# APPLICATIONS OF MACHINE LEARNING IN TEST COST REDUCTION, YIELD ESTIMATION AND FAB-OF-ORIGIN ATTESTATION OF INTEGRATED CIRCUITS

by

Ali Ahmadi


APPROVED BY SUPERVISORY COMMITTEE:

Dr. Yiorgos Makris, Chair

Dr. Carl Sechen

Dr. Mehrdad Nourani

Dr. Jeyavijayan Rajendran

*To my parents*

# APPLICATIONS OF MACHINE LEARNING IN TEST COST REDUCTION, YIELD ESTIMATION AND FAB-OF-ORIGIN ATTESTATION OF INTEGRATED CIRCUITS

by

ALI AHMADI, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

May 2017

# ACKNOWLEDGMENTS

I would like to give special thanks to my advisor, Professor Yiorgos Makris, who guided and supported me throughout my study. His insightful feedback and comments, were essential to make this success happen. I am very grateful for the opportunity to work in his research lab. I would also like to thank Professor Mehrdad Nourani, Professor Carl Sechen and Professor Jeyavijayan Rajendran for serving as my committee members and taking time to review this project and provide valuable comments and suggestions. My sincere thanks also goes to Amit Nahar, Bob Orr and Michael Pas of Texas Instruments for great discussion and feedback as well as providing data for this work. I would like to express my gratitude and thanks to my parents for their constant support and encouragement.

I was blessed to have many amazing friends in Dallas, who were very supportive and fun to be with. Among them, I would like to specifically name Mohammad-Mahdi Bidmeshki, Reza Lotfian, Maziyar Baranpouyan and Constantinos Xanthopoulos.

April 2017

APPLICATIONS OF MACHINE LEARNING IN TEST COST REDUCTION, YIELD ESTIMATION AND FAB-OF-ORIGIN ATTESTATION OF INTEGRATED CIRCUITS

Ali Ahmadi, PhD
The University of Texas at Dallas, 2017

Supervising Professor: Dr. Yiorgos Makris, Chair

The semiconductor manufacturing industry is one of the most technologically advanced and cost-intensive industries. It has been a key driver for economic development and has powered the growth in computers, consumer electronics and the internet industry. Semiconductors are becoming indispensable in health-care, cars, defense and telecommunications. The rapidly growing and dynamically changing electronics market introduces interesting and complex challenges for semiconductor manufacturing companies. One such challenge is identifying production problems and increasing the yield of integrated circuits (ICs), which is getting more difficult due to the complexity of new technology nodes. Another major challenge is the cost that can be devoted to testing each die before it is shipped to a customer. This is important because continuous pressure for superior performance, along with intensified process variations in the latest technology nodes, have resulted in stringent limitations in the test cost. A most recent challenge in the semiconductor industry is security concerns regarding integrity of the electronics supply chain due the globalization of the economy and the gain in pervasiveness of the fab-less paradigm.

To address these challenges, researchers have developed solutions based on statistical techniques and machine learning methods. The range of these solutions are from pre-silicon

simulation-based methods to data analytic techniques that utilize post-silicon high-volume production data. In simulation-based domain, a rich dataset is available to examine and evaluate the proposed solutions. However, these methods are very time-consuming and have a limited view of process statistics, as their grounding to silicon is established only through the variation models reflected in the process design kit (PDK). On the other hand, silicon-based learning methods are often impractical because of extra cost/overhead and new modifications in the production line.

The aim of this work is to address these challenges and provide fast, accurate and feasible solutions using high-volume production data. More specifically, this dissertation introduces an adaptive test cost reduction method that successfully reduces the test cost significantly while abiding the industry principles in order to be readily deployable with minimal test operations support. A fast and accurate yield learning methodology is proposed to forecast high volume manufacturing (HVM) yield of a device based on production datasets from few engineering wafers. Finally, an advanced machine learning approach is proposed to attest the fabrication facility that manufactures a given IC.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

Semiconductor manufacturing is one of the most technologically advanced industrial sectors. The advent of the modern integrated circuit has created an immense market for semiconductor devices, surpassing the $338.9 billion market in 2016, and the year 2017 is predicted to be strong with 6.5% growth to $361 billion. In modern societies, human life extensively depends on electronic devices such as smart phones, wearable devices, laptops, TVs, tablets, cars, etc. These products present dramatically different design constraints. For consumer electronics, low cost is the key driver; despite high manufacturing volume, profit margins are typically small. On the other hand, automotive and defense applications demand high reliability and security, with considerably lower manufacturing volume.

The process of creating integrated circuits (ICs) is called wafer fabrication. It is a sequence of chemical and photographic steps (like lithography, etching, deposition, oxidation and diffusion) in which the circuits are constructed on a semiconductor material typically called a wafer. Figure 1.1 presents an overview of the manufacturing cycle of an IC.

To keep pace with Moore's Law, the semiconductor industry has relied upon many innovations and its complexity has grown extensively. The technological advances have been accompanied by an exponential growth in the size of data collected and stored during the manufacturing process. The semiconductor manufacturing data comprises lot transactions and process and equipment data, in line data, electrical test (e-test), wafer sort data, and final electrical test/performance binning. The granularity of these measurements range from lot level all the way to block level in order to guarantee the fabricated device meets performance, reliability and security requirements. The availability of this data is laden with opportunities for improving the manufacturing flow with statistical and machine learning methods.

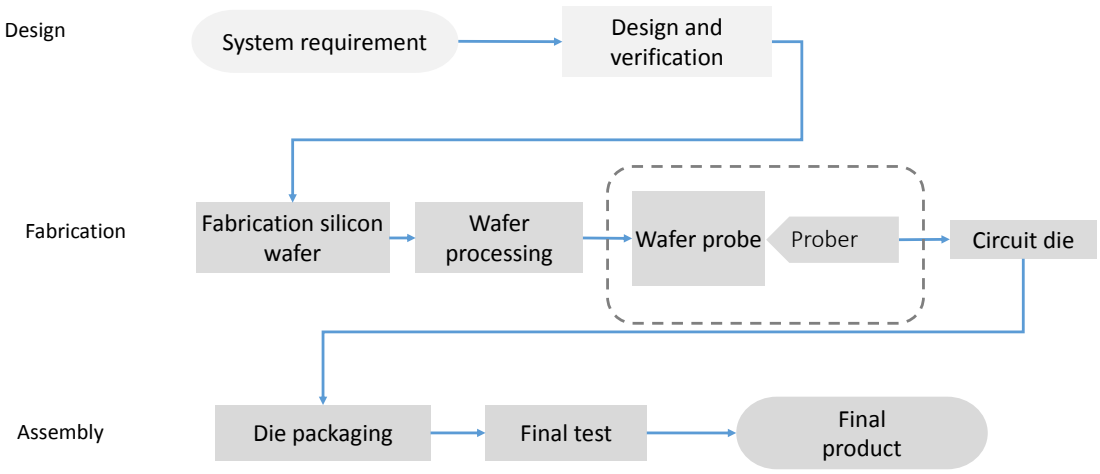



Figure 1.1. IC design and manufacturing cycle.

## 1.1 Motivation

Traditional statistical approaches have been widely used in semiconductor manufacturing industry for many years and will continue to have an important place in production. Often, a simple correlation plot, linear regression, wafer-map plot or variance analysis will tell an important story and enables the needed process discovery and controls. However, the rapidly growing and dynamically changing consumer electronics market introduces new challenges for IC production. New technology nodes are complex and require more than a thousand process steps which jeopardize the production yield. Continuous pressure for superior performance, along with intensified process variations in the latest semiconductor manufacturing technology nodes, are imposing an immense pressure on test and testability and 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. Security is a recent challenge in semiconductor manufacturing industry. As the supply chain grows more complex, with parts being sourced from various suppliers across the globe, ensuring authenticity and

trustworthiness of each part becomes very challenging. Indeed, IC counterfeiting has become a profitable activity and a major headache which poses a significant threat to end applications, especially when deployed in sensitive domains such as military, financial, health, etc.

Addressing these complex challenges by employing traditional statistical and machine learning approaches is not feasible. Therefore, advanced learning methods are needed to process semiconductor data which have high dimensionality, multi-modal distributions and various granularity.

## 1.2 Contribution

This dissertation presents the following methodologies that utilize semiconductor manufacturing data for test cost reduction, yield prediction and fab-of-origin attestation of ICs:

I. Chapter 2 introduces an adaptive test cost reduction methodology for dynamically selecting an optimal probe-test flow which reduces test cost without jeopardizing test quality. The proposed method offers flexibility by optimizing test flow per process variation signature and its implementation is simple and compatible with most commonly used Automatic Test Equipment (ATE). Furthermore, unlike static test elimination approaches, whose agility is limited by the relative importance of the permanently dropped tests, the proposed method is capable of exploring test cost reduction solutions which achieve very low test escape rates. Decisions are made by an intelligent system which maps every point in the e-test signature space to the most appropriate probe-test flow.

II. Yield estimation is an indispensable piece of information at the onset of high-volume production of a device, as it can inform timely process and design refinements in order to achieve high yield, rapid ramp-up and fast time-to-market. To date, yield estimation is generally performed through simulation-based methods. However, such methods are

not only very time-consuming for certain circuit classes, but also limited by the accuracy of the statistical models provided in the process design kits (PDKs). In contrast, this work proposes yield estimation solutions which rely exclusively on silicon measurements and applies them towards predicting yield during (i) production migration from one fabrication facility to another, and (ii) transition from one design generation to the next. These solutions are applicable to any circuit, regardless of process design kit accuracy and transistor-level simulation complexity, and range from rather straightforward to more sophisticated ones, capable of leveraging additional sources of silicon data.

III. A machine learning methodology is introduced for distinguishing between ICs fabricated in a ratified fabrication facility and circuits originating from an unknown or undesired source based on parametric measurements. Unlike earlier approaches, which seek to achieve the same objective in a general, design-independent manner, the proposed method leverages the interaction between the idiosyncrasies of the fabrication facility and a specific design, in order to create a customized fab-of-origin membership test for the circuit in question.

### 1.3 Dissertation Organization

The rest of this dissertation is organized as follows: the details of test cost reduction method and its effectiveness using industrial datasets are explained in Chapter 2. Chapter 3 elaborates the yield estimation methodology along with its experimental evaluation using high volume production data. In Chapter 4, Fab-of-Origin attestation problem is introduced as well as the details of the proposed methodology to address this challenge in semiconductor manufacturing. Finally, this dissertation is concluded in Chapter 5.

# CHAPTER 2

# TEST COST REDUCTION[1]

## 2.1 Overview

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. Various directions have been explored towards reducing test cost such as leveraging spatial and temporal correlation, measuring low cost test and use machine learning algorithms to predict high cost tests and customize the test list per die/wafer/lot. Next Section provides a brief review of recent researches on test cost reduction.

## 2.2 Related Work

There has been an intensified effort from semiconductor industry to reduce the test cost especially for Analog/RF devices due to the expensive and sophisticated automated test equipment that is required. Leveraging spatial and temporal correlation is one of the well-studied research directions toward test cost reduction which has shown great promise in

[1]2016 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Constantinos Xanthopoulos, Amit Nahar, Bob Orr, Michael Pas and Yiorgos Makris, "Harnessing Process Variations for Optimizing Wafer-level Probe-Test Flow", in Proceedings of IEEE International Test Conference ©2016 IEEE

[1]2016 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Amit Nahar, Bob Orr, Michael Pas and Yiorgos Makris, "Wafer-Level Process Variation-Driven Probe-Test Flow Selection for Test Cost Reduction in Analog/RF ICs", in Proceedings of IEEE VLSI Test Symposium ©2016 IEEE

capturing wafer-level spatial variation and, thereby, reducing test cost of electrical measurements [1, 2, 3, 4, 5, 6, 7, 8]. 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, 11]. 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 [12]. 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.

Along a different direction towards leveraging spatial/temporal correlation, various methods have been proposed to customize the test list. 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 [13, 14, 15]. 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 [16, 14, 17, 18]. 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 [19]. Taking adaptation a step further, the method in [20] 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.

This Chapter introduces an adaptive test cost reduction method which reduces the test cost significantly, while abiding by the industry principles in order to be readily deployable with minimal test operations support.

## 2.3 Adaptive Probe-Test Flow Selection

As mentioned in the previous section, there are several industry constraints and principals for any test cost reduction method in order to utilized in semiconductor industry. These principals are as follows:

- 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 set of test groups or a subset thereof.

- Test has to be performed in one pass. In other words, solutions which first apply a reduced test flow and subsequently apply selectively more test items to die for which the decision confidence is low, such as the two-tier test method in [21], 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 as the only source available for capturing the impact of process variations on a particular wafer.

- The ATE supports multiple test flows, where test groups can be dynamically included or excluded based on an input provided before test commences for a 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 2.1. For each wafer, this approach provides a decision as to select the appropriate probe-test flow. Each wafer is subjected to the complete probe-test flow or one of reduced test flows, in which some of the test groups are eliminated. This decision is made at an early stage, before the wafer reaches the probe station, driven through e-test measurements.[2] Indeed, depending on how a wafer has been impacted by process variations, a different reduced test flow may offer the best option. Therefore, this work seeks to investigate the utility of test flow optimization per

[2]The term e-test is referred 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.

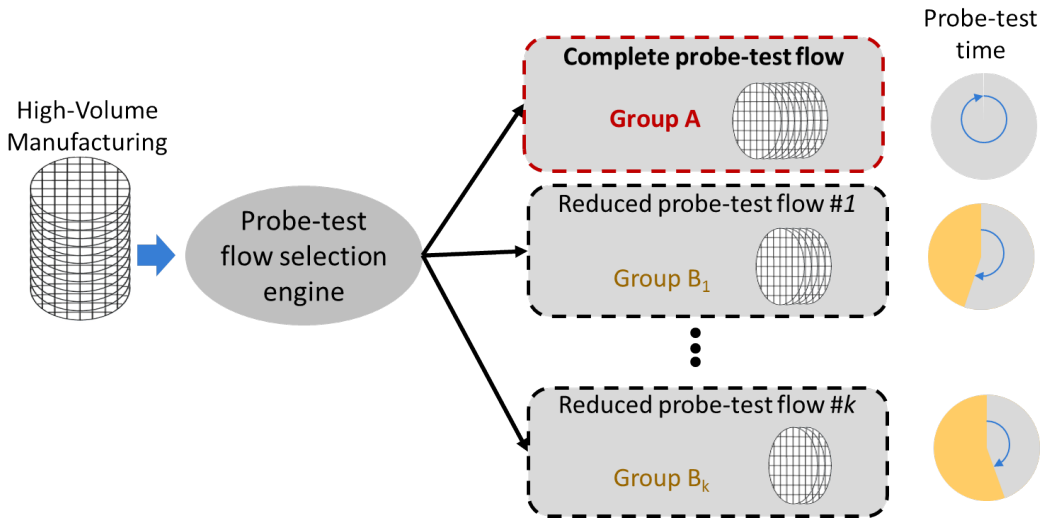



Figure 2.1. An overview of wafer-level probe-test flow selection.

process signature, towards achieving test cost reduction. To accomplish this, an optimization algorithm is employed to statistically select the best test flow for each signature such that test cost reduction is maximized while the required test quality is achieved. The trained test flow selection engine processes the e-test measurements of a wafer, extracts its process signature, and accordingly selects the most appropriate test flow for that signature during probe testing of this wafer. It should be noted that the complete test flow remains one of the possible choices, especially for outlier wafers, i.e., those whose e-test signatures have not been encountered in the past.

### 2.3.1 Preprocessing

Before addressing the problem of deciding an appropriate test flow for a wafer, the initial elements that are required prior to such a decision are discussed. These elements are: (i) identifying an appropriate subset of test groups which could potentially be applied to a wafer, and (ii) crafting a wafer signature from its e-test measurement vector. In the following sections, details of these two components are provided.

### 2.3.2 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 a 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 at the test item level, it may be possible to exhaustively search the space of solutions. For example, experiments of this work dealt with a set of 10 test groups, thus exhaustively searching in the power-set of $2^{10}$ 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 [18], hence it can readily be adopted when exhaustive consideration is infeasible.

For each reduced flow, $j$, the associated cost and the number of test escapes are considered when this reduced flow is applied to all wafers in the training set, and a fitness value is assigned and defined as:

$$index_j = \frac{t_A - t_{B_j}}{t_A} * pctg_{B_j} \tag{2.1}$$

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

### 2.3.3 Wafer Signature Extraction from E-tests

E-test 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 e-test data, prior to crafting a wafer signature based on the e-tests a dimensionality reduction algorithm is applied to transform the data onto a lower count of dimensions. Specifically, this work uses the *t-Distributed Stochastic Neighbor Embedding (t-SNE)* technique [22] which is the state-of-the-art non-linear transformation approach and which is widely used in many applications for unsupervised dimensionality reduction. In general, t-SNE embeds wafers with similar signatures close to each other on a 2-dimensional map.

Figure 2.2, provides an example where a number of wafers are projected to a 2-dimensional space after applying the t-SNE algorithm. The various markers used to represent each point indicate different test escape rates when a randomly selected reduced test flow is applied to all wafers.[3] Wafers with the same marker exhibit a similar level of test escapes. Two key observations 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.

[3]The exact values of B-G are not important for this example.

