Adaptive Reduction of the Frequency Search Space for Multi-Vdd Digital Circuits

Chandra K. H. Suresh
New York University Abu Dhabi

Ender Yilmaz & Sule Ozev
Arizona State University

Ozgur Sinanoglu
New York University Abu Dhabi

***Abstract*— Increasing process variations, coupled with the need for highly adaptable circuits, bring about tough new challenges in terms of circuit testing. Circuit adaptation for process and workload variability require costly characterization/test cycles for each chip, in order to extract particular $V_{\rm dd}/f_{\rm max}$ behavior of the die under test. This paper aims at adaptively reducing the search space for $f_{\rm max}$ at multiple levels by reusing the information previously obtained from the DUT during test-time. The proposed adaptive solution reduces the test/characterization time and costs at no area or test overhead.**

I. Introduction

Increasing process variations result in increasing statistical diversity of manufactured devices. A 9.7nm gate (projected for 2020) contains less than 20 silicon atoms between its source and drain, which limits accuracy in the formation of the channel length and dopant control in the channel. The 2007 International Technology Roadmap for Semiconductors (ITRS) projects that by 2020 variability in performance will rise from 51% in 2010 to 69% in 2020, variability in power will rise from 68% in 2010 to 121% in 2020, and variability in leakage power will rise from 229% in 2010 to 325% in 2020. For future products to be viable, circuit adaptation and tuning strategies will need to be an integral part of the design process.

Most high-end digital circuits include some form of adaptation either due to process variations or due to operating conditions. To increase profits, it is often desirable to separate the manufactured devices into power/performance bins. For instance, for the microprocessor industry, chips manufactured with the same design and by the same manufacturing technology may be sold with different maximum frequencies at different prices. A well-known circuit adaptation approach, particularly for power conscious designs, is dynamic voltage scaling (DVS) [1], [2], where the supply voltage of the device and its frequency is reduced whenever possible to save power. Dynamic voltage scaling has been an effective in-field adaptive computational approach for adjusting performance in return for power savings, and thus, extended battery life.

However, such adaptation capabilities bring about tough new challenges in terms of circuit testing. Circuits must go through potentially iterative testing/tuning cycles, increasing the test cost to an unmanageable level. Both performance binning and dynamic voltage scaling require a costly characterization/test for each chip, in order to extract the particular $V_{\rm dd}$-$f_{\rm max}$ behavior of the die under test. For every $V_{\rm dd}$ mode that the design supports, a time-consuming search is conducted on expensive tester equipment (ATE) so as to identify the maximum frequency ($f_{\rm max}$) that the die under test can operate; the $V_{\rm dd}$-$f_{\rm max}$ information is stored on chip for the application/software layer to refer to in exploiting the performance-power tradeoff. A binary-search like procedure is employed to apply delay patterns, which can be quite a few in number, to the die under test at different voltage levels, and the frequency at which all the patterns pass is the "measured" $f_{\rm max}$ for the die under test at that voltage. Apparently, the granularity in which the $f_{\rm max}$ search is conducted determines the *cost and the accuracy* of these measurements. The accuracy in turn dictates how efficiently the power-performance tradeoff can be explored in mission mode; $f_{\rm max}$ measurements that are off from the actual $f_{\rm max}$ values translate into power/performance waste, defeating the purpose of voltage scaling.

Architectural solutions [3] may be in place to monitor, during runtime, deviations in the operation frequency from the $f_{\rm max}$ measured during the characterization process. If a deviation is detected, the $V_{\rm dd}$-$f_{\rm max}$ information is updated on-chip. Such interventions in mission mode incur significant performance penalties, however, due to pipeline flushes, etc, underlining from another perspective the importance of the *accuracy* of $f_{\rm max}$ measurements during characterization/test.

This paper aims at adaptively reducing the search space for $f_{\rm max}$ at multiple $V_{\rm dd}$ levels using the previous information obtained from the DUT during test-time, an approach that is being explored for the first time to the best of our knowledge. This information can stem from multiple sources, such as scribe-line readings, readings from simple auxiliary circuits embedded with the original design, and previous readings from the original design itself. In this paper, we focus on the responses from the DUT, both from ring oscillators embedded with the original circuit and previous $f_{\rm max}$ measurements from the original design. For each $V_{\rm dd}$ level that the circuit needs to be characterized at, we use a statistical mapping technique to predict the maximum frequency with which the circuit can work. We also predict a viable search space so more efficient search can be conducted with this statistical information. Once the maximum frequency is determined for that $V_{\rm dd}$ level, this measurement becomes another point of information for the next $f_{\rm max}$ search at the next $V_{\rm dd}$ level. Thus, we can iteratively reduce the search space for $f_{\rm max}$ as the testing progresses. Most importantly, the adaptive multi-level $f_{\rm max}$ search we propose is performed non-intrusively to the test process, reducing the test/characterization time and costs without incurring any area or test overhead.

