Wafer-Level Adaptive V_{min} Seed Forecasting

Deepika Neethirajan*, Constantinos Xanthopoulos*, Sirish Boddikurapati†, Amit Nahar† and Yiorgos Makris*
*Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX USA, 75080
†Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, TX 75243

***Abstract*—To combat the effects of process variation in modern, high-performance integrated Circuits (ICs), various post-manufacturing calibrations are performed. These calibrations often aim to bring the device within its specification limits and make sure it abides with current technology standards. Moreover, with the increasing popularity of mobile devices that depend on finite energy sources, energy consumption has introduced another constraint. As a result, post-silicon calibration is often performed in order to identify the optimal operating voltage (V_{min}) of a given Integrated Circuit. This calibration is a very time-consuming process that requires the device to be tested in a big range of voltage inputs. In this paper, we propose a machine learning-based methodology to reduce the cost of performing the V_{min} search, by identifying an optimal starting point (seed) for each wafer.**

***Index Terms*—post-silicon calibration, adaptive, test-cost reduction**

I. Introduction

Recent advancements in semiconductor technology have facilitated the industry to produce high-performance ICs at a relatively low cost, suitable for the consumer market. However, these advancements have also magnified the impact of process variations and their ensued effects in reliability and yield. Therefore, nowadays, post-silicon calibration plays a major role in fine-tuning all the key performance parameters of a fabricated device, thereby reducing the effects of process variation. One major pitfall of performing post-silicon calibration is that it requires numerous test measurements and adjustments that take up a significant chunk of the total test time. These increased test times contribute to the manufacturing cost and hinder the profit margins for new products.

Mainly due the popularization of mobile consumer devices an increased concern for power consumption has been introduced. These devices rely on finite energy sources thus their battery life per charge plays a major factor to their market success. Manufacturers, in order to address this need, while continuing to push the envelope in performance, are forced to employ post-silicon calibration techniques. A common such technique for reducing the power consumption on certain devices involves the identification of the minimum operating voltage V_{min} and the corresponding subsequent tuning. Each Device Under Test (DUT) is tested within a range of allowed operating voltages, until the optimum voltage in terms of power consumption voltage is identified. This calibration process is often referred to as V_{min} search and typically is performed as shown in Figure 1.

The search must start from V_{start} and then it proceeds iteratively, depending on the type of search, until all test patterns have been tested and the minimum acceptable voltage is reached. For every test pattern iteration the DUT is tested against the last known V_{min} and if it passes it moves on to the next pattern otherwise it triggers a V_{min} search for the failing pattern. This is repeated until the optimum V_{min} is reached and stored within the device. Depending on the number of test pattern, the search type and the resolution with witch voltages are tested when a V_{min} search is triggered, the overall testing time can increase significantly.

In this work we will propose a machine learning-based approach that adapts the starting voltage of the V_{min} search per wafer according to its e-test signature. This would allow for significant test time savings without affecting the yield and with a minimal power consumption overhead.

II. Related Work

Several researchers have suggested various post-production calibration techniques that shed light on calibrating the performance parameters to be well within the specification limits. Process variations introduced during various stages of manufacturing (e.g., lithography, thermal treatments, etc.,) propose a great challenge as the industry is moving towards smaller nodes. Hence it becomes the responsibility of post-silicon calibration phase to identify the optimum operating conditions by altering the specification parameters within agreeable limits. Both iterative and adaptive calibration methods have been explored in recent times to help improve yield.

The approach in [7] speeds up the trim code search by using machine learning based methodology to predict the binary trim seed code for each wafer. The predicted trim seed code will function as a starting point for the trimming algorithm. This approach considers the median of trim codes of all dies as an optimal starting point of the search. Post-silicon trimming helps to center the key performance parameters that might have shifted due to process variations. In [8], authors propose an adaptive methodology to cut down trim time using machine learning by effectively predicting the trim lengths of on-chip laser trimmable resistors. This technique focuses on creating a function based on the spatial coordinates of the die which are used in expressing the length of the trim code as a function.