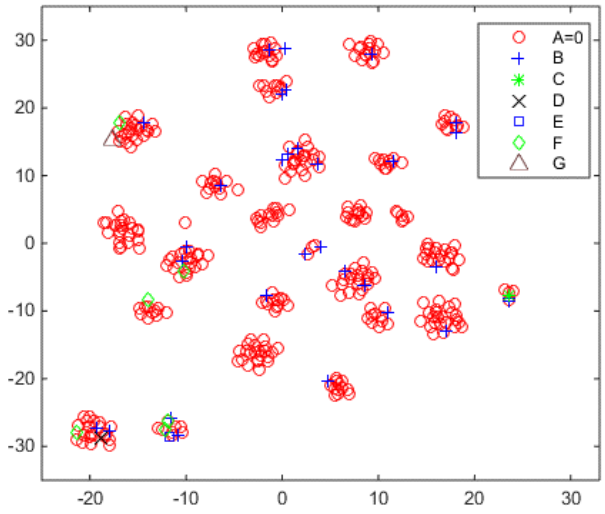



Figure 2.2. Projection of e-test data onto two dimensions using the t-SNE algorithm.

### 2.3.4 Test Flow Optimization

Section 2.3.1 described the process of generating all potential reduced test flows as well as extracting a process signature from the e-test data of a wafer. Now, the objective is to assign the most appropriate reduced test flow to each process signature, such that it maximizes test cost reduction while retaining the test escape rate below a target DPPM level.

#### Bi-Flow Method

This Section explains how to assign a proper probe-test flow to each wafer based on its e-test signature. First, it discusses a simple version of problem, wherein each wafer is subjected either to complete test flow or a reduced test flow (it is called *Bi-Flow* since there are two choices). Figure 2.3 depicts an overview of Bi-Flow method [23]. *Recall that the 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 funnel to the reduced test flow, the higher the test cost reduction can be achieved. Thus, the problem is to map the e-test signature space to the appropriate test flow, such that it meets both of the above objectives.

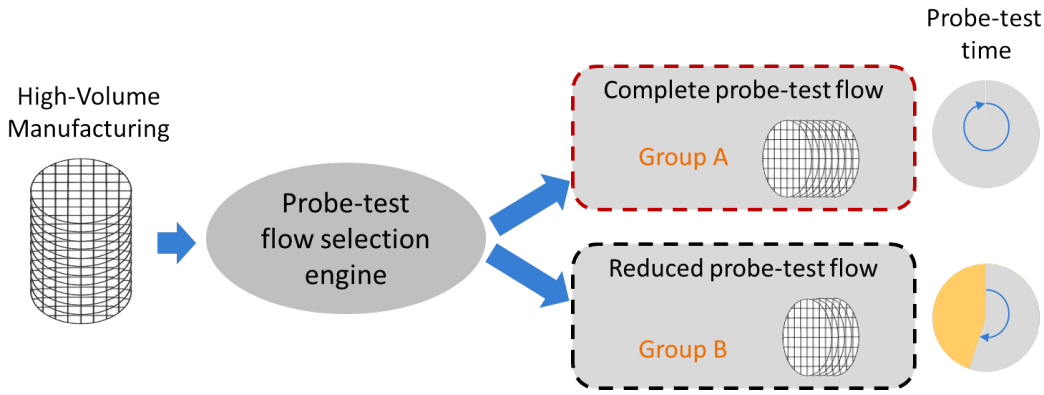



Figure 2.3. Bi-Flow method overview.

This problem is formulated 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 minimized. In this problem, the constraint is on the total number of test escapes, and cost function is the maximization of the number of wafers that go through the reduced flow. The ILP is actually a binary (0-1) version, where the value of each integer variable can only be either 0 or 1. Specifically, in this ILP, the variable $\alpha_i$ 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 I 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 the 0-1 ILP is defined as follows:

$$te_i = \sum_{j \in C_i} TE_B^j \tag{2.2}$$

$C_i$ : all wafers in the cluster i

$$\text{Maximize} \quad \sum_{i=1}^{k} \alpha_i.card_i$$

$$\text{subject to} \quad \sum_{i=1}^{k} \alpha_i.te_i \leq DPPM_t \tag{2.3}$$

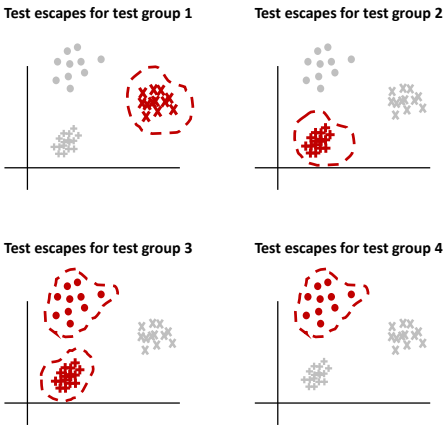
$$\alpha_i \in \{0, 1\},\ i = 1, \ldots, k$$

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 $\alpha_i$ 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.

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, the wafer will undergo only the preselected subset of test groups, otherwise it will be tested by the complete test flow.

### Limits of The Bi-Flow Method

This Bi-Flow method is simple and can be implemented easily. It is also very effective in finding the best reduced test flow and assigning either complete or reduced test flow to each cluster. However, its major limitation is the fact that the same reduced test flow is chosen for all process signatures (clusters) that will not undergo complete test. This simplifies test operations, as only two test flows are maintained, but it is also sub-optimal, since different clusters exhibit dissimilar failure patterns when a test group is removed from the test flow. Indeed, choosing a different test flow for each cluster holds promise for significantly higher test cost reduction.




(a) Test escape rate for selected clusters when test groups 1-4 were individually removed from the test flow.

| Cluster symbol | Cluster id | Bi-Flow approach probe-test flow code $\{TG_1, TG_2, TG_3, TG_4\}$ | This work probe-test flow code $\{TG_1, TG_2, TG_3, TG_4\}$ |
|---|---|---|---|
| ● | 1 | [1 1 0 1] | [1 1 0 0] |
| ✚ | 2 | [1 1 0 1] | [1 0 0 1] |
| ✘ | *3* | [1 1 1 1] | [0 1 1 1] |

(b) Comparison of probe-test flow code corresponding to part (a) generated by Bi-Flow method vs. optimum probe-test code for each cluster.

Figure 2.4. Limitations of the Bi-Flow approach.

To demonstrate this limitation, three clusters from Figure 2.2 are selected and the test escape rate of each cluster is computed when a test group is removed from the test flow. Figure 2.4(a) demonstrates the test escape rate due to elimination of test groups 1-4 for these three clusters. Clusters in red color and enclosed in a red boundary reflect zero test escapes while gray color represents clusters with a non-zero test escape rate. In the table of Figure 2.4 (b), the test flow code of each cluster is represented using a binary vector where inclusion/exclusion of a test group is indicated by value 1/0 respectively. The third column of the table shows the two test flow codes which are generated by the Bi-Flow method [23], while the forth column contains the optimum test flow code if chosen individually per each cluster. As may be observed, the impact of skipping a test group is not identical for all clusters. The figure corroborates the initial conjecture that each cluster requires individualized test flow optimization. Therefore, a dynamic approach is required to generate the most appropriate probe-test flow per process signature, in order to maximize test cost reduction. Next Section introduces a methodology which addresses this limitation.

**Dynamic Test Flow Generation**

Now I proceed to elaborate on how to optimize the test flow per cluster. This methodology consists of two steps: (i) finding the best reduced test flow for each cluster individually for any target DPPM level, and (ii) determining the maximum test escape rate of each cluster through an optimization algorithm. Below I provide details of these two steps [24].

**Test Flow Generation per Cluster** Let consider cluster $C_i$, which includes a set of wafers, and let assume that the goal is to find the best reduced test flow among all $n$ candidates which are generated using exhaustive search. Let $\boldsymbol{TE_i} = [te_1, \cdots, te_n]$ and $\boldsymbol{TTR_i} = [ttr_1, \cdots, ttr_n]$ denote the test escape rate and test cost reduction vectors of the $i$-th cluster, where $te_j$ and $ttr_j$ denote the number of test escapes and the amount of test cost reduction when all wafers in this cluster are tested by the $j$-th reduced test flow. For any DPPM level in the range $[0, DPPM_t]$, where $DPPM_t$ is the target DPPM level, a reduced test flow is selected such that its test escape rate for cluster $C_i$ is lower than the DPPM level, while maximizing the test cost reduction. At the end of this step, each cluster has associated with it a table with multiple rows and three columns. Each row corresponds to a specific DPPM level and the three columns correspond to the test escape rate, test cost reduction and index of selected test flow, respectively.

**Optimization Algorithm** The second part of the proposed method is an optimization algorithm, which selects the best probe-test flow for all $k$ clusters while meeting the required test quality. Let $\boldsymbol{TE} = [\boldsymbol{TE_1}, \cdots, \boldsymbol{TE_k}]^T$ and $\boldsymbol{TTR} = [\boldsymbol{TTR_1}, \cdots, \boldsymbol{TTR_k}]^T$ denote the test escape rate and test cost reduction matrices, where $\boldsymbol{TE_i}$ and $\boldsymbol{TTR_i}$ represent the test escape rate and test cost reduction vectors for the $i$-th cluster, and $te_{ij}$ denotes the test escape rate for the $i$-th cluster for the $j$-th DPPM level. My objective is to distribute the target DPPM level among $k$ clusters so as to maximize test cost reduction. Looked at from a

different angle, the maximum acceptable test escape rate for each cluster need to determined. To do so, this problem is also formulated as an ILP. Similarly, the constraint is on the total number of test escapes, and my cost function is to maximize test cost reduction. The ILP is actually a binary (0-1) version, where the value of each integer variable can only be either 0 or 1. Specifically, the variable $\alpha_{ij} = 1$ is used to indicate that the maximum acceptable test escapes for $i$-th cluster is $te_{ij}$, and therefore the test flow with index $j$ is selected for this cluster. Then, the binary ILP is defined as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{k} \sum_{j=1}^{m} \alpha_{ij}.ttr_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{k} \sum_{j=1}^{m} \alpha_{ij}.te_{ij} \leq DPPM_t \\ & \sum_{j=1}^{m} \alpha_{ij} = 1 \\ & \alpha_{ij} \in \{0,1\},\ i = 1,\ldots,k \ and \ j = 1,\ldots,DPPM_t \end{aligned} \tag{2.4}$$

The constraint $\sum_{j=1}^{m} \alpha_{ij} = 1$ is used to select only one test flow for each cluster.

An additional provision is also incorporated in the proposed methodology, in order to adapt to shifts in the process, which may result in previously unseen wafer signatures in the transformed e-test space. Specifically, as shown in Figure 2.5, for clusters where the ILP selects any probe-test flow other than complete test flow, a boundary is established 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 any reduced test flow, one more check is performed: if its signature is inside the boundary of that cluster, the recommendation is followed. Otherwise, despite being nearest to this cluster, the wafer is sufficiently different

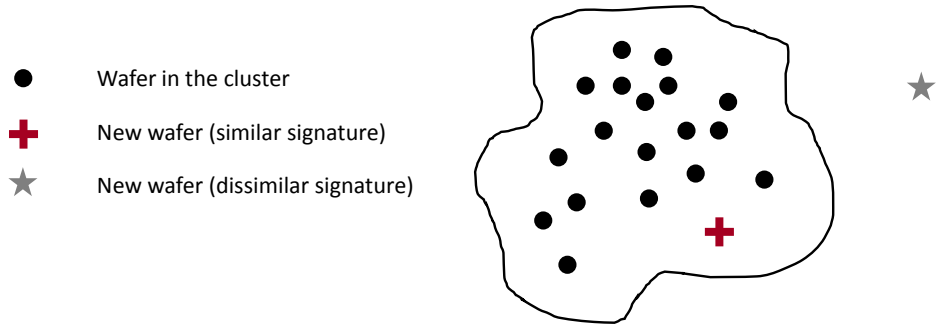



Figure 2.5. Tracking process shifts: signatures of wafers belonging to cluster are enclosed by a boundary. New cluster members with signatures within the boundary are considered equivalent, and new members with signatures outside the boundary are considered outliers.

and it will be sent to the complete test flow. Based on this information, one can periodically enhance the set of clusters and rerun the optimization algorithm to better track the process.

## 2.4 Experimental Results

In order to experimentally evaluate the effectiveness of the proposed methodology, actual production data from a 65nm analog/RF transceiver is used.[4] The dataset comes from 400 wafers, each of which contains approximately 2500 die. E-test is performed on 9 sites across the wafers, with 250 measurements obtained from each site. On each die, 380 parametric probe-test measurements are obtained, organized in 10 groups. The percentage by which each group contributes to the total test cost is also provided. The objective of the proposed method is to find a subset of the 10 test groups as an optimized reduced test flow for each process signature and to train an intelligent system which will use the e-test measurements to select which test flow a wafer should undergo. In these experiments, 5-fold cross validation is employed. Specifically, the data set is divided into 5 folds, where 4 folds are used for training and the remaining fold for validation. The procedure is repeated such that all folds are left

[4]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.

out once as a validation set and, in the end, the average test escapes and test cost reduction across the five iterations is reported. Using this dataset, these 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, therefore adaptivity is required to provide per wafer decision.
- Demonstrate that the effectiveness of the Bi-Flow method, which provides per wafer decision between a complete and a reduced test flow, is rather limited, thus a dynamic test flow generation with wafer-level granularity is required to optimize the test flow per process signature.
- Demonstrate that dynamic test flow generation per wafer based on e-test data can yield significant test cost reduction at realistic low DPPM levels.

### 2.4.1 Limits of Static Test Elimination

Figure 2.6 reflects the number of defective die per million which are uniquely detected by each of the 10 test groups. In other words, this is the number of devices which would escape detection if each of these 10 test groups were to be statically eliminated from the probe-test flow. While I cannot reveal the exact number for $DPPM_{min}$, its order of magnitude is in the several tens. Accordingly, static test elimination cannot be used for test cost reduction when test quality expectations are set below this level. Therefore, exploration of the test cost vs. test quality trade-off in the sub-$DPPM_{min}$ realm requires adaptive test flow selection per wafer.

Figure 2.7 demonstrates the test cost vs. test quality trade-off for various DPPM levels. The two curves on this graph reflect solutions achievable by the static test elimination and Bi-Flow approach, which selects between the complete test flow and a single reduced test flow [23], respectively. Evidently, the Bi-Flow method outperforms static test elimination across

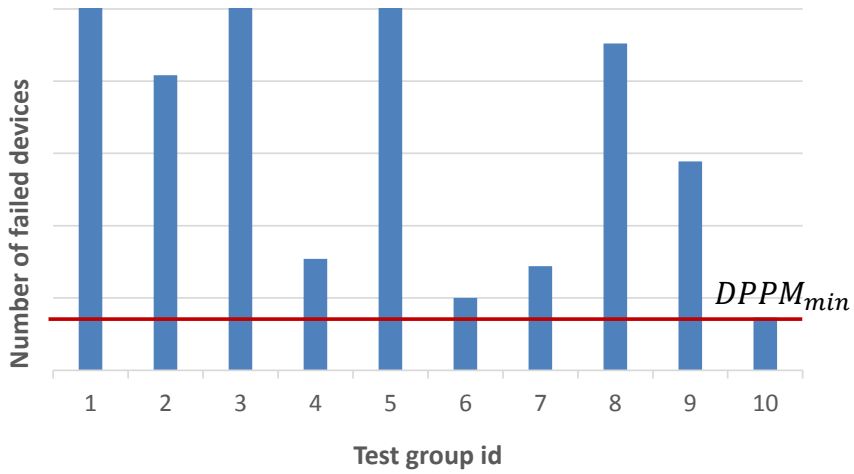



Figure 2.6. Defective die per million which would escape detection if each of the 10 test groups were to be statically eliminated from the probe-test flow.

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 test cost reduction, and progressing at very fine-grained steps towards higher test cost reduction with higher test escape rates.

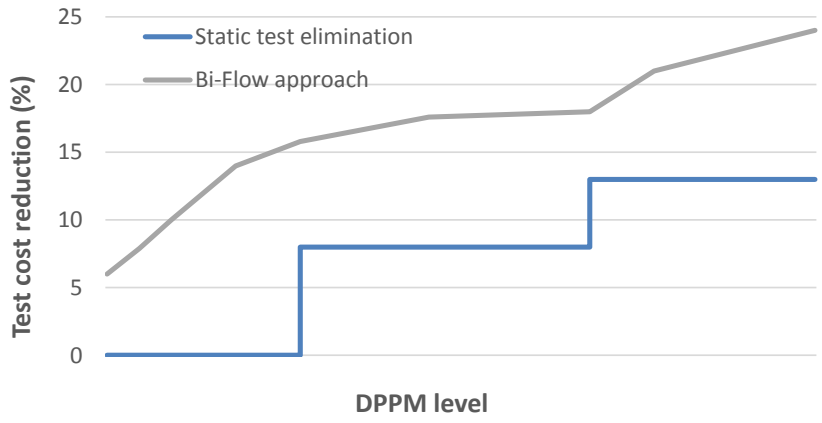



Figure 2.7. Test cost reduction vs. test accuracy for static test elimination and Bi-Flow method for various DPPM levels.

### 2.4.2 Limits of Bi-Flow Technique

In this part, the effectiveness of the Bi-Flow approach, which subjects a wafer either to a complete or a reduced test flow is examined. To do so, Figure 2.8 compares its test cost reduction to the upper bound achievable when an oracle that can perfectly select the appropriate test flow (i.e., the complete or the single reduced test flow) for each wafer is used, for various target DPPM levels. As may be seen from the gap between the two curves, this approach leaves significant potential for test cost reduction on the table. To gain better insight, Figure 2.9 depicts the outcome of the Bi-Flow approach in which the complete test flow assigned to a set of clusters (i.e., clusters with circle marker in red) and a reduced flow is selected for the remaining clusters, when the target test escape rate is set to $DPPM_{min}$. The main disadvantage of this approach is the fact that the reduced test flow is generated collectively for all clusters rather than individually per cluster, based on the process signatures of wafers in a cluster.