This project is funded by SRC/ATIC Contract No. 2012-TJ-2331.
978-3-9815370-0-0/DATE13/©2013 EDAA

II. Prior Work

Increasing process variations have prompted researchers to adapt to the changing characteristics of the devices that come from the production lines. In [4], the authors use known test compaction mechanisms to reduce the overall test time while adaptively changing this test set from one production lot to another. In [5], the authors propose to adapt the test set with respect to the statistical characteristics of each device under test (DUT). They use the online information obtained from each DUT to determine which tests would ideally identify that device as good or bad in the shortest time. They show that compared with static test decisions, test quality and test time can be improved at the same time.

In the digital domain, adaptive test has been widely used for parametric testing. The most common form of statistical adaptation comes in the form of adjusting pass/fail decision criteria for parametric measurements, such as supply current or minimum supply voltage. In, [6]–[10] the authors use tester-based post-processing and neighborhood information to adaptively set the pass/fail limits for parameteric tests. Using neighborhood information reduces the effect of global process variations to that of local process variations within a small location in the wafer. When the variance of the test parameter is reduced with such neighborhood information, smaller deviations in supply current and/or minimum supply voltage can be detected, increasing the overall test quality. Various techniques [11] have been proposed to perform manufacturing test in a process variation aware manner.

The cost of frequency binning [12], and $f_{\max}$/V_{dd} search in terms of testing has been a known problem for some time. The cost of the $f_{\max}$ search can be alleviated by the use of on-chip monitors (ring oscillator). Structures within ring oscillators are subject to the same die-to-die process variations and hence their frequencies can be correlated to the frequency with which the DUT works. A quick reading of the frequency of the ring oscillator provides some information about the process corner that the die under test belongs; the ring oscillator frequency can then be correlated to a predicted $f_{\max}$ for the die, narrowing down the $f_{\max}$ search [13], [14]. The accuracy of this correlation analysis and the $f_{\max}$ predictions determines how quickly the $f_{\max}$ search will converge and terminate, and thus, the costs. Other forms of process monitors can also be used for the same purpose [15]. Measurements from neighboring dies based on the speed clustering expectation [16] and the use of surface response models to map the measurements from test structures onto predicted performance [17] have also been proposed to lower the cost of speed binning.

Another method to narrow down the search space for $f_{\max}$ search is to use the information from structural tests and correlate their response to functional tests. In [18], structural tests are used to initiate the search process. The assumption here is that structural tests are faster to conduct, thus result in lower test time. The close correlation between structural tests, in particular transition and path delay patterns, and system $f_{\max}$ has been shown for industrial designs [19], [20]. This correlation can be further strengthened by data learning methods [21].

While using auxiliary circuit readings to predict the DUT response provides a very effective reduction in the $f_{\max}$ search space, it has been shown [22], [23] using large scale industry data that the correlation between the ring oscillator behavior and the circuit behavior also depends on which process corner the device falls; different ring oscillators have different sensitivities to process parameters. Hence, once there is a process shift (abrubt or gradual), the correlation information becomes invalid and thus needs to be updated. And, within-die variations are also increasing, particularly for the threshold voltage [24], rendering the correlation between ring oscillator response and the circuit response even less reliable.

Two important aspects of prior work in this domain are (a) the readings from the paths that are tested are not utilized in subsequent searches even though this information would capture the within-die variations, and (b) there has been no systemic way that can track and adapt with respect to process shifts, which can occur from lot to lot, but also from wafer to wafer. This work aims to address these two missing pieces.

III. Proposed Method

Our goal in this work is to facilitate the use of information from multiple sources to narrow down the $f_{\max}$ search range for a given supply voltage level and to provide a mechanism for tracking and adapting with respect to process shifts.

To achieve this goal, we need to develop a statistical formulation to model the correlation between a set of measurements that have been conducted (i.e. ring oscillator frequencies, or $f_{\max}$ measurement from other circuits or other supply levels) to a set of measurements that have not yet been conducted.

