Using Transition Fault Test Patterns for Cost Effective Offline Performance Estimation

Mahroo Zandrahimi* Philippe Debaud† Armand Castillejo† Zaid Al-Ars*
*Delft University of Technology, The Netherlands
{m.zandrahimi, z.al-ars}@tudelft.nl
†STMicroelectronics, Grenoble, France
{philippe.debaud, armand.castillejo}@st.com

Abstract—Process variation occurring during fabrication of complex VLSI devices induce uncertainties in operation parameters (e.g., supply voltage) to be applied to each device in order for it to fit within the allowed power budget and get the optimum power efficiency. Therefore, an efficient post manufacturing performance estimation mechanism is needed in order to tune operation parameters for each device during production. The current state-of-the-art approach of using Process Monitoring Boxes (PMBs) have shown some limitations in terms of cost and accuracy that limit their benefit. Simulation results on ISCAS'99 benchmarks using 28nm FD-SOI library show that the accuracy of PMB approaches is design dependent, and requires up to 8.20% added design margin. To overcome those limitations, in this paper we propose an alternative solution using transition fault (TF) test patterns, which is able to eliminate the need for PMBs, while improving the accuracy of performance estimation. The paper discusses a case study on real silicon comparing the performance estimation using functional test patterns and the TF based approach on a 28nm FD-SOI CPU. The results show a very close correlation between TF test patterns and functional patterns.

I. Introduction

As technology scales, integrated circuits become more sensitive to process variations. Due to inter die process variations, each chip has its own characteristics which leads to different speed and power consumption. In order to tune each chip during production, a post manufacturing performance estimation mechanism is needed. Since performance estimation during production should be done as fast as possible, running functional patterns on CPU, which reflects the final application is therefore most of the time not feasible. A standard industrial approach for performance estimation is the use of on-chip Performance Monitor Boxes (PMBs), which are very fast during production. They range from simple inverter based ring oscillators to more complex critical path replicas designed based on the most used cells extracted from the potential critical paths of the design [1]–[6]. The frequency of PMBs is dependent on various silicon parameters such as NMOS and PMOS speeds, capacitances, leakage, etc.

To be able to estimate the circuit performance based on PMB responses during production, the correlation between frequency of PMBs and circuit frequency should be measured during characterization, an earlier stage of manufacturing. Once PMB responses are correlated to application performance, they are ready to be used for performance estimation during production. During production, based on the frequency responses from these monitors, the chip performance will be estimated. According to figure 1, the information could be used to either sort devices based on their speed in order to sell them as a fast or slow device, adapt voltage to enhance yield, or optimize power and battery lifetime, such as voltage scaling and body biasing [7].

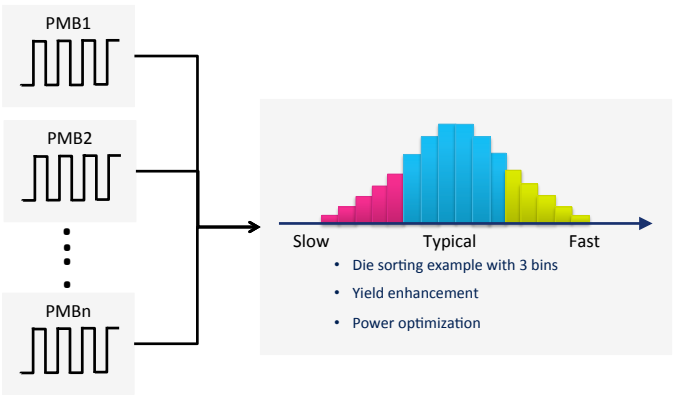

Fig. 1. Performance estimation using PMBs

However, trying to predict performance of the many millions of paths in a given design based on information from a single unique path could be difficult and in many cases inaccurate. This approach might work for very robust technologies and when only very few parameters influence performance, such as voltage, process corner, and temperature. However, in deep sub-micron technologies, as intra-die variation and interconnect capacitances are becoming predominant, it is more complex to estimate the performance of the whole design based on few PMBs. Hence, to improve the accuracy, we should use an alternative approach that increases the number of paths we take into account for performance estimation.

