Wafer-Level Adaptive V_{min} Calibration Seed Forecasting

Constantinos Xanthopoulos*, Deepika Neethirajan*, 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 USA, 75243

***Abstract*—To combat the effects of process variation in modern, high-performance integrated Circuits (ICs), various post-manufacturing calibrations are typically performed. These calibrations aim to bring each device within its specification limits and ensure that it abides by current technology standards. Moreover, with the increasing popularity of mobile devices that usually depend on finite energy sources, power consumption has been introduced as an additional constraint. As a result, post-silicon calibration is often performed to identify the optimal operating voltage (V_{min}) of a given Integrated Circuit. This calibration is time-consuming, as it requires the device to be tested in a wide range of voltage inputs across a large number of tests. In this work, we propose a machine learning-based methodology for reducing the cost of performing the V_{min} calibration search, by identifying the optimal wafer-level search parameters. The effectiveness of the proposed methodology is demonstrated on an industrial dataset.**

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

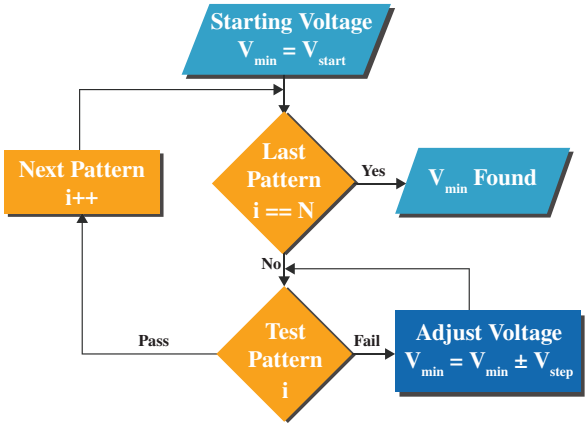


Fig. 1: Typical V_{min} search flow

I. Introduction

The increasing demand for consumer electronics has driven the industry towards producing high-performance ICs at low cost. This has been facilitated by a variety of advancements in all stages of the manufacturing process. Most of these advancements seek to mitigate the impact of process variation and its ensued effects in reliability and yield. Post-fabrication calibration constitutes such a remedy, aiming to fine-tune several critical specification parameters that affect the performances of each device. Post-silicon calibration can often lead to excessive test times due to the numerous test measurements and adjustments that are required. These increased test times contribute to the manufacturing cost and hinder the profit margins for new, high-performance consumer market 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, and their battery life per charge plays a major factor in their market success. Manufacturers, while continuing to push the envelope in terms of performance, are forced to address this need by employing post-silicon calibration techniques. A common such technique for reducing the power consumption on certain devices involves identification of the minimum operating voltage V_{min} and corresponding subsequent tuning. Each Device Under Test (DUT) is tested within a range of allowed operating voltages until the voltage resulting in the minimum power consumption is identified. This calibration process is often referred to as "V_{min} search" and is typically performed as shown in Figure 1.

The search starts from V_{start} and 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 permanently stored in the device. Depending on the number of test patterns, the search type and the resolution with which voltages are tested, the overall testing time can increase significantly.

In this work, we propose a machine learning-based approach that adapts the limits or the starting voltage of the V_{min} search per wafer, according to its e-test signature. The proposed methodology enables significant test time savings without affecting the yield and with minimal power consumption overhead.

The remainder of this paper is organized as follows. In Section II, we discuss some of the details and limitations of the current V_{min} search implementations. An overview of related studies can be found in Section III. Then, in Section IV, we describe the details of the proposed machine learning-based adaptive V_{min} search. Experimental results demonstrating the effectiveness of the proposed methodology on an industrial dataset are presented in Section V and conclusions are drawn in Section VI.

978-3-9819263-2-3/DATE19/©2019 EDAA

II. Current V_{MIN} Search

Depending on the test limitations of a specific DUT, as well as those arising from the Automatic Test Equipment (ATE), either a linear search or a binary search can be used to find the V_{min}. Although implementing a binary search is always preferable with respect to test time, this is not always possible. As previously shown in Figure 1, a search is triggered only when a given test pattern fails, in a nested loop architecture. The first loop iterates between the various test patterns, while the second loop iterates between different voltage values, according to the search algorithm.