In this multi-variate model, we propose to incorporate any measurement taken from the circuit under test so as to make use of all the clues that the circuit provides. This adaptive approach is demonstrated in Figure 1. Initially, we use the information from on-chip sensors (e.g. a ring oscillator), as proposed by other researchers [22], [23], to narrow down the original search space (Figure 1(a)). After the $f_{\max}$ search at the first supply voltage, this information is included and the search space for the second $f_{\max}$ search is narrowed further. In this manner, it is possible to reduce the search overhead for each subsequent datapoint (Figure 1(b)).

Figure 2 provides the results of the proposed multivariate statistical framework on an ISCAS-85 benchmark circuit c1908, visually illustrating the $f_{\max}$ (the figure provides the ranges for T_{min}, which can be reciprocated to compute $f_{\max}$ values) search range reduction delivered by the proposed method. The original search span (blue solid line) is first shifted right once the RO measurement data is fed in, moving the search range closer to the actual $f_{\max}$ value for the fast V_{dd} level (vertical line). This range gradually narrows down upon the use of measurements from the slow V_{dd} level (dashed green line) at first, and subsequently from the measurements of the nominal V_{dd} level (dash-dotted purple line).

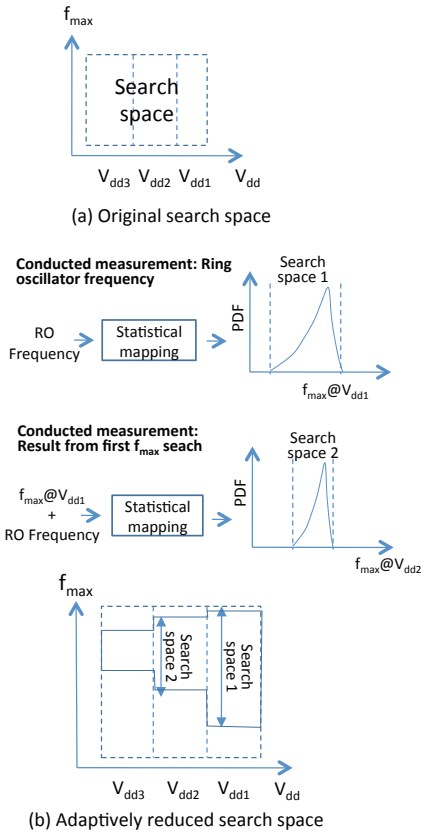

Fig. 1. Proposed concept for fmax search space reduction.

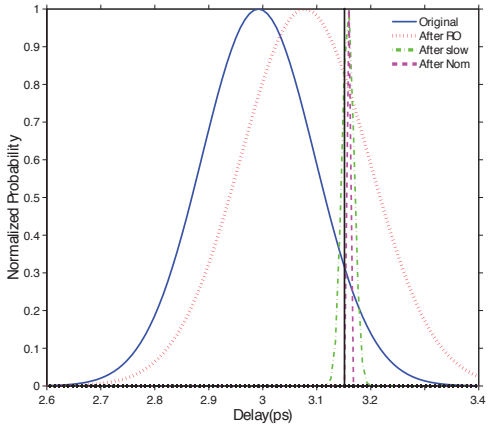

Fig. 2. Adaptive reduction of search space via reuse of previous measurements.

An important challenge in any such correlation technique is that the characterization process, within which we learn the statistical correlations among the parameters of interest, is a snapshot of the manufacturing process. Unfortunately, regardless of how much information is collected on the statistical characteristics of the devices initially, this information eventually becomes invalid, at least partially, due to changes in the underlying process parameters. In order to maintain a high test efficiency and quality level, characterization data should be maintained up-to-date.

	Nom	DtD	WD
L_{eff}	50nm	15nm	1.5nm
V_{th}	180mV	78mV	8mV

TABLE I
NOMINAL VALUES AND VARIATION AMOUNTS FOR THE PROCESS VARIABLES

Our proposed statistical mapping tools based on the concept of growing neural networks and reproducing Hilbert kernel space formulation (RHKS) [25], [26]. The RHKS method is commonly used in information theory to map a nonlinear multi-parameter space to a high dimensional linear space through function, h(x). In this new linear space, estimation can be performed easily with a number of linear transformations [25], [26]. This statistical tool is able to grow with very small overhead as new data points become available.

