Wafer-Level Adaptive $V_{\text{min}}$ Calibration Seed Forecasting using Inter-$V_{\text{min}}$ Correlation

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—High-performance mobile devices have limited power sources and hence, functioning at low power levels is an important constraint for their success. Process variation causes such critical specification parameters of high-performance devices to deviate from their ideal performance. Hence it is necessary to use post-silicon calibration to identify the minimum operating voltage ($V_{\text{min}}$) for a device. The device under study has the capability to operate at four different speeds. Hence, the device has four different $V_{\text{min}}$ values that have a linear relationship to the speeds associated with it. Recent studies have shown that the current $V_{\text{min}}$ search can be improved by modeling the starting point of the search as a function of the e-test signatures per wafer. In this paper, we expand on the $V_{\text{min}}$ search calibration seed forecasting by taking advantage of the relationship between the different operating voltages for the Device Under Test (DUT). The proposed method predicts the starting voltage of the search and the highest possible voltage level as a function of the other operating voltages associated with the device. The effectiveness of the proposed methodology is demonstrated on an industrial dataset provided by Texas Instruments.

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 to 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 $V_{\text{min}}$ 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 $V_{\text{min}}$ search and typically is performed as shown in Figure 1.

The search must start from $V_{\text{start}}$ 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 $V_{\text{min}}$ and if it passes it moves on to the next pattern otherwise it triggers a $V_{\text{min}}$ search for the failing pattern. This is repeated until the optimum $V_{\text{min}}$ is reached and stored within the device. Depending on the number of test patterns, the search type and the resolution with which voltages are tested when a $V_{\text{min}}$ search is triggered, the overall testing time can increase significantly.

Previous research has shown the possibility of using e-test measurements to predict the starting point of the $V_{\text{min}}$ search. In this paper, we extend on that idea and propose a machine learning-based approach that adaptively predict the $V_{\text{min}}$ search starting voltage value and the highest possible voltage value for a device at a specific speed by taking advantage of its correlation with another operating speed’s $V_{\text{min}}$.

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 [5] 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. Post-silicon trimming helps to center the key performance parameters that might have shifted due to process variations. In [6], 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. A midpoint alternate test method has been proposed as a cost effective post-silicon calibration technique by using a single alternate test based model [3]. 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.

In [7], a machine learning-based approach is used to adaptively predict the starting voltage of the $V_{\text{min}}$ search per wafer according to its e-test signature. This method has provided significant test time savings without affecting the yield and with a minimal power consumption overhead. The key difference between the approaches mentioned in [6], [5], [7] and our approach is that the device under study in our approach operates under four different speeds. 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. The device that we studied had the capability to operate under four different speeds. Our goal in this paper is to predict the $V_{\text{min}}$ search range values using the correlation between the different $V_{\text{min}}$ values corresponding to the different speeds of the device.

III. PROPOSED METHODOLOGY

Our methodology aims at reducing the overall $V_{\text{min}}$ 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. In this paper, instead of e-tests our focus is on the relationship between the different $V_{\text{min}}$ values. We make use of the $V_{\text{min}}$ value corresponding to one speed level of the device to predict the starting point and highest possible value of the $V_{\text{min}}$ search for another speed level associated with the device.

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 voltage values corresponding to a specific speed level i.e., speed level 2 or 3 or 4 and the target voltage value related to another speed level i.e., speed level 1. The devices from these early wafers, have been calibrated using current practices. The signature vectors are then used to train a number of regression models, corresponding to each target parameter. During the production phase of the proposed
methodology, the model will be used to predict the target voltages based on the measurements collected in voltage levels corresponding to a speed level of each wafer. These voltages will then be used during the $V_{\text{min}}$ calibration for each device on the same wafer.

