Wafer-Level Process Variation-Driven Probe-Test Flow Selection for Test Cost Reduction in Analog/RF ICs

Ali Ahmadi*, Amit Nahar†, Bob Orr†, Michael Pas† 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*—We introduce a methodology for dynamically selecting whether to subject a wafer to a complete or a reduced probe-test flow, while ensuring that the concomitant test cost savings do not compromise test quality. The granularity of this decision is at the wafer-level and is made before the wafer reaches the probe station, based on an e-test signature which reflects how process variations have affected this particular wafer. While the proposed method may offer less flexibility than approaches that dynamically adapt the test flow on a per-die basis, its implementation is simpler and more compatible with most commonly used Automatic Test Equipment. Furthermore, unlike static test elimination approaches, whose agility is limited by the relative importance of the dropped tests, the proposed method is capable of exploring test cost reduction solutions which maintain very low test escape rates. Decisions are made by an intelligent system which maps every point in the e-test signature space to either the complete or the reduced test flow. Training of the system seeks to maximize the number of wafers subjected to the reduced flow for a given target of test escapes, thereby enabling exploration of the trade-off between test cost reduction and test quality. The proposed method is demonstrated on an industrial dataset of a few million devices from a Texas Instruments RF transceiver.**

I. Introduction

Continuous pressure for superior performance, along with intensified process variations and non-idealities in the latest semiconductor manufacturing technology nodes, have resulted in stringent limitations in the cost that can be devoted to testing each die, in order to ensure that it functions correctly before it is shipped to a customer. Especially in the analog/RF domain, where industrial practice still relies largely on lengthy test procedures and expensive instrumentation to explicitly measure the performances of a device and compare them to its specifications, test cost reduction has become a crucial requirement for maintaining profitability. Among the various directions which have been explored towards reducing test cost, significant effort has been invested in challenging the practice of subjecting every die in production to the exact same set of tests. Generally termed "adaptive test", methods in this category seek to customize the test process to the needs of a target die, wafer, or lot, anticipating that the benefits from a reduced test flow will outweigh the effort and expenditure required for such customization.

A very simple and commonly practiced approach to test cost reduction is to monitor the relative effectiveness of each test and drop the ones which contribute little or not at all to the overall test effectiveness [1]–[3]. Such decisions are usually static and are easy to implement on the ATE by exclusion of the relevant portion of the test program. However, the agility of such methods is insufficient to support solutions which offer savings yet maintain very low test escapes; essentially, they are bound by the percentage of faulty die that the dropped tests uniquely detect. Advanced versions of this idea, wherein statistical correlation between the dropped and retained tests is leveraged to predict the outcome of the former, have also been proposed [2], [4]–[6]. While additional ATE or external support is required to run the statistical models on-the-fly during test, these methods have demonstrated marked improvement in test quality. Still, the decision models remain static or only infrequently retrained to account for major events which can change the statistical profile of the production.

As a first step towards dynamic test adaptation, re-optimization of the test list on a per-lot basis based on the data obtained from the first few wafers, on which the complete flow is applied, was explored in [7]. Taking adaptation a step further, the method in [8] identifies, through sampling and clustering, wafer regions which have been affected similarly by process variations, and customizes the test list and test order to each such region. While this method was demonstrated in the context of final test, it could be readily applied at probe-test. However, it would complicate test floor logistics, as it would require two passes (for sampling and testing) and ATE support for applying different test programs to each region of the wafer. In fact, any adaptive solution at a finer granularity than the wafer-level would require such support, which is often missing or cumbersome to implement in ATE platforms.

Along a different direction towards eliminating items from the test list, various methods have taken advantage of wafer-level spatial correlation. Specifically, these methods identify test items which exhibit high such correlation and only perform these tests on a small sample of die across the wafer, from which they build the correlation model [9], [10]. These tests are, then, omitted for the rest of the die on the wafer and their value is predicted through the learned model, as a function of die coordinates on the wafer. Extensions to spatio-temporal correlation across an entire lot have also been investigated [11]. Besides being limited only to test items which exhibit spatial correlation, such methods also require a two-pass approach (for sampling and testing) and/or may need to delay the die-level test decisions until the entire wafer or the entire lot has been processed, thereby complicating logistics.