In [5] an on-chip self healing methodology using tuning knobs has been proposed. This method relates pre-silicon and post-silicon measurements for the purpose of post-silicon calibration to overcome large scale process variations. A midpoint alternate test method has been proposed as a cost effective post-silicon calibration technique by using a single

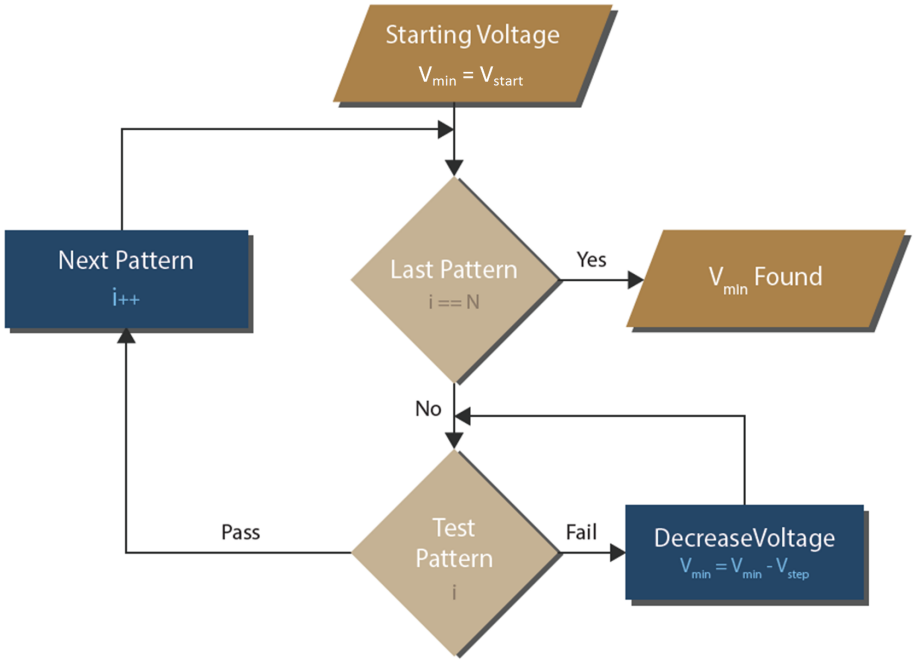

Fig. 1: Typical V_{min} Search

alternate test based model [4]. This method comes with a cost model that compares midpoint alternate test methods alongside other prominent calibration methods in order to establish the effectiveness of the approach. Likewise, in order to substantiate our goal of achieving minimum test cost, we have also developed a cost function to include every step involved in identifying the optimal operating voltage.

When we discuss the cases of post-silicon calibration, we cannot ignore the importance of e-test measurements and their role in understanding the impact of process variations on the manufacturing process. e-tests are electrical measurements extracted at select locations across the wafer by using Process Control Monitors (PCMs) included on wafer scribe lines.In [2] e-test measurements are used to forecast parametric yield to aid in ramping up the production during fab to fab product migration. A regression function models the relationship between e-test and probe test measurements. Similarly, [1] emphasizes on capturing the wafer signature from e-test signature vector which is modelled to predict the suitable test flow for a wafer. On a per-lot basis, the e-test signature vector for each wafer is used to build a model which will eventually predict and dynamically adapt to a suitable test flow process.

Our goal in this paper is to predict the V_{min} seed code using dynamic and fast approach of letting the parameters of the algorithm automatically adapt to the silicon being tested. The key difference between the approaches mentioned in [8], [7] and our approach is that an additional key constraint of power consumption has been introduced. In order to achieve the adaptive search algorithm, we exploit the e-test measurements to identify the search parameters across the wafer without compromising the yield and power consumption.A set of statistical features extracted from e-test measurements and their combinations have been used to predict the starting point of the search.

III. Proposed Methodology

Our methodology aims at reducing the overall V_{min} search time without affecting the production yield. To achieve this, without interfering with current test-floor logistics and processes, we seek to adaptively alter the search parameter values as a function of the silicon's signature. In order to simplify the adoption in production of the proposed methodology, we focused on wafer-level adaptation instead of at die level which would have introduced further complexity.

As in the studies mentioned in Section II, e-tests or Wafer Acceptance Tests (WAT), produce a very characteristic signature for each wafer under test, suitable for wafer-level adaptive methods. Another benefit of utilizing the e-tests is since all calibration steps are performed in a later insertion, there is sufficient time for any adaptive decisions to be made without stalling the production line.

Figure 2 shows an overview of the flow for the proposed approach, where there are two main phases, the training and production phase. During the training phase, a set of wafers is used for the extraction of the model features from the e-tests and the target voltages. The devices from these early wafers, have been calibrated using current practices. Once the feature extraction step is completed, these vectors are then used to train a number of regression models, corresponding to each target parameter.

