Low Cost NBTI Degradation Detection and Masking Approaches

Omana, M., Rossi, D., Bosio, N. and Metra, C.

This is a copy of the author's accepted version of a paper subsequently published in *IEEE Transactions on Computers,* 62 (3), pp. 496-509.

It is available online at:

https://dx.doi.org/10.1109/TC.2011.246

Low Cost NBTI Degradation Detection & Masking Approaches

Martin Omaña, Daniele Rossi, Nicolò Bosio, Cecilia Metra

Abstract— Performance degradation of integrated circuits due to aging effects, such as Negative Bias Temperature Instability (NBTI), is becoming a great concern for current and future CMOS technology. In this paper we propose two monitoring and masking approaches that detect late transitions due to NBTI degradation in the combinational part of critical data-paths and guarantee the correctness of the provided output data by adapting the clock frequency. Compared to recently proposed alternative solutions, one of our approaches (denoted as Low Area and Power (LAP) approach) requires lower area overhead and lower, or comparable, power consumption, while exhibiting the same impact on system performance, while the other proposed approach (denoted as High Performance (HP) approach) allows us to reduce the impact on system performance, at the cost of some increase in area and power consumption.

Index Terms — NBTI performance degradation, aging sensor, transition monitoring, aging effect masking.

1 INTRODUCTION

The design of reliable circuits is becoming increasingly challenging with scaled CMOS technologies. Aggressive scaling of oxide thickness has led to large vertical electric fields in MOSFET devices, making oxide breakdown a critical issue. The high field may also lead to significant threshold voltage shift over time induced by Negative-Bias Temperature-Instability (NBTI), thus creating additional uncertainty in the device behavior [1].

NBTI is recognized as the primary parametric failure mechanism in modern ICs [2]. It is characterized by a positive shift in the absolute value of the pMOS transistor threshold voltage, mainly due to the creation of positively charged interface traps, when the transistor is biased in strong inversion [1, 3]. As a consequence, the absolute threshold voltage can increase by more than 50mV over ten years [4], resulting in more than 20% circuit performance degradation [5]. In case of data-paths, such a threshold voltage increase may lead to a late transition of a flip-flop input signal. If such a transition violates the flip-flop set-up and hold time, an incorrect value is sampled and provided as output of the data-path, possibly compromising the system correct operation. Although such a condition may occur for any data-path, it is more likely to take place in case of critical data-paths, which are therefore considered when developing approaches to minimize its occurrence likelihood [6, 7].

A straightforward approach to fulfill this purpose is to increase the clock period by a time interval (usually referred to as guardband [2]) equal to the expected worst case NBTI performance degradation over the chip lifetime [2]. However, this would introduce an excessive time margin since from the beginning of circuit operation, with a consequent high, and unnecessary negative impact on system performance [2]. As an alternative approach, a smaller time guardband could be adopted together with a proper circuit failure prediction scheme that is able to monitor circuit performance degradation throughout the chip lifetime, to then adapt the clock period according to its provided verdict [2, 4]. Compared to the above mentioned worst case scenario, this latter implies a lower impact on system performance.

More in details, the system will start operating with a small guardband equal to the performance degradation expected for the first weeks of system operation (e.g., 2 weeks [4], or 8 weeks [2]), and aging sensors will be deployed at the outputs of properly selected data-paths. Then, should any aging sensor detect the occurrence of a late transition during a pre-defined guardband, the clock period will be increased by a proper time interval in order to avoid that incorrect data are sampled. This approach can be successfully adopted to allow the system to continue working correctly, at the cost of some performance degradation [1, 8, 4]. In this regard, it should be considered that, since electrical stress on transistors can largely vary for different areas of the chip due to the different transistor work load, NBTI performance degradation will not uniformly affect the chip. Therefore, several aging sensors should be deployed throughout the chip, typically at the input of flip-flops of each critical data-path, thus making the low cost of such sensors a relevant issue.

Several aging sensors have been proposed so far to monitor NBTI degradation (e.g., those in [9, 10, 11, 12, 4, 8, 7, 13, 14]). In [10, 11, 12], the sensors are implemented by ring-oscillators allowing to identify performance degradation by monitoring possible changes in their oscillation frequency. They feature high measurement accuracy, but require considerable area overhead. In [9], the effects of NBTI are monitored by sensing a leakage current reduction using I_{DDQ} testing, thus suffering from well known limitations of I_{DDQ} testing for future technologies

- *M. Omaña, D. Rossi and C. Metra are with the University of Bologna, Bologna, Italy. E-mail: {martin.omana; d.rossi; cecilia.metra}@unibo.it.*
- *N. Bosio is with EFI Technology Srl, Bologna, Italy. E-mail: nbosio@gmail.com*

This work was partially supported by the Italian Education, University and Research Ministry under PRIN 2008K4P7X9_004.

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

[15]. In [14], the sensors measure the performance degradation intentionally induced on a pMOS transistor, which is stressed with a predefined stress signal. Then, the data from the sensors are fitted to a Gaussian distribution to predict the maximum threshold voltage variations for all devices of the core. This solution requires low area and power overheads, but it does not measure the actual circuit degradation associated to the actual circuit workload. In [4, 7, 13], it has been proposed to monitor NBTI by detecting, through proper sensors, NBTI-induced late transitions of signals at the outputs of critical data-paths. The sensors proposed in [4, 7] are signal stability checkers, enabled during a proper guardband. The sensor in [13] is a transition detection circuit connected to the inputs of some flip-flops within a datapath, and is designed to detect delay faults during circuit operation. However, if this sensor is enabled by the same control signal as in the sensors in [4, 7], it may be employed to monitor the effects of NBTI. Finally, sensors comparing the speed of a stressed inverter to that of a non-stressed one have been presented in [8], in order to measure the effect of NBTI. Although the sensors in [4, 8, 7, 13] exhibit a power consumption and an area overhead lower than those of previously published sensors, their required area and power may be non negligible when a large amount of such aging sensors have to be deployed throughout the chip.

Based on these considerations, in this paper we propose two novel monitoring and masking approaches that detect late transitions due to NBTI degradation in the combinational part of critical data-paths and guarantee the correctness of the provided output data by adapting the clock frequency. Particularly, our first proposed monitoring and masking approach, denoted as Low Area and Power (LAP), exploits the idea presented in [16] to transform late signal transitions into code/non-codewords for delay and transient fault detection. As introduced in [17], such an approach can be exploited to monitor also performance degradation induced by NBTI. The outputs of the combinational part of critical data-paths are checked during a proper guardband, and an *alarm* message is produced in case of occurrence of late transitions due to NBTI during such a guardband. We will show that our proposed monitoring scheme continues to detect correctly late transitions even if it is itself affected by NBTI degradation. Upon the generation of the *alarm* message, a clock frequency adaptation (reduction) phase is activated, in order to guarantee the correctness of the data produced at the outputs of the monitored critical data-paths, despite the occurrence of NBTI performance degradation. Compared to the previous low cost NBTI monitors in [4, 8, 7, 13], our LAP approach exhibits lower area and lower or comparable power consumption, while implying the same impact on system performance.

Our second proposed monitoring and masking approach, denoted as High Performance (HP), is based on a new, different implementation of NBTI monitors, which allows to overwrite the possibly produced incorrect data at the output of the monitored flip-flops, thus guaranteeing the correctness of the data produced at their outputs. Our HP approach allows us to run the chip at its maximum clock frequency in its early period of life, thus reducing the impact on system performance, at the cost of small increase in area overhead and power consumption, compared to our LAP approach. Therefore, for a given application, the optimal approach between the two proposed ones can be identified based on area and power budget, as well as on impact on performance constraints.

The rest of this paper is organized as follows. In Section 2, we introduce the basic idea behind our proposed monitoring and masking approaches. In Sections 3 and 4, we present our LAP and HP monitoring and masking approaches, respectively. For both approaches, we show some results of the electrical simulations performed to analyze their behavior, and we prove that they continue to detect NBTI degradation even when they are themselves affected by the same aging mechanism. In Section 5, we evaluate the costs of our proposed LAP and HP schemes and compare them to those of alternative solutions. Finally, some conclusions are drawn in Section 6.

2 Considered System Scenario and Basic Idea

Let us consider a generic critical data-path as shown in Fig. 1. The block C_i denotes a combinational circuit, whose worst case propagation delay is denoted by $t_{pd.}$ FF_{i1} and FF_{i2} are the input and output flip-flops (FFs). They present a set-up time equal to t_{set}, and their outputs reach a final stable value after a time t_{pcq} from the CK rising (sampling) edge. For the circuit correct operation, the output of C_i (S_i) must reach its final stable value before the setup time of FF_{i2}. The clock period T_{CK} is given by:

$$T_{CK} = t_{pcq} + t_{pd} + t_{set} + t_{mar}, \quad (1)$$

where t_{mar} represents a time margin to account for possible parameter variations inducing a speed decrease of the data-path.

In case of NBTI degradation, the performance of C_i can degrade over time, resulting in late transitions of its outputs, possibly no longer satisfying the FF_{i2} setup time constraints. Therefore, incorrect data may be sampled by such flip-flops, possibly compromising the system correct operation. To guarantee that such conditions do not occur, despite NBTI degradation, we propose two detecting and masking approaches, both capable of: i) detecting late transitions of S_i during a proper monitoring interval ($T_{M,i}$); ii) enabling a system level masking procedure, based on in-field adaptive clock period increase, to guarantee that only correct data are provided by the output flip-flops of critical data-paths. Our proposed approaches differ in the required area and power costs, as well as in their impact on system performance. They will be hereafter referred to as Low Area and Power (LAP) and High Performance (HP) approaches. They are described in details in the following sections.

3 Proposed Low Area and Power (LAP)

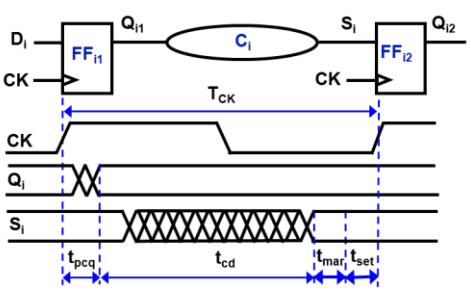

Fig. 1. Representation of the considered data-paths and signals' timing.