In this paper we introduce a cost effective approach for performance estimation during production using transition fault test patterns, which can be used for general logic as well. The contributions of this paper are the following:

- A detailed investigation of PMB approach in terms of accuracy and effectiveness using 29 ISCAS'99 benchmarks with 28nm FD-SOI library for 42 different process corners.
- Proposing the new concept of using transition fault (TF) testing for performance estimation during production.

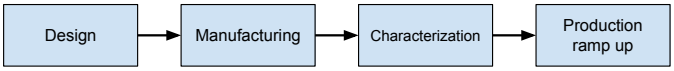

Fig. 2. Stages of the chip design and manufacturing process

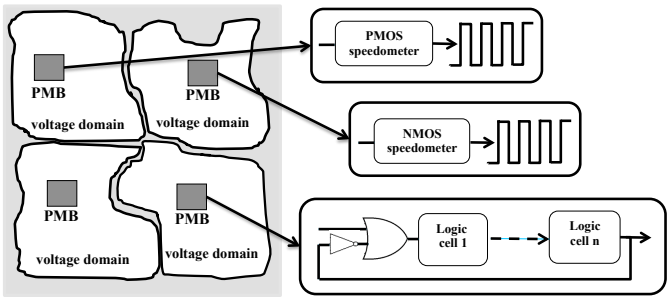

Fig. 3. Performance estimation using PMBs

- A case study on silicon for evaluating the accuracy of performance estimation using TF based approach on a 28nm FD-SOI CPU.

The rest of this paper is organized as follows. Section II introduces the limitations of PMB approaches, which is the reason of investigating new methods for performance estimation during production. Section III proposes the new approach of performance estimation using transition fault test patterns. Evaluation of the proposed approach is presented in Section IV using silicon measurements of a 28nm FD-SOI CPU. Section V concludes the paper and proposes potential solutions for future work.

II. Motivation

Figure 2 shows the various industrial stages of the design and manufacturing process of integrated circuits. The process starts with the design stage, where the circuit structure and functionality is specified based on a given set of specifications. When the design is completed, the manufacturing stage starts where a representative number of chip samples will be manufactured. These chip samples will be used during the characterization stage to find the correlation between PMB responses and the actual performance of the chip. Finally during the production ramp up stage the integrated circuits will be mass produced. In this stage, the PMB correlation measured during the characterization stage will be used to adapt various parameters exclusively to each produced chip.

Figure 3 shows an example of a chip, on which various kinds of PMBs are distributed. The figure shows two PMBs created using PMOS and NMOS speedometers that indicate the speed of PMOS and NMOS transistors. These kind of PMBs are called generic since they can be used for different designs without modifications. The third shown PMB is a critical path replica designed based on the most used logic cells extracted from the potential critical paths of the design, therefore, these kind of PMBs are design dependent. During production based on the frequency responses from these monitors, chip performance is estimated. This information could be used to either sort devices based on their speed (so-called speed binning), adapt voltage to enhance yield, or optimize power and battery lifetime using voltage scaling and body biasing [7].

To be able to estimate the circuit performance based on PMB responses during production, the correlation between frequency of PMBs and circuit frequency should be measured. This process is done during the characterization stage. During this stage, functional patterns are executed on each chip, and the frequency of each PMB and the whole chip are measured. These measurements are repeated for a given amount of test chips representative of the process window to make sure that the information from all process corners have been extracted. Based on this information, the correlation between PMBs and the actual frequency of the circuit is determined. Once PMB responses are correlated to application performance, they are ready to be used for performance estimation during production. However, this correlation process has a negative impact in terms of design effort and time to market, since the process should be repeated for a large amount of test chips to make sure that the calculated correlation reflects the actual chip performance for all manufactured chips. The long correlation process makes these approaches very expensive. Moreover, the fact that functional patterns are used for the correlation process makes PMB approaches not suitable for general logic, since even though using functional patterns for programmable parts of the design such as CPU and GPU is possible, the rest of the design such as interconnects are difficult to be characterized using this approach [8].

