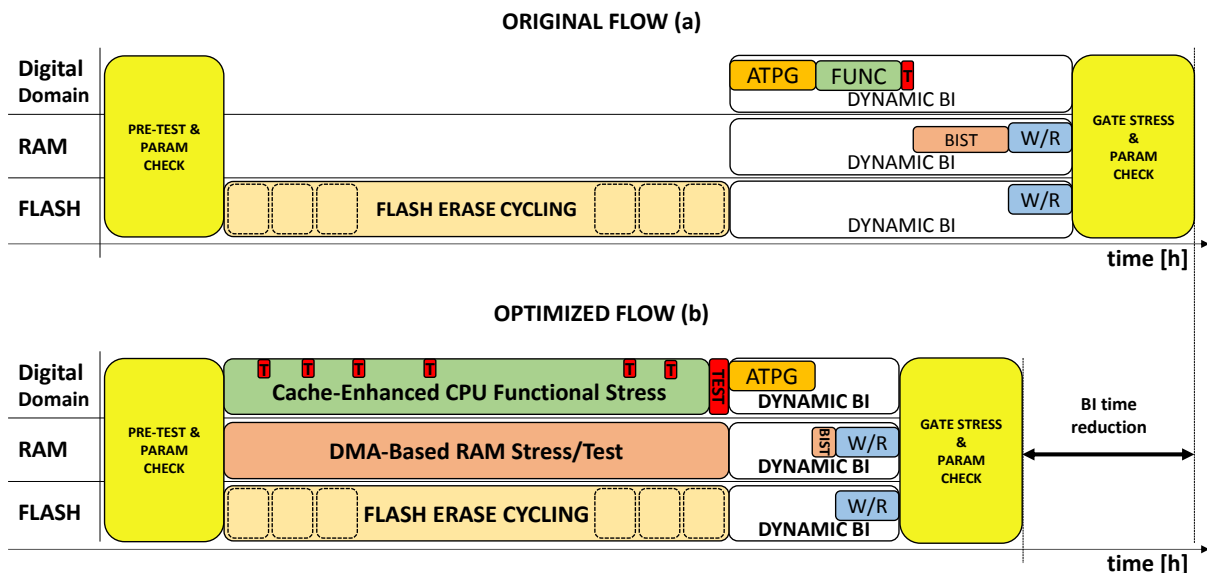


Fig. 2: Original (a) and optimized (b) Burn-In flow

It is important to stress all SoC components along the BI process. Fig. 2(a) shows the typical BI recipe, which takes care of stressing all parts of the SoC by performing an extensive Flash erase cycling followed by a complete dynamic Burn-In phase.

Running a proper stress with functional programs is time consuming. A stress program takes up to seconds of repeated executions to reach the highest temperature and to satisfy the required level of stress. On the other hand, the duration of test programs is usually short because their goal is to detect a misbehavior in the minimum amount of time.

As depicted in Fig. 2(b), the proposed technique addresses BI time reduction by parallelizing several phases, which are currently performed one after the other in distinct moments of the flow. Stress procedures for digital domain, RAM and Flash memories run in parallel.

1. The CPU triggers the Flash erase start then waits for erase completion
2. The CPU programs the Direct Memory Access (DMA) controller to start a DMA-based BIST-like test execution on RAM memory.
3. All independent peripheral cores are programmed to work autonomously (i.e., timers, PWM, etc.)
4. The CPU runs a set of functional programs, which aims at maximizing the CPU activity; the cache memory prevents conflicts between the CPU and the DMA on the bus.

As depicted in Fig. 2, the parallelization approach anticipates the stress contribution of the dynamic Burn-In in the Flash erase cycling. The CPU is also in charge of managing the communication with the ATE.

At the beginning of the BI phase, the tester uploads the program code into the embedded RAM of each device. Then, after the system reset de-assertion, the test code starts its execution. While running, the CPU communicates its status to the ATE through a specific memory-based protocol; the ATE polls a specific memory location to get informed about the Flash, RAM and digital domain stress progression.