APPROACH

The proposed LAP approach consists of two following phases: 1) a monitoring phase, during which the data paths are monitored by a proper scheme capable of detecting late transitions of signals S_i (i = 1..n) due to NBTI and producing an alarm indication if this occurs; 2) a recovery phase, during which adaptive clock period adjustment is activated to mask possible errors. These two phases are implemented as described in the following subsections.

3.1 LAP Monitoring Phase

We monitor the inputs of FFs at the output of critical data paths as represented in Fig. 2. Each signal S_i is monitored during a proper time interval $T_{M,i}$. In order to avoid that a late transition of S_i results in an incorrect sampling, $T_{M,i}$ must be larger than the flip-flop set-up time $t_{set,i}$. Particularly, it is:

$$T_{M,i} = t_{set,i} + \Delta t_{GB,i}\,, \qquad (2)$$

where $\Delta t_{GB,i}$ is a guardband to be chosen based on estimation of the NBTI degradation in the first 2 weeks [4], or 8 weeks [2] of chip lifetime. For simplicity, we assume that all flip-flops have the same set-up time t_{set}. Similarly, we consider the same guardband, denoted as Δt_{GB}, for all monitored S_i. Thus, also the monitoring interval $T_{M,i}$ is the same for all flip-flops, and will be denoted by T_M. To enable our monitoring scheme only during the predetermined monitoring interval T_M, we generate a time window control signal (*TWC)* which is asserted only during T_M. This yields our monitoring scheme to be immune to signal glitching originated in logic blocks. In fact, as shown in Fig. 2, T_M is a time interval which, by design, starts after the time margin t_{mar} that is included in the clock cycle T_{CK} to allow that critical timing paths of logic blocks can reach, even in presence of process parameter variations, their final stable value before the setup time of the sampling flip-flops.

Figure 3 shows the internal block structure of the proposed LAP monitoring scheme.

Each monitored signal S_i (i = 1..n) is connected to a *Transition Detector*, giving on its outputs O_{i1}, O_{i2} (i = 1..n) an alarm indication in case of S_i late transitions. Namely, if a late transition of S_i occurs during T_M, the respective detector produces a non two-rail codeword ($O_{i1}O_{i2}$= 00 or *11*), while it produces a two-rail codeword ($O_{1i}O_{2i}$= 10 or *01*) otherwise. The outputs

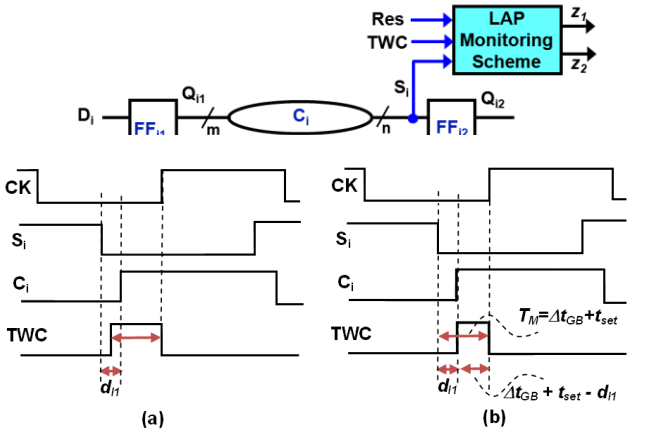

Fig. 5. (a) Licit S_i transition producing a wrong alarm indication. (b) TWC identification to produce alarm indications only in case of illicit S_i transitions.

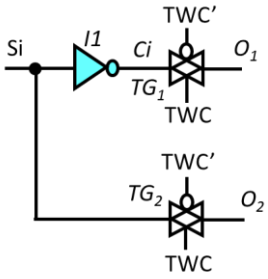

Fig. 4. Internal structure of the transition detector of our LAP monitoring scheme.

produced by *n* transition detectors are gathered by the Error Indicator (EI) block. This implies an area overhead also due to the required routing, similarly to the alternative solutions in [4, 8, 7, 13]. If at least one of the detectors generates an *alarm* message ($O_{i1}O_{i2}$ = 00 or 11), EI will produce an *alarm* indication on its outputs Z_1 and Z_2 (i.e., Z_1Z_2 = 00 or 11) which will be maintained till the assertion of the reset signal *Res*. The number *n* of detectors distributed through the chip is given by the number of critical data-paths in the chip, as shown in [6]. However, such a number is generally limited by the chip available area.

3.1.1 Transition Detector

The internal structure of our proposed transition detector is shown in Fig. 4. Starting from the monitored signal S_i, by means of a proper inverting delay block (e.g., a simple inverter), the detector generates an additional signal C_i that, together with S_i, provides a word belonging to the two-rail code (O_1O_2 = 10 or 01), if S_i is stable while TWC = 1, while it gives a two-rail non codeword (O_1O_2 = 00 or 11), if late transitions of S_i occur when TWC=1.

Instead, when TWC = 0, the switching of S_i is allowed, and the transfer gates disconnect C_i and S_i from the detector outputs O_1 and O_2, respectively, which maintain the previous indication.

Let us analyze in more details the behavior of our transition detector when TWC = 1. During this time interval (T_M), the transfer gates are conductive, and O_1 and O_2 are connected to C_i and S_i, respectively. In this case, two conditions may occur, depending on whether the monitored signal is stable, or presents a late transition. In the first case, the logic values (01) or (10), which are present at the inputs of the transfer gates after the latter legal signal transition (occurring when TWC=0), are given to (O_1O_2) when TWC switches from 0 to 1. Instead, in the second case, a (00) or (11) configuration is produced on (O_1O_2) for a time interval equal to the input-output delay d_{I1} of the in-

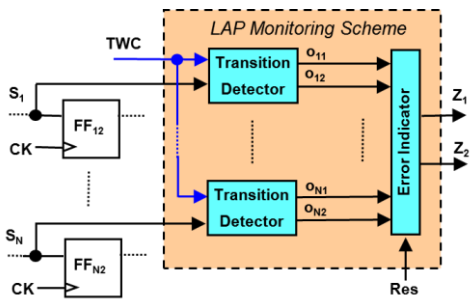

Fig. 3. LAP monitoring scheme internal block structure.

verter I1. The duration of the interval d_{I1} can be adjusted to the required value by properly sizing the inverter I1, in order to make it long enough to allow the alarm indication to propagate up to the outputs Z_1Z_2 of the error indicator in Fig. 3. Therefore, assuming the presence of a (01) or (10) on (O_1O_2) as the occurrence of a *no alarm* message, and that of a (00) or (11) as an *alarm* indication, this circuit can be used to detect on-line late transitions of S_i due to NBTI, to inhibit the sampling of incorrect data by the output flip-flop FF_{i2}.

It should be noticed that, due to the delay of the inverter I1 (Fig. 4), the effective monitoring time interval is slightly larger than the time interval during which TWC=1. In fact, as depicted in Fig. 5(a), if signal S_i licitly switches before the rising edge of TWC by a time interval lower than d_{I1}, then signal C_i switches while TWC=1. In this case, since S_i and C_i present the same logic value (i.e., $S_i = C_i = 0$) while TWC=1, an incorrect *alarm* indication is generated by the detector.

To avoid this misbehavior, the delay of the inverter I1 should be taken into account when sizing the time interval during which TWC = 1. Once the monitoring interval $T_M = \Delta t_{GB} + t_{set}$ has been chosen, the time interval during which the signal TWC has to be asserted should be such that $T_M - d_{I1} = \Delta t_1 + t_{set} - d_{I1}$. This way, as easily derived from Fig. 5(b), only illicit S_i transitions occurring during T_M will make S_i and C_i produce an *alarm* indication. As an example, TWC can be generated by utilizing the circuit proposed in [7] that is able to generate a pulse with a programmable width.

3.1.2. Error Indicator

The error indicator block (EI) consists of an *n+1* variable two-rail code checker, which can be implemented by means of the low-cost high-speed two-rail code checker presented in [18]. It gathers the *alarm* indications $O_{i1}O_{i2}$ ($i=1..n$) produced by *n* transition detectors deployed throughout the chip, as shown in Fig. 6. The outputs of EI (Z_1 and Z_2) are feedbacked to its inputs. This way, if at least one of the transition detectors generates an *alarm* message (i.e., $O_{i1}O_{i2}$ = 00 or 11), the error indicator produces an *alarm* indication Z_1Z_2 = 00 or 11, which is maintained till the assertion of the reset signal *Res*. This signal is generated at the system level and is asserted after the activation of the clock adjustment procedure, as will be described in the following subsection. It is worth noticing that the nMOS (pMOS) transistor driven by the *Res* (*Res'*) signal must be dominant over the pull-up (pull-down) network driving the output of the *(n+1)*-variable TRC.

3.2 LAP Recovery Phase

As introduced in the Section 2, upon the detection of a late S_i transition and the generation of an alarm message at the output of EI, a masking procedure based on in-field clock adjustment is activated to guarantee the sampling of correct values by the data-path output flip-flops. Let us describe in details the proposed masking procedure.

As previously discussed, our LAP monitoring scheme initially detects late transitions of S_i occurring during a monitoring time interval $T_M = t_{set} + \Delta t_G$,. Once the value of Δt_{GB} is selected (as described in Subsection 3.1), the period $T_{CK\text{-}I}$ of the clock at which the monitored circuit runs at the beginning of the chip operation (Fig. 7(a)) is chosen according to the following expression:

$$T_{CK\text{-}I} = \Delta t_{GB} + t_{pcq} + t_{pd} + t_{set} + t_{mar}. \quad (3)$$

If no late transition of S_i occurs, no alarm indication is produced by our LAP monitoring scheme, and no masking procedure needs to be activated. Instead, once S_i presents its first transition within T_M (time instant denoted as t_{AL} in Fig. 7(b)), our LAP scheme produces an alarm message, upon which no masking procedure needs to be immediately activated. In fact, after t_{AL}, the performance of the combinational block C_i will continue to degrade due to NBTI. However, as long as the delayed transitions fall within Δt_{GB}, the flip-flop FF_{i2} (Fig. 2) will continue to sample the correct value of S_i, so that the system will keep on working properly. We can estimate the time interval (t_{INC}) from the occurrence of the alarm message (denoted by t_{AL}) during which a delayed transition of S_i will fall within the guardband Δt_{GB} by employing the model in [6].