Another parameter that depends on the test setup is the direction of the search. Due to test stability issues, for some devices and ATEs it is preferred to start the search from a low voltage value and gradually increase it until the search criterion is met; for other devices, starting from a higher voltage and descending is, instead, required. In some cases, the direction of the search does not affect the quality of the measurements and a random order search can be performed. For the first type of setups, some form of linear search is conducted while for the second, binary search is commonly used.

A. Linear Search

Several parameters affect the test time when a linear search is used. These are carefully selected during the characterization of each device to ensure high yield and reduced cost. The first set of such parameters is the voltage range ($[V_{low}, V_{high}]$), which is usually determined by the specifications of each device. In modern devices, this range is often decided based on industry standards that the device has to abide by (e.g., USB and HDMI). Another parameter is the starting voltage of the search, which is usually defined as either the lower or the higher limit of the search; thus, it also determines the direction of the search.

For example, when a search is triggered with V_{start} equal to V_{high}, for each iteration of the sweep the voltage is decreased by V_{step} until the test fails and returns the last working voltage. In order to save time, a well-known technique is to split the search into two stages, namely the *coarse* and the *fine*. During the coarse stage, the stepping voltage value (V_{coarse_step}) is larger as compared to the fine stage, in order to quickly determine a smaller range wherein the higher-resolution fine search will be performed. The fine stepping voltage (V_{fine_step}) is chosen according to the desired fidelity and the resolution limits of the ATE.

In order to speed-up the V_{min} search, one can adjust the above parameters to reduce the number of search steps performed. This adjustment can be made either once, in a static way, using historical statistics after an early set of wafers has been produced and characterized, or periodically, to adjust for process shifts. When the adjustments do not reflect the silicon being tested, this can either result in sub-optimal V_{min} identification if only the V_{start} is adjusted, or in yield loss when the limits are also adjusted. *In this work, we seek to adaptively predict these parameters as a function of the wafer signature and, thus, limit the power consumption limit while guaranteeing no yield loss.*

B. Binary Search

In the case of a binary search, the only parameters that are determined are the voltage range ($[V_{low}, V_{high}]$) and the minimum voltage step which is used as a stopping criterion.

To speed the binary search up, we can also adjust the range of the search, thus eliminating a number of the halfing steps. As in the case of the linear search, the risk of adjusting the search limits at the wafer level is that a subset of die may fall outside that range and consequently affect power consumption or yield loss.

III. 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.) pose a substantial challenge as the industry is moving towards smaller nodes. Hence, it becomes the responsibility of the post-silicon calibration phase to identify the optimum operating conditions by altering the parameters within agreeable limits. Authors in [9] have shown the importance of taking into account process variation when preparing the test patterns and adapting the overall test process according to these variations.

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. In case of complex analog and mixed-signal circuits, solutions involving Bayesian Model Fusion (BMF) aim to reduce test cost by minimizing the measurements involved in pre-silicon and post-silicon stages [4].

When we discuss the case of post-silicon calibration, we cannot ignore the importance of e-test measurements and their role in understanding the impact of variation on the semiconductor manufacturing process. E-tests are electrical measurements performed 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 and 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] focuses on capturing wafer-level variation from an e-test signature which is then used to predict the most 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, eventually, predicts and dynamically adapts the test flow process.

Post-silicon trimming helps to center the critical performance parameters that might have shifted due to process variations. The approach in [7] speeds up the trim code search by using a machine learning-based methodology to predict the trim seed code for each wafer. The predicted trim seed code