2.2. SoC to ATE communication issues analysis

Devices losing the communication with the tester cannot provide the confirmation of the execution of the complete stress sequence. Communication issues are thus producing “suspect”

failing chips; it is not known whether a suspect chip fully completed the stress process or it was run only partially. Suspect failing chips are usually recycled as shown in Fig. 3(a) where a traditional scenario is illustrated with a flowchart. In case a recycled chip is again responding with a fail, it is actually discarded. This system is fair, but it may also ask a device to do double the BI stress while already completed once. Our proposed method tackles this uncertainty by using SoC functional resources to manage and trace the stress execution in a non-volatile memory space; the introduced mechanisms also permit the chip itself to early detect a disconnection.

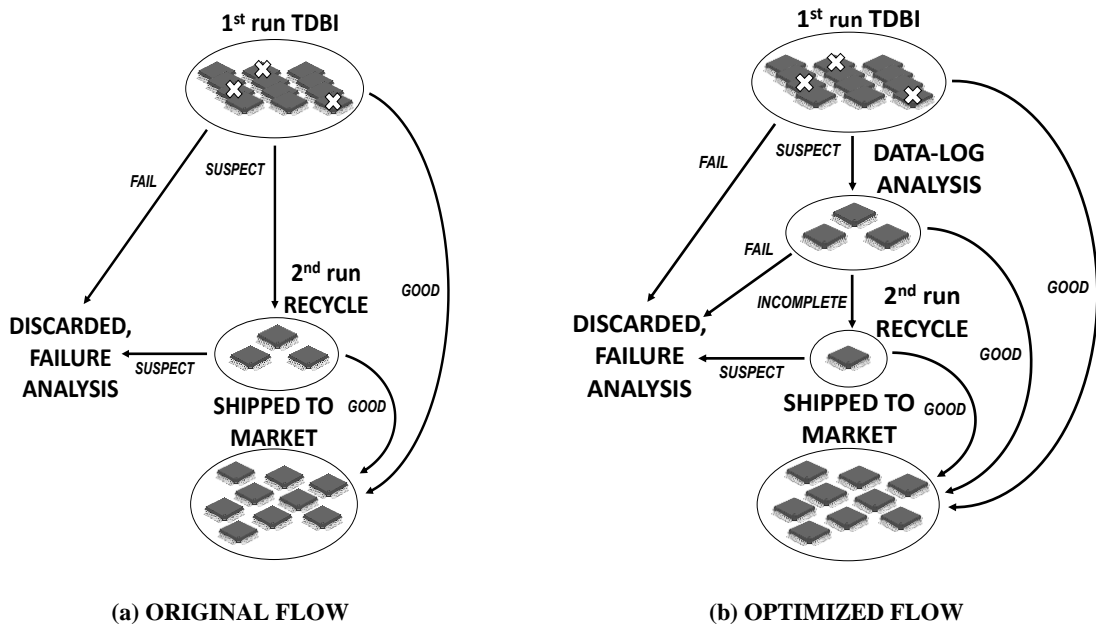


Fig. 3: Fails and possible recycles in TDBI without (a) and with (b) data-log capability.

In our framework, the CPU control has been devised to intervene in case of unexpected malfunctioning of the communication to the tester. It requires:

- Data structures and SW procedures to store data permanently and to support analysis
- Strategies ensuring the detection of communication loss by the DUT.

Concerning data structures, Flash Memory blocks suitable for this purpose are the so-called *shadow flash blocks*. Users in normal functional mode cannot access to erase or program this part of memory because it stores sensitive data such as device passwords, Digital Rights

Management (DRM) keys, calibration constants, etc. For this reason, shadow blocks do not need to undergo the erase cycling process.

If the tester could not read the execution trace during communication issues, such trace is anyway stored in the shadow flash block to be read in a successive connection. The analysis of this trace permits to fully distinguish among real fail, good and incomplete; the last category chips are those where the stress was not entirely performed, and they need to be recycled. The readout operation is done at the end of the BI experiment when, at low temperature, the communication is usually resumed; if not, the device is classified as incomplete and recycled. The proposed modification of the original TDBI flow allows a refined classification of suspect failing chips, as illustrated in Fig. 3(b).