978-1-4673-8454-4/16/$31.00 ©2016 IEEE

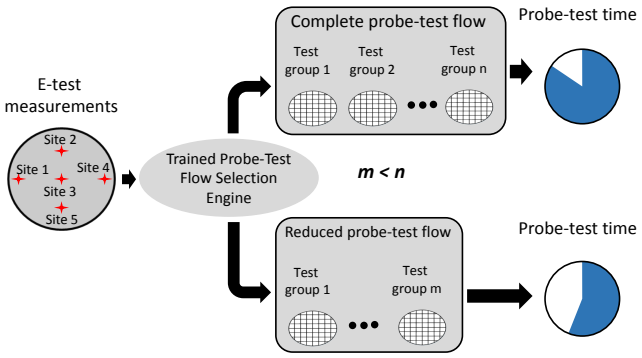

Fig. 1: Wafer-level probe-test flow selection

In this work, we seek to develop an adaptive test flow which combines the advantages of the aforementioned methods, while abiding by the following principles in order to be readily deployable with minimal test operations support:

- Adaptation is limited to a very small number of options: in our case, the choice is between a complete test flow and a reduced version, wherein any number of tests may be omitted.
- The granularity at which test elimination decisions are made is at the test group level. The underlying assumption here is that the bulk of the cost incurred by a test group is related to switching into the appropriate test configuration. Accordingly, the incremental savings of eliminating a few measurements within a group are negligible.
- The granularity of the adaptation decision is at the wafer level, i.e., all die on a wafer are subjected to the same test flow, either the complete or the reduced version.
- Test has to be performed in one pass. In other words, solutions which first apply the reduced test flow and subsequently apply selectively the remaining test items to die for which the decision confidence is low, such as the two-tier test method in [12], are not within scope.
- The decision has to be driven by a signature which reflects how process variations have affected a particular wafer. This is justified by historical evidence documenting that the necessity of a test group is strongly correlated with the operating point of the fabrication process.
- The decision has to be available prior to insertion of the wafer in the probe station and cannot be informed by measurements taken at probe. Inevitably, this leaves e-test[1] as the only source available for capturing the impact of process variations on a particular wafer.

Consistent with the above constraints, an overview of the proposed wafer-level process variation-driven probe-test flow selection method is depicted in Figure 1. The key component of this method is an intelligent system, which is trained to map the e-test measurements obtained from a wafer to a decision regarding application of the complete or a reduced test flow to every die of this wafer. A detailed description of the proposed

[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.

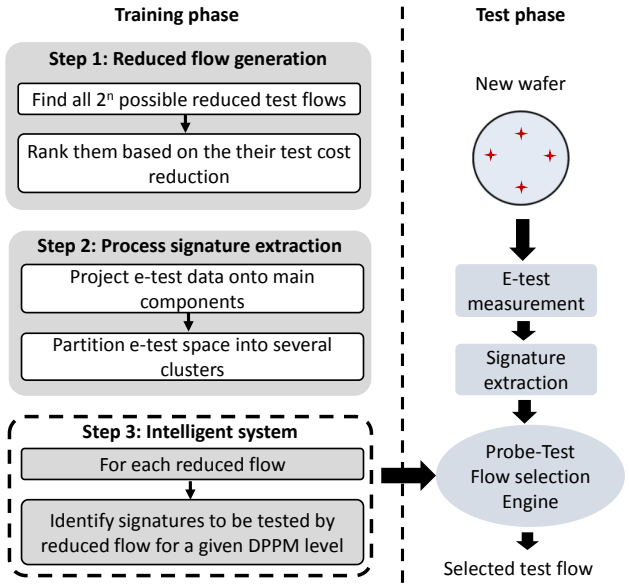

Fig. 2: Steps of proposed method: training involves reduced test flow selection, wafer signature creation, and intelligent system training. Testing of new wafer involves signature computation and processing by trained intelligent system for selection of appropriate test flow

methodology, including identification of the most appropriate reduced test flow and training of the intelligent system to achieve a target test escape rate is presented in Section II. Experimental results demonstrating the effectiveness of the proposed method on a large industrial dataset are presented in Section III and conclusions are drawn in Section IV.

II. Proposed Methodology

As depicted in Figure 2, the three key elements of the proposed method are: (i) identifying an appropriate subset of test groups which will serve as the reduced test flow, (ii) crafting a wafer signature from its e-test measurement vector, and (iii) training an intelligent module to map these wafer signatures to either the complete or the reduced probe-test flow while maintaining test quality within a given DPPM target. Once the training phase is finished, the e-test signature for each new wafer is computed and fed into the trained intelligent system, which selects the appropriate test flow for this wafer. Details of these three components are provided below.