IV. Experimental Results

A. Process Model

We have used Synopsys Hspice Monte Carlo simulations (MCS). The process parameters are assigned based on the 45nm predictive model [27]. Length (L_{eff}) and Threshold Voltage (V_{th}) were varied globally and locally to depict the die-to-die (DtD) and within-die (WD) variations. The nominal values and variations in length and threshold voltage are given in Table I. The circuits are characterized for three supply voltage levels. The nominal supply level corresponds to 1V whereas the slow V_{dd} level corresponds to 0.82V and the fast V_{dd} level corresponds to 1.19V.

Critical path delay for ISCAS-85 circuits and ring oscillator period were measured for each MCS; to compute the passing frequency (f_{max}), we have used path delay patterns generated by Synopsys Tetramax (ATPG tool) for the longest paths identified by Synopsys Primetime (static timing analysis tool). The ring oscillator that we used has a NAND gate, followed by $2n$ inverter stages. MCS were done for three levels of Vdd: nominal, slow and fast. The estimated speed of the circuit was calculated by a Matlab code, which implements our statistical framework, based on the ring oscillator data and previous readings from MCS. Actual speed from MCS and the estimated speed are used to compute the RMS error.

B. Results

In this section, we present the search range reduction results delivered by the proposed adaptive technique that can reuse information from the previous measurements. We have picked a few ISCAS-85 combinational benchmark circuits, and ran extensive Hspice Monte Carlo simulations, some of which resulted in convergence problems and prolonged the experimentation even further; we utilized the results of 90 MCS for training our statistical tool and 10 MCS for computing the results that we present herein. For the final version of the paper, we are planning to extend our experimentation to a larger set of benchmark circuits and provide more results.

We present the results in Table II. The first column provides the benchmark circuit name, while the second and the third columns provide the RMS error in the predictions by the

Circuit	Linear Fit [23]	Proposed	Reduction
c432	29	16	1.8x
c1908	51	27	1.9x
c3540	60	16	3.8x

TABLE II
RMS ERROR IN PREDICTING MAX DELAY (PS) AND REDUCTION IN SEARCH SPACE.

commonly used linear fit approach [23] and the proposed progressive search method, respectively, for the fast $V_{\rm dd}$ level. The results are presented for the maximum delay in terms of ps, which can be reciprocated to compute the RMS error for $f_{\rm max}$. Finally, the last column provides the reduction in search range offered by the proposed progressive method over the linear fit method.

It should be noted that the errors in predicting $f_{\rm max}$, however large, do not result in incorrect characterization of the circuit. However, a large error results in longer iterations for the $f_{\rm max}$ search, thus longer test times. For the benchmarks whose results we have evaluated, the proposed statistical model based on multiple readings from the CUT consistently presents with smaller RMS error compared with a linear fit. This improvement is almost 2x for the smaller c432 and c1908, and 4x for our largest benchmark c3540. Note that the proposed method becomes more effective for larger benchmark circuits, which bodes well for cost savings on much larger industrial designs. Most importantly, this benefit can be reaped free of any overhead in test development/application or area.

V. CONCLUSIONS

We propose an adaptive approach where previously collected readings and measurements are reused to predict the $f_{\rm max}$ value of the same part. For this purpose, we have developed a multi-variate statistical framework that is capable of taking in the information regarding readings and measurements from the previous $V_{\rm dd}$ levels, and performing the statistical mapping via the RHKS method. The proposed framework runs in the background without incurring any test overhead.

We show that the proposed cost-free flow-nonintrusive approach is capable of narrowing down the search range, providing commensurate savings in test time/cost for binning the tested parts or for computing the $V_{\rm dd}/f_{\rm max}$ information for circuits that support DVS. We expect even higher savings for larger-sized industrial designs. Furthermore, the proposed framework supports a quick removal of outdated information, and is thus capable of recovering from process shifts.

REFERENCES

[1] S. Dighe, S. Vangal, P. Aseron, S. Kumar, T. Jacob, K. Bowman, J. Howard, J. Tschanz, V. Erraguntla, N. Borkar, V. De, and S. Borkar, "Within-die variation-aware dynamic-voltage-frequency scaling core mapping and thread hopping for an 80-core processor," in *IEEE ISSCC*, 2010, pp. 174–175.

[2] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," in *CM/IEEE International Symposium on Low Power Electronics and Design*, 2007, pp. 38–43.

[3] D. Ernst, S. Das, S. Lee, D. Blaauw, T. Austin, T. Mudge, N. S. Kim, and K. Flautner, "Razor: circuit-level correction of timing errors for low-power operation," *IEEE Micro*, vol. 24, no. 6, pp. 10 –20, 2004.

[4] S. Benner and O. Boroffice, "Optimal production test times through adaptive test programming," in *IEEE International Test Conference*, 2001, pp. 908–915.