Based on such an estimation, the time instant $t=t_{AL}+t_{INC}$ (Fig. 7(c)) at which the proposed system level masking procedure should be activated can be derived. Such a masking procedure consists of changing the clock period from $T_{CK\text{-}I}$ to $T_{CK\text{-}II}$, where $T_{CK\text{-}II}$ is:

$$T_{CK\text{-}II} = 2\Delta t_{GB} + t_{pcq} + t_{pd} + t_{set} + t_{mar} = T_{CK\text{-}I} + \Delta t_{GB} \quad (4)$$

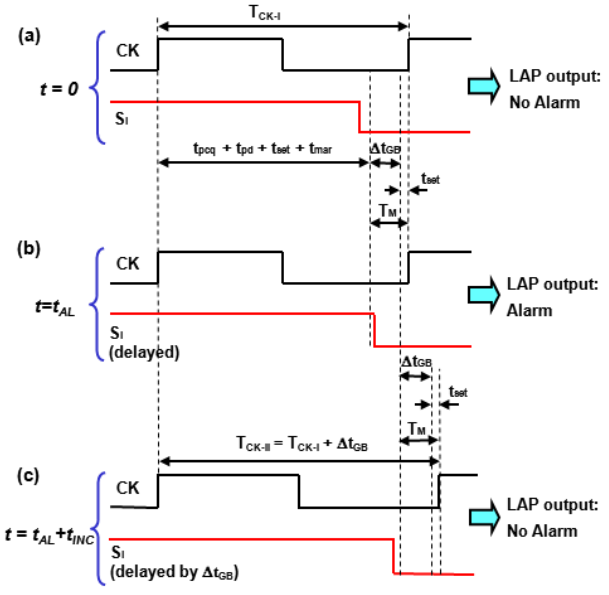

Fig. 7. Representation of the masking procedure of our LAP approach.

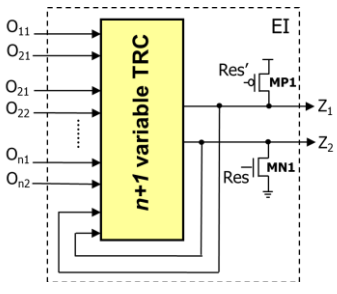

Fig. 6. Possible error indicator used in our LAP monitoring scheme to maintain till reset the alarm indications of the transition detectors.

After increasing the clock period, a reset signal *Res* is activated to restore a *no alarm* indication at the output of the EI block (Fig. 6).

The procedure described above is iterated each time a successive alarm indication is received, for a maximum of *M* times. After the receipt of the *i-th* alarm indication, the $T_{CK\text{-}i}$ is:

$$T_{CK\text{-}i} = (i+1)\Delta t_{GB} + t_{pcq} + t_{pd} + t_{set} + t_{mar} = T_{CK\text{-}(i\text{-}1)} + \Delta t_{GB} \quad (5)$$

while the value of *M* is given by:

$$M = \left\lceil \frac{pd - \Delta t_{GB}}{\Delta t_{GB}} \right\rceil, \quad (6)$$

where *pd* represents the circuit performance degradation after its whole life time (here considered to be 10 years) estimated by the model in [6]. It is worth noting that, if after the *i-th* increment of the clock period, it is $T_{CK\text{-}i} > T_{CK\text{-}worst}$ (where $T_{CK\text{-}worst}$ is the clock period guaranteeing the system correct operation in case of worst case NBTI degradation during the whole circuit life time), our masking procedure sets $T_{CK\text{-}i} = T_{CK\text{-}worst}$ in order to avoid unnecessary performance loss.

3.3 LAP Monitoring Scheme Implementation and Verification

We have implemented our proposed LAP monitoring scheme in Fig. 3, with the transition detector circuit in Fig. 4, and the EI scheme in Fig. 6**.** We have considered the 45nm CMOS technology by PTM [19], with V_{dd} = 1V and clock frequency of 3GHz. All nMOS transistors have been sized with a shape factor *(W/L)* = 1, while all pMOS transistors have *(W/L)* = 2. As for the flip-flops FF_{i1} and FF_{i2} (Fig. 11), they have been implemented as a cascade of two minimum sized standard latches in a master-slave fashion. For these flip-flops, we obtained t_{set}=15ps. The guardband Δt_{GB} has been chosen equal to Δt_{GB}=*30ps*. As an example, we have considered the case of 32 transition monitors.

The behavior of our proposed LAP monitoring scheme (including the pulse generation circuit) has been verified by means of Monte Carlo electrical simulations, performed considering PVT variations (with uniform distribution) up to 20%. Fig. 8 shows the obtained simulation results. As an example, we have simulated the case in which the outputs O_{51} and O_{52} of the transition detector #5 present an indication of late transition, while all other 31 detectors provide no late transition indications. Particularly, two situations are represented: no late transition when TWC = 1 at time t_1; a late transition due to NBTI occurring at time t_2, while TWC=1. In the first case, EI provides a no alarm indication (Z_1Z_2 = 01), while, in the second case, the outputs of EI present the alarm indication Z_1Z_2 = 00 till the activation of *Res* at time t_3.

3.4 Robustness to NBTI Effects

The NBTI phenomenon described before may degrade also the performance of our proposed LAP monitoring scheme. In this section, we prove that, similarly to the sensors in [4, 8, 7], our LAP approach keeps on detecting correctly late transitions of the signal S_i, even though it is itself affected by NBTI degradation.

We evaluated the increase of the absolute value of the pMOS threshold voltage (ΔV_{th}) due to NBTI degradation by utilizing the model presented in [6]. Such a model allows us to estimate the voltage shift ΔV_{th} due to NBTI after a given period of time Δt of chip operation (or chip lifetime). Besides the value of the electric field and the junction temperature, the value of ΔV_{th} depends on the parameter $\alpha = t_{on}/\Delta t$ [6], where t_{on} is the total time in which the considered pMOS is under a stress condition (i.e., conductive). It is $0 \leq \alpha \leq 1$, where α=0 if the considered pMOS transistor is always off, while α=1 if it is always on.

As described in [6], for specific and constant environmental conditions, such as operating voltage and temperature, ΔV_{th} can be accurately expressed as a function of the time Δt of chip operation, and of the parameter α introduced above. The value of α for each pMOS transistor composing our scheme has been estimated as follows:

1) The pMOS transistors of transfer gates TG_1 and TG_2 in Fig. 4 are conductive only during the monitoring interval T_M of each clock cycle, so that for these transistors it is $\alpha = T_M/T_{CK} = 45ps/333ps = 0.13$.
2) The pMOS transistors composing the circuit generating TWC are conductive for half of the clock cycle, so that for

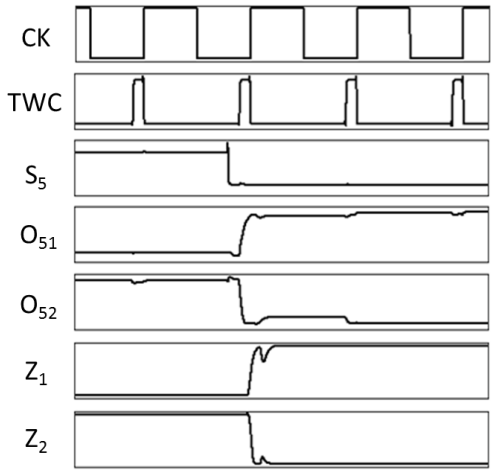

Fig. 9. Simulation results obtained in case of no late transition of S_5 occurring during T_M and a 10 year NBTI performance degradation of our LAP.

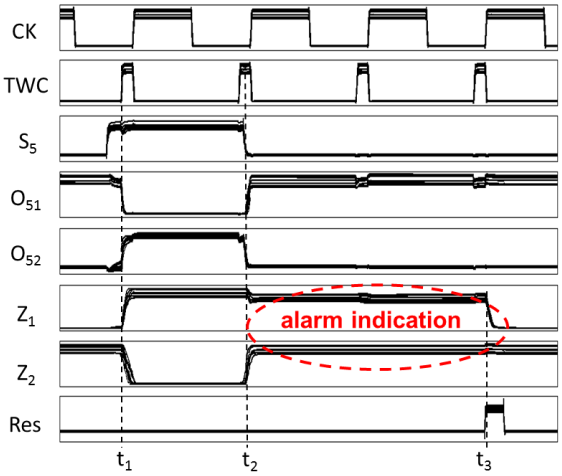

Fig. 8. Monte-Carlo simulation results obtained for the case of PVT variations up to 20% and for signal S_5 not presenting (at t_1) and presenting (at t_2) a late transition during T_M.

these transistors it is $\alpha=T_{CK}/2T_{CK}=1/2$.

3) The pMOS transistors of the inverter I1 in Fig. 4(a) and of the error indicator in Fig. 6 are conductive for a time period that depends on the input statistics. Considering a signal S_i switching activity equal to 50%, we obtain $\alpha=t_{on}/\Delta t=0.5$.

The respective threshold voltage shifts of the pMOS transistors have been estimated by means of the model in [6] considering $\Delta t=10$ *years* and the values of α estimated in cases 1), 2) and 3) above. The derived voltage shifts ΔV_{th1}, ΔV_{th2} and ΔV_{th3} have been utilized to build three different device models, allowing us to simulate each pMOS transistor of our proposed LAP monitoring scheme with the proper NBTI degradation. Apart from these customized device models, the simulation set up is the same as that described in the previous subsection.

Fig. 9 shows the simulation results obtained in case of no transition of S_5 occurring during T_M. We can observe that no *alarm* indication is generated at the outputs of the error indicator (Z_1 and Z_2), thus verifying the correct operation of our monitor, despite its NBTI degradation.