A. Reduced Test Flow Selection

A reduced test flow is a subset of the complete flow, wherein one or more test groups are eliminated. The first challenge that naturally arises is the selection of the test groups which should be omitted in the reduced flow, such that the attained test cost reduction does not compromise test quality beyond a target level of acceptable test escapes. Since the granularity of elimination is at the test group rather than the test item level, it may be possible to exhaustively search the space of solutions. For example, in our experiments we dealt with a set of 14 test groups, thus exhaustively searching in the power-set of 2^{14}

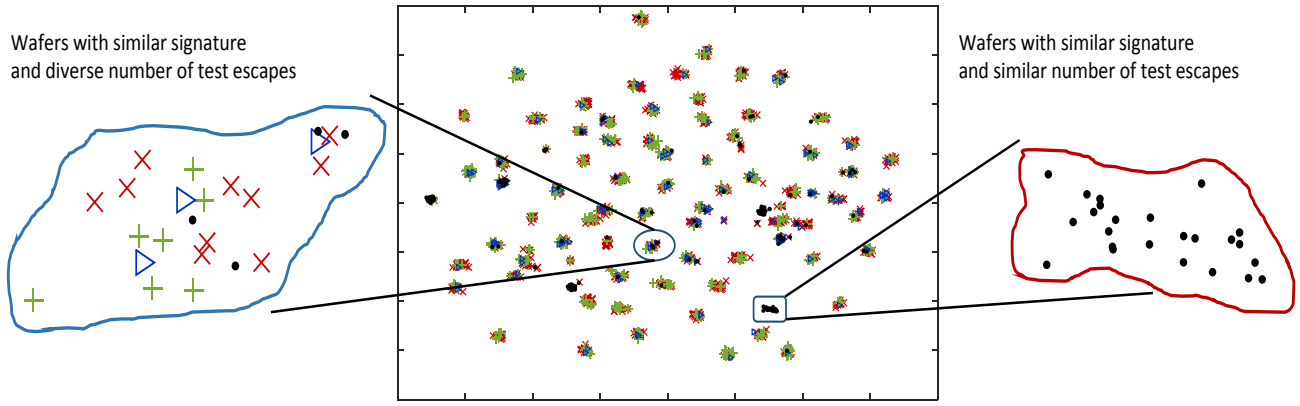

Fig. 3: Projection of e-test data onto the top two principal components

subsets of the complete test flow to find the optimum subset was feasible and chosen due to its simplicity. In case of a large number of test groups, however, this approach will not scale. In this case, heuristic search methods can be employed for effectively searching this space. The use of Genetic Algorithms has been popular in the literature and very successful when applied to this task [6], hence we can readily adopt it when exhaustive consideration is infeasible.

For each reduced flow, j, we consider the associated cost and the number of test escapes when this reduced flow is applied to all wafers in our training set, and we assign a fitness value using:

$$index_j = \frac{t_A - t_{B_j}}{t_A} * pctg_{B_j} \quad (1)$$

where, $\boldsymbol{t_{B_j}}$ denotes the test cost of the j-th reduced flow, $\boldsymbol{t_A}$ denotes test cost of complete test flow and $\boldsymbol{pctg_{B_j}}$ represents the percentage of wafers that can be tested using the j-th reduced test flow, while the total number of test escapes remains below a target DPPM level.

B. Wafer Signatures Based on E-tests

E-tests data contain many types of parameters, mainly focusing on simple physical/electrical characteristics reflecting the position of a wafer in the process space. For some of these measurements, there is no physical connection or reason why they should be correlated with probe-test outcomes or the necessity thereof. Accordingly, to avoid spurious autocorrelations and to gain better insight from our e-test data, prior to crafting a wafer signature based on the e-tests we apply a dimensionality reduction and filtering stage. Specifically, we perform principal component analysis (PCA), which is a commonly used technique for unsupervised dimensionality reduction. PCA projects the data onto a new set of orthonormal components, each of which captures part of the variability of the data. We then retain only the principal components that capture 90% of the information content of the data.