On the other hand, since there are discrepancies in the responses of same PMBs from different test chips, the estimated correlation between the frequency of PMBs and the actual performance of the circuit could be very pessimistic, which results in wasting power and performance. In [9], a silicon measurement on 625 devices manufactured using nanometric FD-SOI technology had been done. 12 PMBs are embedded in each device. Figure 4 shows an example of Vmin discrepancy for one of the 12 PMBs. The Y axis shows the frequency responses of the PMB on all 625 devices, while the X axis shows the optimal voltage of each chip where the corresponding PMB is located. The optimal minimum voltage for each chip is measured using test patterns. To quantify the amount of Vmin discrepancy in this figure, for each value of frequency response, Vmin variation is measured (the red arrow). The maximum amount of this variation is considered as the Vmin discrepancy for that PMB. This inaccuracy in Vmin

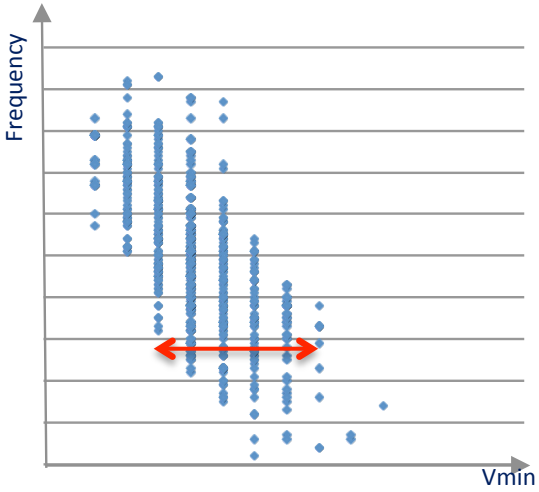

Fig. 4. Example of Vmin discrepancy for one PMB on all 625 devices measured [9]

TABLE I. FEATURES OF DIFFERENT CORNERS OF 28NM FD-SOI LIBRARY USED IN SIMULATIONS

Corner	Voltage [V]	Temperature [°C]	Biasing	Aging	Corner	Voltage [V]	Temperature [°C]	Biasing	Aging
SS	0.7	-40	no	no	SS	0.7	0	no	no
SS	0.7	125	no	no	SS	0.7	-40	no	yes
SS	0.7	0	no	yes	SS	0.7	125	no	yes
SS	0.8	-40	no	no	SS	0.8	0	no	no
SS	0.8	125	no	no	SS	0.8	-40	no	yes
SS	0.8	0	no	yes	SS	0.8	125	no	yes
TT	0.8	25	no	no	TT	0.8	125	no	no
SS	0.85	-40	no	no	SS	0.85	0	no	no
SS	0.85	125	no	no	SS	0.85	-40	no	yes
SS	0.85	0	no	yes	SS	0.85	125	no	yes
TT	**0.85**	**25**	**no**	**no**	TT	0.85	125	no	no
SS	0.9	-40	yes	no	SS	0.9	125	yes	no
SS	0.9	-40	no	no	SS	0.9	0	no	no
SS	0.9	125	no	no	SS	0.9	-40	no	yes
SS	0.9	0	no	yes	SS	0.9	125	no	yes
TT	0.9	125	no	no	TT	0.9	25	no	no
FF	0.9	-40	no	no	FF	0.9	125	no	no
SS	0.95	-40	no	no	SS	0.95	0	no	no
SS	0.95	125	no	no	SS	0.95	-40	no	yes
SS	0.95	0	no	yes	SS	0.95	125	no	yes
TT	0.95	25	no	no	TT	0.95	125	no	no

T and S stand for typical and slow corners, respectively.