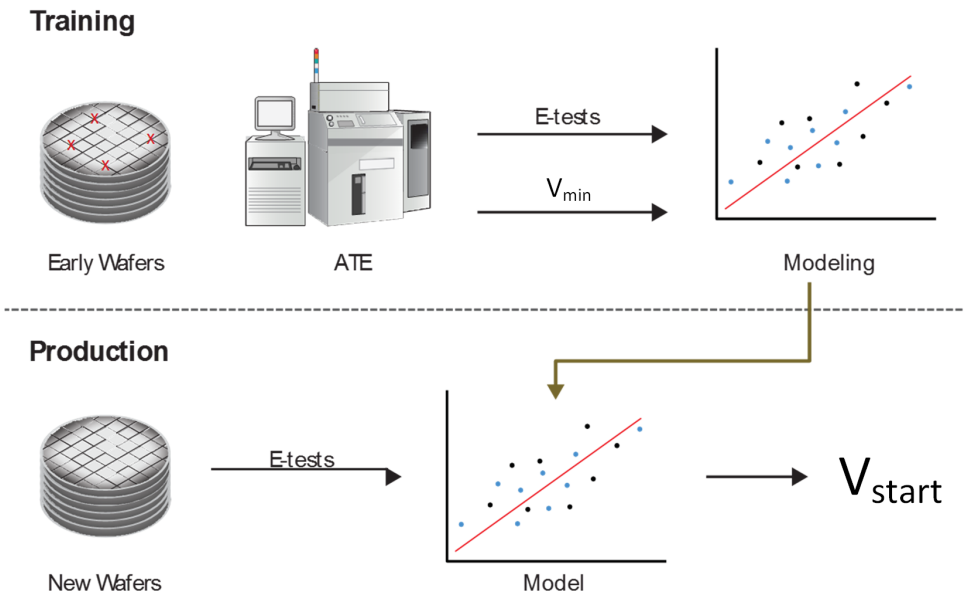

Fig. 2: Proposed Approach

During the production phase of the proposed methodology, the model will be used to predict the target voltages based on the measurements collected in e-test insertion of each wafer. These voltages will then be used during the V_{min} calibration for each device on the same wafer.

A. Feature Extraction

The first step in both phases of the proposed methodology is feature extraction, where the goal is to generate the features with which we will train our model. As mentioned above, these features are generated using the e-test measurements for each particular wafer. To compact the feature vector length and sufficiently represent the complete wafer, the e-test measurements across all wafer sites are aggregated using statistics. To extract the central tendency, dispersion and skewness of each e-test measurement the mean, variance and skewness statistics from all e-test sites are computed. This feature vector serves as a signature of the effects process variations had in the production of each wafer.

During the training phase of the regression model the target voltage values also need to be generated according to the V_{min} calibration that was performed for each die in the early wafers that are used for training. The selection of the target value affects the performance of the proposed approach both in terms of savings as well as power consumption overhead.

For linear search we predicted the V_{start} of the search process. If a device fails at V_{start}, the search will begin from the highest possible voltage and proceed downwards until the device passes. If the device fails at a specific step down voltage value, then we identify the voltage value at the previous step as the V_{min}.

B. V_{start} Selection

For the linear search Figure 3 shows how test time and power consumption is affected by predicting the various parameters in relation to the V_{min}. As shown, for a given die in a wafer, if the predicted V_{start} is below the actual V_{min} the search is the same, starting from the V_{high} and decrease. The reason for this is that the resulting V_{min} would remain the same, thus no power consumption overhead and it will take the same number of steps. On the other hand, when the predicted V_{start} is over the actual V_{min} the search will return the provided V_{start} at a cost of one step, since that will be a passing voltage and the V_{min} search will not get triggered. The difference between the actual V_{min} and the resulting sub-optimal is translated to power consumption overhead. As shown, power consumption overhead and test time savings of the proposed method are directly related to each other as well as to the selection of the V_{start}.

