Wafer-Level Adaptive Trim Seed Forecasting Based on E-Tests

Constantinos Xanthopoulos*, Ali Ahmadi*, Sirish Boddikurapati†, Amit Nahar†,
Bob Orr†, and Yiorgos Makris*

*Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080
†Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, TX 75243

***Abstract*—Post silicon trimming is extensively used to counter the effects of manufacturing process variation on certain critical electrical parameters of an integrated circuit (IC). Usually, trimming is performed iteratively by adjusting the resistance value of a trim circuit to specific discrete values. Test programs represent those values by codes and apply common search algorithms in order to find a code which makes a device (optimally) compliant to its design specifications. Consequently, manufacturing yield is increased significantly, yet at the expense of added test time and complexity. In this work, we introduce a novel methodology wherein a trained multivariate model is used to predict, adaptively for each wafer, the optimal starting point of the algorithm that searches for the trim code. Thereby, we seek to minimize the number of code changes that the search algorithm has to perform and, by extension, the overall trim time. In order to provide this prediction prior to wafer sort, so that simplicity of test-floor logistics does not get compromised, the predictive model is built using electrical test (e-test) measurements, which are available before wafer sort, and is trained through measurements from a set of early wafers. Effectiveness of the proposed method in reducing trim time is demonstrated on 370 wafers of an high performance device manufactured by Texas Instruments.**

I. Introduction

As IC manufacturing nodes shrink to allow for better area utilization and power consumption, process variations increasingly affect the performance parameters of fabricated devices, often forcing them outside their specification range. In many cases, these fluctuations can affect compliance to industry standards, such as HDMI, USB or electromagnetic guidelines set by regulatory agencies. In an effort to counter these effects, analog IC designers in collaboration with test engineers introduce adjustable structures in various locations of the design. These adjustable structures are capable of centering the performance distributions and, as a result, increasing production yield. Such structures often include registers whose value controls specific electrical parameters (i.e., voltages, currents or capacitances), depending on the design requirements.

From a test program perspective, the binary values of those registers are represented by their decimal notation, commonly termed *trim codes*. In order to perform post-fabrication calibration, also known as *trimming*, the trim code space needs to be explored, starting from an initial code called *seed*, and until an acceptable value for the performance parameter of interest is achieved. In the event that none of the possible trim codes causes the desired IC performance to fall inside its specification range, the device is marked as failing. Although trimming significantly improves yield and enables high volume manufacturing of high performance devices, it also drastically increases test time.

Recently, significant effort has been invested in challenging the current practices of the trimming procedure, during which every die is trimmed independently, either in a sequence or in parallel when multi-probing is used. These methods seek to utilize existing spatial correlation of a trimming parameter code across die on the same wafer, as well as interdependent correlation among different trimming parameters on the same die, in order to predict trim codes rather than search for them. In [1] the authors leveraged wafer-level spatial correlation, by training an intelligent model from a sample of die locations across each wafer and predicting the laser trim lengths for the untrimmed die. Alternatively, the authors of [2] introduced a methodology for exploiting the correlation between interdependent trims in order to expedite the trimming process. In another direction, authors in [3] introduced techniques to optimize the search algorithm by using parallelism and on-die measurements to cut down trim time. Similarly, in [4], trim time was reduced by employing a hybrid search algorithm to reach the target trim code in a small number of steps. In many cases, test-engineers are able to apply more sophisticated techniques in order to further reduce test time. Some of these techniques include the implementation of a binary search, linear fit with only a few die-level measurements and others. However, the utilization of such techniques is not possible for applications where the electrical parameter of interest needs to be very accurately matched (e.g. high performance amplifiers) or the trim is designed to exhibit hysteresis. In these cases, a simple linear search has to be used instead, starting from the seed, towards one direction, until an acceptable value is reached.