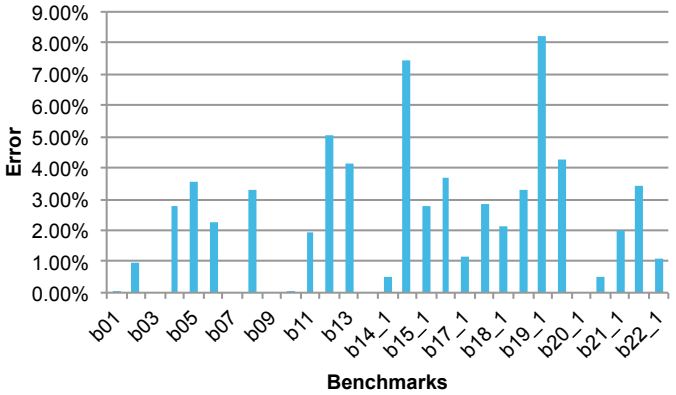

Fig. 5. Percentage of *error* for ISCAS'99 benchmarks using 28nm FD-SOI library

TABLE II. ERROR IN PERFORMANCE ESTIMATION USING ONE PMB FOR ISCAS'99 BENCHMARKS WITH 28NM FD-SOI LIBRARY

Benchmark	# Cells	*error*	Benchmark	# Cells	*error*
*b*01	30	0.02%	*b*15	3142	7.45%
*b*02	21	0.96%	*b*15_1	3141	2.77%
*b*03	76	0.00%	*b*17	9559	3.67%
*b*04	196	2.80%	*b*17_1	9584	1.14%
*b*05	390	3.53%	*b*18	22175	2.86%
*b*06	29	2.27%	*b*18_1	22093	2.14%
*b*07	179	0.00%	*b*19	43916	3.31%
*b*08	71	3.31%	*b*19_1	43822	8.20%
*b*09	94	0.00%	*b*20	3970	4.25%
*b*10	110	0.07%	*b*20_1	4025	0.00%
*b*11	326	1.96%	*b*21	4022	0.48%
*b*12	547	5.04%	*b*21_1	4082	2.02%
*b*13	154	4.12%	*b*22	6102	3.45%
*b*14	1967	0.00%	*b*22_1	6164	1.08%
*b*14_1	2043	0.49%	-	-	-

measurement results in wasting power. The same procedure is done for all 12 PMBs, and the results show that minimum voltage estimation based on PMBs lead to nearly 10% of wasted power on average and 7.6% in the best case, when a single PMB is used for performance estimation.

To further investigate the accuracy and effectiveness of PMB approaches, we performed static timing analysis (STA) with Primetime (SYNOPSYS tool for STA [15]) on ISCAS'99 benchmarks [14] using 28nm FD-SOI library. ISCAS'99 contains 29 benchmarks from small circuits with 21 cells to more complicated benchmarks with almost 44K cells. Table I lists the characteristics of the 42 different corners used in the STA simulation for the 28nm FD-SOI library with voltage, body biasing, temperature, transistor speed and aging parameters.

The results of the simulation are expressed in terms of the performance error in the PMB estimation. We assume that the PMB performance estimation for each benchmark is represented by the critical path reported by STA in the typical corner for that benchmark. The characteristics of the typical corner simulation are (TT, 0.85, 25, no, no), as highlighted as the bold row in Table I. Then, we estimate the performance of the design in the 41 other corners using that PMB (represented by the typical corner simulation). In order to quantify the results, we define a parameter named *error* which is measured for each benchmark. The concept relates to how much margin should be taken into account due to inaccuracies as a result of performance estimation using PMBs. To be able to measure *error* for each benchmark, first we check if the critical path in each corner is different from the critical path of the typical corner (PMB for each benchmark). In the case of critical path difference, we measure $error_{corner}$ for the process corner by:

$$error_{corner} = (P_{corner} - \text{PMB})/P_{corner} \quad (1)$$

where P_{corner} is the delay of the critical path measured in *corner*, and PMB is the delay of the critical path identified in the typical corner but measured in *corner*. Once $error_{corner}$ is calculated for all process corners, *error* can be obtained for each benchmark by:

$$error = \max_{all\ corners} [error_{corner}] \quad (2)$$

Figure 5 illustrates the *error* for all 29 ISCAS'99 benchmarks. As shown in this figure, although for some designs the error is zero or negligible, for some other designs the error is rather high and for one case, *b*19_1, it even reaches a maximum of 8.20%. Table II presents the detailed simulation results for all 29 ISCAS'99 benchmarks. According to this table, it is not possible to find a unique critical path for most designs, which stays critical in all 42 corners. Hence, we can conclude that trying to predict performance of the many millions of paths in

a given design based on information from a single unique path could be difficult and in many cases grossly inaccurate. This approach might work for very robust technologies and when only very few parameters influence performance, such as voltage, process corner, and temperature. However, in deep submicron technologies, as intra-die variations and interconnect capacitances are becoming predominant, it is more complex to estimate the performance of the whole design based on one or a couple of PMBs. Hence, to improve the accuracy, we should use an alternative approach that increases the number of paths we take into account for performance estimation.

III. TF BASED PERFORMANCE ESTIMATION

In this paper, we propose an innovative new approach for performance estimation using delay testing during production. Since delay testing covers many path-segments of the design, it can be a better performance representative than a PMB. Such an approach has a number of unique advantages as compared to PMB-based approaches.

1) First, this approach can be performed *at no extra cost*, since delay tests are routinely performed during production to test for chip functionality.
2) In addition, since delay testing is performed to explicitly test for actual chip performance, the expensive phase of correlating PMB responses to chip performance is not needed anymore, which reduces the length of the characterization stage (see Figure 2), and subsequently dramatically reduces cost and time to market.
3) Moreover, as functional patterns are not used anymore, the delay testing approach could be a solution for general logic, and not only for CPU and GPU components.
4) And last but not least, this approach makes using PMBs redundant, which saves silicon area as well as PMB design time.

There are three different types of delay test patterns: TF tests, small delay defect tests, and path delay tests [10]. TF test patterns target all gates and indirectly cover all path-segments. Hence, it covers all different kinds of gates and interconnect structures. Since several faults can be tested in parallel, we can achieve a high coverage with few patterns. However, ATPG choices are based on heuristics like SCOAP [11], which tend to minimize computational effort. Thus, when several solutions are available for path sensitization, ATPG will use the easiest, which means that the tool tends to target short paths and not critical paths of the design [12]. On the other hand, we can alternatively use small delay defect testing, which sensitizes paths with smallest slacks, as well as path delay testing, which sensitizes a selected path. Among these two delay testing methods, path delay seems more promising since it sensitizes functional, long paths, which is an advantage over TF testing. However, in path delay testing the objective is to obtain a transition along critical paths which are on average longer and more complex than the paths targeted in transition fault, thus reducing parallel testing capability and thereby reduces the overall coverage achieved. Therefore, we target TF test patterns in this paper for performance estimation during production since these give the highest path coverage of the three delay test alternatives.