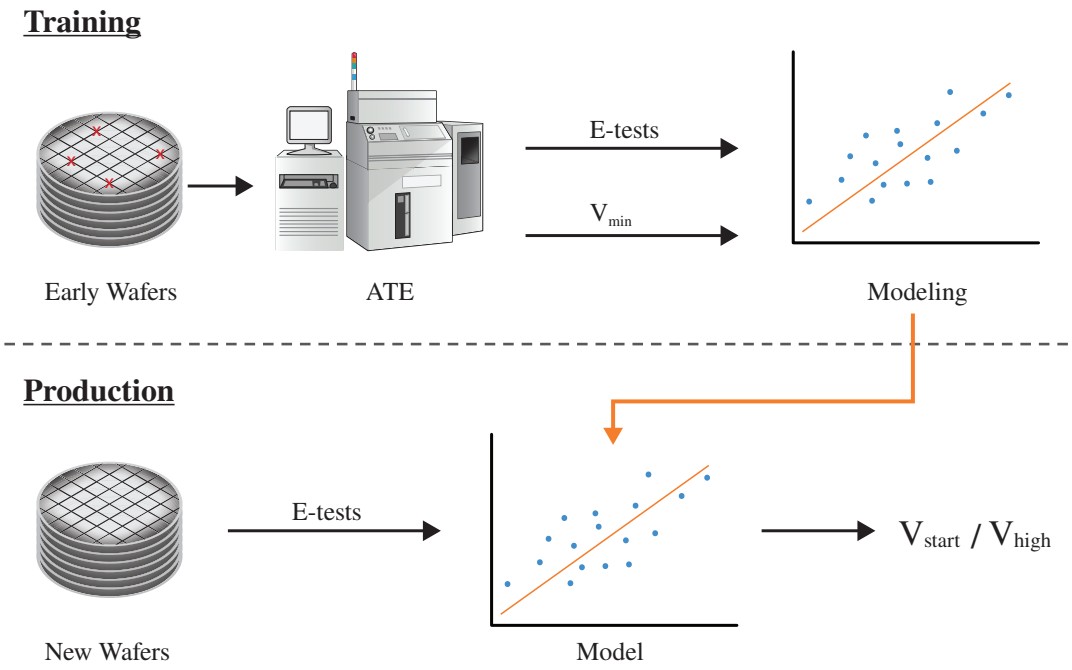


Fig. 2: Proposed Approach

functions as a starting point for the trimming algorithm. This approach considers the median of trim codes of all die in a wafer as an optimal starting point of the search, as it seeks to minimize the overall number of steps per wafer. In [8], the authors propose an adaptive methodology to cut down the trim time using machine learning by effectively predicting the trim lengths of on-chip laser trimmable resistors. This technique focuses on utilizing the wafer-level spatial correlations to extrapolate from the trim length of the sampled locations to the remaining die on the wafer.

In this paper, our goal is to predict the V_{min} seed code using a fast, dynamic 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 V_{min} search algorithm, we exploit the e-test measurements to identify the search parameters across the wafer without compromising yield.

IV. 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 process signature. In order to simplify the adoption of our proposed methodology in production, we focused on wafer-level adaptation instead of die-level adaptation, which would have introduced further complexity.

As in the studies mentioned in Section III, e-tests or Wafer Acceptance Tests (WAT), produce a characteristic signature for each wafer under test, suitable for driving wafer-level adaptive methods. This signature reflects how a specific wafer has been affected by process variations. A key benefit of utilizing e-tests is due to the fact that all calibration steps are performed at a later test insertion (i.e. probe or final test), thus allowing 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, which includes two main phases, the training phase and the production phase. During the training phase, a set of wafers is used for extraction of model features from the e-tests and the target voltages. The devices from these early wafers have been calibrated using current practices, as described in Section II. Once the feature extraction step is completed, these 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 is used to predict the target voltages based on the measurements collected during e-test insertion of each wafer. These voltages are then 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