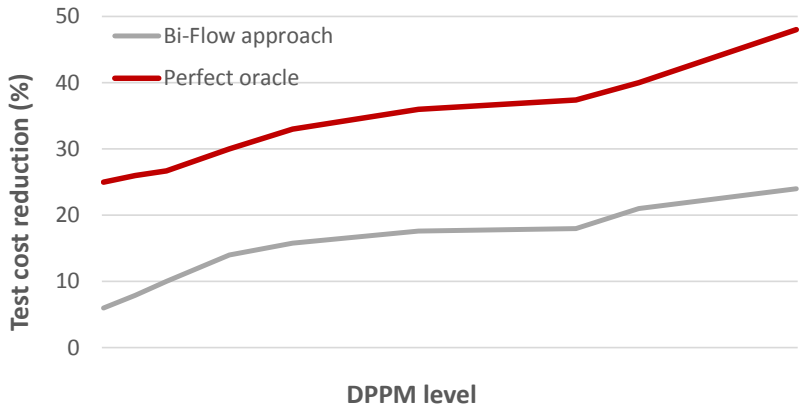
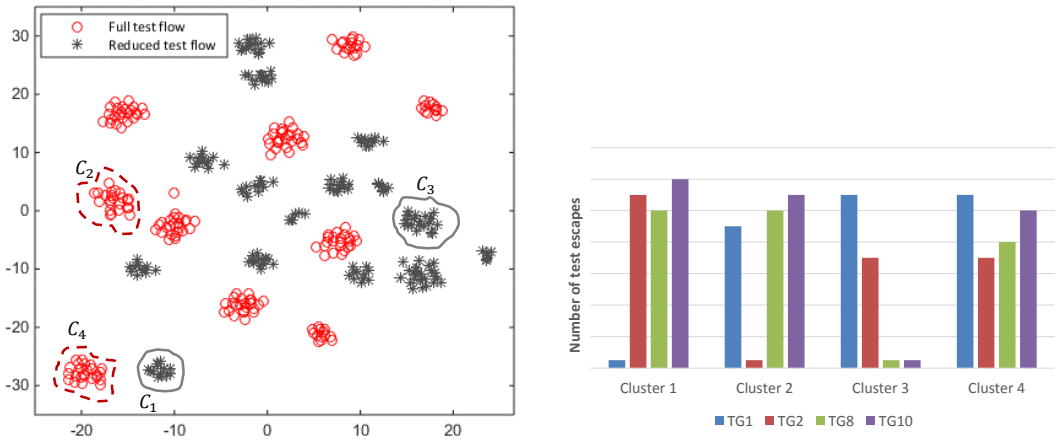



Figure 2.8. Achieved test cost reduction using the Bi-Flow method vs. maximum possible test cost reduction for various DPPM levels.

To demonstrate the unique characteristics and test flow needs of each cluster, clusters $C_1 - C_4$ in Figure 2.9 (a) are selected. Then, the test escapes of each cluster is computed




(a) Test flow assignment. (b) Test escape profile for selected clusters.

Figure 2.9. (a) Test flow assignment (either complete or reduced flow) for each cluster in the e-test space. (b) Number of test escapes for selected clusters when each of the four test groups are eliminated from the probe-test flow of the cluster.

when a test group is eliminated from the test flow (either complete or reduced test flow) which is assigned to that cluster. Figure 2.9 (b) shows the number of test escapes for these clusters when test groups 1, 2, 8 and 10 were removed from the test flow individually. As it can be seen, the test escape profile of each cluster varies drastically. More specifically, based on this information, wafers in cluster $C_1$ can skip test group 1, while those in cluster $C_2$ require test groups 1 and 2, yet test groups 8 and 10 can be eliminated from their test flow. This experiment confirms that a dynamically optimized test flow generation per cluster is needed to maximize test cost reduction for any target DPPM level.

### 2.4.3 Dynamic Test Flow Optimization

Figure 2.10 depicts the outcome of the proposed dynamic test flow optimization technique when the target test escape rate is set to $DPPM_{min}$. In this graph, clusters with identical probe-test flow are represented by the same color; for example, clusters in blue, such as $C_4$, require the complete test flow. On the bottom right of this graph, the optimized test flow

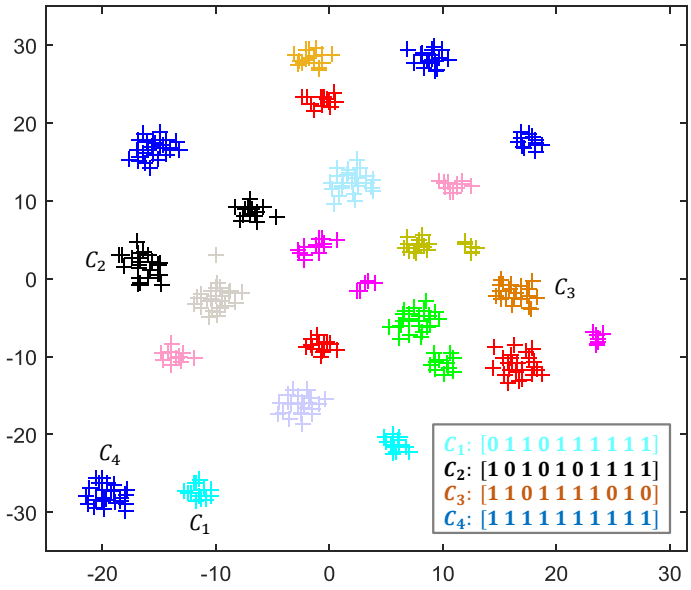
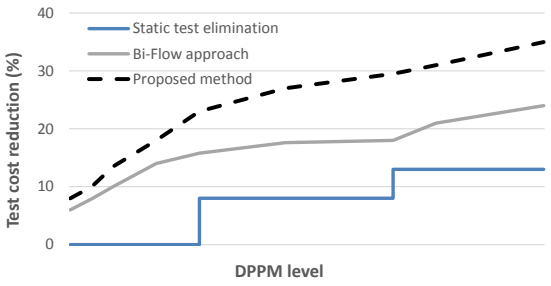



Figure 2.10. Final assignment of the optimized probe-test flow code for each cluster (process signature).

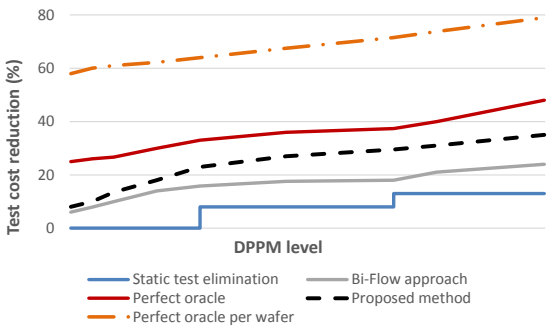
code for clusters $C_1 - C_4$ is presented. In comparison to Figure 2.9, which shows the outcome of the Bi-Flow method for the same target DPPM level, the new method provides more flexibility for test cost savings.

The ability of the proposed dynamic test flow generation method to explore the trade-off between test cost reduction and test quality, even in the region of very low DPPM, is demonstrated in Figure 2.11 (a). The three curves on this graph reflect solutions achievable by static test elimination (blue curve), the Bi-Flow method (gray curve), and the proposed dynamic test flow generation (dotted black line) for various target DPPM levels. It is evident that the proposed dynamic test flow optimization approach significantly outperforms the other two approaches for any DPPM level. This is expected, since the dynamic approach successfully generates an optimized probe-test flow for each process signature.

Finally, to gain better insight as to how well the proposed method works, Figure 2.11 (b) compares its test cost reduction to the upper bound achievable when an oracle is used,




(a) Test cost reduction vs. test accuracy.




(b) Test cost reduction vs. test accuracy.

Figure 2.11. (a) Test cost reduction vs. test accuracy of three approaches for various DPPM levels. (b) Test cost reduction vs. test accuracy for three approaches and maximum possible test cost reduction for various DPPM levels.

for various target DPPM levels. It should be noted that the maximum achievable test cost reduction, which is demonstrated in Figure 2.11 (b) by the red solid curve, is the upper bound when only two test flows are allowed (complete or reduced). However, in the new scenario where several test flows can be handled, the upper bound would be achieved by an oracle which can perfectly select the best test flow for each wafer. In Figure 2.11 (b), the new upper bound is represented by the dotted line which is above all other curves.

Note that the gap between the proposed method and the upper bound shrinks as the targeted DPPM increases. This is explained by the fact that, at very low DPPM levels, incorrectly channeling a wafer to a reduced instead of a 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 reduced test flows and, thereby, limiting the achieved test cost reduction. This gap indicates what is still left on the table as possible further test cost reduction, which more advanced methods and better statistics may be able to potentially achieve. Therefore, future research efforts can be directed towards further reducing this gap.

## 2.5 Conclusion

Test cost becomes a significant portion of an IC especially for analog/RF devices. Semiconductor industry invests and supports several researches that target test cost reduction while comply with industry principals and constraints. In this chapter, an adaptive test cost reduction method is proposed to reduce wafer-level probe-test time. Specifically, judicious harnessing of process variations is utilized to optimize probe-test flow which demonstrates great promise towards test cost reduction in analog/RF ICs. As presented herein, each signature in the process space may require its own optimized test flow. The signature of a wafer can be obtained at an early stage through e-test, reflecting how process variations have affected a given wafer. Deployment of the proposed 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.

# CHAPTER 3

# YIELD LEARNING [1]

## 3.1 Overview

The inherent variation of the semiconductor manufacturing process is a fundamental obstacle towards achieving high yield, especially for contemporary mixed-signal System-on-Chip (SoC) designs, wherein digital, analog and RF circuits are integrated together in advanced technology nodes. Indeed, understanding the complex interaction between design and manufacturing, and accurately estimating the expected yield prior to high-volume manufacturing (HVM) of a device in light of such variation, constitutes a challenging yet highly desirable task towards production and yield ramp-up. To this end, a large number of methods have been proposed in the past to estimate and optimize yield of a device [25, 26]. The vast majority of these methods concern yield estimation prior to fabrication and are based on simulation. Therefore, besides being very time-consuming and, often, impractical for large and complex circuits, they have a limited view of process statistics, as their grounding to silicon is established only through the variation models reflected in the PDK.

---

[1] 2017 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Haralampos-G. Stratigopoulos, Ke Huang, Amit Nahar, Bob Orr, Michael Pas, John M. Carulli Jr. and Yiorgos Makris, "Yield Forecasting Across Semiconductor Fabrication Plants and Design Generations", in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, accepted for publication ©2017 IEEE

[1] 2016 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Haralampos-G. Stratigopoulos, Amit Nahar, Bob Orr, Michael Pas and Yiorgos Makris, "Harnessing Fabrication Process Signature for Predicting Yield Across Designs", in Proceedings of the IEEE International Symposium on Circuits and Systems ©2016 IEEE

[1] 2015 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Haralampos-G. Stratigopoulos, Amit Nahar, Bob Orr, Michael Pas and Yiorgos Makris, "Yield Forecasting in Fab-to-Fab production migration Based on Bayesian Model Fusion", in Proceedings of the IEEE International Conference on Computer-Aided Design ©2015 IEEE

[1] 2015 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Haralampos-G. Stratigopoulos, Ke Huang, Amit Nahar, Bob Orr, Michael Pas, John M. Carulli Jr. and Yiorgos Makris, "Yield Prognosis for Fab-to-Fab Product Migration", in Proceedings of the IEEE VLSI Test Symposium ©2015 IEEE

## 3.2 Related Work

In this section, a brief overview of well-known simulation-based techniques for yield estimation is provided.

### 3.2.1 Monte Carlo

Monte Carlo (MC) simulation [27, 28] has been the most popular technique for yield estimation. In the MC method, a large number of random circuit samples are generated based on expected process variations defined in the PDK; thereafter, these circuit samples are simulated to estimate yield based on relative frequencies. Simplicity and generality are the advantages of the MC method. However, it is a time-consuming procedure which makes it prohibitive for large and complex circuits, as well as for circuits with long simulation times. Even for circuits with reasonable simulation times, MC ends up being too slow or inaccurate, especially when yield is very high. Furthermore, its accuracy is often limited due to insufficient process variation modeling in the PDK. Therefore, the MC method is not always practical for yield estimation.

### 3.2.2 Monte Carlo with speed enhancement

Several methods can be used to speed up MC, including Latin hypercube sampling (LHS) [29], quasi-Monte-Carlo (QMC) [30], and importance sampling [31, 32]. Compared to MC, which is purely random and requires many samples to cover the design space, LHS and QMC produce quasi-random sequences of samples that cover the design space much faster, thus allowing expedited and more accurate estimation of yield. However, LHS and QMC may still not produce enough samples at the tails of the design distribution where yield loss events typically occur. By focusing precisely on these distribution tails, importance sampling can produce better yield estimates with smaller variance. However, importance sampling requires definition of an optimal sampling distribution which, in general, is very challenging.

### 3.2.3 Statistical Blockade

Statistical blockade is a method that also offers significant speedup, as compared to the classical MC simulation, by focusing the simulation effort on the tails of the design distribution [33]. Unlike importance sampling, however, it only relies on the PDK and does not impose any *a priori* assumptions on the form of process parameter statistics, device models, or performance metrics. The underlying observation is that sampling a circuit instance is not time-consuming. What is time-consuming is performing an actual electrical simulation of the circuit instance. Statistical blockade is, in essence, a MC method, wherein simulation is blocked for circuit instances that are unlikely to exhibit performances far from the nominal design point and, thereby, are unlikely to lie at the tails of the design distribution. This decision of whether to block a simulation or not is taken based on a classifier which is trained in the space of process parameters. In the end, the simulated "extreme" circuit instances can be used to estimate yield probabilistically based on extreme value theory [34, 33, 35]. In [36], a recursive strategy is proposed to further accelerate the simulation effort.

### 3.2.4 Response surface and symbolic performance modeling

Another popular method for yield estimation is based on performance modeling [37, 38, 39, 40]. The underlying idea is to approximate the mappings between circuit performances and process parameters. These mappings can, then, replace electrical simulations. In particular, the process parameter space is sampled, with each sample corresponding to a circuit instance. Then, the mappings are used to predict the performances of these circuit instances instead of directly simulating them.

### 3.2.5 Behavioral modeling

For circuits such as data converters, phase locked loops (PLLs), complete RF transceivers, etc., a single transistor-level simulation may take hours or days to complete. In this case, none

of the above methods is practical since they require simulating at least hundreds of circuit samples at the transistor level. For circuits with long simulation times, yield estimation is typically carried out by first developing a behavioral model that captures effectively the circuit functionality and then applying any of the above methods by considering the behavioral-level description of the circuit instead of the transistor-level or layout-level description [41, 42]. A behavioral model is constructed by decomposing the circuit into independent sub-circuits, creating a separate behavioral model for each sub-circuit to reflect its functionality, and then linking these behavioral models and manipulating the data flow so as to compute the circuit performances. The key is to capture the correlation amongst the behavioral parameters that correspond to sub-circuit performances, such that this correlation draws upon the correlation that exists amongst the low-level process parameters, as these are expressed in the PDK.

## 3.3 Yield Forecasting Across Semiconductor Fabrication Plants and Design Generations

The focus of this chapter of dissertation is on yield estimation in two specific scenarios wherein much more silicon data reflecting process statistics is available:

- **Fab-to-Fab Production Migration:** Demand fluctuations and other financial, geographical or political reasons often cause a production to be migrated from one fabrication plant to another, wherein a device may have never been fabricated before [43, 44]. Forecasting how well a device will yield in the target plant is extremely valuable for production planning and yield ramp-up purposes.

- **Transition to New Design Generation:** In order to remain competitive, offer new features, and deal with production quality issues, designs are, sometimes, subjected to re-spins where minor modifications and tweaks are introduced to enhance performance and robustness [45]. Estimating how well the new device generation will yield when

it replaces the prior one in HVM production is, again, an indispensable piece of information.

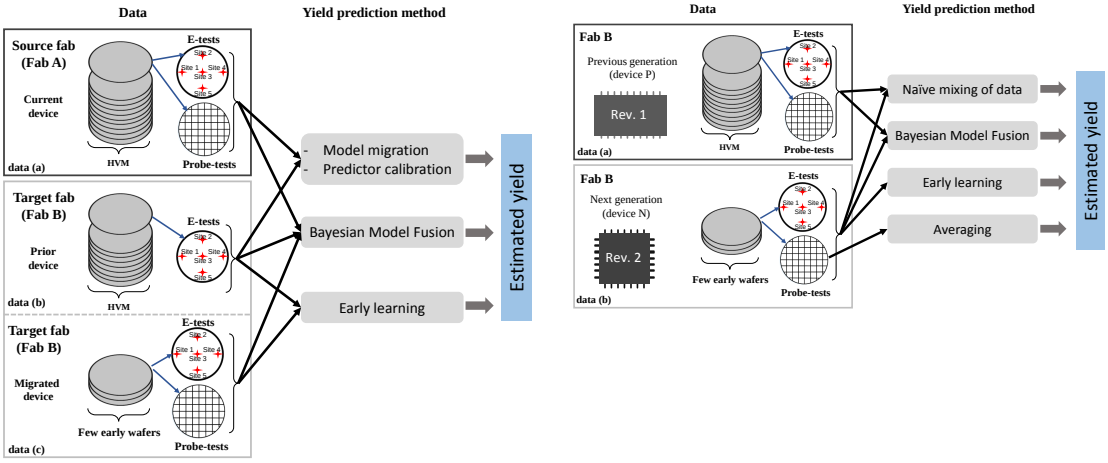
In principle, these two yield estimation problems may be solved by relying on existing simulation-based methods. However, in both scenarios, a large volume of relevant silicon data, such as measurements on devices produced in the source fab, or measurements from the prior generation of a device, is already available. Therefore, this work seeks to develop yield forecasting solutions which rely solely on such silicon measurements; thereby these solutions are not susceptible to PDK accuracy limitations and are applicable regardless of size, complexity and simulation time of a design.