A meaningful selection of the data collected during the TDBI process reduces the number of bytes in the shadow block reserved for the data-log. We propose to save the following TDBI parameters:

- SEAL: indicates whether the device has already run the Burn-In flow at least once;
- TEST FAIL FLAG: indicates whether the device failed at least one of the functional test performed in parallel with the Flash erase phase;
- FAILING TEST SIGNATURE: predetermined location in which, for each test, we eventually store the signature computed at the end of the execution of a failing test;
- INDIVIDUAL TEST COUNT: predetermined location in which, for each test, we store the number of successful test executions, stopping the count when the test fails the first time;
- GLOBAL ERASE COUNT: counter of performed erases;
- GLOBAL TEST COUNT: cumulative count of performed tests, meant to stop counting as soon as the first FAIL occurs;

-
- **COMMUNICATION FAIL FLAG:** indicates whether the device was not able to establish a communication with the ATE within a predetermined amount of time (e.g. a DUT-ATE disconnection occurred).

This set of information related to the TDBI flow execution are not only stored in the shadow blocks but also periodically read out by the tester at regular intervals through the JTAG port. An End Of Test flag is polled by the tester at a specific location in RAM in order to identify a device with available data-log to be transmitted.

Concerning the **COMMUNICATION FAIL FLAG**, it is necessary to implement a technique that allows the DUT to spot the occurrence of a failure in the communication with the tester. This flag manages possible disconnection issues when downloading data-log information. The implementation of the mechanism to set such a flag relies on the Real Time Interrupt (RTI) module, and it can be considered as the mechanism of a watchdog timer. A top initialized count starts in the DUT firmware after the execution of the last test. In case the interrupt occurs, a proper ISR is invoked to set the **COMMUNICATION FAIL FLAG** in the shadow block for future reads.

After data-log retrieval, different sub-cases for suspect failing chips may happen:

- asserted **COMMUNICATIONFAIL FLAG:** the processor capabilities were not completely compromised when the communication was lost
 - asserted **TEST FAIL FLAG:** the chip is not passing all tests and it has to be discarded as a **FAIL**;
 - not asserted **TEST FAIL FLAG** and incomplete BI flow spotted if **GLOBAL ERASE COUNT** is different from the expected number of erases: the chip needs to recycle;

-
- not asserted TEST FAIL FLAG and complete BI flow spotted with GLOBAL ERASE COUNT equal to the expected value: the chip is good and BI flow completed successfully, meaning that the chip can be classified as GOOD;
 - not asserted COMMUNICATION FAIL FLAG:
 - asserted TEST FAIL FLAG: the chip is FAIL
 - not asserted TEST FAIL FLAG: data-log collection mechanisms failed due to unexpected behaviors and the chip needs RECYCLE;

3. An industrial case study

The herein introduced methodology has been applied to a SoC powered by a 32-bit, pipelined, dual-issue microprocessor based on the Power Architecture™ surrounded by several peripheral blocks used in mission mode and testing. The SoC is equipped with 4MB of embedded Flash and 192KB of embedded RAM with its own RAM BIST engine. This SoC is employed in safety-critical automotive embedded systems, such as airbag controllers. It is currently being manufactured by STMicroelectronics and undergoes a BI process. For this mature product, the requirement for Flash quality mandates for 500 erase operations.

The next paragraphs describe how an optimization of the TDBI process allows a significant time reduction and a mitigation of the disconnection problem.