Similarly, Fig. 10 shows the simulation results obtained in case of late transitions of S_5, that is in case of transitions occurring during the monitoring time interval T_M. We can observe that, also in this case, the monitoring scheme behaves properly: an *alarm* indication is generated at the outputs of EI (Z_1 and Z_2), which is maintained till the assertion of the reset signal *Res*. Thus, we can state that our degraded monitors keep on detecting correctly late transitions due to NBTI of the monitored signals.

From a qualitative point of view, the robustness of our LAP monitoring scheme to the aging effects of NBTI can be explained by analyzing the internal structure of the transition detector in Fig. 4. During circuit operation, NBTI degrades the pMOS composing: i) the transfer gates TG1 and TG2; ii) the inverter I1; iii) the circuit generating TWC.

As for the pMOS transistors in i), they are each connected in parallel with an nMOS transistor that does not exhibit any NBTI degradation. Therefore, TG_1 and TG_2 continue to work properly even in case of NBTI, and consequently our LAP monitoring scheme is not affected by the NBTI degradation of the pMOS in i).

As for the pMOS in ii), its degradation causes an 8.9% increase in the propagation delay of the $0\rightarrow1$ transitions of inverter I1. This produces only a longer duration of the alarm indication on O_1O_2=00 if an illegal S_i $1\rightarrow0$ transition occurs while TWC=1. Therefore, also in this case, the correct operation of our LAP monitoring scheme is not compromised.

Finally, as for pMOS transistors in iii), their NBTI degradation do not affect the operation of the circuit generating TWC, since it is resilient by design to the aging effects produced by NBTI [7].

4 Proposed High Performance (HP) Approach

An alternative solution to the LAP approach described in the previous section is here introduced. It is denoted as High Performance (HP) approach, since it exhibits a lower impact on system performance compared to our LAP scheme and to other aging sensors proposed in literature [4, 5, 7, 13]. This is achieved at the cost of an increase in area overhead and power consumption.

Analogously to the LAP approach, our HP approach consists of two successive phases, namely a monitoring phase followed by a recovery phase. However, differently from the LAP approach, during the monitoring phase, the HP monitoring scheme is also able to correct possible errors at the output of the monitored datapaths, thus allowing to reduce the impact on performance of the following recovery procedure, as described in details in the next subsections.

4.1 HP Monitoring Phase

Our proposed HP monitoring scheme is inserted in the monitored data-paths as shown in Fig. 11. Analogously to the LAP scheme, the HP scheme checks the signals S_i (i= 1..n) for possible transitions during a proper monitoring time interval denoted by $T_{M\text{-}COR}$, which is determined by the time window control signal TWC. However, the monitoring time interval $T_{M\text{-}COR}$ is narrower than that (T_M) used for our LAP scheme. Particularly, it is $T_{M\text{-}COR}=t_{set}$, where t_{set} is the set-up time of the flip-flop FF_{i2} (assuming for simplicity $t_{seti} = t_{set}$ $\forall$i = 1..n).

If a delayed transition of S_i occurs during $T_{M\text{-}COR}$, the flip-flop FF_{i2} will likely sample an incorrect value. Therefore, differently from the LAP approach, our HP scheme must be able to correct the logic value of the output Q_{i2} provided by FF_{i2}. This is achieved by providing the correct logic value on signal Q_{iC} (i = 1..n), which is then short-circuited to the respective Q_{i2} signal, as shown in Fig. 11. Our HP scheme must be able to force the correct logic value on $Q_{iC} \equiv Q_{i2}$ till the next sampling instant of FF_{i2}, that is for a time interval equal to (Fig. 11):

$$t_{force} = T_{CK} - t_{set} = T_{CK} - T_{M\text{-}COR}, \quad (7)$$

which corresponds to the time interval during which TWC=0.

The internal block structure of the proposed HP monitoring scheme is shown in Fig 12. For each monitored signal S_i (i = 1..n), the HP monitoring scheme consists of two component blocks: a *transition detector*, and a *correction block*.

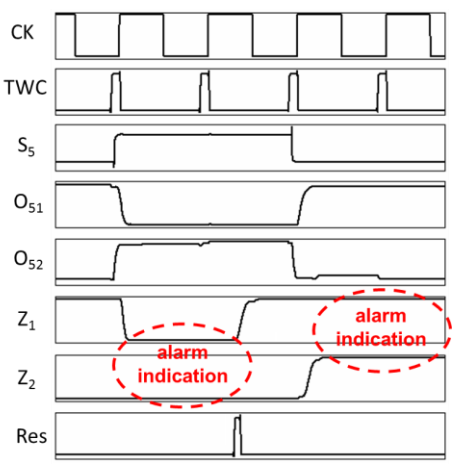

Fig. 10. Simulation results obtained in case of late transitions of S_5 occurring during T_M and a 10 year NBTI performance degradation of our LAP monitoring scheme.

Each *transition detector* checks S_i during the monitoring interval $T_{M\text{-}COR}$ and produces two-rail encoded outputs ($O_{i1}O_{i2}$ = 01, or 10, i = 1..n) if no transition of S_i occurs during $T_{M\text{-}COR}$, and a non two-rail codeword ($O_{i1}O_{i2}$= 00, or 11, i = 1..n), considered as an *alarm* indication, if S_i switches during $T_{M\text{-}COR}$. Each signal couple $O_{i1}O_{i2}$ is given as input to the respective *correction block* (Fig. 12) producing signal Q_{iC} as output. Q_{iC} is then short-circuited to the output Q_{i2} of the flip-flop FF_{i2}.

In case of no late transition at the input S_i of FF_{i2}, the voltage value provided by FF_{i2} on output Q_{i2} is correct and must not be altered by the correction block. This can be obtained by leaving the node Q_{iC} in a high impedance state, thus acting only as an extra load connected to Q_{i2}.

Instead, in case of a S_i late transition, the voltage value sampled by FF_{i2} and provided as output on node Q_{i2} is incorrect. The correction block should give on node Q_{iC} the (correct) logic value that FF_{i2} should have sampled in case of a late transition of S_i. Since the node Q_{iC} is short-circuited to the FF_{i2} output node Q_{i2}, an electrical conflict may originate. In this case, in order to perform correction properly, the correction block should win the electrical conflict and force the node Q_{i2} to the correct value.

Analogously to the LAP case, the outputs produced by the *n* transition detectors are gathered by an error indicator EI, which produces an alarm indication on its outputs Z_1 and Z_2 (i.e., Z_1 Z_2 = 00, or 11) in case of a late transition of at least one signal S_i. The alarm is maintained active till the assertion of the reset signal *Res*.

We can notice that our HP scheme requires an initial monitoring interval $T_{M\text{-}COR} = t_{set}$ which, differently from that required by the LAP monitoring scheme (i.e., $T_M = t_{set} + \Delta t_{GB}$) does not impact system performance. However, since the combinational block C_i will progressively degrade in time due to NBTI, transitions on S_i will be increasingly delayed. Eventually, transitions on S_i will become delayed by more than $T_{M\text{-}COR} = t_{set}$, and our scheme will be no longer able to perform correction. To avoid this to occur, upon the detection of a late S_i transition and the generation of an *alarm* at the output of EI, a masking procedure based on in-field clock adjustment could be activated to guarantee the generation of correct values at the outputs of the data-path output flip-flops.

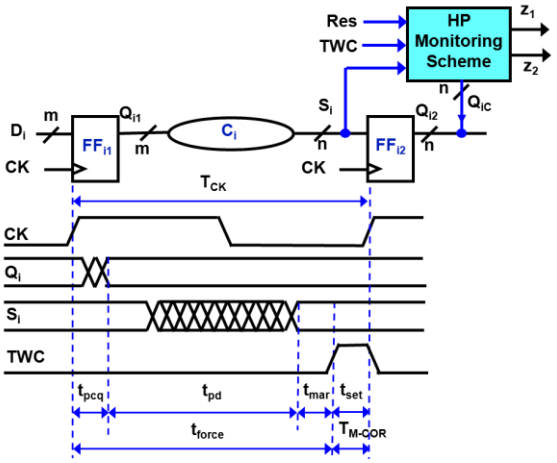

Fig. 11. HP monitoring scheme insertion within the considered critical data-paths.

4.1.1. Transition Detector and Error Indicator

The internal structure of our *transition detector* is shown in Fig. 13(a). It is similar to that of our monitoring circuit presented in Section 2 (Fig. 4), but for two main differences: i) the control signal TWC is now equal to 1 only during $T_{M\text{-}COR} = t_{set}$ (Fig. 13(b)), as discussed in the previous subsection; ii) the delay element D1 now presents an inverting delay d_1 that should be equal to $T_{M\text{-}COR} = t_{set}$. To satisfy condition i), we can generate the signal TWC by means of the circuit in [7] that provides a pulse of programmable width. This way, after determining the variation in the nominal value of t_{set} due to process parameter variations (which, as an eample, may be derived from the measurements performed by ring oscillators of the kind in [20], that are usually integrated on-die to measure parameter variations), the TWC signal generator can be calibrated in order to make TWC=1 only during the actual t_{set} of the flip-flops of fabricated chips.

Apart from the two abovementioned differences, the behavior of the transition detector is analogous to that described in Subsect. 3.1.1: it provides a two-rail codeword on O_1 and O_2 only if S_i is stable during $T_{M\text{-}COR}$ (that is, while TWC = 1), while it produces a two-rail non-codeword if S_i switches during $T_{M\text{-}COR}$ (Fig. 13(a)). The two-rail non-codeword on O_1 and O_2 is maintained till the following clock period, when TWC switches again to 1 and the transfer gates become conductive. Particularly, since it is $d_1 = t_{set}$, signals O_1 and O_2 give a two-rail non-codeword for a time interval equal to:

$$T_{CK} - t_{set} = T_{CK} - T_{M\text{-}COR} = t_{force}, \quad (8)$$

which corresponds to the time interval t_{force} required by the correction block to force the output Q_{i2} at the correct logic values (Fig. 11).