The type of silicon measurements that the proposed methods are based on are the typical *e-test* and *probe-test* data that is obtained and logged as part of a production. E-tests are electrical measurements performed on simple structures known as process control monitors (PCMs), which are typically placed in the scribe lines of the wafer. Probe-tests, on the other hand, are the measurements performed through standard functional or structural tests on every die at wafer level.

In the fab-to-fab production migration scenario, consider a device currently being produced in HVM in a source fab A, whose production will be migrated to a target fab B of the same technology node. In order to predict how well the device will yield in fab B, various methods are experimented which make use of one or more of the following data sources: (a) e-test and probe-test data from HVM production of the device in source fab A; (b) e-test data from HVM production of a *prior* device fabricated recently in the same technology node in fab B; and (c) limited e-test and probe-test data from production of the device in target fab B, originating from a very small number of characterization wafers, which are typically produced prior to ramping-up HVM production. In particular, I examine four different methods, namely *model migration*, *predictor calibration*, *early learning*, and *Bayesian Model Fusion* (BMF). As illustrated in Figure 3.1 (a), the model migration and predictor calibration

methods make use of data sources (a) and (b), the early learning method makes use of data sources (b) and (c), while the BMF method makes use of all three data sources (a)-(c).

In the transition to a new generation scenario, assume a device N, which stems from minor modifications to a previous generation device P, and which is to be produced in HVM in the same fab and technology node as its predecessor. In order to predict how well the device N will yield, an experiment with various methods is performed which make use of one or more of the following data sources: (a) e-test and probe-test data from HVM production of device P; and (b) limited e-test and probe-test data from device N, originating from the few characterization wafers which are typically produced prior to ramping-up HVM production. In particular, I consider four different methods, namely *averaging*, *early learning*, *naive mixing of data*, and *Bayesian Model Fusion* (BMF). As shown in Figure 3.1 (b), the averaging method uses only probe-tests from (b), while all other methods make use of e-test and probe-test data from both (a) and (b).




(a) During fab-to-fab product migration.

(b) Across design generations.

Figure 3.1. Yield prediction overview.

All aforementioned methods, except for the averaging method in the scenario of yield estimation across design generations, establish a model which predicts **wafer yield** (i.e., the fraction of devices on a wafer which pass all their specifications) or **parametric yield** (i.e.,

the fraction of devices on a wafer which pass a given specification) from the e-test profile of the wafer. The underlying conjecture is that there exists sufficient correlation between e-tests and device performances, as they are subject to the same process variations experienced by the wafer. Therefore, variation of device performances and, by extension, wafer or parametric yield, can be predicted sufficiently well through the e-test measurements of a wafer. Such correlations are very intricate and, most often, it is impossible to analyze and explain why they are in force. For this reason, they are extracted using machine learning.

It is important to stress that the proposed methods can expose yield loss whenever its root-cause is also reflected by the e-tests. Yield loss can be due to random defects (e.g., particle contamination) or process variations, which can be further classified into systematic inter-die variations (e.g., lithography-related gate-length variation) and random within-die variations (e.g., random dopant fluctuation) [46]. Evidently, random defects affecting a device do not necessarily affect simultaneously the PCMs. To detect such defects, one could rely, for example, on Iddq measurements or on dedicated on-chip, compact, non-intrusive temperature sensors [47, 48]; yet it is unlikely that such defect-oriented tests can cover the entire design. Thus, similarly to the simulation-based methods, the proposed methods do not concern yield loss due to random defects. On the other hand, there exist numerous PCMs that provide e-tests which can capture effectively both inter-die and within-die variations [49, 50, 51]. Multiple copies of such PCMs are typically dispersed across a wafer, in order to reflect the spatial aspects of process variation, and, collectively, offer valuable information so that process engineers may monitor and adjust the fabrication process. E-test data contain various types of measurements reflecting physical, electrical, and mismatch characteristics of simple layout components (i.e., transistors, resistors, capacitors, etc.) and basic circuits (i.e., ring oscillators, current mirrors, etc.). Thus, as is the case with the simulation-based methods, the focus of the proposed methods is to expose the yield loss component that is due to process variations. Finally, existence of correlation between e-tests and yield should be

verified on a case-by-case basis before the methods can be applied. This can be done based on high-volume silicon data from the source fab or based on a previous-generation device.

### 3.3.1 Yield/E-test correlation

Before attempting to use e-tests as a yield predictor when migrating production across fabs and when transitioning to new design generations, first the use of e-tests as a yield predictor for a specific device fabricated in a specific fab is discussed. Given the nature of e-tests, whose role is to reflect process variations that lead to yield loss and to drive yield learning, the conjecture is that they are correlated with and can serve as an accurate predictor of parametric yield and wafer yield. Such correlations are intricate, and do not have known closed-form mathematical expressions. Therefore, by regression functions one can learn how to approximate them.

Consider a device that is currently in production. Assume that the e-test measurements from $w$ wafers that contain this device and the probe-test measurements from all $n$ devices contained in each of these wafers are available. Let $\boldsymbol{ET}^i = [ET_1^i, \cdots, ET_l^i]$ denote the $l$-dimensional e-test measurement pattern of the $i$-th wafer, where $ET_k^i$ denotes the $k$-th e-test measurement in the $i$-th wafer. Let $\boldsymbol{PT}^{ij} = [PT_1^{ij}, \cdots, PT_d^{ij}]^T$ denote the $d$-dimensional probe-test measurement pattern obtained on the $j$-th device contained in the $i$-th wafer, where $PT_k^{ij}$ denotes the $k$-th probe-test measurement on the $j$-th device in the $i$-th wafer. Let also $\boldsymbol{PT}^i = [\boldsymbol{PT}^{i1} \cdots \boldsymbol{PT}^{in}]$ denote the $d \times n$ matrix of probe-test measurements on the $i$-th wafer.

By knowing the specification limits for the $k$-th probe-test measurement, parametric yield of the $k$-th probe-test measurement for the $i$-th wafer is computed and is denoted by $y_k^i$, as the percentage of devices in the $i$-th wafer that comply with these limits. Let $\boldsymbol{y}^i = [y_1^i, \cdots, y_d^i]$ denote the $d$-dimensional parametric yield vector of the probe-test measurements for the $i$-th wafer. $\boldsymbol{y}^i$ is directly computed from $\boldsymbol{PT}^i$ in conjunction with the specifications of the

probe-test measurements. Let us also consider the wafer yield for the $i$-th wafer, denoted by $Y^i$, which is defined as the percentage of die on a wafer that comply with the specification limits for all probe-tests. In summary, the information available on this device includes

$$\text{wafer}^i = [\boldsymbol{ET}^i, \boldsymbol{y}^i, Y^i], \qquad i = 1, \cdots, w \tag{3.1}$$

The training data in (3.1) is used to learn the regression functions which predict the parametric yield of the $k$-th probe-test measurement or the wafer yield for the $i$-th wafer from its e-test measurement pattern.

$$y_k^i \approx f_k\left(\boldsymbol{ET}^i\right) \tag{3.2}$$

$$Y^i \approx f\left(\boldsymbol{ET}^i\right). \tag{3.3}$$

Once the regression functions are learned and their generalization accuracy is validated, they are ready to be used to estimate the parametric yield $\hat{\boldsymbol{y}}^i$ and the wafer yield $\hat{Y}^i$ for future wafers, i.e., $i > w$, based solely on their e-test profile. I will show that these estimates approximate accurately the ground truth values $\boldsymbol{y}^i$ and $Y^i$, respectively. Accordingly, significant cost savings can be obtained when computing parametric or wafer yield, since only the e-test measurements need to be obtained rather than all probe-test measurements for all devices on a wafer.

### 3.3.2 Regression models

Several methods exist in the literature for multivariate regression, including *Multivariate Adaptive Regression Splines* (MARS), *Least-Angle Regression Splines* (LARS), *Projection Pursuit Regression*, *Feed-Forward Neural Networks* (FFNN), and *Support Vector Machines* [52, 53]. In this work, MARS[53] is used, which has also been successfully used in several other test cost reduction methods in the past [54, 55].

MARS is a non-parametric regression method which is capable of modeling complex non-linear relationships and considers interactions between variables during model construction. MARS builds the regression using basis functions as predictors in place of the original input variables. Generally, it fits the data to the following model.

$$\hat{f}(X) = a_0 + \sum_{m=1}^{M} a_m \cdot B_m(X), \tag{3.4}$$

where $a_0$ is the intercept, $a_m$ denotes the slope parameter, and $B_m(X)$ represents the $m$-th basis function which may include the interaction effect between the original input variables $X$. The basis function transformation enables MARS to blank out certain regions of data and focus on specific sub-regions. When the number of predictors is very high and disproportional to the size of the training set, this capability is used to select a subset of predictors to improve the quality of the regression model. MARS constructs the regression in two phases. In the forward phase, MARS starts with an empty model and enhances it by adding basis functions to overfit the data. Then, in the backward phase, MARS removes basis functions associated with the smallest increase in generalized cross-validation error. MARS models are built using e-tests as input variables and yield vectors as the dependent output variables. The piecewise-cubic basis functions is utilized, and the maximum number of which is set to half of the number of input variables.

### 3.3.3 Model improvement through feature selection

While typically many e-tests are performed, not all of them may be necessary for learning the regression models that estimate yield. In fact, for many of e-tests, there may exist no physical underlying reason why they should be correlated with some probe-test outcomes. Therefore, including them in the model will not only offer no additional value but may even deteriorate its quality due to the curse of dimensionality. Indeed, learning a model in a low dimensional space improves its robustness.

Selecting a subset of e-tests that best correlates to probe-tests and, thereby, to parametric and wafer yield values, is essentially a feature selection problem. Since the number of possible subsets of a set of $n$ features (i.e., e-tests) is $2^n - 1$, exhaustive search is not feasible even for a moderate number of features. In general, as explained in a review presented in [56], feature selection methods are categorized into greedy and heuristic. In the context of semiconductor testing, solutions from both categories have been employed for test compaction [18, 16] and machine learning-based test[17].

In this work, a heuristic-based technique is employed to select a subset of e-test parameters. More specifically, a multi-objective GA, called NSGA-II [57] is used. GAs are evolutionary algorithms attempting to emulate the biological natural selection. The GA starts with an initial random population of solutions (i.e., feature subsets). Mating and mutation operations are repeatedly applied to the current population in order to generate a new population which, hopefully, contains better solutions. In each iteration, the fitness of every instance of the population is evaluated using two objective functions and the best solutions are retained. These two objective functions reflect the goals of employing the smallest possible number of features while achieving the highest possible prediction quality. Evidently, these can be competing objectives, hence the NSGA-II algorithm explores the trade-off space.

Figure 3.2 depicts an overview of the GA-based feature selection method. A bit-string specifies the corresponding e-test subset that will be included in the correlation model (i.e., "1" indicates inclusion, whereas "0" indicates exclusion). The fitness of an e-test subset is assessed by constructing the MARS model using a training dataset and, then, evaluating its prediction accuracy on an independent validation dataset. Fitness, in this case, is the prediction error on the validation dataset, computed as the average difference between true yield values and predicted values by the correlation model. Yield, in this context, could be either the parametric yield for a specific probe-test or the overall wafer yield. It should be noted that different optimal e-test subsets may be selected for each probe-test. The algorithm

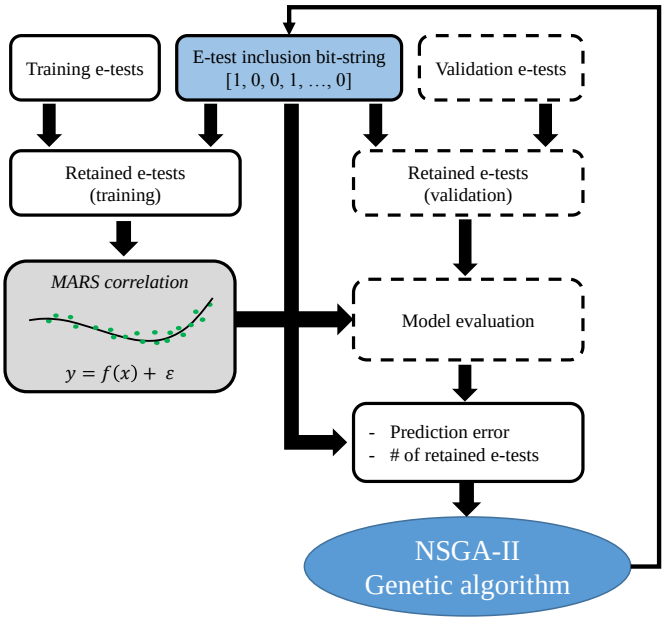



Figure 3.2. GA-based feature selection method (NSGA-II).

stops when there is no significant improvement in the fitness values of a population over a window of the last five generations. Also in each iteration of the GA, the same settings are used in the MARS models.

### 3.3.4 Yield prediction during production migration

Let now consider a device which is currently being fabricated in HVM in fab A and whose production is planned to be migrated to fab B. The goal is to build a model that predicts the HVM parametric yield of each probe-test and of the overall wafer yield in fab B. To this end, different methods will be discussed, exploring a trade-off between simplicity, required input data, and accuracy. Without loss of generality, the formulation considers only parametric yield; overall wafer yield is dealt with in a similar fashion. Each method may make use of one or more input data sources among the ones listed below [58].

- E-test and probe-test measurements from $w_A$ wafers fabricated in fab A, containing the device whose production is being migrated. Following similar notation as in Section 3.3.1, the available information from fab A includes:

$$\text{wafer}_A^i = [\boldsymbol{ET}_A^i, \boldsymbol{y}_A^i, Y_A^i], \qquad i = 1, \cdots, w_A. \tag{3.5}$$

- E-test and probe-test measurements from the first $w_B$ wafers ($w_B \ll w_A$) fabricated in fab B, containing the device whose production is being migrated. In short, information from fab B includes:

$$\text{wafer}_B^i = [\boldsymbol{ET}_B^i, \boldsymbol{y}_B^i, Y_B^i], \qquad i = 1, \cdots, w_B. \tag{3.6}$$

- E-test from a large number, $w_0$, of wafers fabricated in the same technology node in fab B, containing a *prior* device, different than the one whose production is being migrated from fab A to fab B. The only assumption for this prior device is that, since it is fabricated in the same technology, its wafers contain the same e-test PCM structures as the wafers of the device being migrated. The e-test profile of the $i$-th fabricated wafer of this prior device is denoted as $\boldsymbol{ET'}_B^i$, $i = 1, \cdots, w_0$.

**Model migration**

A straightforward approach for predicting yield in fab B is model migration. In this method, a model is first trained in fab A to express parametric yield of a wafer as a function of its e-test profile, $y_{A,k}^i \approx f_{A,k}\left(\boldsymbol{ET}_A^i\right)$. Then, the trained regression function is applied directly to the e-test profile of wafers produced in fab B containing the *prior* device, in order to predict HVM parametric yield as

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0} \sum_{i=1}^{w_0} f_{A,k}\left(\boldsymbol{ET'}_B^i\right). \tag{3.7}$$

Model migration success relies on two assumptions:

1. E-tests in the source fab A and target fab B must come from the same distribution.

2. If a wafer from fab A and a wafer from fab B have the same parametric yield, then they must also have similar e-test profiles, i.e., $\boldsymbol{p}_A(\boldsymbol{y}_A^i \mid \boldsymbol{ET}_A^i) = \boldsymbol{p}_B(\boldsymbol{y}_B^j \mid \boldsymbol{ET}_B^j) \Rightarrow \boldsymbol{ET}_A^i \approx \boldsymbol{ET}_B^j$.

As these assumptions may not necessarily hold true in a semiconductor manufacturing context, the accuracy of model migration is expected to be limited.

### Predictor calibration

Another approach, which does not rely on any of the two aforementioned assumptions, is predictor calibration. The distribution of each e-test (i.e., predictor) in fab B is calibrated based on the distribution of the same e-test in fab A, $\hat{\boldsymbol{ET'}}_{B,j} = h_j(\boldsymbol{ET'}_{B,j}, \boldsymbol{ET}_{A,j})$, where $\boldsymbol{ET}_{A,j} = [ET_{A,j}^1, \cdots, ET_{A,j}^{w_A}]$ and $\boldsymbol{ET'}_{B,j} = [ET_{B,j}^1, \cdots, ET_{B,j}^{w_0}]$ represent the profile of the $j$-th e-test in fab A and fab B, respectively. A simple way of achieving this would be *mean calibration*, which subtracts the mean shift $\Delta(\mu_j)$

$$\hat{\boldsymbol{ET'}}_{B,j} = \boldsymbol{ET'}_{B,j} - \Delta(\mu_j), \tag{3.8}$$

$$\Delta(\mu_j) = \mu(\boldsymbol{ET'}_{B,j}) - \mu(\boldsymbol{ET}_{A,j}). \tag{3.9}$$

However, in order to achieve better precision, other parameters of the distribution, such as variance, skewness and kurtosis, also need to be calibrated. To accomplish this, a two-step procedure is employed. First, using the cumulative distribution function (CDF) of the $j$-th e-test in fab B, $F_{B,j}$, the cumulative probability associated with each sample is identified, $x_j^i = F_{B,j}(ET'^i_{B,j})$. Then, using the inverse CDF of fab A, the e-test value associated with

cumulative probability $x_j^i$ is determined, $\hat{ET'}_{B,j}^{\,i} = F_{A,j}^{-1}(x_j^i)$, where $F_{A,j}^{-1}$ is the inverse CDF of the $j$-th e-test for fab A. The kernel density estimation (KDE) [59] is employed to estimate the CDF of each e-test.