3.1. Optimization of stress procedures

The optimized flow uses the parallelization principles described in the Sections 3.1. Once the erase cycle has started, a set of functional initializations run and start the digital domain stress. More detailed, the SoC performs the following functional sequence:

1. The Erase operation is started;
2. The DMA controller is configured to mimic the behavior of the RAM BIST and CRC compresses the signature;
3. Timers, PWM and some coprocessor (eMIOs) are activated;

4. A stress program runs cyclically from the 4KB Instruction CACHE memory.

The CPU monitors the erase operations and manages the application time of a stress program by means of a scheduling software layer. The management software schedules eight stress programs, which excite different modules inside the processor core. Stress programs need to be applied separately in order to optimize the stress effects to all significant CPU parts. A functional program requires usually around 10 minutes of cyclic execution to reach stable junction temperature. Even very strong functional procedure are not leading to issues related to thermal behaviors, such as the thermal runaway effect [10].

ATPG patterns are applied at higher than nominal voltage and a run of RAM BIST is executed in the reduced Dynamic BI phase for testing purposes.

The original flow (with no parallelization) requires up to 12 hours. Flash erase cycling takes about 7 hours, dynamic stress/test requires 2 hours and 30 minutes and the remaining time is distributed among pretest and parametric stress/tests.

The parallelization of Flash cycling and functional stress reduces the length of the overall flow by 1 hour and 30 minutes, allowing the elimination of the digital domain and RAM stress activities from the dynamic BI part. This means a gain of 12.5% in terms of BI time, which leads to a gain of fab throughput.

3.2. Communication robustness evaluation

The complete data-log, recorded according to the techniques described in 3.2, occupies 132 bytes out of the available 32KB in the shadow flash block. The BI management is performed by an on-chip software that utilizes RTI-based watchdog to identify deadlocks and takes autonomous actions to preserve the flow of information in absence of communication with the tester.

The described strategy was experimented over a population of around two thousand DUTs and the results were compared with the original flow. Table II reports the comparison of the

percentage of suspects in the original flow and the incomplete fails provided in the proposed methodology. The arrows highlight the suspect fail reduction achieved by data-log analysis.

Table II: First run fails analysis.

SEGMENT	PHASE	SUSPECT FAILS ORIGINAL [%]	INCOMPLETE FAILS PROPOSED [%]
FLASH ERASE CYCLING	Stand-Alone Flash Cycling	0.14 %	-
	Flash Cycling + Functional + RAM stress	-	0.26 % → 0 %
DYNAMIC BURN-IN	ATPG stress	0.99 %	0.84 %
	RAM memory BIST	0.11 %	0 %
TOTAL		1.24 %	1.10% → 0.84%

For the sake of a fair analysis, we report that the used population was not showing any real failure, as yield is very high. For all devices disconnecting during the parallelized Flash cycling the communication was resumed afterward, enabling the download of the logged information at the current insertion. Data-log analysis provided the evidence that the BI flow was successfully completed also for disconnecting devices; the 0.26% of such suspect devices of this phase are classified as good and do not need further recycling. Concerning the RAM stress, reducing the BIST to a single execution has abated to 0% the disconnection rate during this phase. Therefore, the final figures for the proposed flow shows a reduction of recycled chips going from 1.24% in the original flow to 0.84% with the proposed methodology, which can be attributed uniquely to the ATPG phase of the dynamic BI part.

4. Conclusions

Test time optimization and techniques aiming at the reduction of the number of recycled chips are producing a significant efficiency benefit for the BI process over the production volume. In the traditional scenario, having a 12 hours BI time and around 1,600 chips as tester parallelism capability, the ideal number of devices stressed per year by a tester is about 1.16 millions of devices. In this set, around 15 thousand chips (e.g., 1.24% of the production) suffer from disconnection issues and need to be recycled, soaking about 4.5 days of tester productivity over a year. With the introduced optimizations, the BI time was reduced by the 12.5%, thus

leading to 10.5 hours per experiment and bringing a volume increase up to 1.33 million of chips processed per year. The adoption of the proposed methods to avoid suspect recycling lowers to 11 thousand the recycles per year (e.g., 0.84%) and reduce to 3 days the tester usage per recycle. In percentage, the gain introduced in terms of tester throughput is the 14.7% and the test time inefficiency is reduced by the 32.2%.