In Figure 3, we provide an example where we project a number of wafers to a 2-dimensional space whose two axis correspond to the two main principal components of the e-tests of these wafers after performing PCA. The various markers used to represent each point indicate different test escape rates when a randomly selected reduced test flow is applied for all wafers. Wafers with the same marker exhibit similar level of test escapes. Two key observation can be made using this figure:

1) Projection of wafers on the e-test space is discontinuous, with most wafers being part of small clusters in this 2-dimensional space. This reflects the fact that the process jumps between a finite number of points.
2) Wafers within each cluster, i.e., with similar e-test signature, do not necessarily exhibit the same test escape rate. This implies that the correlation between device specifications and e-test parameters is complex and there is no simple boundary to separate wafers with high test escapes from wafers with low or zero test escapes. A more elaborate approach is, consequently, required for mapping e-test signatures to the appropriate test flow.

Accordingly, our method partitions the projected e-test space into k clusters. For this purpose, k-means clustering is applied and Gap statistics [13] is used to estimate the number of clusters. Wafer signatures are, then, mapped to the closest cluster and decisions regarding complete vs. reduced test flow are made at the cluster level.

C. Mapping Wafer Signatures to Probe-Test Flows

We now proceed to elaborate on how the intelligent system is trained to map the e-test signature of a wafer to either the complete or the reduced test flow. *Recall that our objective is to save test cost by applying a reduced test flow to a subset of wafers, while keeping test escapes below a given DPPM level.* Evidently, the more wafers we funnel to the reduced test flow, the higher the test cost reduction we can achieve. Thus, our problem is to map the e-test signature space to the appropriate test flow, such that we meet both of the above objectives.

We formulate this problem as an integer linear program (ILP). An ILP consists of a set of variables, which can only assume integer values, a set of linear constraints on these variables, and a cost function which is to be maximized or

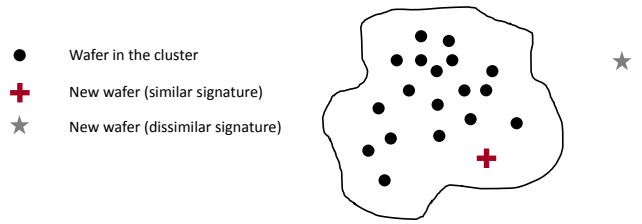

Fig. 4: Tracking process shifts: signatures of wafers belonging to cluster are enclosed by a boundary. New cluster members with signatures within the boundary are considered equivalent, and new members with signature outside the boundary are considered outliers

minimized. In our problem, our constraint is on the total number of test escapes, and our cost function is the maximization of the number of wafers that go through the reduced flow. Our ILP is actually a binary (0-1) version, where the value of each integer variable can only be either 0 or 1. Specifically, in our ILP, the variable α_i is used to indicate whether the wafers that belong to cluster i should be subjected to the complete test (i.e., $\alpha_i = 0$) or to the reduced flow, $\alpha_i = 1$. Suppose that we have a reduced test flow, TF_B, whose test escape vector for training wafers is, TE_B, and whose test cost is t_B. Let also denote the targeted DPPM level as $DPPM_t$. Then our 0-1 ILP is defined as follows:

$$te_i = \sum_{j \in C_i} TE_B^j \tag{2}$$

$$\text{C}_\text{i} : \text{all wafers in the cluster i}$$

$$\begin{aligned} &\text{Maximize} \quad \sum_{i=1}^{k} \alpha_i.card_i \\ &\text{subject to} \quad \sum_{i=1}^{k} \alpha_i.te_i \leq DPPM_t \\ &\alpha_i \in \{0,1\},\ i = 1,\ldots,k \end{aligned} \tag{3}$$

where k is the number of clusters, and te_i and $card_i$ are the total number of test escapes and the cardinality of the i-th cluster, respectively. This procedure is repeated for all candidate reduced flows, each time resulting a mapping between clusters in the e-test space and the appropriate test flow, through the chosen values for the α_i variables. This mapping is learned based on a training set of wafers, on which it ensures maximal test cost reduction while meeting the required test quality.

An additional provision is also incorporated in the intelligent system, in order to adapt to shifts in the process, which may result in previously unseen wafer signatures in the projected e-test space. Specifically, as shown in Figure 4, for clusters which the ILP maps to the reduced probe-test flow, we establish a boundary around the e-test signatures that belong to the cluster. For a new wafer, the distance of its e-test signature from the centers of the clusters is first computed, and the wafer is assigned to the nearest cluster. If the decision for this cluster is to apply the reduced test flow, we perform one more check: if its signature is inside the boundary of that cluster, we follow the recommendation. Otherwise, we assume that despite being nearest to this cluster, the wafer is sufficiently different and we send it to the complete test flow. However, once the tests are performed, we also examine whether the reduced test flow (which is a subset of the complete flow and, therefore, available) would have resulted in test escapes below the acceptable DPPM level. Based on this information, we periodically enhance the set of clusters and retrain the intelligent system to better track the process.