A. Target Voltages

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

For the linear search, Figure 3a. shows how test time and power consumption are affected by predicting the $V_{\text{start}}$ for which the search for $V_{\text{min}}$ commences, compared to the current approach. As shown, for a given die in a wafer, if the actual $V_{\text{min}}$ is above the predicted $V_{\text{start}}$, the search time remains the same, as the search starts from $V_{\text{high}}$ and decreases the voltage until we reach the same $V_{\text{min}}$. Since this will result in the same $V_{\text{min}}$, there will not be any power consumption overhead. On the other hand, when the predicted $V_{\text{start}}$ is over the actual $V_{\text{min}}$, the search will return the provided $V_{\text{start}}$ at the cost of one step, since this will be a passing voltage and the $V_{\text{min}}$ search will never get triggered. The difference between the actual $V_{\text{min}}$ and the sub-optimal $V_{\text{start}}$ to which the device will be calibrated, will also induce some power consumption overhead.

When both $V_{\text{start}}$ and $V_{\text{high}}$ are predicted, as shown in Figure 3b., the location of the $V_{\text{start}}$ relatively to the actual $V_{\text{min}}$ has similar behavior as above. The difference here is that when the actual $V_{\text{min}}$ is between the predicted $V_{\text{start}}$ and $V_{\text{high}}$, test cost savings are attained by the reduction in the number of steps needed when starting from the adjusted $V_{\text{high}}$. Moreover, if the actual $V_{\text{min}}$ is higher than the predicted $V_{\text{high}}$, a provision can be implemented in the test program to ensure that the proposed approach will not affect yield, at the cost of two extra steps. This exception can be triggered when both the predicted $V_{\text{start}}$ and $V_{\text{high}}$ fail the first test, thus resulting in a rollback to the current static approach for that particular die.

B. Modeling: Multiple Adaptive Regression Splines

One of the key component of building the model to predict the $V_{\text{start}}$ and $V_{\text{high}}$ of the search algorithm is the implementation of Multivariate Adaptive Regression Splines (MARS) algorithm [2]. 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
TABLE I: Experiments Performed

<table>
<thead>
<tr>
<th>Search Type</th>
<th>Adapt</th>
<th>Alias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$V_{\text{start}}$ (Proposed)</td>
<td>$L_0$</td>
</tr>
<tr>
<td></td>
<td>$V_{\text{start}} &amp; V_{\text{high}}$ (Proposed)</td>
<td>$L_2P$</td>
</tr>
</tbody>
</table>

Fig. 4: $V_{\text{min}}$ Search Savings for Speed 1 when predicted using correlation with Speed 2

variables to represent a high dimensional dataset. Due to the aforementioned advantages, MARS algorithm has been used in many test cost reduction approaches [1] [4].

IV. 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 the four different $V_{\text{min}}$ values corresponding to each speed level. The industrial dataset was split into training and testing sets. From the available dataset, the $V_{\text{min}}$ value of speed levels 2, 3 and 4 were separately used to train 3 different machine learning models with the independent variable for all three models being the $V_{\text{min}}$ value of speed level 1 during the training phase. This order was followed based on the industrial recommendations.

Table I shows the list of experiments we performed in order to evaluate the effectiveness of the proposed method for the linear and binary search. The first set of experiments seeks to identify the cost savings and power consumption overhead for the linear search. This includes only the adjustment of $V_{\text{start}}$ (L1P) as well as both the adjustment of $V_{\text{start}}$ and $V_{\text{high}}$ (L2P).

From the preliminary results, it is evident from Figure 4 that the proposed adaptive methodology of predicting the starting point ($V_{\text{start}}$) of the $V_{\text{min}}$ search and the $V_{\text{high}}$ in the L2P method shows considerable improvement with respect to test time savings. We were able to see approximately 80% test time savings with a 10% 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. Similarly from the Figure 5, we can see that L2P approach shows a savings of 35% test time savings with a 10% power overhead. 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 as well as the highest possible voltage value 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.

ACKNOWLEDGMENT

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

REFERENCES