It is worth noticing that, since D1 is an inverting delay, in case of late transitions on S_i, it is $(O_1O_2) = (11)$ if S_i switches from 0 to 1, and $(O_1O_2) = (00)$, if S_i switches from 1 to 0. Therefore, the logic value present on both O_1O_2 is the same as the one the flip-flop FF_{i2} would have produced in case of correct sampling of S_i. This property will be exploited by the correction block to perform correction. Finally, as for EI, the same structure and implementation as described in Subsection 3.1.2 has been considered.

4.1.2. Correction Block

The *correction block* must force the logic value at the output Q_{i2} of flip-flop FF_{i2} (Fig. 11) for a time interval equal to t_{force} =

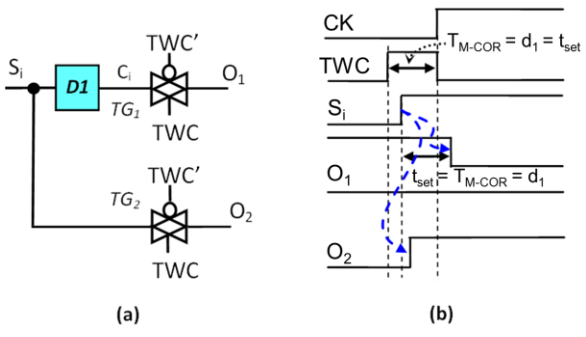

Fig. 13. (a) Internal structure of the transition detector; (b) example of the timing of its signals in case of a transition of S_i while TWC=1.

$T_{CK} - t_{set}$, during which it is TWC=0. A possible implementation is shown in Fig. 14. It consists of a C-element that receives as inputs the signals O_{1R} and O_{2R} (which are the inverted versions of O_1 and O_2), and produces Q_C as output, which is connected to Q_{i2} (Fig. 12). In order to work properly, our *correction block* requires, as unique design constraint, that the series of pMOS and nMOS of the C-element are dominant over the transistors of FF_{i2} driving the output Q_{i2}.

The transfer gates TG_3 and TG_4 connect the signals O_1 and O_2 from the *transition detector* to the inputs of the C-element O_{1R} and O_{2R}, respectively, only when O_1 and O_2 present a stable value. In fact, TG_3 and TG_4 are off during the monitoring time interval $T_{M\text{-}COR}=t_{set}$ (TWC=1), during which O_1 and O_2 may change, while they switch on (thus connecting O_1 and O_2 to O_{1R} and O_{2R}, respectively) during the time interval $t_{force} = T_{CK} - t_{set}$ (TWC=0), during which O_1 and O_2 are stable.

As for the output C-element, it is conductive only if its two inputs (O_{1R} and O_{2R} in Fig. 14) present the same logic value. Otherwise the output of the C-element (Q_C) is left in a high impedance state. This way, during the time interval t_{force}, the output Q_C of the C-element is left in a high impedance state if O_1 and O_2 are two-rail encoded, while it is $Q_C = O_1 = O_2$ if O_1 and O_2 are non two rail encoded because of a late transition of S_i. In the former case, no correction is needed, and the logic value at Q_{i2} is imposed by FF_{i2}, while in the latter case, the C-element forces the output Q_{i2} of FF_{i2} to assume the correct logic value present on Q_C.

4.2 HP Recovery Phase

Similarly to our LAP approach, upon the detection of a late S_i transition and the generation of an alarm message at the outputs Z_1Z_2 of the error indicator EI, a masking procedure based on in-field clock adjustment is activated, in order to guarantee the generation of correct values at the outputs of the monitored data-paths. Let us describe in details the proposed recovery procedure for the HP approach.

As previously discussed, our HP monitoring scheme initially detects late transitions of S_i occurring during a monitoring time interval $T_{M\text{-}COR} = t_{set}$. Thus, as shown in Fig. 15(a), the initial clock period $T_{CK\text{-}I}$ is given by:

$$T_{CK\text{-}I} = t_{pcq} + t_{pd} + t_{set} + t_{mar}. \quad (9)$$

In the represented case, no late transition of S_i occurs, and no alarm indication is produced by our HP monitoring scheme. Therefore, no masking procedure is activated. By comparing Eqs. (9) and (3), we can observe that our HP approach allows us to reduce by Δt_{GB} the value of the initial $T_{CK\text{-}I}$ with respect to our LAP approach.

Fig. 15(b) represents the time instant, denoted by t_{AL}, in which S_i presents the first transition within $T_{M\text{-}COR}$ due to NBTI degradation. At this time instant, our HP scheme produces an alarm message on Z_1Z_2, that will be used to activate a masking procedure at system level. In particular, after t_{AL} the performance of the combinational block C_i will continue to degrade in time due to NBTI, but as long as the delayed transitions fall within $T_{M\text{-}COR}$, our approach continues to force the output of flip-flop FF_{i2} (Fig. 12) to assume the correct logic value, so that the system will keep on working properly.

By means of the model in [6], we estimate the time interval t_{INC} from the alarm message occurrence (at t_{AL}), during which a delayed transition of signal S_i falls within t_{set}. Then, as shown in Fig. 15(c), at time $t=t_{AL}+t_{INC}$ our approach activates the masking procedure (at the system level) that commutes the clock period from $T_{CK\text{-}I}$ to $T_{CK\text{-}II}$, where $T_{CK\text{-}II} = T_{CK\text{-}I} + \Delta t_{GB}$. The guardband Δt_{GB} can be chosen using the same criteria as for our LAP approach to preserve the FF_{i2} correct sampling. The clock period $T_{CK\text{-}II}$ is therefore:

$$T_{CK\text{-}II} = \Delta t_{GB} + t_{pcq} + t_{pd} + t_{set} + t_{mar}. \quad (10)$$

After the commutation to clock period $T_{CK\text{-}II}$, a reset signal *Res* is activated to restore a *no alarm* indication at the output of EI. Again, by comparing Eqs. (10) and (4), we can observe that our HP approach allows a reduction of Δt_{GB} on the value of the increased $T_{CK\text{-}II}$, with respect to our LAP approach.

Similarly to our LAP approach, the recovery procedure of our HP approach described above is repeated each time a successive alarm indication is generated. Thus, after the generation of the *i-th* alarm indication, $T_{CK\text{-}i}$ will be:

$$T_{CK\text{-}i} = i\Delta t_{GB} + t_{pcq} + t_{pd} + t_{set} + t_{mar} = T_{CK\text{-}(i\text{-}1)} + \Delta t_{GB}. \quad (11)$$

By comparing Eqs. (11) and (5), we can observe that also after the *i-th* increment of the clock cycle, our HP approach allows a reduction of Δt_{GB} on the value of the increased $T_{CK\text{-}i}$, with respect to our LAP approach.

The recovery procedure of our HP approach is repeated till a maximum of *M* times given by:

$$M = \left\lceil \frac{pd - t_{set}}{\Delta t_{GB}} \right\rceil, \quad (12)$$

where the symbols are the same as those in Eq. 6. If after the *i-th* increment of clock period, it is $T_{CK\text{-}i} > T_{CK\text{-}worst}$ (where $T_{CK\text{-}worst}$ is the clock period guaranteeing the system correct operation in case of worst case NBTI degradation during the whole circuit life time), our masking procedure sets $T_{CK\text{-}i} = T_{CK\text{-}worst}$ in order to avoid unnecessary performance degradations.

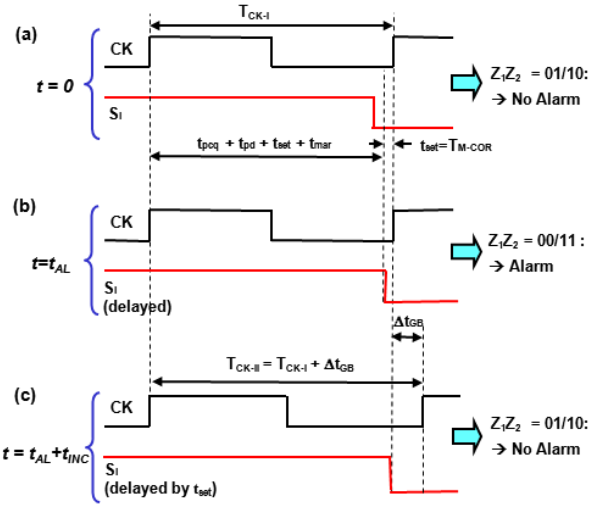

Fig. 15. Schematic representation of the proposed masking procedure of our HP approach based on in-field clock adjustment.

4.3 HP Monitoring Scheme Implementation and Verification

We implemented our proposed HP monitoring scheme by means of the same 45nm CMOS technology as for the LAP scheme, with V_{dd} = 1V and clock frequency of 3GHz. In particular, we designed the *transition detector* shown in Fig. 13(a), the *correction* block represented in Fig. 14, and the pulse generation circuit proposed in [7] to generate the TWC signal, considering the following transistor aspect ratios: (i) *(W/L) = 1 (W/L = 2)*, for the nMOS (pMOS) transistors of TG_1, TG_2, TG_3, TG_4, IR_1 and IR_2; (ii) *(W/L) = 4 (W/L=10)* for the nMOS (pMOS) transistors of the C-element. As for the flip-flops FF_{i1} and FF_{i2} (Fig. 11), they have been implemented in a master-slave fashion (as described in Subsection 3.3) and feature a set-up time t_{set} = 15ps. As for the delay element D1 of the *transition monitor*, it has been implemented by means of a programmable delay element of the kind in [21], in order to set its delay *d1* equal to t_{set} of FF_{i2}.

The behavior of our HP monitoring scheme has been verified by means of Monte Carlo electrical simulations, performed considering PVT variations (with uniform distribution) up to 20%. Fig. 16 shows the simulation results obtained for the case of no late transition of S_i. We can observe that, as expected, the *transition detector* always gives two-railed encoded outputs O_1 and O_2. In this case, the output of our correction block is left in a high impedance state and the output Q_{i2} is driven to the correct logic value by flip-flop FF_{i2}.