III. Experimental Results

In order to experimentally evaluate the effectiveness of the proposed methodology, we use actual production data from a 65nm analog/RF transceiver currently in high volume manufacturing (HVM) production by Texas Instruments[2]. The dataset comes from 1800 wafers, each of which contains approximately 1500 die. E-test is performed on 9 sites across the wafers, with 54 measurements obtained from each site. On each die, 168 parametric probe-test measurements are obtained, organized in 14 groups. The percentage by which each group contributes to the total test cost is also provided. The objective of our method is to select a subset of the 14 test groups as the reduced test flow and to train an intelligent system which will use the 9 sets of 54 e-test measurements to predict whether a wafer should undergo the complete or the reduced probe-test flow. In our experiments, we use 50% of the data for training and the remaining 50% for validation. Using this dataset, our experiments seek to:

- Confirm that static test group elimination does not have the agility to support reduced test flows while maintaining a test escape rate in the very low DPPM region.
- Determine the upper bound of test cost savings which can be expected through adaptive per-wafer selection between the complete and the reduced test flow, in the ideal scenario where an oracle is used to make the decision.
- Evaluate the adequacy of the information contained in the e-test measurements for driving the decision between the complete and the reduced test flow.
- Demonstrate that the proposed method enables better exploration of the trade-off between test cost and test quality, thereby yielding viable test cost reduction solutions in the very low DPPM region.

A. Limits of Static Test Elimination

Figure 5 reflects the number of defective die per million which are uniquely detected by each of the 14 test groups. In other words, this is the number of devices which would escape detection if each of these 14 test groups were to be statically eliminated from the probe-test flow. While we cannot reveal the exact number for $DPPM_{min}$, its order of magnitude is in the few hundreds. Accordingly, static test elimination cannot be used for test cost reduction when test

[2]Details regarding the device and exact test escape numbers and DPPM levels may not be released due to an NDA under which this data has been provided to us.

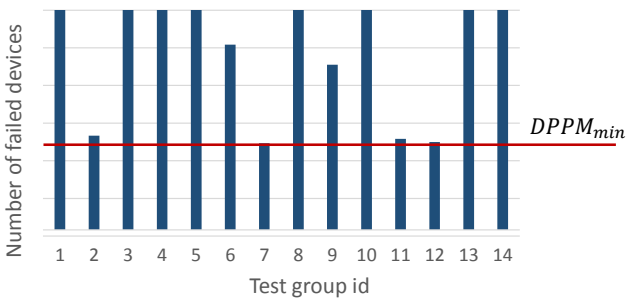

Fig. 5: Defective die per million which would escape detection if each of the 14 test groups were to be statically eliminated from the probe-test flow

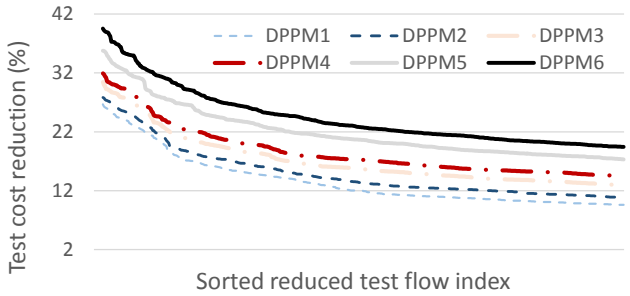

Fig. 6: Maximum test cost reduction achievable for various DPPM levels if oracle is used for test flow selection

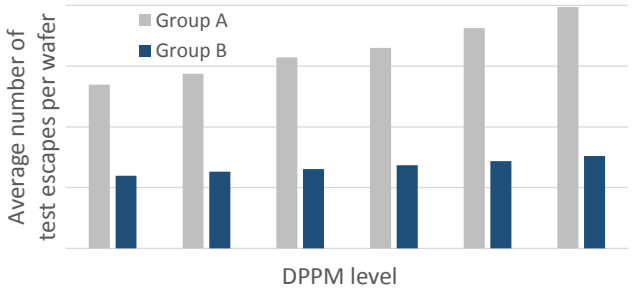

Fig. 7: Average number of test escapes when wafers in Groups A and B are subjected to the reduced test flow

quality expectations are set below this level. Exploration of the test cost vs. test quality trade-off in the sub-$DPPM_{min}$ realm requires dynamic per-wafer adaptation of the test flow.