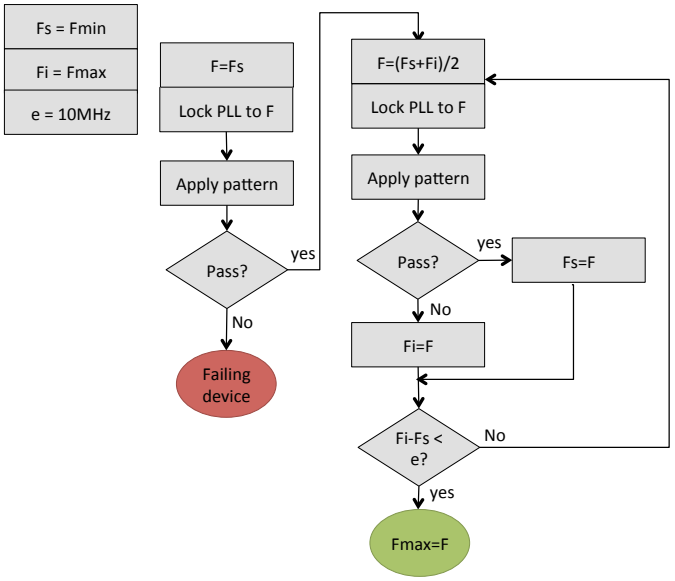

Fig. 6. Proposed flow for performance estimation using TF test patterns

Figure 6 proposes a flow of the TF based approach that could be used during production. The proposed flow performs a binary search to identify the maximum frequency (Fmax) the chip can attain while passing all TF test patterns. The following steps are performed for each operation point of the chip: 1. apply chip setup at nominal values and initialize variables, 2. set PLL to Fmin and wait for stabilization, 3. apply transition fault at speed test, 4. if the chip fails the test, discard it, otherwise, 5. compute new values and do a binary search to find Fmax. Conversion from Fmax to Vmin might be required depending on either performance estimation is done for yield enhancement or power optimization. "e" is an arbitrary value which is up to the users to define the resolution they want.

IV. EVALUATION RESULTS

The basic requirement of using TF-based AVS is that there should be a reasonable correlation between TF frequency the chip can attain while passing all TF test patterns and the actual frequency of the chip. In this case, TF frequency could be a representative of actual chip performance. In order to investigate if such correlation exits, we performed measurements on-silicon using both TF test patterns and functional patterns. Since running functional patterns on CPU reflects the final application, and thus the actual performance of the chip, we used functional frequency as a reference for comparison versus TF frequency. It is important to note that since performance estimation during production should be done as fast as possible, running functional patterns on CPU is therefore most of the time not feasible.

The device under test is a high speed 28nm FD-SOI CPU. This device is equipped with an Adaptive Voltage Scaling system (AVS), which means whenever maximum performance is not required, supply voltage can be scaled so that power can be saved while the system can still meet the timing constraints. Therefore, during production, the optimal voltage should be measured for each frequency point of the chip. We have performed the following steps to compare TF frequency

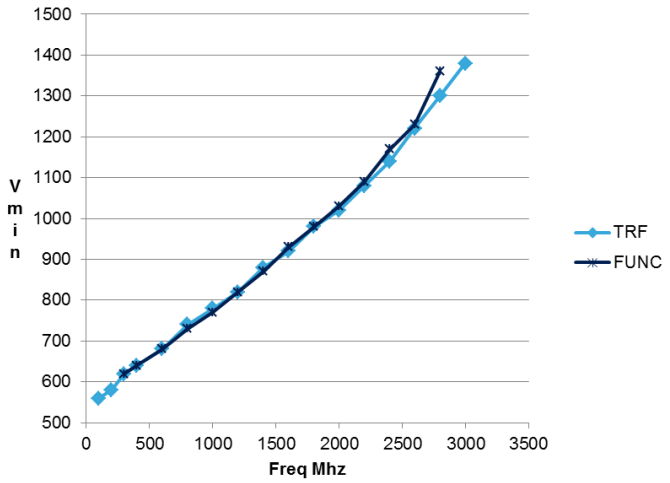

Fig. 7. Correlation of TF and functional patterns on a 28nm FD-SOI CPU

versus functional frequency, which reflects the actual frequency of the chip:

1) We first performed functional test patterns on CPU, and measured the optimal voltage for each frequency point of the chip.
2) Then we have done the same flow discussed in Figure 6, which performs a binary search to identify the minimum voltage (Vmin), at which the chip can pass all TF test patterns for each operating point.

Results are shown in Figure 7. In this figure, the light blue line represents the minimum voltage (y-axis) for each operating point (x-axis) estimated using TF test patterns. The dark blue line represents the minimum voltage (y-axis) measured for each frequency settings (x-axis) of the chip using functional patterns. According to this figure, there is a very close correlation between TF test patterns and functional patterns, which indicates that TF frequency is a very accurate indicator of performance, and therefore can be used for performance estimation during production as an alternative for PMB approach.