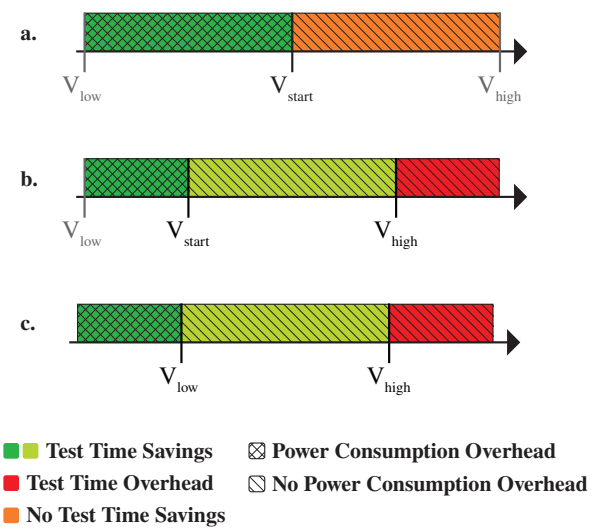


Fig. 3: Wafer-level V_{start} and V_{high} selection

length and sufficiently represent the complete wafer, the e-test measurements across all wafer sites are aggregated using several statistics. To extract the central tendency, dispersion, and skewness of each e-test measurement, we compute the mean, variance and skewness statistics from all e-test sites on a wafer. This feature vector then serves as a signature of the effects that process variations had in the production of each wafer.

B. Target Voltages

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

When both V_{start} and V_{high} are predicted, as shown in Figure 3b., the location of the V_{start} relatively to the actual V_{min} has similar behavior as above. The difference here is that when the actual V_{min} is between the predicted V_{start} and V_{high}, test cost savings are attained by the reduction in the number of steps needed when starting from the adjusted V_{high}. Moreover, if the actual V_{min} is higher than the predicted V_{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_{start} and V_{high} fail the first test, thus resulting in a rollback to the current static approach for that particular die.

When a binary search is implemented, there is no risk of affecting power consumption by adjusting the search range. Rather test cost savings may be reduced, as compared to the static binary search implementation. Figure 3c. shows the case where adjusted V_{low} and V_{high} have been chosen. Any V_{min} between these two limits will be found in a smaller number of steps, due to the reduced search range. On the other hand, when the actual V_{min} is outside the predicted limits, this would result in a two-step penalty before the test program rolls back to the static limits, similarly to the above-mentioned case.

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} and V_{low}, depending on the search type. Given that the objective of our proposed methodology is to reduce the test cost by maintaining certain levels of power consumption overhead, the selection of the targets has to be made a function of the power consumption overhead. To achieve this, we can select maximum power consumption overhead limits, ranging from 0.1% to 19%, and identify the corresponding V_{start} or V_{low} per wafer in the training set. Modeling these power consumption overhead bounded V_{min} values as a function of the e-test features, as explained above, would enable product engineers to reduce the test cost by selecting an acceptable power overhead risk when either linear or binary search is used.

C. Modeling: Multiple Adaptive Regression Splines

One of the key components of building the model to predict the V_{start} of the search algorithm is the implementation of the Multivariate Adaptive Regression Splines (MARS) algorithm [3]. MARS is a powerful and flexible regression model that helps in representing relationships between a 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, the MARS algorithm has been used in many test cost reduction approaches (e.g., [1] [6]). The MARS algorithm supports the proposed methodology by modelling the wafer-level V_{start} value as a function of the e-test signature vector.

V. Experimental Results

We used an industrial dataset of more than 800 wafers with more than 500 devices each to evaluate our method. This particular product-ATE setup combination does not allow the implementation of a binary V_{min} search; thus, for each device in the dataset, the traditional linear V_{min} search was performed and the resulting voltages were recorded. The V_{min} search is

TABLE I: Experiments Performed

Search Type	Adapt	Alias
Linear	-	L0
	V_{start} (Historical)	L1H
	V_{start} (Proposed)	L1P
	V_{start} & V_{high} (Proposed)	L2P
Binary	-	B0
	V_{low} & V_{high}	B2P

performed in two stages, with two vastly different step sizes in order to limit the overall number of search steps. In addition to the dataset, the V_{low}, V_{high}, V_{coarse_step}, and V_{fine_step} voltages were used in order to be able to define realistic test cost and power consumption functions. For the test cost, the number of search steps was used as a proxy for the actual test time due to its low computational complexity.

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_{start} (L1P) as well as both the adjustment of V_{start} and V_{high} (L2P). Similarly, for the second set of experiments the test cost and power consumption is calculated when binary search is employed (B2P).

For all experiments presented below, the dataset was split into two sets, the training set, composed of 75% of the wafers, and the validation set, composed of the remaining 25%. The historical search parameter values were computed based on the training set and applied on the validation set. The L2P and B2P methods require two MARS models to be trained in order to predict the V_{start} and V_{high} or the V_{low} and the V_{high}, respectively.