B. Test Cost Reduction Potential

The upper bound of test cost reduction which can be achieved through the proposed probe-test flow selection method depends on the level of acceptable test escapes. To obtain a better feel for this potential, we first selected 6 different DPPM levels, $DPPM_1$ through $DPPM_6$, ranging in increasing order from a few tens to a few hundreds. As an additional point of reference, $DPPM_5$ was set to the value of $DPPM_{min}$ in the previous subsection. Then, for each of these targets, we considered each of the possible subsets of the 14 test groups as the reduced test flow, and we identified the maximum number of wafers in our dataset which could be subjected to the reduced flow, without the overall test quality (across all wafers) falling below this target. Using the relative test costs of the 14 test groups, we then calculated the maximum test cost reduction possible for each such solution.

Figure 6 presents the results. Each of the 6 curves corresponds to a different DPPM level. The x-axis reflects the list of the possible reduced test flows, rank ordered through Equation 1, while the y-axis reflects the test cost reduction achievable for the target DPPM level. Evidently, the leftmost points of these curves are the ones of the highest interest, as they maximize savings for a given DPPM level. These numbers reveal that significant test cost reduction may be possible, even for very low test escape rates. We emphasize, however, that this test cost reduction is an upper bound, as it assumes availability of an oracle that can perfectly select the appropriate test flow for each wafer. In reality, depending on what mechanism is used for making this decision, only a fraction of this upper bound may be achievable. Finally, as expected, these curves confirm that the larger the targeted DPPM level, the higher the test cost reduction that may be achieved, since more wafers can be channeled to the reduced test flow.

C. Adequacy of E-test for Flow Selection

Since our proposed method relies on e-test for deciding whether to apply the complete or the reduced test flow, we need to evaluate the adequacy of the information reflected in e-test for accurately making this decision. To this end, we applied the methodology described in Section II for each of the 6 targeted DPPM levels mentioned in the previous subsection, $DPPM_1$, through $DPPM_6$. In each case, the method returned a reduced test flow, as well as a trained intelligent system, which we used to split the wafers in the validation set into two groups, with wafers in group A slated to be subjected to the complete flow, and wafers in group B slated to be subjected to the reduced flow. We, then, computed the average test escapes that would occur for wafers in each of these two groups, if they were subjected to the reduced test flow. The results are depicted in Figure 7. As may be observed, wafers in Group A consistently exhibit much higher test escape rate than wafers in Group B. Thereby, the decision to channel them to the complete test flow is well justified, demonstrating that the e-test of a wafer can drive an informed choice regarding the test flow that the wafer should be subjected to.

D. Test Cost vs. Test Quality Trade-off Exploration

The ability of the proposed method to facilitate exploration of the trade-off between test cost reduction and test quality, even in the region of very low DPPM, is demonstrated in Figure 8. The two curves on this graph reflect solutions achievable by the proposed method and by static test elimination, respectively. Evidently, the adaptive nature of the proposed test flow selection method enables it to outperform static test elimination across the board. More importantly, it allows higher fidelity in the selection of a desirable point on this trade-off, starting from solutions with very low DPPM and small

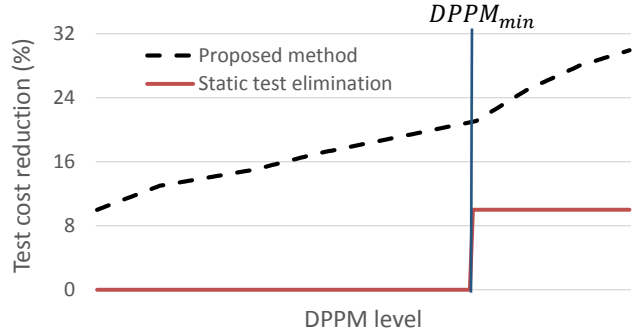

Fig. 8: Test cost reduction vs. test accuracy for the proposed method and static test elimination