This procedure is applied to all instances of the e-test profile of fab B (i.e., for $i = 1, \cdots, w_0$), and to all e-tests for each instance (i.e., for $j = 1, \cdots, l$).

In order to utilize predictor calibration in yield prediction during production migration, a regression function is trained to express parametric yield in fab A as a function of the e-test profile, i.e., $y_{A,k}^i \approx f_{A,k}\left(\boldsymbol{ET}_A^i\right)$. Then, the trained regression model is applied to the calibrated e-test profile of wafers produced in fab B containing the *prior* device, in order to predict HVM parametric yield as

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0}\sum_{i=1}^{w_0} f_{A,k}\left(\hat{\boldsymbol{ET'}}_B^{\,i}\right). \tag{3.10}$$

Since predictor calibration does not make any of the two assumptions stated earlier, it is expected to outperform model migration. This method is very successful in mapping the distribution of fab B into that of fab A and is capable of predicting yield without requiring probe-test measurements from fab B.

### Early learning

Model migration and predictor calibration were developed in the context of yield prognosis when migrating a device from fab A to fab B, while assuming that no probe-tests are available for this device from fab B. Now consider the scenario where probe-tests are available from a small number $w_B$ of early silicon wafers from fab B, containing this device. This data can be used to train a regression model to express parametric yield as a function of the e-test profile, relying only on the information from fab B, i.e. $y_{B,k}^i \approx f_{B,k}\left(\boldsymbol{ET}_B^i\right)$. Subsequently, this model can be applied to the available e-test profile from the *prior* device produced in fab B, in order to predict HVM parametric yield as

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0} \sum_{i=1}^{w_0} f_{B,k} \left( \boldsymbol{ET'}^{i}_{B} \right). \tag{3.11}$$

**Bayesian Model Fusion (BMF)**

The accuracy of the early learning method may be limited because the regression model is trained using limited, possibly not representative, data from a few initial wafers in fab B. Another more elaborate technique is BMF, which intelligently fuses the limited data from fab B with the rich readily available data from fab A, in order to enhance the prediction accuracy of the early learning method. BMF is a very powerful technique which has been used successfully for model improvement in various contexts [60, 61, 62, 63, 64, 65].

The training data in (3.5) allows to learn an accurate regression function for predicting parametric yield of the $k$-th probe-test in fab A

$$y^{i}_{A,k} \approx f_{A,k} \left( \boldsymbol{ET}^{i}_{A} \right) = \sum_{m=1}^{M} a_{A,k,m} \cdot b_{k,m} \left( \boldsymbol{ET}^{i}_{A} \right). \tag{3.12}$$

This is relied on a general expression of a regression function based on $M$ basis functions, where $b_{k,m}$ is the $m$-th basis function for the $k$-th probe-test and $a_{A,k,m}$ corresponds to the coefficient of the $m$-th basis function for the $k$-th probe-test, $m = 1, \cdots, M$. This general expression can accommodate any regression approach mentioned in Section 3.3.2.

For small $w_B$, given the limited training data in (3.6), the objective is to learn an accurate regression function for fab B

$$y^{i}_{B,k} \approx f^{'}_{B,k} \left( \boldsymbol{ET}^{i}_{B} \right) = \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m} \left( \boldsymbol{ET}^{i}_{B} \right), \tag{3.13}$$

where $a_{B,k,m}$ is the coefficient of the $m$-th basis function for the $k$-th probe-test corresponding to fab B.

The conventional learning procedure is to use a fraction of the data in (3.6) for training and the rest for assessing the generalization ability of the regression function on previously

unseen wafers. However, since the ultimate goal is to learn the regression function based on the very first few wafers, the data in (3.6) is not representative enough to learn a regression function that accurately predicts the parametric yield of future wafers. The aim of the BMF technique is to learn the regression function in (3.13) by leveraging information from the data in (3.5), which was produced in fab A.

The BMF learning procedure consists of solving for the coefficients $\boldsymbol{a}_{B,k} = [a_{B,k,1}, \cdots, a_{B,k,M}]$ that maximize the *posterior* distribution $\text{pdf}(\boldsymbol{a}_{B,k}|\mathbf{wafer}_B)$, that is,

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}|\mathbf{wafer}_B), \tag{3.14}$$

where $\mathbf{wafer}_B = [\text{wafer}_B^1, \cdots, \text{wafer}_B^{w_B}]$. In this way, the "agreement" of the selected coefficients is maximized with the limited observed data from fab B.

By applying Bayes' theorem, it becomes

$$\text{pdf}(\boldsymbol{a}_{B,k}|\mathbf{wafer}_B) \propto \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k}). \tag{3.15}$$

Thus, the problem boils down to

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k}). \tag{3.16}$$

Next, expressions for the *prior* distribution $\text{pdf}(\boldsymbol{a}_{B,k})$ and the *likelihood function* $\text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k})$ are developed.

Assuming that the coefficients $a_{B,k,m}$ are independent, one can write

$$\text{pdf}(\boldsymbol{a}_{B,k}) = \prod_{m=1}^{M} \text{pdf}(a_{B,k,m}). \tag{3.17}$$

I define the *prior* distribution $\text{pdf}(a_{B,k,m})$ by involving the prior knowledge from fab A. Specifically, $\text{pdf}(a_{B,k,m})$ is assumed to follow a Gaussian distribution with mean $a_{A,k,m}$ and standard deviation $\lambda|a_{A,k,m}|$

$$\text{pdf}(a_{B,k,m}) = \frac{1}{\sqrt{2\pi}\lambda|a_{A,k,m}|} \cdot \exp\left[-\frac{(a_{B,k,m} - a_{A,k,m})^2}{2\lambda^2 a_{A,k,m}^2}\right]. \tag{3.18}$$

This approach accounts for the fact that $a_{B,k,m}$ is expected to be similar to $a_{A,k,m}$ and deviate from it according to the absolute magnitude of $a_{A,k,m}$.

The *likelihood function* $\text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k})$ is expressed in terms of the data in (3.6). Specifically, since the data from each wafer is independent, it can written

$$\text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k}) = \prod_{i=1}^{w_B} \text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right). \tag{3.19}$$

Furthermore,

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \text{pdf}(\varepsilon^i), \tag{3.20}$$

where $\varepsilon^i$ is the prediction error introduced by the regression for the $i$-th wafer in fab B

$$\varepsilon^i = y_{B,k}^i - f_{B,k}\left(\boldsymbol{ET}_B^i\right). \tag{3.21}$$

This error is a random variable that is assumed to follow a zero-mean Gaussian distribution with some standard deviation $\sigma_0$

$$\text{pdf}(\varepsilon^i) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left(-\frac{\left(\varepsilon^i\right)^2}{2\sigma_0^2}\right). \tag{3.22}$$

Therefore, combining (3.20), (3.21), (3.22), and (3.13), one can write

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \cdot \exp\left\{-\frac{1}{2\sigma_0^2} \cdot \left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2\right\}. \tag{3.23}$$

By combining (3.17), (3.18), (3.19), and (3.23), the expression of $\text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k})$ is obtained. By taking the natural logarithm of this expression, the maximization problem in (3.16), after eliminating constant terms, becomes

$$\max_{\boldsymbol{a}_{B,k}} -\left(\frac{\sigma_0}{\lambda}\right)^2 \sum_{m=1}^{M} \frac{\left(a_{B,k,m} - a_{A,k,m}\right)^2}{a_{A,k,m}^2} - \sum_{i=1}^{w_B}\left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2. \tag{3.24}$$

The optimal values of $\sigma_0$ and $\lambda$ are determined by $k$-fold cross-validation [52, 53].

Finally, the HVM parametric yield of each $k$ probe-test is computed as

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0} \sum_{i=1}^{w_0} f'_{B,k} \left( \boldsymbol{ET'}^{i}_{B} \right). \tag{3.25}$$

### 3.3.5 Yield prediction across design generations

Consider a device N, which is the new generation of a previously designed device P, introducing slight modifications and improvements, and let assume that device N is planned to be produced in HVM in the same technology node and fabrication facility where device P was produced. Finally, suppose that for device P the e-test and probe-test data from $w_P$ wafers are in hand. Using similar notation as in Section 3.3.1, information from device P includes

$$\text{wafer}^{i}_{P} = [\boldsymbol{ET}^{i}_{P}, \boldsymbol{y}^{i}_{P}, Y^{i}_{P}], \qquad i = 1, \cdots, w_P. \tag{3.26}$$

Let also assume the availability of the e-test measurements from the first $w_n$ wafers which contain device N as well as the probe-tests from all devices contained in each of these wafers. This information includes

$$\text{wafer}^{i}_{N} = [\boldsymbol{ET}^{i}_{N}, \boldsymbol{y}^{i}_{N}, Y^{i}_{N}], \qquad i = 1, \cdots, w_n. \tag{3.27}$$

Given the above information, below four solutions to the problem of yield prediction across design generations are discussed. Without loss of generality, the focus is on estimating wafer yield, accounting for the fact that devices N and P may not necessarily have the exact same probe-tests.

### 3.3.6 Averaging

A simple and straightforward approach is to compute the average yield of the $w_n$ early wafers and use it as an estimation of HVM wafer yield of device N

$$\hat{\bar{Y}}_N = \frac{1}{w_n} \sum_{i=1}^{w_n} Y^{i}_{N}. \tag{3.28}$$

**Early learning**

Another approach is to use the data in (3.27) as a training set and learn a regression model to express wafer yield as a function of the e-tests for device N

$$Y^i \approx f_N\left(\boldsymbol{ET}^i_N\right). \tag{3.29}$$

The HVM wafer yield of device N can, then, be predicted by employing the e-test profile of device P

$$\hat{\bar{Y}}_N = \frac{1}{w_P}\sum_{i=1}^{w_P} f_N\left(\boldsymbol{ET}^i_P\right). \tag{3.30}$$

**Naive mixing of data**

A third approach is to naively mix data in (3.26) and (3.27), use the combined data as a training set, and learn a regression model to express wafer yield as a function of the e-tests

$$Y^i \approx f_{PN}\left(\boldsymbol{ET}^i\right). \tag{3.31}$$

The HVM wafer yield of device N can, then, be predicted as

$$\hat{\bar{Y}}_N = \frac{1}{w_P}\sum_{i=1}^{w_P} f_{PN}\left(\boldsymbol{ET}^i_P\right). \tag{3.32}$$

**Bayesian Model Fusion**

Finally, similar to Section 3.3.4, one can intelligently combine the information from the prior generation device P with the new generation device N using BMF. In particular, for devices P and N regression models can be learned

$$Y^i_P \approx f_P\left(\boldsymbol{ET}^i_P\right) = \sum_{m=1}^{M} a_{P,m}\cdot b_m\left(\boldsymbol{ET}^i_P\right) \tag{3.33}$$

and

$$Y_N^i \approx f^{'}{}_N\left(\boldsymbol{ET}_N^i\right) = \sum_{m=1}^{M} a_{N,m} \cdot b_m\left(\boldsymbol{ET}_N^i\right), \tag{3.34}$$

respectively. These regression models are based on $M$ basis functions, where $b_m$ is the $m$-th basis function, and $a_{P,m}$ and $a_{N,m}$ correspond to the coefficient of the $m$-th basis function for devices P and N, respectively. The coefficients $\boldsymbol{a}_P = [a_{P,1}, \cdots, a_{P,M}]$ of regression model $f_P$ can be learned accurately based on the rich dataset in (3.26). The coefficients $\boldsymbol{a}_N = [a_{N,1}, \cdots, a_{N,M}]$ of regression model $f_N^{'}$ are learned by maximizing the posterior distribution

$$\max_{\boldsymbol{a}_N} \text{pdf}(\boldsymbol{a}_N|\mathbf{wafer}_N), \tag{3.35}$$

where $\text{pdf}(\boldsymbol{a}_N|\mathbf{wafer}_N) \propto \text{pdf}(\boldsymbol{a}_N)\text{pdf}(\mathbf{wafer}_N|\boldsymbol{a}_N)$, $\text{pdf}(\boldsymbol{a}_N)$ is the *prior* distribution, $\text{pdf}(\mathbf{wafer}_N|\boldsymbol{a}_N)$ is the *likelihood function*, and $\mathbf{wafer}_N = [\text{wafer}_N^1, \cdots, \text{wafer}_N^{w_n}]$. Similar steps as in Section 3.3.4 can be applied to refine the regression functions for the new-generation device N.

The HVM wafer yield of device N can now be predicted as

$$\hat{Y}_N \approx= \frac{1}{w_P} \sum_{i=1}^{w_P} f_N^{'}\left(\boldsymbol{ET}_P^i\right). \tag{3.36}$$

## 3.4 Experimental Results

### 3.4.1 Case study and datasets

In order to experimentally evaluate the various yield prediction methods during fab-to-fab production migration and during transition to a new design generation, actual HVM production datasets from two consecutive design generations of a Texas Instruments 65nm RF transceiver are used. These two design generations are referred as device P and device

N, respectively, emphasizing that device N is the new-generation of device P with slight enhancements. These datasets originate from two geographically dispersed fabs, which will be referred to as fab A and fab B. Device P is produced only in fab B, while device N is produced in both fabs. The dataset for device N from both fabs and the dataset for device P from fab B will be used for yield prediction during fab-to-fab production migration. The dataset of device N from fab B and the dataset from device P from fab B will be used for yield prediction across design generations.

As illustrated in Figure 3.3, the dataset for device N from fab A includes $l$=54 e-tests and $d$=200 probe-tests from a total of $w_A$=500 wafers. Each wafer has 5 e-test measurement sites and approximately 1500 dies per wafer. The dataset for device N from fab B includes the same e-tests and probe-tests from a total of $W_B$=1600 wafers, with the only difference being that e-tests are obtained on 9 instead of 5 e-test measurement sites. These two datasets were obtained from the two fabs at approximately the same time period. The dataset for device P from fab B includes $l$ =54 e-tests (i.e., the same as for device N) and $d_P$=160 probe-tests (i.e., fewer and different than those for device N) from a total of $w_P$=700 wafers. Each wafer has 9 e-test sites and approximately 1500 dies per wafer.

Since several e-test measurement sites are available across each wafer (i.e., 5 e-test measurement sites across wafers produced in fab A and 9 e-test measurement sites across wafers produced in fab B), its e-test signature is generated as the means and standard deviations of the 54 e-tests, as computed across all the available e-test measurements sites. Thus, in all cases, the e-test signature of a wafer has a total of 108 features.

Probe-tests include both structural tests (i.e., open/short circuit, IDDQ, input voltage threshold, etc.) and functional tests (i.e., BER, EVM, CMMR, etc.). E-test measurements include gate-oxide quality, leakage current, threshold voltage, effective channel length, etc. The specification limits for the probe-tests are also available, hence for each of the two fabs the parametric yield of each probe-test on every wafer, as well as the overall yield of each wafer are computed.

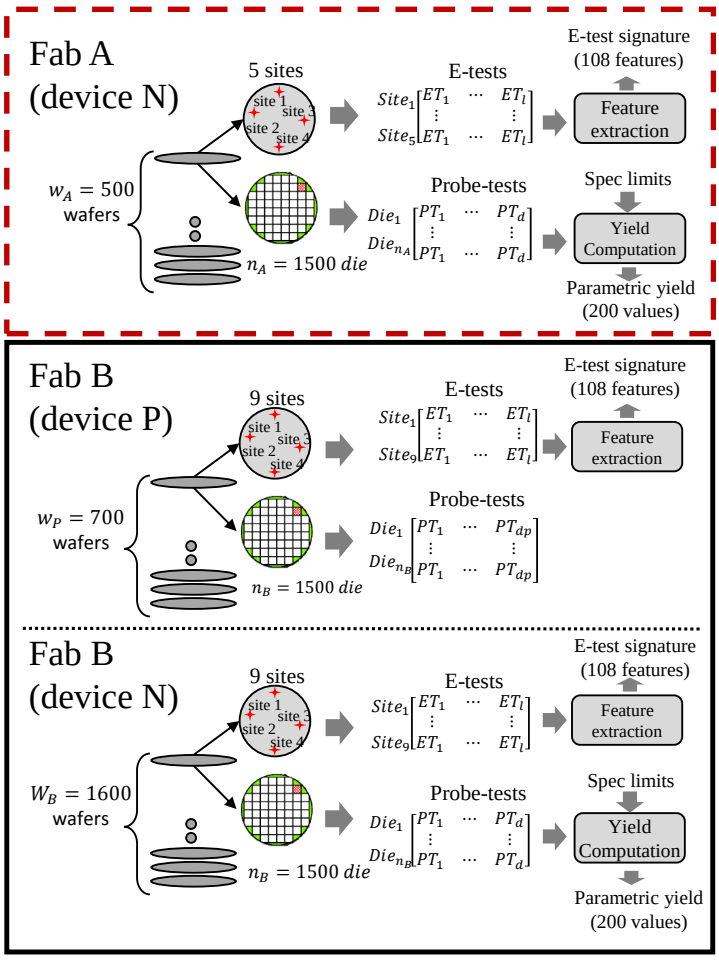



Figure 3.3. Datasets from fab A and fab B.

These datasets are used to:

- Quantify the accuracy of predicting parametric yield of probe-tests and overall wafer yield from the e-test signature of the wafer.
- Demonstrate that this prediction accuracy is improved when employing dimensionality reduction through a GA-based feature selection algorithm.
- Quantify the accuracy of the discussed methods for predicting yield during fab-to-fab production migration.

- Quantify the accuracy of the discussed methods for predicting yield across design generations.