In this work, we introduce a novel methodology for speeding up the trimming procedure through intelligent adaptation of the trim seed code. Our proposed approach is orthogonal to the above-mentioned techniques and can be easily combined with them. More specifically, we predict the trim seed code of each wafer by examining its e-test[1] measurement vector. The

[1]By the term e-test we refer to electrical measurements, which are typically performed on a few select locations across the wafer, using process control monitors (PCMs) included on the wafer scribe lines.

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

trimming seed code is predicted per wafer and is used as the starting point of the trimming procedure for every die on this wafer. Starting with an intelligently chosen seed code reduces the number of steps required to reach a target performance. The key component of our methodology is a predictive model which correlates multivariate e-test data to trim codes. Since e-test data has been used successfully for parametric yield estimation [5], final test outcome prediction [6] and test flow selection [7], our conjecture is that it comprises sufficient information regarding the impact of process variation on the performances of a device to allow for accurate selection of the optimal trim seed of a wafer.

II. Proposed Approach

Consider a device that is currently being manufactured and for which we have collected the e-test measurements from several wafers, as well as the final trimming code corresponding to the performance parameters of interest from all devices contained in each of these wafers.

We are interested in building a methodology that uses the data above to reduce the trimming time by intelligently determining the starting point (i.e., trimming seed) of the trim search algorithm, rather than starting from 0. The ideal case would be to have one trim seed per die; however, this complicates test logistics and requires an understanding of intra-wafer variation. A more practical approach would be to have one trimming seed per wafer for each trimming parameter, thus the trim process for that parameter and for all die across the wafer would start with that seed. To do so, the first challenge is to find the best trimming seed for a wafer, such that it minimizes the trimming steps of all die across that wafer. Statistically, the *median* of a distribution is the point which minimizes the sum of distances to all its points. Therefore, the median of the trim codes of all die on a wafer is the optimal starting point for the trimming procedure for this wafer.

A simple and straightforward approach is to compute the median of the trimming code across all training data. In this approach, for each performance parameter, the trim codes of all devices in the training data are considered and their median is used as the trimming seed for all future wafers.

Although the historic median approach is simple to use and implement, it has limited accuracy due to process variations and process shifts. Moreover, relying on a *single* historical estimate for the entire future production does not take into account the lot-to-lot or wafer-to-wafer variations. Therefore, we are interested in utilizing e-test data, which reflects process variations, to handle process shifts, drifts and jumps. More importantly, this enables us to provide *per wafer* adaptive trimming seed codes and use them in our proposed methodology.

Our conjecture for this work is that there exists a relationship between the trimming seed and the e-test measurement patterns, since the purpose of e-test is to reflect process variations that lead to the need for a trimming procedure. This relationship does not have a known closed-form mathematical expression. For this reason, it is approximated using