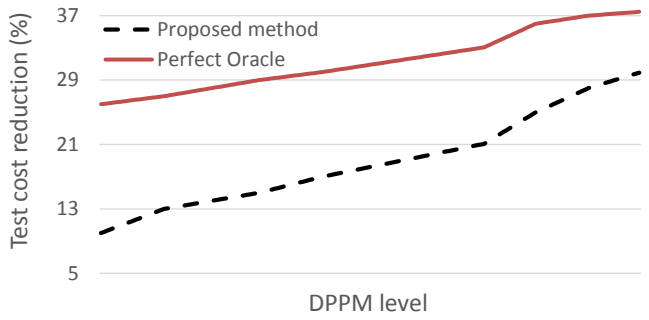

Fig. 9: Achieved vs. maximum possible test cost reduction for various DPPM levels

test cost reduction, and progressing at very fine-grained steps towards higher test cost reduction with higher test escape rate. In contrast, static test elimination does not offer any solution with test escapes below $DPPM_{min}$ and progresses at very coarse steps.

Finally, to gain better insight as to how well our method works, in Figure 9 we compare its test cost reduction to the upper bound achievable when an oracle is used, for various target DPPM levels. While it is evident that in the realm of very low DPPM the proposed method leaves significant potential for test cost reduction on the table, the gap continuously shrinks as the targeted DPPM increases. This is explained by the fact that at very low DPPM levels, incorrectly channeling a wafer to the reduced instead of the complete flow can be detrimental and very difficult to recover from. In other words, very low DPPM leaves little room for error, hence the proposed method acts conservatively, selecting very few e-test signatures for the reduced test flow and, thereby, limiting the achieved test cost reduction.

IV. Conclusion

Judicious wafer-level selection between the complete probe-test flow and a carefully reduced version shows great promise towards test cost reduction in analog/RF ICs. As demonstrated herein, this selection may be effectively driven by an early signature obtained through e-test, reflecting how process variations have affected a given wafer. Deployment of the proposed test flow selection method requires minimal test infrastructure support, yet is capable of identifying solutions with very low test escape rates, which is not possible through static test elimination. Experimental results using a large dataset of actual test measurements from a 65nm Texas Instruments RF transceiver confirmed the aptitude of the proposed method in effectively exploring the trade-off space between test quality and test cost.

V. Acknowledgment

This research has been partially supported by the Semiconductor Research Corporation (SRC) Task 1836.131.

References

[1] P. Drineas and Y. Makris, "Independent test sequence compaction through integer programming," in *IEEE International Conference on Computer-Aided Design*, 2003, pp. 380–386.

[2] H.-G. D. Stratigopoulos, P. Drineas, M. Slamani, and Y. Makris, "Non-RF to RF test correlation using learning machines: A case study," in *IEEE VLSI Test Symposium*, 2007, pp. 9–14.

[3] S. Biswas and R. D. Blanton, "Test compaction for mixed-signal circuits using pass-fail test data," in *IEEE VLSI Test Symposium*, 2008, pp. 299–308.

[4] S. Biswas, P. Li, R. Blanton, and L. T. Pileggi, "Specification test compaction for Analog circuits and MEMS," in *IEEE Design, Automation and Test in Europe*, 2005, pp. 164–169.

[5] H.-G. D. Stratigopoulos and Y. Makris, "Nonlinear decision boundaries for testing Analog circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 11, pp. 1760–1773, 2005.

[6] H.-G. D. Stratigopoulos, P. Drineas, M. Slamani, and Y. Makris, "RF specification test compaction using learning machines," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 18, no. 6, pp. 998–1002, 2010.

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

[8] 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*, vol. 32, no. 12, pp. 1934–1942, 2013.

[9] N. Kupp, K. Huang, J. M. Carulli Jr, and Y. Makris, "Spatial correlation modeling for probe test cost reduction in RF devices," in *ACM International Conference on Computer-Aided Design*, 2012, pp. 23–29.

[10] X. Li, R. R. Rutenbar, and R. D. Blanton, "Virtual probe: a statistically optimal framework for minimum-cost silicon characterization of nanoscale integrated circuits," in *ACM International Conference on Computer-Aided Design*, 2009, pp. 433–440.

[11] A. Ahmadi, K. Huang, S. Natarajan, J. M Carulli, and Y. Makris, "Spatio-temporal wafer-level correlation modeling with progressive sampling: A pathway to HVM yield estimation," in *IEEE International Test Conference*, 2014, pp. 1–10.

[12] H.-G. D. Stratigopoulos and Y. Makris, "Error moderation in low-cost machine-learning-based Analog/RF testing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 2, pp. 339–351, 2008.

[13] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.