A. Linear Search Test Cost Savings

Figure 4 shows the percentage of cost savings achieved for each of the selected power consumption bounds, ranging from 0.1% to 19%. The L1H curve shows that the use of the historical V_{start} and V_{high} values can achieve up to 90% cost savings compared to the current static methodology (L0). Unfortunately, these savings start after the 3-4% power consumption overhead bound and slowly ramp-up to 80% when more than 8% overhead is permitted.

On the other hand, our L1P proposed methodology, which adaptively predicts the V_{start} voltage for each wafer according to its e-test signature, offers significant savings at much low percentages of power consumption overhead. Moreover, the 80% test time savings point is reached at approximately 4% power overhead; compared to the static approach, this allows greater flexibility to adapt to market needs and product specifications. The red line in Figure 4 shows even more significant savings, for the lowest power consumption overhead bound at 0.1%, with a jump to 48% compared to the current approach. Although, adjustment of both V_{start} and V_{high} significantly reduces the cost for power consumption overhead percentages

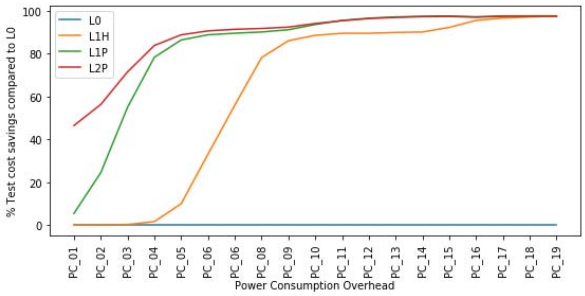


Fig. 4: Linear V_{min} search savings

below 4%, the benefits against L1P approach get diminished for higher percentages and thus the L1P might be preferable due to its relative simplicity.

B. Binary Search Test Cost Savings

Due to test setup and ATE-related limitations, the device used in this study did not permit implementation of a binary V_{min} search. Nevertheless, we studied theoretically how the proposed method would have affected the performance of a binary V_{min} search.

Figure 5 shows the relative cost savings of the proposed approach when predicting the V_{low} and V_{high} voltages based on e-tests, compared to a static range binary search (B0).

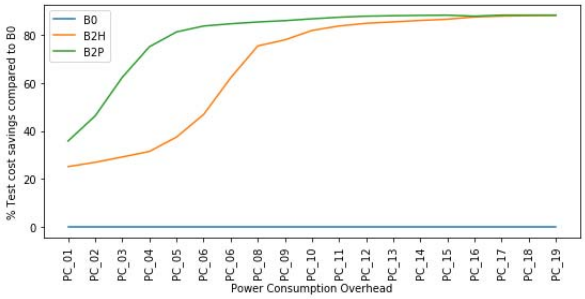


Fig. 5: Binary V_{min} search savings

Similarly to the linear search, a comparison was made against historical V_{low} and V_{high} voltages (B2H), where savings starting from 24% were achieved when power consumption overhead was bounded to 1%. For more substantial power consumption overhead rates, these savings continued to increase until they flattened out at approximately 10% overhead. Comparatively, for the proposed approach where the search parameters where adaptively predicted based on the e-tests (B2P), similar order savings where achieved at 4% additional power consumption.

C. Power consumption

To evaluate the effectiveness of the methods listed in Table I, the power consumption overhead was calculated based on the outcome of each V_{min} search for all die in our validation set. Figure 6 shows the comparison of these methods with the power consumption of the current linear V_{min} search as

the baseline. The purple line shows the expected worst-case overhead for which the search parameters where calculated or predicted. The historical based linear search results in significantly less power consumption on average, compared to the expected line, partly due to its limited cost savings, as presented in Figure 4. Parallel to that, but with relatively higher power consumption penalty, the proposed adaptive linear and binary searches achieve approximately an average of 2% less power consumption than the respective target.