V. Conclusions and Future Work

Process variation occurring in deep sub-micron technologies limit PMB effectiveness in silicon performance estimation leading to unnecessary power and yield loss. Simulation results on ISCAS'99 benchmarks using 28nm FD-SOI library show that the accuracy of PMB approaches is design dependent, and requires up to 8.20% added design margin. Thus, we can conclude that estimation of overall application performance from one or few oscillating paths is becoming more and more challenging in nanoscale technologies where parameters such as intra-die variation and interconnect capacitances are becoming predominant. All those efforts have a negative impact in terms of cost and time to market. Finally the fact that functional patterns are used for the correlation process makes PMB approaches not suitable for general logic.

Alternatively, this paper proposes a new approach that uses transition fault testing for performance estimation during production. Since transition fault test patterns target all gates and indirectly cover all path-segments, it can be a better performance representative than a PMB. This approach can be performed at no extra cost, remove the expensive correlation phase and reduces time to market dramatically. Moreover, as functional patterns are not used anymore, testing approach could be a solution for general logic, not only for CPU and GPU. Based on silicon measurements on a high speed 28 nm FD-SOI CPU, there is a very close correlation between TF test patterns and functional patterns proving the relevancy of the TF based approach.

Acknowledgements

This work is carried out under the BENEFIC project (CA505), a project labelled within the framework of CATRENE, the EUREKA cluster for Application and Technology Research in Europe on NanoElectronics.

References

[1] T. Chan and A.B. Kahng, *Tunable Sensors for Process-Aware Voltage Scaling*, in ICCAD, pp. 7-14, 2012.

[2] T. Chan, et al., *DDRO: A Novel Performance Monitoring Methodology Based on Design-Dependent Ring Oscillators*, in ISQED, pp. 633-640, 2012.

[3] A. Drake, et al., *A Distributed Critical-Path Timing Monitor for a 65nm High-Performance Microprocessor*, in ISSCC, pp. 398-399, 2007.

[4] TD. Burd, et al., *A dynamic voltage scaled microprocessor system*, in ISSCC, pp. 294-295, 2000.

[5] J. Kim and M.A. Horowitz, *An efficient digital sliding controller for adaptive power-supply regulation*, in IJSSC, vol. 37, no. 5, pp. 639-647, 2002.

[6] Q. Liu and S.S. Sapatnekar, *Capturing Post-Silicon Variations Using a Representative Critical Path*, in TCAD, vol. 29, no. 2, pp. 211-222, 2010.

[7] M. Zandrahimi and Z. Al-Ars, *A Survey on Low-power Techniques for Single and Multicore Systems*, in ICCASA, pp. 69-74, 2014.

[8] M. Zandrahimi, Z. Al-Ars, P. Debaud, and A. Castillejo, *Industrial Approaches for Performance Evaluation Using On-Chip Monitors*, in IDT, 2016.

[9] M. Zandrahimi, Z. Al-Ars, P. Debaud, and A. Castillejo, *Challenges of Using On-Chip Performance Monitors for Process and Environmental Variation Compensation*, in DATE, 2016.

[10] M. Tehranipoor et al., Test and Diagnosis for Small-Delay Defects, in Springer Science+Business Media, LLC, 2011.

[11] L.H. Goldstein, E.L. Thigpen, SCOAP: Sandia Controllability/Observability Analysis Program, in DAC, 1980.

[12] B. Kruseman, A. Majhi, and G. Gronthoud, On Performance Testing with Path Delay Patterns, in VTS, 2007.

[13] http://www.st.com/content/st_com/en/about/innovation—technology/FD-SOI.html

[14] http://www.cad.polito.it/downloads/tools/itc99.html

[15] http://www.synopsys.com/tools/pages/default.aspx