[5] E. Yilmaz, S. Ozev, and K. M. Butler, "Adaptive test flow for mixed-signal/RF circuits using learned information from device under test," in *IEEE International Test Conference*, 2010, pp. 1–10.

[6] R. Daasch, J. McNames, D. Bockelman, K. Cota, and R. Madge, "Variance reduction using wafer patterns in IDDQ data," in *IEEE International Test Conference*, 2000, pp. 189–198.

[7] R. Daasch, K. Cota, J. McNames, and R. Madge, "Neighbor selection for variance reduction in IDDQ and other parametric data," in *IEEE International Test Conference*, 2001, pp. 92–100.

[8] M. Rehani, R.Madge, K. Cota, and R.Daasch, "Statistical post processing at wafersort," in *IEEE VLSI Test Symposium*, 2002, pp. 69–74.

[9] R. Madge, C. Macchetto, V. Rajagopalan, B.H.Goh, R. Daasch, and C. Schuemeyer, "Screening minVDD outliers using feedforward voltage testing," in *IEEE VLSI Test Symposium*, 2002, pp. 673–682.

[10] R. Daasch, K. Cota, J. McNames, and R. Madge, "In search of the optimum test set," in *IEEE International Test Conference*, 2004, pp. 203–212.

[11] U. Ingelsson and B. Al-Hashimi, "Investigation into voltage and process variation-aware manufacturing test," in *IEEE International Test Conference*, 2011, pp. 1 –10.

[12] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Design Automation Conference*, 2003, pp. 338 – 342.

[13] A. Datta, S. Bhunia, J. H. Choi, S. Mukhopadhyay, and K. Roy, "Profit aware circuit design under process variations considering speed binning," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 7, pp. 10–18, July 2008.

[14] S. Idgunji, V. Chandra, C. Pietrzyk, I. Iqbal, R. Aitken, and G. Yeric, "An embedded process monitor test chip architecture," in *IEEE International Conference on Microelectronic Test Structures*, 2010, pp. 121–127.

[15] F. Klass, A. Jain, and G. Hess, "An embedded process monitor test chip architecture," in *IEEE International Conference IC Design and Technology*, 2009, pp. 203–206.

[16] K. Brand, S. Mitra, E. Volkerink, and E. McCluskey, "Speed clustering of integrated circuits," in *International Test Conference, 2004*, 2004, pp. 1128 – 1137.

[17] J. Lee, D. Walker, L. Milor, Y. Peng, and G. Hill, "IC performance prediction for test cost reduction," in *IEEE Semiconductor Manufacturing Conference Proceedings*, 1999, pp. 111 –114.

[18] J. Zeng, M. Abadir, G. Vandling, L. Wang, A. Kolhatkar, and J. Abraham, "On correlating structural tests with functional tests for speed binning of high performance design," in *International Test Conference*, 2004, pp. 31 – 37.

[19] B. Cory, R. Kapur, and B. Underwood, "Speed binning with path delay test in 150-nm technology," *IEEE Design Test of Computers*, vol. 20, no. 5, pp. 41 – 45, 2003.

[20] T. McLaurin, "Creating structural patterns for at-speed testing: A case study," *IEEE Design Test of Computers*, vol. PP, no. 99, p. 1, 2012.

[21] J. Chen, J. Zeng, L.-C. Wang, J. Rearick, and M. Mateja, "Selecting the most relevant structural fmax for system fmax correlation," in *VLSI Test Symposium*, 2010, pp. 99 –104.

[22] A. Gattiker et. al., "Data analysis techniques for CMOS technology characterization and product impact assessment," in *IEEE International Test Conference*, 2006, pp. 1–10.

[23] M. Bhushan, A. Gattiker, M. B. Ketchen, and K. K. Das, "Ring oscillators for CMOS process tuning and variability control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10–18, February 2006.

[24] N. Drego, A. Chandrakasan, and D. Boning, "Lack of spatial correlation in MOSFET threshold voltage variation and implications for voltage scaling," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 2, pp. 245 –255, 2009.

[25] J.-W. Xu, A. Paiva, I. Park, and J. Principe, "A reproducing kernel hilbert space framework for information-theoretic learning," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 5891 –5902, 2008.

[26] P. Bouboulis and S. Theodoridis, "Extension of wirtinger's calculus to Reproducing Kernel Hilbert Spaces and the complex kernel LMS," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 964 –978, 2011.

[27] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, 2006.