5. References

- [1] A. Birolini, "Reliability Engineering Theory and Practice," Heidelberg: Springer, 2007
- [2] D. Appello, et al. "A Comprehensive Methodology for Stress Procedures Evaluation and Comparison for Burn-In of Automotive SoC," Design Automation and Test in Europe, 2017.
- [3] T.C.Ong, A.Fazio, N.Mielke, S.Pan, N.Righos, G.Atwood and S.JA "Erratic Erase in *ETOXTM* Flash Memory Array" 1993: VLSI Symp. on Tech., 7A-2, pp. 83-82
- [4] C.Dunn, C.Kay, T.Lewis, T.Strauss, J.Schreck, P.Hefley, M.Middendorf, T. San "Flash EPROM Disturb Mechanisms" 1994: IEEE/IRPS Proc., pp. 299-308
- [5] Wook H. Lee, Chan-Kwang Park, and Kinam Kim "Temperature Dependence of Endurance Characteristics in NOR Flash Memory Cells" 2006: IEEE 44th Annual International Reliability Physics Symposium, San Jose.
- [6] N. Aghaee, Z. Peng; P. Eles, "An Efficient Temperature-Gradient Based Burn-In Technique for 3D Stacked ICs," in Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014 , pp.1-4, 24-28 March 2014
- [7] A. Benso et al., "ATPG for Dynamic Burn-In Test in Full-Scan Circuits", IEEE ATS, 2006, pp. 75-82
- [8] S. Bahukudumbi, K. Chakrabarty, R. Kacprowicz, "Test Scheduling for Wafer-Level Test-During-Burn-In of Core-Based SoCs," in Design, Automation and Test in Europe, 2008. DATE '08 , pp.1103-1106, 10-14 March 2008
- [9] M. F. Zakaria, Z. A. Kassim, M. P. L. Ooi and S. Demidenko, "Reducing Burn-In Time Through High-Voltage Stress Test and Weibull Statistical Analysis," in IEEE Design & Test of Computers, vol. 23, no. 2, pp. 88-98, March-April 2006.
- [10] J. Alt et al., "Thermal issues in test: An overview of the significant aspects and industrial practice," 2016 IEEE 34th VLSI Test Symposium (VTS), Las Vegas, NV, 2016, pp. 1-4.

6. Biographies

Daide Appello holds a degree in Electronics Engineering from the Università di Pavia. He is with STMicroelectronics since 1994 where he is concerned with testability and testing and is currently product-engineering director for the automotive digital products. He is active within TTTC and TTEP groups of IEEE.

Paolo Bernardi received his PhD degrees in computer science from Politecnico di Torino, Torino, Italy, in 2006. He is currently an associate professor at the Department of Computer Engineering, Politecnico di Torino. His interests cover the areas of testing of electronic circuits and systems and the design of fault-tolerant electronic systems. He is a member of the IEEE Computer Society and IEEE.

Conrad Bugeja holds a Diploma in Management Studies from the University of Malta. Conrad has joined STMicroelectronics in 2000, where he is a Product Engineering following Production & Engineering activities at Burn-In for ADG.

Riccardo Cantoro received the Ph. D. degree in Computer Engineering from Politecnico di Torino, Torino, Italy in 2017. His research interests include test and stress of microprocessor based systems, and IEEE Std 1687.

Giorgio Pollaccia holds a degree in Electronics Engineering from the Università di Palermo. He is with STMicroelectronics since 2001 where he is concerned with testability and testing and is currently burn-In test engineering for the automotive digital products.

Marco Restifo received the MS degree in Computer Engineering from Politecnico di Torino, Torino, Italy in 2015. His research focuses on microprocessor testing, functional safety and reliability.

Ernesto Sanchez received his Ph. D. degree in computer engineering from Politecnico di Torino, Italy, in 2006. He is currently an associate professor at the Department of Control and

Computer Engineering, Politecnico di Torino. His main research interests include microprocessor testing and evolutionary computation. He is a senior member of the IEEE.

Federico Venini received the MSc degree in electronic engineering from Politecnico di Torino, Torino, Italy in 2015. His research focuses on manufacturing test optimization, microprocessor testing and functional safety for automotive devices.