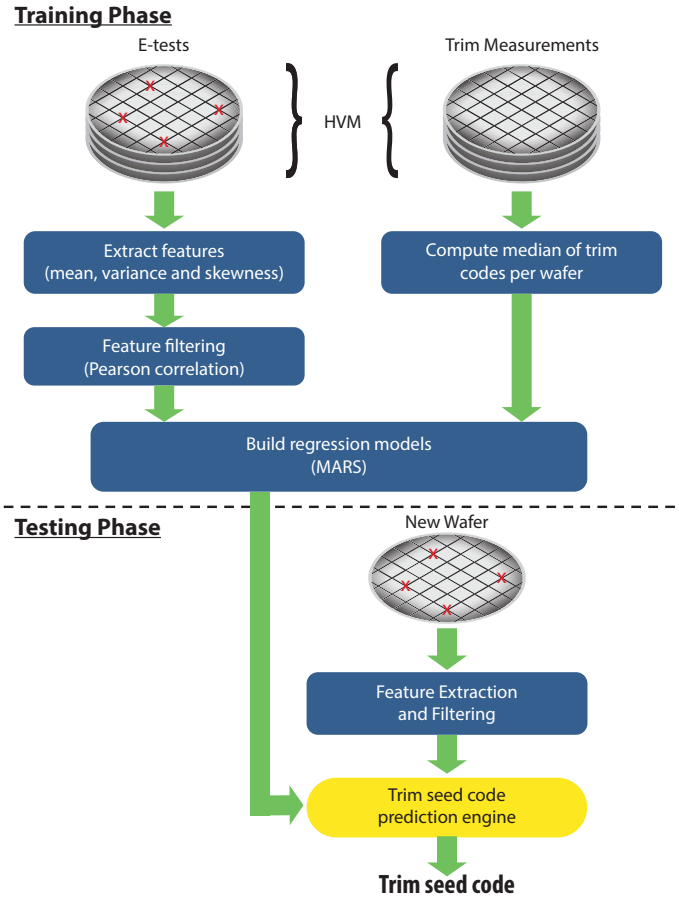

Fig. 1: Overview of proposed method: training phase involves e-test feature vector generation, median trim computation and construction/training of regression models. Testing of a new wafer involves e-test feature vector generation and processing by trained regression models for predicting trim seed code

a regression function. The data collected and mentioned in the beginning of this section is used to build this regression function that predicts the trimming seed code for each wafer from its e-test measurement pattern.

One of the challenges in building these regression functions is the dimensionality of e-test data. E-tests contain many types of parameters, mainly focusing on simple physical/electrical characteristics. For some of these measurements, there is no correlation with the trimming codes. Accordingly, to avoid spurious autocorrelations and gain better insight from the e-test data, we apply a filtering stage prior to training the regression models. To do so, we compute the Pearson correlation coefficient (PCC) of all e-test measurements to the median trim code of each wafer and we only keep e-tests which show significant correlation. Next, a multivariate regression model is used to capture the underlying correlation between e-tests and trim codes. Once the training phase is finished, the e-test feature vector for each new wafer is computed and fed into the trained system, which predicts the appropriate trim seed code for this wafer. Figure 1 shows the steps of this method.

A key component of our proposed approach is building the regression functions which model the wafer-level trim seed

codes as a function of the multivariate e-test feature vector. in this work, we use Multivariate Adaptive Regression Splines (MARS) [8], which was also used in [6], [7] and several other test cost reduction methods in the past [9]. The MARS model is a regression model using basis functions as predictors in place of original data. We build a MARS regression model with e-tests as our predictor matrix and the optimal trimming seeds as the dependent variables.

III. Experimental results

In order to experimentally evaluate the effectiveness of the proposed methodology, we use actual production data from a high performance device currently in high volume manufacturing (HVM) production by Texas Instruments[2]. The dataset comes from 370 wafers (15 lots), each of which contains approximately 500 die. E-test is performed on 21 sites across the wafers, with 500 measurements obtained from each site. For every die across the wafer, final trim codes for three different high precision trims are also provided. All these trims satisfy the criteria of monotonicity and are performed using linear search, therefore constitute perfect candidates for the proposed methodology. Additionally, for each e-test measurement we compute the mean, variance and skewness across all 21 e-test sites. The objective of our method is to train three regression models that use the e-test feature vector to predict the ideal initial point of the trimming procedure for all die across the wafer. In order to assess the accuracy of the trained model, we split our dataset into 8 lots for training and 7 for testing, maintaining their chronological order of production.

A. Trim Code Analysis

To gain better insight on the provided dataset we visualized the trim code distribution for each of the available trim parameters. The blue bars in Figure 2 show the complete distribution of codes for each trim, respectively, and for all die within our data set. As can be seen from these histograms, significant variation exists in the trim codes, indicating the high impact of process variation on the corresponding electrical parameters. Likewise, it is evident that using the historic median, marked with the yellow vertical lines, would decrease the number of trimming steps, as compared to the naive approach of starting with zero. The most important observation is that the median trim codes per wafer also exhibit high variation, as represented by the green bars. It is also important to notice that, by having the ability to accurately predict the optimal seed code for each wafer, we can reduce the number of search steps, as compared to the historic median. This certainly holds true for trims A and B; however, given the limited variation of the wafer medians for trim C, any wafer-level decision will hardly have any advantage against the historic median. *Given the above, our expectation is that the proposed approach, which aims to predict the wafer optimal seed code, will outperform the historic median based approach.*

[2]Details regarding the device, trim code range and distribution may not be released due to an NDA under which this data has been provided to us.