### 3.4.2 Predicting yield from the e-test signature of the wafer

In order to quantify the accuracy of predicting parametric yield of probe-tests based solely on e-tests collected on the wafer, the entire datasets of device N from both fab A and fab B are used to perform two independent experiments, one for each fab. The regression models are trained using MARS and 5-fold cross validation is used to report robust prediction error values. Specifically, for a given fab, the dataset is divided into 5 folds, where 4 folds are used for training and the remaining fold is used for validation. The procedure is repeated such that all folds are left out once as a validation set and, in the end, the average prediction error across the 5 iterations is reported.

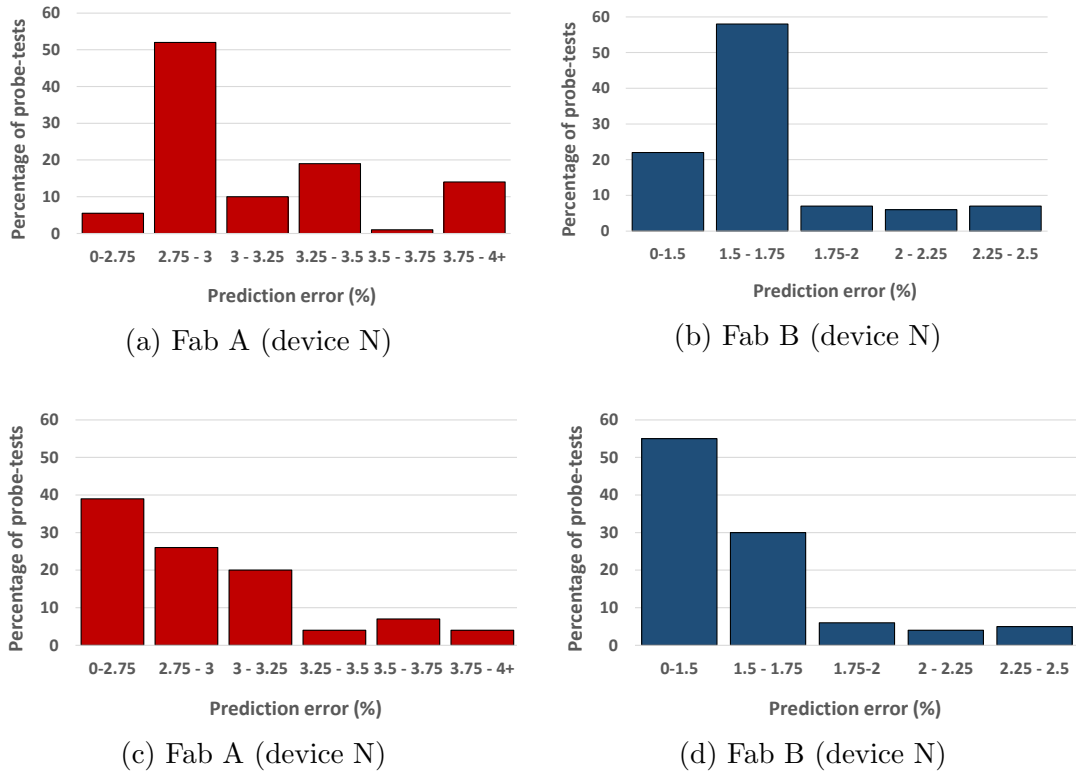
The following expression is used for calculating the error in predicting the parametric yield of the $k$-th probe test

$$\delta_k = 100 \cdot \frac{1}{w} \sum_{i=1}^{w} \frac{|\hat{y}_k^i - y_k^i|}{y_k^i}, \tag{3.37}$$

where $w$ is the number of wafers in the validation set, while $\hat{y}_k^i$ and $y_k^i$ are the predicted and the actual parametric yield values of the $k$-th probe-test on the $i$-th wafer, respectively.

Figures 3.4(a)-(b) present the parametric yield prediction results for the datasets of device N from fab A and fab B, respectively. In this experiment, all 108 e-test features are considered. In each histogram, the horizontal axis is the prediction error, while the vertical axis shows the percentage of probe-tests that are predicted within a given error range. For example, the first bar of Figure 3.4(a) shows the percentage of probe-tests for which the parametric yield prediction error is below 2.75%, with the corresponding value being 5%. As may be observed for both fabs, the parametric yield of the majority of probe-tests can be predicted using

e-tests with an error of less than 3%, corroborating that parametric yield can be predicted very accurately from the e-tests of a wafer.




(a) Fab A (device N)

(b) Fab B (device N)

(c) Fab A (device N)

(d) Fab B (device N)

Figure 3.4. Average parametric yield prediction error for fabs A and B. In (a) and (b) all e-test features are used while in (c) and (d) a subset of e-tests are selected by GA prior to building regression models.

Figures 3.4(c)-(d) present the same results as in Figures 3.4(a)-(b), but this time using only the subset of e-test features that are selected by the GA-based feature selection method of Section 3.3.3. Feature selection is performed individually for each probe-test, thus each probe-test has its own subset of e-tests to build a regression model from. Figures 3.4(c)-(d) show that, for both fabs, most of the weight of the histograms is further towards the left side, i.e., towards smaller prediction errors, as compared to the histograms of Figures 3.4(a)-(b). These results corroborate that, by reducing the dimensionality of the e-test signature, feature selection improves significantly the quality of predictions. It should be noted that the MARS

algorithm does have its own internal feature selection method, which picks a subset of the most relevant e-tests; nevertheless, performing an *a priori* feature selection using a GA appears to be improving further the quality of the prediction models.

Next, the use of e-tests for predicting wafer yield is examined. As before, the regression models are trained using MARS, and 5-fold cross validation is used to report robust prediction errors values, and a similar expression is employed for evaluating the prediction error of the overall wafer yield

$$\delta = 100 \cdot \frac{1}{w} \sum_{i=1}^{w} \frac{|\hat{Y}^i - Y^i|}{Y^i} \tag{3.38}$$

where $w$ is the number of wafers in the validation set, while $\hat{Y}^i$ and $Y^i$ are the predicted and the actual wafer yield values of the $i$-th wafer, respectively. Table 3.1 presents the wafer yield prediction error for both fabs, first when training regression models using all e-test features, and then when training regression models using only the subset of e-tests chosen by the GA-based feature selection method. As may be observed, the prediction error for both fabs is very low and confirms that e-tests of a wafer carry sufficient information regarding quality of the fabricated silicon, thus, they can be successfully used for wafer yield prediction. Similar to parametric yield prediction, incorporating the feature selection method to reduce the cardinality of the e-test signature results in lower prediction error. In order to quantitatively demonstrate this improvement, the metric $\Delta\epsilon$ is used, and defined as

$$\boldsymbol{\Delta}\epsilon = |\frac{\text{All e-tests error - Subset of e-tests error}}{\text{All e-tests error}} \times 100|. \tag{3.39}$$

Using this metric, the GA-based feature selection method reduces the wafer yield prediction error by 12% and 17% for fab A and fab B, respectively.

Since GA-based feature selection improves the quality of the regression models, as demonstrated in Figure 3.4 and Table 3.1, for the rest of experiments all regression models are trained with the subset of e-tests selected by this method.

Table 3.1. Wafer yield prediction error

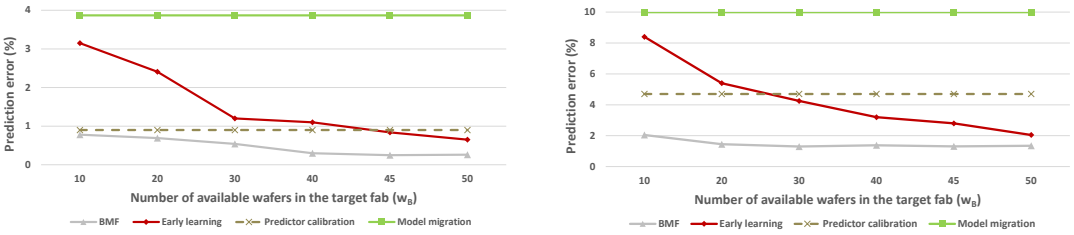
| Parameter | All e-tests | Subset of e-tests | improvement ($\Delta\epsilon$) |
|---|---|---|---|
| Fab A (device N) | 6.12% | 5.41% | 12% |
| Fab B (device N) | 4.9% | 4.05% | 17% |

### 3.4.3 Yield prediction during migration from fab A to fab B

In order to quantify yield prediction accuracy during fab-to-fab production migration using the methods discussed in Section 3.3.4, the following experiment is performed using fab A as the source fab and fab B as the target fab. The model migration and predictor calibration methods assume access to both e-tests and probe-tests of device N in fab A, as well as to the e-tests of device P in fab B. In other words, device P is used as the *prior* device in these methods. The BMF and early learning methods assume, in addition, access to both e-tests and probe-tests for device N in fab B from a small number of $w_B$ early engineering wafers, where $w_B \ll W_B$. $w_B$ is varied in the range [10, 50], in order to study the influence of the size of this training set on BMF and early learning.

Since $w_B$ is small, the results for the BMF and early learning methods may vary with respect to the subset of $w_B$ out of $W_B$ wafers that is being used. For this reason, bootstrapping is employed to report robust prediction errors, smoothen them, and assist with the interpretation of the overall results. In total, 10 bootstrap iterations are performed and, in each iteration, $w_B$ wafers are sampled uniformly at random from the $W_B$ wafers using 5-fold cross validation. The reported prediction errors are averaged over these 50 iterations. In each iteration, the following expressions are used for evaluating the prediction error of the HVM parametric yield of the $k$-th probe-test and the HVM wafer yield

$$\delta_k = 100 \cdot \frac{|\hat{\bar{y}}_{B,k} - \bar{y}_{B,k}|}{\bar{y}_{B,k}}, \tag{3.40}$$




(a) A randomly-selected probe-test (b) Results for overall wafer yield

Figure 3.5. Yield prediction error during production migration.

$$\delta = 100 \cdot \frac{|\hat{\bar{Y}}_B - \bar{Y}_B|}{\bar{Y}_B}, \tag{3.41}$$

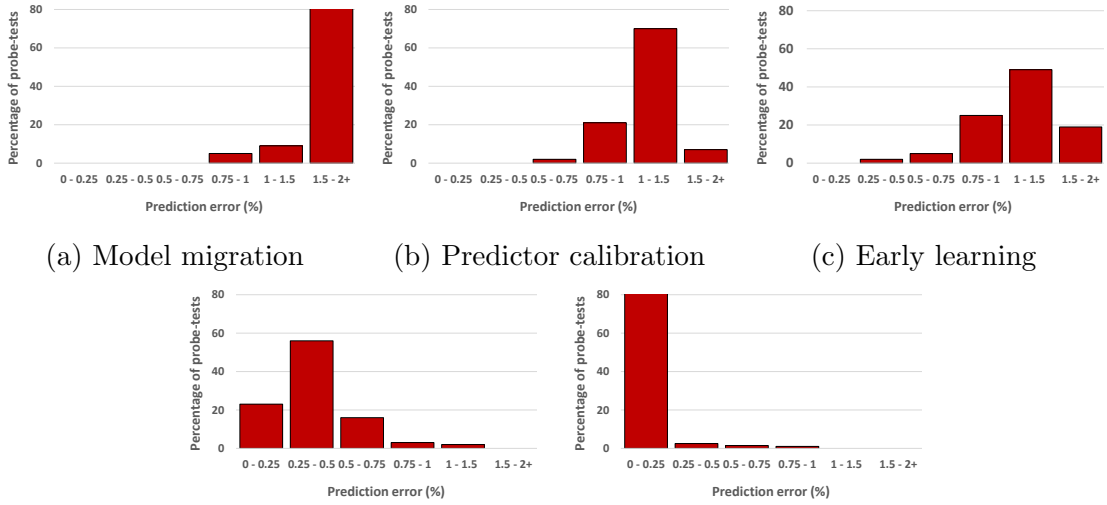
where $\hat{\bar{y}}_{B,k}$ and $\bar{y}_{B,k}$ are the predicted and actual HVM parametric yield values of the $k$-th probe-test, respectively, while $\hat{\bar{Y}}_B$ and $\bar{Y}_B$ are the predicted and actual HVM wafer yield values in fab B, respectively.

The accuracy of the yield prediction methods of Section 3.3.4 is demonstrated in Figures 3.5(a) and (b), for one randomly-chosen probe-test and for the overall wafer yield, respectively. These plots show the prediction error as a function of the training set size $w_B$. The model migration and predictor calibration methods do not utilize any information from fab B for training purposes. They only rely on the e-tests of the *prior* device P in fab B. Therefore, the corresponding curves for these two methods are flat and independent of $w_B$.

As may be seen in Figure 3.5, model migration shows the worst performance, which is expected since it naively uses the model that is learned on data from fab A for predicting yield in fab B. Early learning strongly depends on the size of the training set. The prediction error is small for large $w_B$ and increases exponentially as the training size becomes smaller. This is expected, since the information available for training is weakened and the ability to extrapolate the regression towards the tails of the distribution deteriorates, resulting in large prediction error on the validation set. Predictor calibration outperforms model migration and, in the case of small $w_B$, it also outperforms early learning, despite the fact that it does not use any information from fab B.

BMF outperforms all other methods regardless of the size of training set $w_B$. It shows a remarkably stable behavior, maintaining nearly constant prediction error even when the training set size is very small. This implies that, by incorporating prior knowledge from fab A, BMF is capable of generating accurate prediction models for fab B based only on a few early wafers from fab B. Thus, BMF can be used to quickly estimate yield from a few engineering wafers or from the first few wafers in HVM, without having to wait until a large volume of data is collected. This result, showing that the BMF method reduces the burden of collecting large datasets for yield estimation, is consistent with the outcome of other studies that employ the BMF method in different contexts [60, 61, 62, 63, 64, 65].

Finally, Figure 3.6 compares the cumulative results for all 200 probe-tests, in the scenario where production is migrated from fab A to fab B and $w_B = 30$. Individual histograms are provided for each method. For comparison purposes, a "lower bound" result is also included where the early learning method is applied by employing all available $W_B$ wafers. This corresponds to having sufficient statistics for the distribution of e-tests and probe-tests in the target fab, hence the quality of prediction depends only on the correlation between e-tests and probe-tests and the ability of the regression functions to capture it. In these histograms, each bar shows the percentage of probe-tests that have a yield prediction error within a specific range. As may be seen, the histogram of the BMF method has most of its weight on the left side, i.e. towards smaller prediction errors, as compared to the histograms of the other three methods. The yield prediction results for the BMF method are also closer to the lower bound results. Therefore, the BMF method provides the best option for predicting parametric yield, provided that a few early characterization wafers are available. If such wafers are not readily available, then between the two applicable methods, i.e., model migration and predictor calibration, the latter provides the best parametric yield prediction results.




(a) Model migration (b) Predictor calibration (c) Early learning

(d) BMF (e) Lower bound ($w_B$=$W_B$)

Figure 3.6. Yield prediction error across all 200 probe measurements during fab A to fab B production migration with $w_B = 30$.

### 3.4.4 Yield prediction across design generations

In order to quantify yield prediction accuracy across design generations using the methods discussed in Section 3.3.5, the following experiment is performed using the datasets of devices N and P from fab B. The averaging method assumes access only to $w_n \ll W_B$ early characterization wafers of the next-generation device N; in this experiment, wafers from the first two lots in the dataset are used. In addition, the rest of the methods assume access to the entire dataset of the previous-generation device P. Also 10 bootstrap iterations are performed and, in each iteration, $w_B$ wafers are sampled uniformly at random from the available $w_n$ wafers and 5-fold cross validation is performed. The reported prediction errors are averaged over these 50 iterations. The experiment is repeated by varying $w_B$ in the range [10, 50]. The following expression is used for evaluating prediction error of the HVM overall wafer yield of device N

$$\delta = 100 \cdot \frac{|\hat{\bar{Y}}_N - \bar{Y}_N|}{\bar{Y}_N}, \tag{3.42}$$

where $\hat{\bar{Y}}_N$ and $\bar{Y}_N$ are the predicted and actual HVM wafer yield values for device N, respectively.

Figure 3.7 shows the yield prediction error as a function of the number of available wafers $w_B$ in the training set. As may be seen, BMF again outperforms the other methods, regardless of the training set size. It shows a remarkably stable behavior, maintaining steady HVM yield prediction error even when the training set size is as small as 10 wafers. This shows that, by statistically fusing prior knowledge from the previous-generation device P, BMF is capable of providing a very accurate HVM yield prediction model for the new-generation device N, based on only a few early characterization wafers. Therefore, BMF can be used for fast and precise forecasting of HVM wafer yield, without having to wait until a large volume of data is collected. The second best method is the averaging method. Its stable behavior implies that the wafer yield in the first two lots that are included each time in the training set is very similar. Averaging is outperformed by BMF, since the wafers in the first two lots are not necessarily representative of HVM statistics. Success of early learning depends strongly on the size of the training set. The prediction error is low for large $w_B$ and exponentially increases as $w_B$ becomes smaller. This is anticipated, since the information content of the training set is weakened, becoming biased and non-representative of HVM, and the regression model is unable to extrapolate towards the tails of the distribution, resulting in large prediction error. The accuracy of naive mixing improves slightly as the number of training samples from device N increases. The fact that the accuracy of this method is inferior implies that the datasets from devices P and N do not exhibit strong similarity and/or that the rich dataset from device P overshadows the limited dataset from device N.