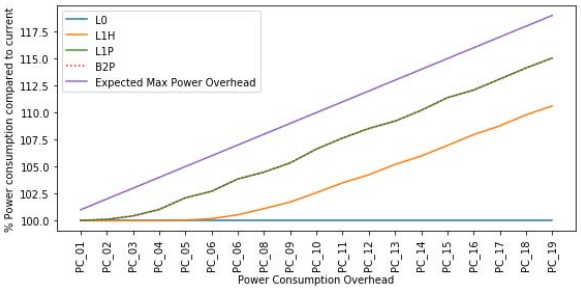
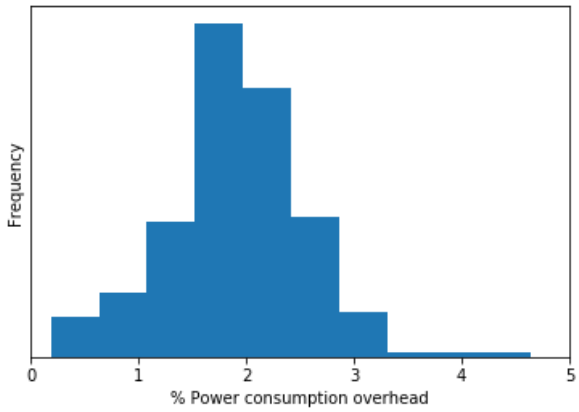
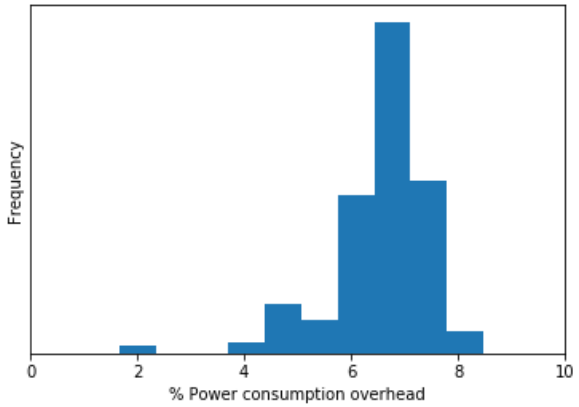


Fig. 6: Power Consumption

To assess how the prediction error affects the power consumption, the distributions of two different bound rates are shown in Figure 7. These plots are generated for the single adaptive parameter linear search approach (L1P). For Figures 7a. and 7b., the distributions are bounded to 5% and 10% of power consumption overhead, respectively. As shown, the power consumption for both rates is below the respective bound values which ensures that the proposed methodology will not impose power consumption overhead above the expected threshold.



(a) Power consumption capped at 5%



(b) Power consumption capped at 10%

Fig. 7: Distribution of Power Consumption for two different caps

VI. Conclusion

We presented an adaptive, wafer-level, machine learning-based approach to reduce the V_{min} calibration costs, based on the e-test signature of a wafer. This is achieved by adjusting the V_{min} search parameters to reduce the number of search steps without affecting the production yield, while at the same time controlling the incurred power consumption overhead. The effectiveness of the proposed method was demonstrated on an industrial dataset of more than 400,000 devices, with significant savings, even under the strictest power consumption overhead constraints.

VII. Acknowledgement

This work is supported by Semiconductor Research Corporation (SRC) task 2712.031 through The University of Texas at Dallas Texas Analog Center of Excellence (TxACE).

References

[1] A. Ahmadi, A. Nahar, B. Orr, M. Pas, and Y. Makris. Wafer-Level Process Variation-Driven Probe-Test Flow Selection for Test Cost Reduction in Analog/RF ICs. In *IEEE VLSI Test Symposium (VTS)*, 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 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2015.

[3] J. H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 1991.

[4] X. Li, F. Wang, S. Sun, and C. Gu. Bayesian Model Fusion: A Statistical Framework for Efficient Pre-silicon Validation and Post-silicon Tuning of Complex Analog and Mixed-signal Circuits. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2013.

[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*, 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*, 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 *IEEE International Symposium on Circuits and Systems (ISCAS)*, 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 *IEEE International Test Conference (ITC)*, 2014.

[9] E. Yilmaz, S. Ozev, and K. M. Butler. Efficient Process Shift Detection and Test Realignment. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2013.