B. Training

As mentioned in Section II, in order to prepare the training set, the first task is to calculate the wafer medians that will be used as the dependent variable of the MARS model. Then, the e-tests need to be preprocessed as well, via feature extraction and selection. Specifically, feature extraction involves calculation of the first three statistical moments (i.e., mean, variance and skewness) for each wafer by aggregating all 21 e-test sites. Although this method reduces the number of features per e-test to three, as compared to the multi-site vector, it ensures that the distribution characteristics are preserved. Once the aggregation process has been completed, the absolute value of PCC between the e-test and trim codes is calculated and is used to limit the number of predictors for the MARS model. The absolute PCC threshold is set to 25%, such that only the grossly uncorrelated e-tests are excluded. We, then, rely on the MARS training process, which includes a two-stage feature selection procedure, to only retain the highly correlated ones.

C. Results

To evaluate the accuracy of the trained models, we calculated the number of steps that are required in order for the pre-recorded trim code (i.e. ground truth) to be achieved. Along with the proposed approach, the we also evaluated the simple linear search with seed code 0 and the historic median approach, for which the seed has been calculated using the median code of all the wafers in the training set. Additionally, as a baseline, we calculated the number of code changes that it would require if a perfect oracle was used to predict the optimal seed for each wafer in our test set. This is particularly useful in this analysis, since the granularity of our decisions is at the wafer-level; hence, intra-wafer variation will inevitably cause some die to deviate from the median.

The results of this study are shown in Figure 3, where the relative number of code changes, as compared to the Linear search, is visualized for each method across all available trims. The first and rather expected observation is that the savings achieved by using an intelligent trim seed code, as compared to starting from 0, are very profound. This phenomenon is well-known in the industry and shows why the use of an estimate based on historical data is currently the standard approach. With this in mind, the savings achieved by our predictive model, as compared to the standard historic median approach, indicate that we were able to capture and exploit the wafer-to-wafer variability (which the historic median approach is oblivious to) and to more accurately predict the trim seed codes, thereby achieving fewer code changes. Another key point to note is the high accuracy of the proposed model, as demonstrated by the proximity of our savings to the ones achieved by the perfect oracle.

In more detail, the savings of our proposed approach, as compared to the historical median, reach 20% for Trim B and over 5% for Trim A. In contrast, Trim C does not show any tangible improvement relatively to the historical median, which was foreseen due the limited variability of the seed codes. The above-mentioned results are summarized in Table

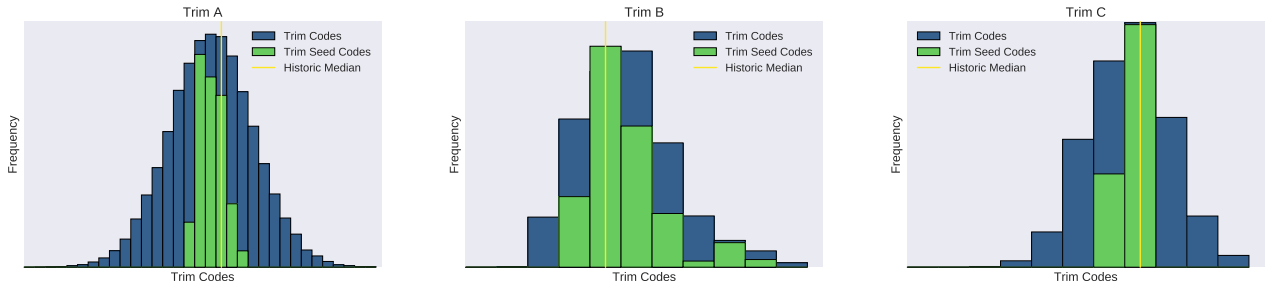

Fig. 2: Complete distribution of trim codes, trim seed code distribution for the training set and historic median for the three considered performances