Instead, Fig. 17 depicts the results of the Monte Carlo simulations regarding to the case of late S_i transitions occurring at time instants t_1 and t_4 (while TWC=1). As expected, the *transition detector* produces equal logic values on O_1 and O_2, which are maintained till TWC=0 (instant t_6). As for the output signal Q_{i2}, the figure reports the waveforms obtained with our proposed HP monitoring scheme (solid line), and without it (dashed line). It can be seen that, without any correction (dashed line), late transitions of S_i due to NBTI occurring at times t_1 and t_4 are incorrectly sampled by FF_{i2}, and wrong logic values are produced on Q_{i2} during the time intervals $\Delta t_I = t_3 - t_2$ and $\Delta t_{II} = t_7 - t_5$.

When our HP monitoring scheme is employed (solid line), we can observe that Q_{i2} presents the correct logic vales during

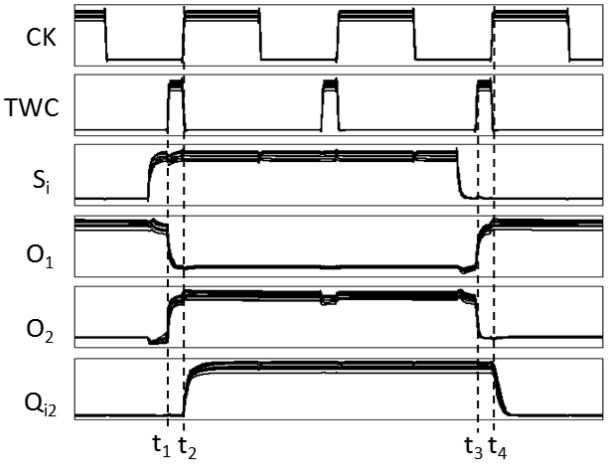

Fig. 16. Monte Carlo simulation results obtained for the case of PVT variations up to 20% and no late transition of S_i occurring during $T_{M\text{-}COR}$.

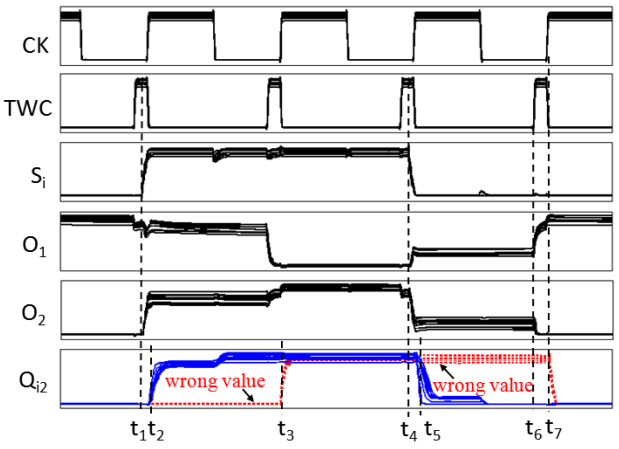

Fig. 17. Monte-Carlo simulation results obtained for the case of PVT variations up to 20% and late transitions of S_i occurring during $T_{M\text{-}COR}$.

both the time intervals Δt_I and Δt_{II}, since our scheme forces Q_{i2} to assume the correct logic value in case of late transitions of S_i. As can be noted, Q_{i2} does not exhibit a full swing transition. This is because an electrical conflict arises between the pMOS transistors of the C-element of our correction block (Fig. 12) and the nMOS transistor of the output of flip-flop FF_{i2}, which has sampled an incorrect logic value. However, we have verified that the voltage values reached by Q_{i2} are very close to V_{dd} and ground and are correctly recognized as high and low logic values.

4.4 Robustness to NBTI Effects

We have verified that our proposed HP monitoring scheme keeps on working properly even when degraded by NBTI, that is it keeps on properly detecting late S_i transitions and correcting the value provided by FF_{i2}.

As described in Subsection 3.4 for the LAP monitoring scheme, the increase in the absolute value of the pMOS threshold voltage (ΔV_{th}) due to NBTI degradation has been evaluated by means of the model presented in [6]. The value of the parameter α accounting for the time interval during which each pMOS transistor composing our HP monitoring scheme is under a stress condition has been estimated as follows.

1) The pMOS transistors of the transfer gates TG_1 and TG_2 in Fig. 13 are conductive only during the interval $T_{M\text{-}COR}$ of each clock cycle. For these transistors it is: $\alpha = T_{M\text{-}COR}/T_{CK} = 15ps/333ps = 0.045$.
2) The pMOS transistors of the transfer gates TG_3 and TG_4 in Fig. 14 conduce only in the time interval during which TWC=0, so that, in every clock cycle, they are conductive for a time interval T_{CK}-$T_{M\text{-}COR}$=t_{force} (Fig. 11). Therefore, for these transistors it is: $\alpha = t_{force}/T_{CK} = 318ps/333ps = 0.954$.
3) The pMOS transistors composing the circuit generating TWC are conductive for half the clock cycle. For these transistors it is: $\alpha = T_{CK}/2T_{CK} = 1/2$.
4) The pMOS transistors in the delay element D1 (Fig. 13), the inverters IR_1 and IR_2 (Fig. 14), and the C-element (Fig. 14) conduce for a time interval depending on the input statistics. By considering a signal S_i switching activity equal to 50%, we obtain: $\alpha = t_{on}/\Delta t = 0.5$.

The previous values of the parameter α have been used to

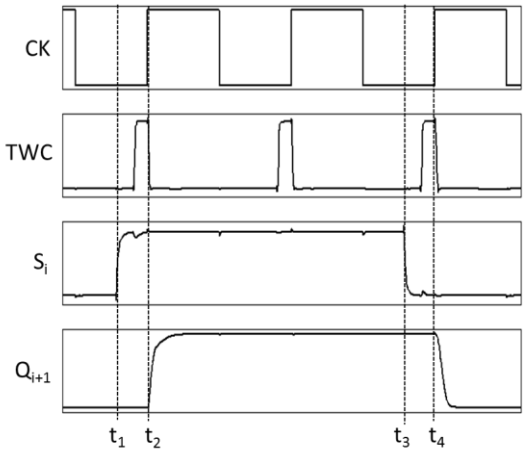

Fig. 18. Simulation results obtained in case of no late transition of S_i occurring during $T_{M\text{-}COR}$ and 10 year NBTI performance degradation.

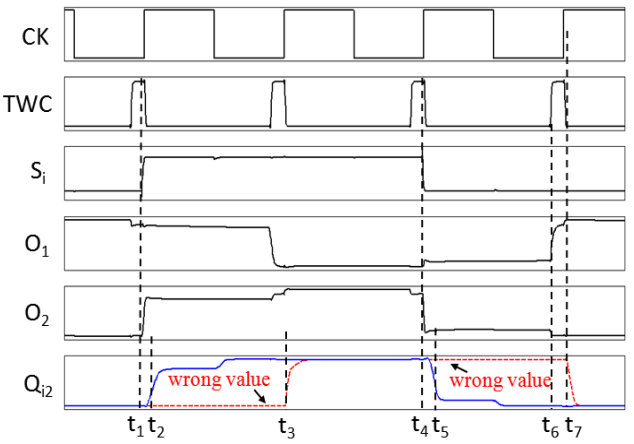

Fig. 19. Simulation results obtained in case of a late transition of S_i occurring during $T_{M\text{-}COR}$ and 10 year NBTI performance degradation.

evaluate the threshold voltage shift of the pMOS transistors composing our HP monitoring scheme and the monitored datapath, considering a circuit life time of 10 years. Finally, a proper device model accounting for the correct threshold voltage degradation has been created for each pMOS transistor of our correcting scheme.

In Fig. 18 we present the simulation results obtained when the monitored signal S_i switches correctly before the interval $T_{M\text{-}COR}$ (at time instants t_1 and t_3). As can be seen, the correct value is provided on output Q_{i2} and, as expected, our HP monitoring scheme keeps on correctly operating, despite its being affected by NBTI.

Instead, Fig. 19 presents the simulation results in case of signal S_i late rising (at the time instant t_1), and falling (at the time instant t_4) transitions. As in the previous subsection, we report the output signal Q_{i2} obtained with our HP monitoring scheme (solid line waveform), and without it (dashed line waveform). When no correction is performed, a wrong Q_{i2} value is produced during the clock period after the clock sampling edges occurring at instants t_1 (wrong low value) and t_4 (wrong high value).

In case of our HP monitoring scheme, instead, the logic value on node Q_{i2} is forced to assume the correct value but, as in the previous subsection with no NBTI degradation, the commutation on Q_{i2} is not a full swing transition. However, we have verified that the voltage values on Q_{i2} are still very close to V_{dd} and ground, and are recognized as correct logic values by fan-out gates. Therefore, our degraded scheme keeps on working properly also in case of late transitions of S_i due to NBTI.

The robustness of our HP monitoring scheme to NBTI effects can be qualitatively assessed by observing the circuits in Fig. 13(a) and 14. As previously stated, during the circuit operation time, NBTI causes the degradation of the pMOS transistors composing: i) the transfer gates TG_1, TG_2, TG_3 and TG_4; ii) the delay element D1, iii) the inverters IR_1 and IR_2; iv) the C-element; v) the circuit generating TWC.

As for the pMOS transistors in i), the same considerations as for the LAP monitoring scheme hold true, and their NBTI degradation does not affect the correct operation of our HP monitoring scheme.

As for the pMOS in ii), we have verified that when they are degraded, the consequent increase of delay *d1* of the block D1 is approximately equal to that of the setup time t_{set} of the flip-flops connected to our HP scheme, which also degrade due to NBTI. Therefore, the required condition $d1 = t_{set}$ is still approximately satisfied, and the correct operation of our HP monitoring scheme is not impacted by NBTI degradation of the pMOS in ii).

The degradation of pMOS transistors in iii) causes an 8.9% increase of the propagation delay of inverters IR_1 and IR_2 in case of 0→1 transitions. However, since their outputs are given to the C-element only after the following falling edge of TWC, an increase in their propagation delay does not affect the correct operation of our HP scheme.