To gain better insight, consider $w_B = 20$ and Figure 3.8 illustrates the distribution of wafer level prediction error for all wafers in the validation set for the BMF and early learning methods. The prediction error is expressed as

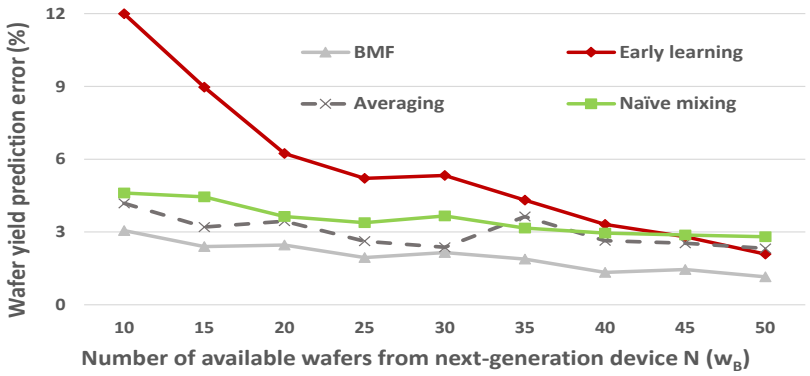
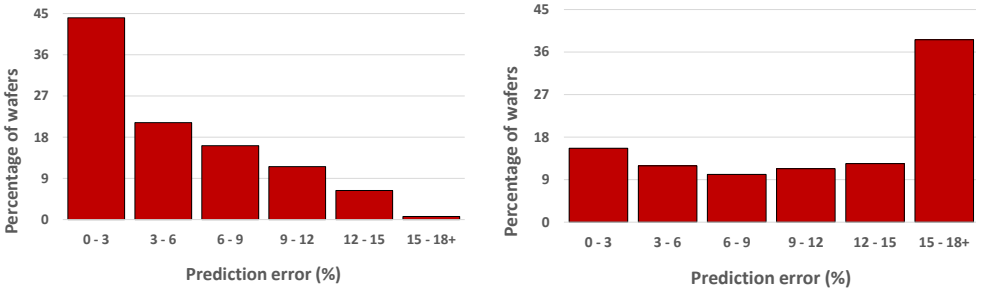



Figure 3.7. Error in predicting device N yield from early wafers.

$$\delta_i = 100 \cdot \frac{|\hat{Y}_N^i - Y_N^i|}{Y_N^i}, \tag{3.43}$$

where $\hat{Y}_N^i$ and $Y_N^i$ are the predicted and actual wafer yield values for the $i$-th wafer, respectively. In each histogram, the horizontal axis represents the prediction error range and the vertical axis represents the percentage of the wafers in the validation set whose wafer yield is predicted within a given error range. As may be seen, for the BMF method the histogram is skewed to the left, showing that the wafer yield of the majority of the wafers is predicted accurately, whereas for the early learning method the histogram is skewed to the right, showing that the wafer yield of about half of the wafers is predicted with error greater than 12%.




(a) BMF (b) Early learning

Figure 3.8. Wafer yield prediction error of device N with $w_B = 20$.

## 3.5 Conclusion

This chapter introduced and compared several methods for yield prediction during fab-to-fab production migration and during transition to a new design generation. In these two yield prediction scenarios, plenty of silicon data is already available, therefore making the use of simulation-based methods, which may be time-consuming and of limited accuracy, unnecessary. The proposed methods span a range of sophistication levels and make use of increasingly rich datasets, including HVM silicon data from the source fab or the previous-generation device, as well as silicon data from a few early characterization wafers from the target fab or the new-generation device, respectively. All methods, except for the simplest ones, capitalize on the existence of correlation between the e-test profile of a wafer and its yield. Effectiveness of the proposed methods was evaluated using large datasets obtained from two different fabs which produced two generations of a Texas Instruments 65nm RF transceiver device. Among the options discussed, the most advanced BMF method which intelligently combines data from the source and target fab or from the previous-generation and next-generation devices, outperforms all other more straightforward methods and offers a highly accurate yield prediction solution during production migration and design generation transition, respectively.

# CHAPTER 4

# FAB-OF-ORIGIN ATTESTATION[1]

## 4.1 Overview

As the semiconductor industry has largely adopted the fab-less paradigm and as globalization has amplified concerns regarding integrity of the electronics supply chain, the ability to definitively identify the fabrication facility wherein an IC was manufactured has become imperative. Such a fab-of-origin attestation ability could constitute the cornerstone for numerous applications in the electronics industry, including intellectual property (IP) protection, licensing enforcement, quality and hardware integrity assurance, supply chain risk management, counterfeit IC detection and failure analysis, among others.

The importance of fab-of-origin attestation is highlighted by a recent US government research initiative whose objective is *to devise methodologies which use measurable electronic or physical characteristics for determining the specific fabrication facility of origin of a given electronic component* [66]. The various methods developed under this initiative seek to leverage the specifics of a manufacturing process, such as the use of particular materials or geometric rules during fabrication, in order to identify the fab-of-origin. Utilizing on-die laser markings during fabrication, atomic force microscopy (AFT), nanoscale structural, mechanical and electrical characterization based on transmission electron microscopy, device characterization, and using features of spectroscopic chemical signals from electronic components for identifying the source fab are among the explored directions [66]. All of these approaches, however, require additional complicated steps during manufacturing or specialized and expensive equipment during characterization in order to perform fab-of-origin attestation.

[1]2016 IEEE Adapted/Reprinted, with permission, from Ali Ahmadi, Mohamad-Mahdi Bidmeshki, Amit Nahar, Bob Orr, Michael Pas and Yiorgos Makris, "A machine learning approach to fab-of-origin attestation", in Proceedings of IEEE International Conference on Computer-Aided Design ©2016 IEEE

In contrast to the aforementioned design-independent approaches to fab-of-origin attestation, this Chapter introduces a learning-based methodology which leverages the *interaction* between the idiosyncrasies of a fabrication facility and a particular design [67]. Specifically, solutions for four variants of the fab-of-origin attestation problem are developed. The first two variants assume availability of the test data profile from the ratified fabrication facility *only*, and seek to attest whether a single chip or a batch of chips, respectively, has been fabricated therein or not. The other two variants assume availability of the test data profile from *all* facilities which fabricate this chip and seek to identify whether a single chip or a batch of chips, respectively, were fabricated in the ratified fab or not. The proposed solutions rely only on the typical parametric test measurements of a fabricated IC and require neither knowledge of the design, nor any additional provisions during manufacturing or any specialized measurement equipment.[2]

Effectiveness of the proposed solutions is demonstrated using two large industrial datasets from a 65nm Texas Instruments RF transceiver produced in two geographically dispersed fabrication facilities. Considering that alternate fabrication facilities within the same company are highly tuned to resemble each other as much as possible, I point out that the evaluation is performed not only using realistic datasets but also ones that are very hard to tell apart.

## 4.2 Applications of Fab-of-Origin Attestation

In semiconductor manufacturing, performance parameters of a device varies during high-volume production due to process variations. These variations appear at different scales in time and space: intra-die, die-to-die, wafer-to-wafer, and lot-to-lot. Layout and topography interaction of design with process results in intra-die variation. The wafer-level variation is caused by equipment non-uniformity and other physical effects such as wafer spinning and

[2]While in this work only probe-test data is considered, should on-die process control measurement (PCM) data be available, they can be seamlessly integrated into these solutions.

thermal gradient. Wafer-to-wafer variation is typically caused by drift in process equipment operations from one wafer to the next. The lot-to-lot variation which is significantly larger than previously-mentioned variations occurs due to process control and maintenance operations between lots.

The impact of the process variations during high-volume production is significant such that fabricated devices express dissimilar performance/power profile. The dissimilarity is even more evident when the comparison is between profile of ICs for the same design and process which are manufactured in two geographically dispersed fabrication facilities. This is anticipated, as equipment age, manufacturing tool installations, chemical sources, as well as altitude and geomagnetic location of the fabrication facility lead to systematic disparities in the resulting products of different fabs. Moreover, fabrication companies typically manufacture several products in the same production line. Thus, it might be the case that the current hardware/software setup as well as process table in one facility result in superior performances, while in another cause performance degradation. In summary, when a design fabricated in two different fabrication facilities, manufactured ICs have dissimilarities in the following aspects:

- **Performance parameters**: profile of parameters such as delay, power, etc. are dissimilar due to difference in equipment, tools and process material of production line.

- **Certificates and standards**: devices produced in a fabrication facility fail to meet certain certifications and standards due to equipment age, tools or process materials, and the training of personnel.

- **Security concerns**: the geographical location of a fabrication facility may introduce security concerns, such as intellectual property theft or the insertion of hardware Trojans, for a specific customer.

Considering the above-mentioned dissimilarities, the next part discusses a set of applications in which identifying the fabrication facility where a component was manufactured is crucial. These applications are divided in two broad categories as the following:

### 4.2.1 Risk management

As the semiconductor industry has largely adopted the fab-less paradigm, design, fabrication and distribution of today's electronic components has amplified concerns regarding performance, reliability, integrity and security of ICs. Consider a customer that needs a specific IC product which is produced in several fabrication facilities, and this product will be integrated with other modules in a system. During the integration process, customer has experienced that ICs of a specific facility will result in better performance/yield and reliability of end application. Therefore, the preference is ICs which were fabricated in that fab. Another example could be a customer who wants devices which are manufactured in facilities that comply with medical/automotive/military standards and certificates. Similar concerns need to be considered when fab-less customers want to fabricate their own design, and ask a foundry to produce it in a specific fabrication facility. In this case, geographical location of a fab might be important due to trust concerns.

In all these situations, customers are highly interested to attest the fab-of-origin of ICs, in order to meet the end application requirements.

### 4.2.2 Litigation

Another application of fab-of-Origin is to handle litigation challenges for both the IC manufacturer and the IC customers. A manufacturer might be challenged with remarked or cloned devices that are claimed were fabricated by this company, and is interested in a methodology to prove these devices were not fabricated in its facilities. Foundries also might be asked by customers to verify the facility that a device was produced. Such requests can be initiated

from a customer of a specific product from this manufacturer, or a design owner whose design is fabricated therein.

It is evident that a reliable methodology for attesting the fab-of-origin of a device without relying on the markings on it (as such markings can be easily forged) is greatly helpful in solving new challenges facing the IC industry such as the ones mentioned above, and can increase the security of the IC supply chain. As the proposed methodology uses the interaction between the idiosyncrasies of a fabrication facility and a design, as captured by the process variation, next section reviews some of the existing applications of process variation modeling in the literature, and related methodologies introduced for foundry identification.

## 4.3 Related Work

Post-silicon process variation modeling has been employed in various contexts, including: (i) decomposition for identifying prominent sources of variance [68], (ii) spatial or spatio-temporal correlation modeling for test cost reduction [9, 69], (iii) post-silicon diagnosis for identifying design sensitivity to process parameters, (iv) yield learning and forecasting [43], and (v) outlier detection [70]. The corresponding statistical methods leverage correlations to broadly separate chips to a few classes (good/bad, sensitive/robust, typical/outlier, etc.). Recently researchers have started to leverage process variations modeling for counterfeit IC detection and foundry identification. This section briefly review the state-of-the-art methods which leverage the process variations for foundry identification and similar problems along with their advantages and limitations.

### 4.3.1 Counterfeit detection using process variation modeling

Counterfeit ICs have become an issue for semiconductor manufacturing. One source of counterfeit devices is legitimate products that are extracted from electronic waste, i.e., reselling aged devices as brand new. Measuring performance parameters of a device and

approximating its age can be used to detect this class of counterfeit ICs. In [71], authors introduced a method to identify aged devices from brand new devices solely based on parametric measurements. To do so, they trained a one-class SVM classifier using distribution of brand new devices (affected by process variations), to build a boundary which encloses the population of fresh devices. The conjecture was that the aged devices exhibit different distribution in the space of parametric measurements. In [72] a similar methodology was proposed for counterfeit IC detection in which they used information of both fresh and aged devices to train a two-class classifier. They used simulation models to approximate the aging process and extract information of aged devices.

### 4.3.2 Foundry identification by reverse engineering

In [73], a methodology introduced which leverages intrinsic variation of the semiconductor manufacturing process for foundry identification purposes. They were the first to demonstrate the utility of process variations in this context. The base of their methodology was reverse engineering of process parameters such as threshold voltages and effective channel length of CMOS devices. To accomplish this, they used gate delay measurements which are obtained through an elegant path decomposition formulation to extract process parameters. Statistical tests such as Kolmogorov-Smirnov test is used to compare the distribution of these parameters to the profiles of known foundries in order to identify which foundry fabricated the IC in question. While this method is design-independent, it requires access to the gate level implementation of the fabricated IC in order to reverse engineer these process parameters, which may pose an obstacle due to IP protection issues. Moreover, as they explained in the paper, reverse engineering of these parameters can become quite complicated in practice.

### 4.3.3 Manufacturer attribution through electronic forensic

A methodology which uses embedded circuits to measure manufacturing characteristics of a device was introduced in [74]. They proposed to fabricate PCM-like structures (i.e., resistors,

capacitors, ring oscillators) along with the design to capture the manufacturing characteristics of an IC in a given fab. They fabricated 159 silicon ICs in two fabs and demonstrated separable distributions for two fabs for some measurements. By employing a threshold-based classifier, they were able to identify the manufacturer of ICs with 98% accuracy. Although, they showed the effectiveness of methodology using silicon measurements, their methodology has several limitations. First, using measurements of 80 chips from 2 wafers is not a representative sample for a foundry, and is statistically insufficient to draw a conclusion. As the number of wafers increase, wafer-to-wafer as well as lot-to-lot variations result in overlapped distributions which is difficult to distinguish where the device was fabricated. Second, this method requires additional structures to be embedded into the design which adds extra cost and effort and challenge the time to market of a product.

In contrast to the aforementioned design-independent approaches to fab-of-origin attestation, this work introduces a learning-based methodology which leverages the *interaction* between the idiosyncrasies of a fabrication facility and a particular design [67].

## 4.4 Machine-Learning Based Method for Fab-of-Origin Attestation

The methods proposed in this work seek to identify whether an IC was manufactured in a ratified fabrication facility based solely on the parametric measurements obtained during post-manufacturing testing. Note that these measurements have predefined acceptable ranges; any IC whose values fall outside these ranges is considered faulty and is discarded. Hence, the objective is to distinguish between the footprints of healthy chips from the ratified fab and the footprints of healthy chips from other fabs *within* the hyper-dimensional parametric space of acceptable performances. The conjecture here is that, for the same design and process, certain idiosyncrasies stemming from manufacturing tool installations, chemical sources, as well as altitude and geomagnetic location of the fabrication facility, lead to minor, yet systematic disparities in the resulting products of different fabs. These disparities may,

therefore, be leveraged through machine learning methods in order to attest the source of origin of a given IC [66].

Four variants of the fab-of-origin attestation problem are considered herein:

- **AttestMe-I:** In this variant of the fab-of-origin attestation problem the only available data is the parametric test data profile from a statistically significant number of chips manufactured in the ratified fab. Given this profile and the parametric tests of a single IC, the goal is to decide whether it was manufactured in the ratified fab or not.

- **AttestUs-I:** This variant assumes availability of the same information as above; instead of making a decision for a single IC, however, it considers the parametric tests of an entire batch of ICs and seeks to make a collective decision for the batch, assuming that they were all manufactured in the same fabrication facility.

- **AttestMe-II:** The assumption in this variant, is the availability of the parametric test data profile from a statistically significant number of chips manufactured in each of the fabs wherein a given design could have been produced. Given these profiles and the parametric tests of a single IC, the objective is to decide whether it was produced by the ratified fab or any other fab.

- **AttestUs-II:** Using the same information as above, this variant seeks to decide whether an entire batch of ICs, originating from the same facility, was manufactured in the ratified fab or any other fab.

Note that the *Attest(Me/Us)-I* variants require less training data, since they only rely on the profile of the ratified fab, as opposed to all fabs, yet are more difficult than their *Attest(Me/Us)-II* counterparts. Similarly, the *AttestMe-(I/II)* variants require less test data, since they make decisions for individual ICs, as opposed to batches of ICs, yet are more difficult than their *AttestUs-(I/II)* counterparts. Figure 4.1 summarizes these four attestation scenarios.

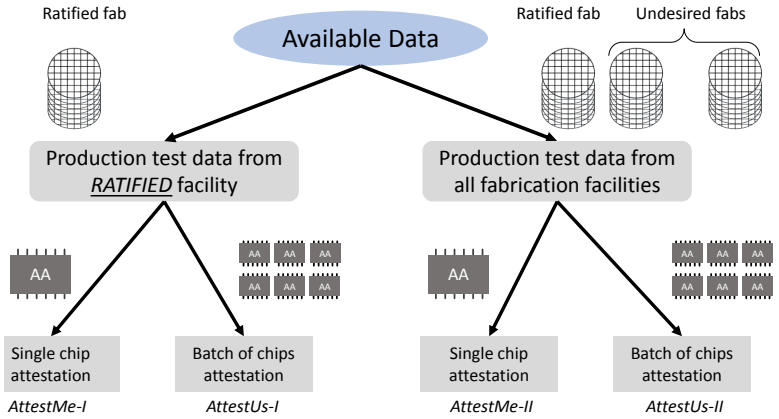



Figure 4.1. Fab-of-Origin attestation scenarios.

### 4.4.1 Proposed Solutions

This Section presents the proposed solutions for the four variants of the fab-of-origin attestation problem, which were introduced in previous Section.

#### AttestMe-I

The *AttestMe-I* variant is, essentially, a one-class classification problem, for which numerous solutions exist in the literature [75]. Specifically, given a statistically significant set of parametric test data from the ratified fab, the objective is to learn a boundary that encloses this population in the multidimensional space of these measurements. The trained one-class classifier then compares the footprint of a new IC to this boundary, in order to decide whether it came from the ratified fabrication facility or not.