C. Modeling: Multiple Adaptive Regression Splines

One of the key component of building the model to predict the V_{start} of the search algorithm is the implementation of Multivariate Adaptive Regression Splines (MARS) algorithm [3]. MARS algorithm helps the methodology by modelling the wafer level search seed code as a function of e-test signature vector. The MARS model is a powerful and flexible regression model that helps in modelling the relationships between using few variables in high dimensional datasets. It takes advantage of additive and interactive relationships between variables thereby resulting in using fewer variables to represent a high dimensional dataset. Due to the aforementioned advantages, MARS algorithm has been used in many test cost reduction approaches [1] [6].

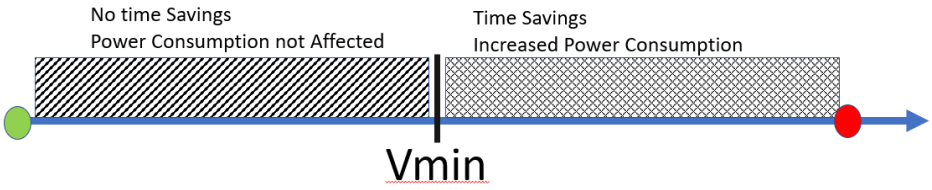

Fig. 3: Wafer-level V_{start} Selection

IV. Expected Results

An industrial dataset consisting of high performance devices was provided by Texas Instruments Inc. The devices provided in the dataset were calibrated based on the current methodology. The devices consisted of their e-test measurements and their respective optimal operating voltages (V_{min}). The industrial dataset was split into training and testing sets. From the available e-test measurements, the statistical measurements mean, variance and skew were extracted to train the machine learning model during the training phase. The effectiveness of such model was evaluated by performing a leave-one-out experiment for all wafers in the dataset.

From the preliminary results, it is evident that the proposed adaptive methodology of identifying the starting point (V_{start}) of the V_{min} search shows considerable improvement with respect to test time savings. We were able to see approximately 80% test time savings with only a 5% power overhead. This is a significant improvement when compared to the current approach setting the default high voltage as the starting point of the search. Based on the preliminary results conducted on the linear search technique, there are ways that we can extend this approach to be applied for other popular search algorithms. It might provide us with more test time savings with a minimal power consumption overhead.

V. Conclusion

We have analyzed a machine learning based intelligent approach to predict the starting point of the optimum voltage search. This approach is capable of being combined with several other post-silicon calibration techniques. By applying this technique, increase in test time and cost in terms of the Automatic Test Equipment (ATE) usage can be minimized. This approach once again proves that the e-test measurements contain considerable amount of key information that can be used to improve the yield of devices and reduce manufacturing costs at the same time.

Acknowledgment

This work was supported in part by Semiconductor Research Corporation (SRC).

References

[1] A. Ahmadi, A. Nahar, B. Orr, M. Past, and Y. Makris. Wafer-level process variation-driven probe-test flow selection for test cost reduction in analog/rf ics. In *2016 IEEE 34th VLSI Test Symposium (VTS)*, pages 1–6, April 2016.

[2] A. Ahmadi, H. G. Stratigopoulos, A. Nahar, B. Orr, M. Pas, and Y. Makris. Yield forecasting in fab-to-fab production migration based on bayesian model fusion. In *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 9–14, Nov 2015.

[3] J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[4] N. Kupp, H. Huang, P. Drineas, and Y. Makris. Post-production performance calibration in analog/rf devices. In *2010 IEEE International Test Conference*, pages 1–10, Nov 2010.

[5] S. Sun, F. Wang, S. Yaldiz, X. Li, L. Pileggi, A. Natarajan, M. Ferriss, J. O. Plouchart, B. Sadhu, B. Parker, A. Valdes-Garcia, M. A. T. Sanduleanu, J. Tierno, and D. Friedman. Indirect performance sensing for on-chip self-healing of analog and rf circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 61(8):2243–2252, Aug 2014.

[6] P. N. Variyam, S. Cherubal, and A. Chatterjee. Prediction of analog performance parameters using fast transient testing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 21(3):349–361, March 2002.

[7] C. Xanthopoulos, A. Ahmadi, S. Boddikurapati, A. Nahar, B. Orr, and Y. Makris. Wafer-level adaptive trim seed forecasting based on e-tests. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, May 2017.

[8] C. Xanthopoulos, K. Huang, A. Poonawala, A. Nahar, B. Orr, J. M. Carulli, and Y. Makris. Ic laser trimming speed-up through wafer-level spatial correlation modeling. In *2014 International Test Conference*, pages 1–7, Oct 2014.