As for the pMOS transistors in iv), they must force the output of the flip-flop at which our scheme is connected to assume the correct logic value. NBTI weakens these transistors over time, thus increasing the correction time. However, they have been sized in order to allow proper correction, even for worst case NBTI degradation.

Finally, the degradation of pMOS transistors in v) does not affect the operation of the circuit generating TWC, since it is resilient by design to the aging effects produced by NBTI [7].

5 Cost Evaluation and Comparison

We have evaluated the costs of both our proposed LAP and HP NBTI monitoring schemes in terms of area overhead, power consumption and impact on system performance, and we have compared them to those of the aging sensors recently published in [4, 8, 7, 13]. Electrical simulations of all compared solutions have been performed considering a standard 45nm CMOS technology [19], a power supply V_{dd} = 1V and clock frequency of 3GHz. All solutions have been implemented assuming the minimum transistor sizes making them work properly. As for the solution in [8], we have not included the cost of the circuitry required to generate its control signal, since it was not specified, while we have considered the circuitry employed to generate the TWC signal in our LAP and HP schemes, and the control signals in the solutions [4, 7, 13].

5.1 Area Overhead and Power Consumption

For the purpose of comparison, we have evaluated the costs of all compared monitoring schemes for the case of a single monitored signal. Area overhead has been roughly estimated in terms of squares, while the power consumption has been as-

sessed as the average power consumed by each solution, considering the monitored signal with no late transition and with a switching activity of 25%. The static power consumption due to leakage has been accounted as well. Additionally, the signal identifying the monitoring time interval has been generated as described in [7] for all compared schemes.

Table 1 reports the area and power costs, as well as the relative variations of the compared solutions over our LAP and HP monitoring schemes (Δ= *100·([4, 8, 7, 13] – our)/ our*). As can be seen, our LAP scheme presents the lowest area and power consumption. Compared to it, the alternative solutions in [4, 8, 7, 13], which induce the same impact on performance (as clarified in the next subsection), exhibit an increase in area ranging from +13.8% of the solution in [7] to +48.3% of the scheme in [13], and a power consumption increment ranging from +4.9% of the solution in [13] to +25% of the scheme in [8].

As for the proposed HP monitoring scheme, it requires the highest area among the considered solutions, while its power consumption is comparable. This cost increase is counterbalanced by a reduction in the impact on performance, as clarified in the next subsection. As previously mentioned, the value of power consumption reported in Table 1 has been estimated considering a single monitored signal with no late transition. In this regard, it is worth reminding that, if correction is required (i.e., in case of a late transition), an electrical conflicts is originated between our HP scheme, which forces the correct output value, and its connected flip-flop. In this case, the power consumption rises up to 35μW. However, according to the HP recovery phase presented in Subsection 4.2, the output of the flip-flop needs to be corrected only during short time intervals. Therefore, being the correction a rare event, we can expect that the actual average power consumed by our HP scheme is slightly higher than reported in Table 1.

Let us now evaluate the absolute area overhead (AO) of our proposed LAP and HP approaches, in case of *n* critical data-paths to be monitored. For both approaches, AO is given by the sum of the area of each monitoring scheme (reported in Table 1 and hereafter denoted by $A_{mon\text{-}LAP}$ or $A_{mon\text{-}HP}$, respectively) and the area of the *n*-input error indicator (denoted by $A_{n\text{-}in\text{-}EI}$) gathering the outputs of the *n* monitors. Therefore, the AO of our LAP and HP approaches, expressed in squares (Sq) as an estimate, is given by:

$$AO_{LAP}(Sq) = nA_{mon\text{-}LAP} + A_{n\text{-}in\text{-}EI} = 58n(Sq) + (54n + 86)Sq \quad (13)$$

$$AO_{HP}(Sq) = nA_{mon\text{-}HP} + A_{n\text{-}in\text{-}EI} = 98n(Sq) + (54n + 86)Sq \quad (14)$$

Finally, we have analyzed the area required by all EIs in the chip and the area of the routing resources as a function of the number of transition detectors connected to a single error indicator. As an example, for this analysis we have considered the case of 256 transition detectors distributed simmetrically throughout the chip. Fig. 20 reports the area required by all EIs and routing resources as a function of the number of transition detectors per EI (i.e., number of inputs of the EIs), normalized with respect to the cases with the lowest EI and routing area. As can be observed, the area required by all EIs in the chip decreases, while the area of the routing resources increases, with the number of transition detectors connected to each EI. The total area overhead presents a minimum when 8 transition detectors are connected to each EI. Of course, the existing tradeoff between the area of all EIs and the routing area depends on circuit functionality and layout and could be estimated during the design phase.

TABLE 1
AREA AND POWER CONSUMPTION COMPARISON.

Scheme	Area (Sq)	ΔA (%)		Power (μW)	ΔP (%)	
		LAP	HP		LAP	HP
monitor [13]	86	+48.3%	-12.2%	15.0	+4.2%	-7.4%
monitor [7]	66	+13.8%	-32.6%	15.9	+10.4%	-1.85%
monitor [4]	78	+34.5%	-20.4%	16.0	+11.1%	-1.23%
monitor [8]	69	+19.0%	-29.6%	18.0	+25.0%	+11.1%
LAP scheme	**58**	-	-40.8%	**14.4**	-	-11.1%
HP scheme	98	+69.0%	-	16.2	+12.5%	-

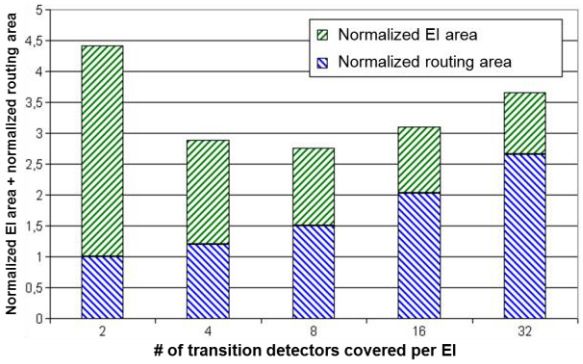

Fig. 20. Normalized area required by the EIs in the chip and by the routing resources to cover 256 transition detectors as a function of the number of transition detectors connected to each EI.

5.2 Impact on Performance

We have evaluated the impact on system performance of our LAP and HP monitoring schemes, and we have compared it to that of the solutions in [4, 7, 8, 13]. The maximum operating frequency allowed by each considered scheme has been evaluated as a function of the circuit life time considering, as realistic assumption, a maximum circuit lifetime of 10 years [6].

As an example, we have considered a critical data-path C_i (in Fig. 1) composed by 29 min-sized inverters, and flip-flops FF_{i1} and FF_{i2} implemented as described in the previous sections. For the considered 45nm CMOS technology, such a data-path implementation has allowed an initial (i.e., without performance degradation of the block C_i due to NBTI) maximum operating frequency of 3.44 GHz.

We have connected our LAP and HP monitoring schemes to the input of FF_{i2} and applied the masking procedures described in Subsections 3.2 and 4.2, respectively, to evaluate the maximum operating frequency as a function of circuit lifetime. For the schemes in [4, 8, 7, 13] we have applied the same masking procedure as developed for our LAP scheme and carried out the same evaluation. For all the compared solutions, we have considered six different values for the guardband Δt_{GB}. Of course, the introduction of the monitoring circuits increases the capacitive load of the circuit, causing a slight increase in the monitored signal propagation delay. This additional delay is approx-

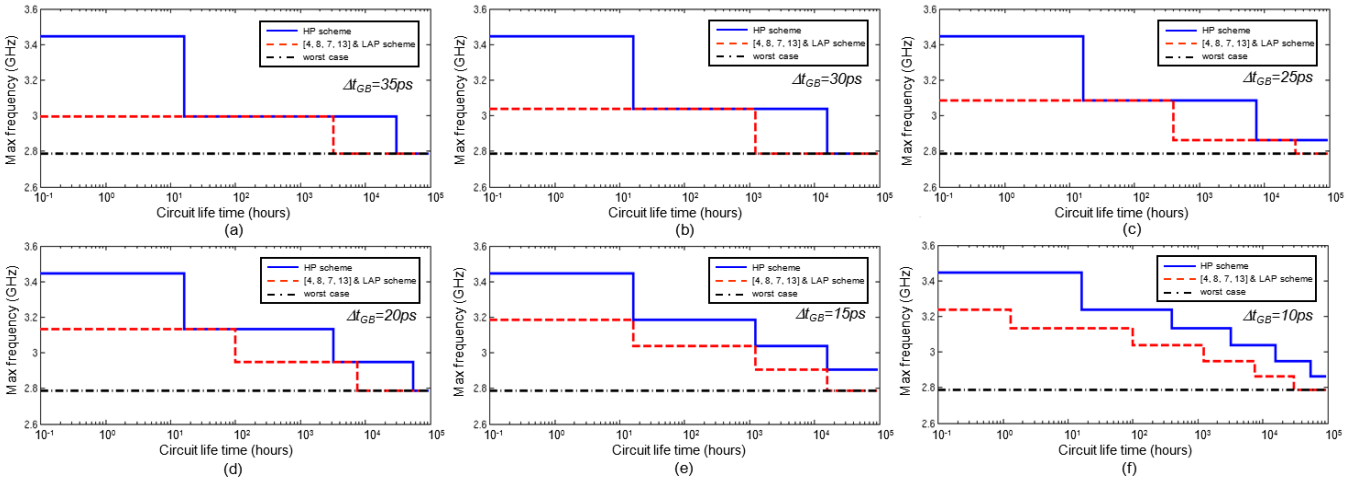

Fig. 21. Maximum operating frequency allowed by our LAP and HP schemes and by the solutions in [4, 8, 7, 13] as a function of the time elapsing from the fabrication of the circuit, for the case of: (a) Δt_{GB}=35ps; (b) Δt_{GB}=30ps; (c) Δt_{GB}=25ps; (d) Δt_{GB}=20ps; (e) Δt_{GB}=15ps; (f) Δt_{GB}=10ps.

imately the same for all compared solutions, and we have verified that it is in the order of only 3% of the initial system clock cycle (3.44GHz).