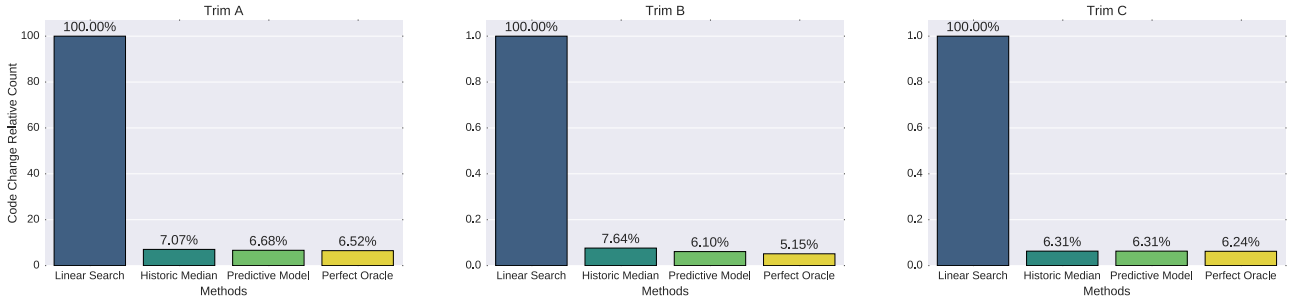

Fig. 3: Relative trim savings of the historic median and the proposed approach in comparison to linear search.

TABLE I: Trim Code Change Savings

Trim	From Linear Search		From Historic Median	
	Regression	Perfect Oracle	Regression	Perfect Oracle
A	93.32%	93.48%	5.51%	7.77%
B	93.90%	94.85%	20.15%	32.59%
C	93.69%	93.76%	0%	1.10%

I, where the relative trim code savings for all three available trims are shown.

IV. Conclusion

We presented a novel methodology for adaptively predicting the optimal seed codes that should be used for trimming a wafer based on the e-test measurements obtained before wafer-sort. As a result of applying our methodology, the ATE minimizes the number of trim code changes required during ATE-based trimming is minimizes and the trimming process is significantly sped up. The underlying conjecture motivating this approach is that the e-test measurements obtained from process control monitors on the scribes of each wafer can provide enough information to predict the distribution of the trim codes and adaptively adjust the search algorithm starting points per wafer. Experimental results on a sizeable number of wafers of an high performance device manufactured by Texas Instruments corroborate this conjecture by demonstrating savings of up to $\approx 20\%$, as compared to the static approach of using the historic median of the trim codes as the seed.

References

[1] 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*, 2014, pp. 1–7.

[2] P. Bongale, V. Sundaresan, P. Ghosh, and R. Parekhji, "A novel technique for interdependent trim code optimization," in *IEEE VLSI Test Symposium*, 2016, pp. 1–6.

[3] R. Mittal, L. Balasubramanian, V. Devanathan, M. Kawoosa, and R. A. Parekhji, "Towards adaptive test of multi-core rf socs," in *IEEE/ACM Design, Automation & Test in Europe Conference & Exhibition*, 2013, pp. 743–748.

[4] R. Mittal, M. Kawoosa, and R. A. Parekhji, "Systematic approach for trim test time optimization: Case study on a multi-core RF SoC," in *IEEE International Test Conference*, 2014, pp. 1–9.

[5] 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*, 2015, pp. 9–14.

[6] N. Kupp, M. Slamani, and Y. Makris, "Correlating inline data with final test outcomes in Analog/RF devices," in *IEEE/ACM Design, Automation & Test in Europe Conference & Exhibition*, 2011, pp. 1–6.

[7] 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*, 2016, pp. 1–6.

[8] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[9] 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*, vol. 21, no. 3, pp. 349–361, 2002.