Fig. 21 shows the maximum operating frequency allowed by the compared schemes as a function of circuit life time, for the cases of: (a) Δt_{GB}=35ps, (b) Δt_{GB}=30ps, (c) Δt_{GB}=25ps, (d) Δt_{GB}=20ps, (e) Δt_{GB}=15ps, and (f) Δt_{GB}=10ps. It is also shown the operating frequency guaranteeing the system correct operation in the presence of worst case NBTI degradation during the whole circuit life time (reported as worst case). We can observe that, for all considered cases and for the entire circuit lifetime, our LAP scheme features the same impact on performance as the solutions in [4, 8, 7, 13].

On the other hand, during the early stage of the circuit life time, our HP monitoring scheme features the highest operation frequency, with relative improvements over our LAP scheme and those in [4, 8, 7, 13] ranging from 15% to 7%, depending on the Δt_{GB} value.

Additionally, from Figs. 21(a)-(f) it can be observed that, the higher the value chosen for Δt_{GB}, the larger the improvement in maximum operation frequency allowed by our HP scheme over the other solutions during the early stage of circuit life time. We can also note that, for Δt_{GB} > 15ps (Figs. 21(a), (b), (c) and (d)), there are time intervals during which the maximum operating frequencies allowed by our HP and LAP schemes and those in [4, 8, 7, 13] coincide. Instead, for $\Delta t_{GB} \leq$ 15ps (Figs. 21(e) and (f)), our HP scheme allows always a higher maximum operating frequency. In this latter case, we observed that our HP scheme exhibits a relative increase in the maximum operation frequency ranging from 8% to 6% over the other compared solutions.

Finally, we can observe that, during the early stage of the circuit lifetime, our HP and LAP schemes, as well as the solutions in [4, 8, 7, 13] allow a considerable increase in the operating frequency compared to the worst case in Fig. 21, where the circuit operates from the beginning with a clock period increased by the maximum NBTI degradation expected for its entire lifetime.

6 Conclusions

We have proposed two monitoring and masking approaches, denoted as Low Area and Power (LAP) and High Performance (HP), which are able to detect late transitions due to NBTI degradation in the combinational part of critical data-paths and guarantee the correctness of the provided output data by properly adapting the clock frequency. We have shown that, compared to recently proposed alternative solutions, our LAP approach requires lower area overhead and lower, or comparable, power consumption, while featuring the same impact on system performance. As for our HP approach, it allows to reduce the impact on system performance, at the cost of a limited increase in area and power consumption. Therefore, between the proposed approaches, the better solution for any considered application could be devised based on area, power and impact on performance constraints.

References

[1] J. Keane, T-H. Kim, C. H. Kim, "An On-Chip NBTI Sensor for Measuring pMOS Threshold Voltage Degradation", IEEE Trans. On Very Large Scale Integration (VLSI) Syst., 2009.

[2] M. Agarwal, B.C. Paul, M. Zhang, S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging", in Proc. of IEEE VLSI Test Symp., pp. 277-286, 2007.

[3] V. Huard, M. Denais, "Hole Trapping Effect on Methodology for DC and AC Negative Bias Temperature Instability Measurements in PMOS Transistors", in Proc. of IEEE Int. Rel. Physics Symp., pp 40-45, 2004.

[4] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B.C. Paul, W. Wang, B. Yang, Y. Cao, S. Mitra, "Optimized Circuit Failure Prediction for Aging: Practicality and Promise", in Proc. *of IEEE Int. Test Conf*., pp. 1-10, 2008.

[5] S. Borkar, "Electronics Beyond Nano-Scale CMOS", ACM/IEEE Design Automation Conf., 2006.

[6] W. Wang, Z. Wei, S. Yang, Y. Cao, "An Efficient Method to Identify Critical Gates under Circuit Aging", in Proc. of IEEE/ACM Int. Conf. on Computer-Aided Design, pp. 735-740, 2007.

[7] J.C. Vazquez, V. Champac, A.M. Ziesemer, R. Reis, J. Semiao, I.C. Teixeira, "Predictive Error Detection by On-Line Aging Monitoring", in Proc. of IEEE Int. On-Line Testing Symp., pp. 9-14, 2010.

[8] A. C. Cabe et al., "Small Embeddable NBTI Sensors (SENS) for Tracking On-Chip Performance Decay", in Proc. of Symp. on Quality Electronic Design, pp. 1-6, 2009.

[9] K. Kang, et al., "Characterization and Estimation of Circuit Reliability

Degradation Under NBTI Using On-Line IDDQ Measurement“, in Proc. of Design Automation Conf., pp. 358-363, 2007.

[10] T. H. Kim et al., “Silicon Odometer: an On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits“, IEEE J. Solid State Circuits, Vol. 3, No. 4, pp. 874-880, April 2008.

[11] E. Karl, et al., “Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation“, in Proc. of Solid State Circ. Conf., pp. 410-411, 2008.

[12] K. K. Kim, W. Wang, K. Choi, “On-Chip Aging Sensor Circuits for Reliable Nanometer MOSFET Digital Circuits”, IEEE Trans. on Circuits and Systems – II: Express Briefs, Vol. 57, No. 10, pp. 798-802, October 2010.

[13] K.A. Bowman, J.W. Tschanz, N.S. Kim, J.C. Lee, C.B. Wilkerson, S.L. Lu, T. Kernik, V.K. De, “Energy-Efficient and Metastability-Immune Resilient Circuits for Dynamic Variation Tolerance”, IEEE J. of Solid-State Circuits, Vol. 44, No. 1, January 2009.

[14] P. Singh, E. Karl, D. Sylvester, D. Blaauw, “Dynamic NBTI management using a 45nm multi-degradation sensor “, in Proc. of IEEE Custom Integrated Circuits Conference (CICC), 2010.

[15] C. Thibeault, “On the Comparison of ΔI_{DDQ} and I_{DDQ} Testing”, in Proc. of IEEE VLSI Test Symp., pp. 143-150, 1999.

[16] C. Metra, et al. “Self-Checking Detection and Diagnosis of Transient, Delay, and Crosstalk Faults Affecting Bus Lines“, IEEE Trans. on Comp., Vol. 49, No. 6, pp.560-574, June 2000.

[17] M. Omaña, D. Rossi, N. Bosio, C. Metra, “Self-Checking Monitor for NBTI Due Degradation”, in Proc. of IEEE 16th International Mixed-Signals, Sensors and Systems Test Workshop (IMS3TW), 2010.

[18] M. Omaña, D. Rossi, C. Metra, “Low Cost and High Speed Embedded Two-Rail Code Checker”, IEEE Transactions on Computers, Vol. 54, Issue 2, pp. 153-164, February 2005.

[19] http://ptm.asu.edu/

[20] M. Bhushan, A. Gattiker, M. B. Ketchen, K. K. Das, “Ring Oscillators for CMOS Process Tuning andVariability Control ”, IEEE Transactions on Semiconductor Manufacturing, Vol. 19, No. 1 February 2006. pp. 10-18.

[21] S. Tam, et al., “Clock Generation and Distribution for the 130-nm Itanium® 2 Processor With 6-MB On-Die L3 Cache”, *IEEE J. of Solid-State Circuits*, Vol. 39, No. 4, pp. 636-642, April 2004.

Martin Omaña received the degree in Electronic Engineering from the University of Buenos Aires (Argentina) in 2000. In 2002 he was awared a MADESS grant and joined the Electronics Department of the University of Bologna, Italy, where he obtained the PhD in Electronic Engineering and Computer Science in 2005. He is currently a Post Doctoral fellow at the same University. His research interests are in the field of design and test of digital systems, reliable and error resilient systems, fault tolerance, on-line testing, fault modeling and diagnosis and debug.

Daniele Rossi received the degree in electronic engineering and the PhD degree in electronic engineering and computer science from the University of Bologna in 2001 and 2005, respectively. He is currently a Post Doctoral fellow at the same University. His research interests include fault modeling, online testing and fault tolerance techniques, with particular focus on coding techniques for fault-tolerant and low-power buses, signal integrity for VDSM communication infrastructures, and robust design for soft error resiliency. He holds one patent. He is a member of the IEEE Computer Society.

Nicolò Bosio obtained his BSc and MSc degree in electronic engineering from the University of Bologna in 2007 and 2010, respectively. He is currently with EFI Technology Srl (Bologna) as an hardware/software design engineer. His research interests include battery modeling and power management in automotive applications.

Cecilia Metra is an Associate Professor in Electronics in the Department of Electronic, Computer Science and Systems (DEIS) of the Univ. of Bologna. She is also affiliated with the Advanced Research Center on Electronic Systems for Information and Communication Technologies E. De Castro (ARCES) of the Univ. of Bologna. She is the General Chair of the IEEE Int'l VLSI Test Symp. 2012, and she has been General Chair/Co-Chair of the IEEE Int'l VLSI Test Symp. 2011, The IEEE Int'l Symp. on Defect and Fault Tolerance in VLSI Systems 2005 and 1999, the IEEE Int'l On-Line Testing Symp. 2006 and the IEEE Int'l On-Line Testing Workshop 2001, and Program Chair/Co-Chair of the IEEE Int'l VLSI Test Symp. 2009 and 2008, the IEEE International Workshop on Design and Test of Nano Devices, Circuits and Systems (NDCS) 2008, The IEEE Int'l Symp. on Defect and Fault Tolerance in VLSI Systems 1998, the IEEE Int'l On-Line Testing Symp. 2005, 2004 and 2003, and the IEEE Int'l On-Line Testing Workshop 2002. She serves/served as Topic Chair and as Member of the Organizing Committee and/or Technical Program Committee of several international conferences. Her research interests are in the field of Design and Test of Integrated Digital Systems, Reliable and Error Resilient Systems, Fault Tolerance, On-Line Testing, Fault Modeling, Diagnosis and Debug, Emergent Technologies, Energy Harvesting and Security. She is Associate Editor in Chief of the IEEE Transactions on Computers, and Member of the Editorial Board of the Journal of Electronic Testing: Theory and Applications and the International Journal of Highly Reliable Electronic System Design. She is a Senior Member and a Golden Core member of the IEEE Computer Society.