The key challenge in the context of *AttestMe-I*, however, is the high dimensionality of the data, which is typically in the few hundreds (i.e., number of probe-tests). Indeed, due to the curse of dimensionality, it is practically impossible to capture the underlying interaction between the design and the idiosyncrasies of a specific fab and to establish any meaningful boundary in the raw data space. Instead, the proposed method employs the following steps:

**Dimensionality Reduction:** In order to reduce the dimensionality of the test data, the t-Distributed Stochastic Neighbor Embedding (t-SNE) [22] technique is employed. t-SNE is a non-linear transformation of the parametric test data into a lower-dimensional feature space, wherein enough discriminative power exists for learning the boundary that encloses the population.

**Clustering:** Once the data is projected in the transformed space, the GAP statistic method [76] is used to estimate the number of clusters that the data consists of, followed by k-means clustering to separate the data into the corresponding number of clusters.

**Boundary Identification:** A simple one-class classifier (i.e., a convex hull) is then trained to enclose the data of each cluster. Collectively, the acceptance region of the trained one-class classifiers for all the clusters, define the space where ICs from the ratified fab are expected to reside.

**Decision Making:** Given the test data of a new IC, its footprint in the transformed space is computed and compared to the acceptance region. The IC is considered as originating from the ratified fab *if and only if* this footprint falls within any of the learned clusters.

For sake of comparison, the simpler and very popular Principal Component Analysis (PCA) [77] method for dimensionality reduction is also considered. However, as will be demonstrated later, the variance of the data appears to be highly non-linear; therefore, PCA, which linearly transforms the original data to a lower dimensional subspace, while retaining most of its variance, performs poorly. An advanced one-class classifier (i.e., SVM) is also trained to directly learn a single boundary in the reduced feature space. However, the data in this space is highly discontinuous, with the vast majority of the points congregating in small clusters. Therefore, as will be shown in Section 4.5.1, learning a single boundary to successfully include all these discontinuous regions while excluding the rest of the space is of limited effectiveness.

**AttestUs-I**

The solution to the *AttestUs-I* variant seeks to take advantage of the fact that process variations are expected to affect ICs produced within the same fab in a correlated way. Accordingly, this correlation can be leveraged to improve fab-of-origin attestation effectiveness for a batch of ICs, all of which originate from the same fab. To achieve this, the underlying distribution of performance parameters for this batch is assessed against the profile of the ratified fab using non-parametric statistical tests. In particular, this solution employs the Anderson-Darling (AD) test [78], and Kolmogorov-Smirnov test [79] which are the well-known procedures for determining whether a sample of $k$ observations comes from a given distribution or not. In order to utilize these two tests in the fab-of-origin attestation context, the following procedure is applied:

**Density Estimation:** For every performance parameter $t$ of the device under attestation, the parametric measurements in the statistically significant training set from the ratified fab are used, to estimate the underlying distribution of that parameter. To do so, Kernel Density Estimation (KDE) [59] is employed which has been successfully used in the past for density estimation and synthetic population generation [80]. This method relies on the estimation of the densities $f(\vec{t})$, using the available observations $\vec{t_i}$, $i = 1, \cdots, M$, where $M$ is the number of available samples used to build the density. There is no assumption regarding its parametric form (e.g., normal). Instead, the non-parametric KDE is used, which allows the observations to speak for themselves. The kernel density estimate is defined as [59]

$$\hat{f}(\vec{t}) = \frac{1}{M \times h^d} \sum_{i=1}^{M} K_e(\frac{1}{h}(\vec{t} - \vec{t_i})) \tag{4.1}$$

where $h$ is a parameter called bandwidth, $d = 1$ is the dimension of $\vec{t}$, and $K_e(m)$ is the Epanechnikov kernel.

**Membership Test:** Consider $m_t$ as the measurement vector of performance parameter $t$ from all ICs in the batch under attestation. The objective of this test is to compare the

parameter distributions of this unknown batch of chips with the parameters of ratified fab. Kolmogorov-Smirnov (KS) test was first introduced in [81, 82] as a non-parametric test which can be used to decide if a sample comes from a population with a specific distribution. The KS statistic is based on the largest distance between empirical cumulative distribution value of the sample (i.e., $m_t$) and cumulative distribution of the hypothesized distribution. The null hypothesis is that *the sample follow the specified distribution*, and *data do not follow the specified distribution* is alternate hypothesis. Output of the KS test is an asymptotic $p$-value in the range 0 to 1. For a $p$-value less than a chosen threshold (usually 0.05), the null hypothesis is rejected and I deduce that the distribution of the measured data, $m_t$, is dissimilar to the estimated density (i.e., this batch of chips does not originate from the ratified fab). Anderson-Darling (AD) test is another statistical test that is employed in this work in order to accomplish the membership test. AD test is very well-known procedure for determining whether a sample of $k$ observations come from a given distribution or not. It was developed in 1952 by Anderson and Darling [78] and has several advantages such as its sensitivity to the shape of a distribution and applicability for small sample sizes. The $p$-value output of AD test is evaluated similar to that of KS test to reject or accept the null hypothesis. Various statistical packages in R and MATLAB support the AD and KS tests.

**Decision Making:** This procedure is repeated individually for each performance parameter. A majority vote is, then, employed to provide the final decision for the batch.

### AttestMe-II

The *AttestMe-II* variant of attesting an individual chip, when parametric measurements from a statistically significant number of chips from both the ratified and all other (i.e., undesired) fabs are available, boils down to a two-class classification problem. Availability of populations from both classes simplifies the problem drastically and eliminates the need for clustering. Instead, the solution to this variant involves the following steps:

**Classifier Training:** The available training data is used to train a classifier which will be used to determine whether an unknown device originating from the ratified fab or an untrusted fab. This work explores a set of well-known classifiers and compare their classification performances.

- *Naive Bayes:* Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with independence assumptions [83]. These independence assumptions of features make the dimensionality of features irrelevant, thereby the presence of one feature does not affect other features. Thus, it is particularly suited when the dimensionality of the input is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods [84]. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. Bayesian classification approach arrives at the correct classification as long as the correct category is more probable than the others.

- *K-Nearest Neighbors (KNN)*: KNN is an instance-based learning algorithm, and is used to test the degree of similarity between an unseen instance and $K$ instances of training data, in order to determine the label of the unseen instance [85]. It is based on the principal that instances within a dataset will generally reside in close proximity to other instances that have similar properties/label. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular instance. Euclidean Distance is a widely used metric to measure the distance between the vectors. In the training phase of KNN, training vectors along with their class labels are stored. Then, in the classification phase, distances from a new vector, representing an unseen instance, to all training vectors are computed, and $K$ closest samples are selected. Finally, the class label of the new instance will be the most frequent class label of these $K$ instances. In this work, a hold-out set of data is used to assign $K = 9$ and Euclidean distance is used as a distance metric to find nearest instances.

- *Support Vector Machine (SVM)*: SVMs are one of the discriminative classification methods that are well-recognized for their accurate performances in real-world applications [86]. The principal idea is the separation of two classes through a hyperplane that is specified by a vector and a bias term. The optimal separating hyperplane is the one that maximizes the distance between the hyperplane and the nearest points of both classes (known as the margin). SVM tries to find out the linear separating hyperplane which maximize the margin, i.e., the optimal separating hyperplane and maximizes the margin between the two data sets. Kernel functions can be used in conjunction with the SVM formulation to allow non-linear decision boundaries. In this sense, the nonlinearity of the classification solution is included via a kernel function. SVM has several advantages such as good generalization properties, and insensitivity to overtraining and the curse-of-dimensionality [87, 88].

- *Linear Discriminant Analysis (LDA)*: LDA is a statistical, multivariate method used in statistics and machine learning to find a linear combination of features that separates two or more classes of objects [89]. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby, guaranteeing maximal separability. This technique has a very low computational requirement which makes it suitable for real-time systems. Moreover, this classifier is simple to use and generally provides good results.

- *Deep Neural Networks (DNNs)*: DNNs have recently achieved state-of-the-art performance in a wide range of classification tasks of high dimensionality in speech recognition, computer vision and text processing [90]. A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs. They became more successful in recent years due to the availability of inexpensive, parallel hardware (GPUs, computer clusters) and massive amounts of data. Deep

learning discovers intricate structures in large data sets by using the back-propagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer, from the representation in the previous layer. For this two-class classification problem, a five-layer DNN with two output neurons is trained using measurement data from both the ratified and the undesired fabs. To train the entire network, a generative pre-training step is applied to train one layer at a time. Then, the whole network is fine-tunned using the back-propagation learning algorithm.

**Decision Making:** Given a new IC whose source of origin needs to be attested, its performance parameters are measured and provided to the trained classifier, that determines which of the two classes the IC belongs to, i.e., whether it was produced in the ratified fab or in an undesired fab.

#### AttestUs-II

The solution to the *AttestUs-II* variant follows the general principles of what was described in Section 4.4.1 and consists of the following steps:

**Density Estimation:** For every performance parameter of the IC batch under attestation, e its probability density function (PDF) in both the ratified fab and the undesired fab(s) is computed by applying KDE on the corresponding training sets.

**Membership Test:** For every performance parameter, the AD and KS tests are applied using the measurement vector from all ICs in the batch under attestation and the estimated densities of the ratified fab and the undesired fab(s). For each test, the combination of the two $p$-values determines whether, with respect to this performance parameter, the ICs in the batch were manufactured in the ratified fab or in an undesired fab.

**Decision Making:** This procedure is repeated individually for each performance parameter. A majority vote is, then, employed to provide the final decision for the batch.

Table 4.1. Attestation scenarios.

| Attestation scenario | Attestation approach | Training set | Validation set | Attestation granularity |
|---|---|---|---|---|
| AttestMe-I | **Dimensionality reduction + clustering + boundary identification** | Data from RATIFIED fab | **Devices manufactured in ratified and undesired fabs** | Single IC |
| AttestUs-I | **Distribution test (AD/KS test)** | Data from RATIFIED fab | | Batch of ICs |
| AttestMe-II | **Two-class classifier (LDA, KNN, SVM, DNN, NB)** | Data from RATIFIED and undesired fabs | | Single IC |
| AttestUs-II | **Distribution test (AD/KS test)** | Data from RATIFIED and undesired fabs | | Batch of ICs |

Table 4.1 summarizes the learning approach, training data, validation data and the granularity of attestation for all the above-mentioned attestation scenarios.

## 4.5 Experimental Results

This section presents the evaluation of the effectiveness of the proposed solutions using actual production test data from a 65nm RF transceiver currently in high volume manufacturing (HVM) by Texas Instruments.

This dataset comprises devices from two geographically dispersed fabs wherein this RF transceiver is fabricated. For the purpose of this study, one of these facilities is considered as the ratified fab and the other one as the unknown or undesired fab. The dataset for the ratified fab includes 600 wafers from 20 lots, with approximately 1500 die per wafer. For each die, 276 probe-test measurements are provided.

These tests are the typical measurements performed at wafer probe to ensure compliance of the performances of an RF transceiver design to its specifications (i.e., production tests). They include both structural tests (open/short circuit, power consumption, $I_{DDQ}$, input voltage threshold, output voltage level, etc.) and functional tests (BER, EVM, CMMR,

receiver sensitivity, output power, phase noise, etc.) and indirectly cover a broad range of process parameters.

The dataset of the undesired fab includes the same 276 probe-test measurements from 500 wafers in 20 lots. These two datasets were obtained from the two fabs at approximately the same period. Using this dataset, this work seeks to:

- Visualize the overlap of the two populations in the raw data space and in the linearly transformed PCA space, as well as the effectiveness of the non-linear t-SNE transformation in increasing discrimination, and demonstrate the limited effectiveness of training a one-class classifier (i.e., SVM) to separate the populations through a single boundary, due to data discontinuity.

- Quantify the effectiveness of *AttestMe-I* and *AttestUs-I*, which use data solely from the ratified fab for learning the underlying model, in distinguishing between ICs produced in the ratified and in an unknown fab.

- Assess the attestation accuracy improvement achieved by *AttestMe-II* and *AttestUs-II*, which are trained with datasets from both the ratified and the undesired fabs.

- Demonstrate the effectiveness of the proposed solutions in handling process variations by assessing attestation accuracy on ICs from future production.

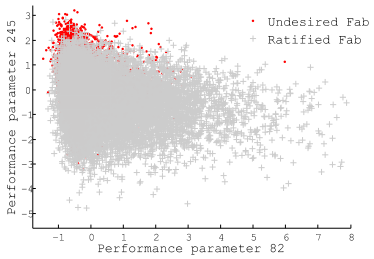
### 4.5.1 Population overlap

To demonstrate population overlap, 5 wafers are randomly selected from each of the 20 lots in the ratified fab and all probe-test data of all die on these 100 wafers is used as the training set. Then, a one-class SVM is trained to learn the boundary that encloses the population originating from the ratified fab in three different spaces: (i) in the raw data space which includes all 276 dimensions, (ii) in a PCA transformed space where the data is linearly

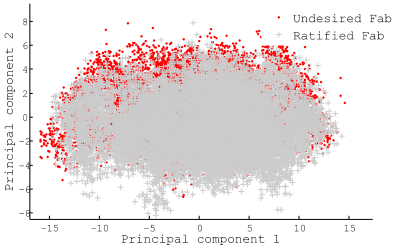
projected on the first 30 principal components, and (iii) in the t-SNE transformed space where the retained data is non-linearly projected on 3 dimensions. The validation set includes all die from a randomly selected wafer from each of the 20 lots of the ratified fab (excluding the wafers used for training) and from each of the 20 lots of the undesired fab. The trained SVMs are, then, used to individually decide whether each die in the validation set originated from the ratified fab or not.

Figures 4.2 (a)-(c) visualize the training and validation data on the space of the two most discriminative raw measurements, on the two main components of the linearly transformed PCA space, and on the two components of the non-linearly transformed t-SNE space, respectively. As may be observed, there is an almost complete population overlap in the first case, which is only slightly reduced after linear transformation in the second case, because the variability of the data is non-linear. The non-linear transformation of the third case, however, performs significantly better in separating the two populations. While this is visualized only in a two-dimensional space, an extensive experimentation with multiple dimensions has confirmed this observation, justifying the use of t-SNE as the method of choice for enhancing discrimination via dimensionality reduction in this context.

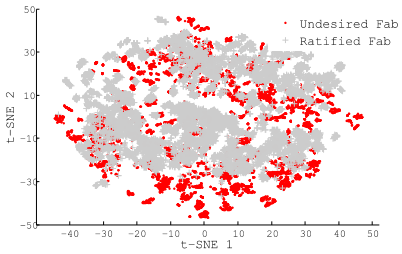
The results reported in the table of Figure 4.2 (d), which quantify the effectiveness of a single boundary established by training a one-class SVM in each of the three spaces mentioned earlier, are also consistent with this observation. Indeed, attestation accuracy of a single IC in the raw data space is only 57.3%, barely higher than a coin-toss. Learning the boundary in the 30-dimensional PCA space only slightly improves accuracy to 61%, while doing so in the 3-dimensional t-SNE space boosts accuracy to 71.4%. This rather low accuracy is attributed to the highly discontinuous nature of the data in the projected space, which calls for a clustering-based classification approach, as shows next.




(a) Raw data




(b) PCA




(c) t-SNE

| Method | Attestation Accuracy |
|---|---|
| Raw data + One-class SVM | 57.3% |
| PCA + One-class SVM | 61.0% |
| t-SNE + One-class SVM | 71.4% |

(d) One-class SVM results

Figure 4.2. Population overlap and single boundary classification accuracy in raw and transformed measurement spaces.

### 4.5.2 Learning only from ratified fab

In order to assess the effectiveness of *AttestMe-I*, clustering and boundary identification is applied on the t-SNE transformed space of the training data, as detailed in Section 4.4.1. Then, for each IC in the validation set, the decision making step is performed, which examines whether its footprint in this space lies within the boundary of any of the clusters assigned to the ratified fab. Table 4.2 reports the attestation accuracy for *AttestMe-I*, noting that *positive*, $(P)$, is considered as a chip originating from the ratified fab and as *negative*, $(N)$, a chip originating from an undesired source. In this confusion matrix, True Positive Rate $(TPR)$ denotes the percentage of ICs that are correctly identified as originating from the ratified fab, while True Negative Rate $(TNR)$ refers to the percentage of ICs that are correctly labeled as originating from an undesired fab. False Positive Rate $(FPR)$ and False Negative Rate $(FNR)$ are defined similarly. As may be observed, the overall attestation accuracy is 85%,

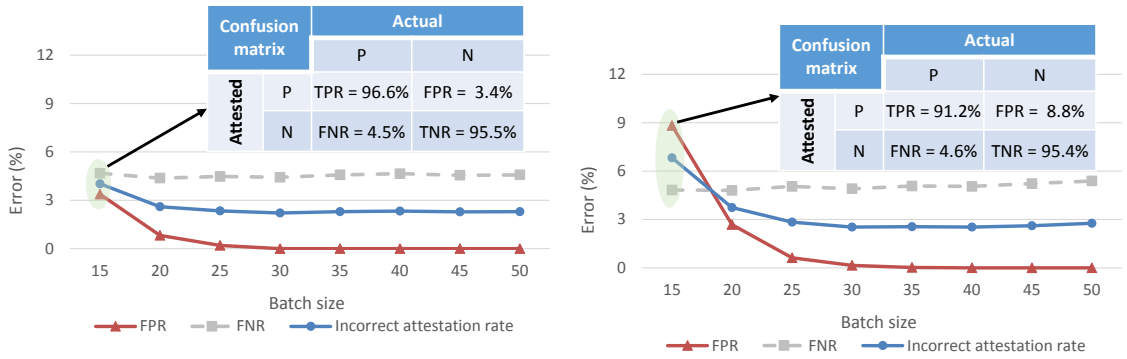
Table 4.2. *AttestMe-I* results.

| Confusion matrix | | Actual | |
|---|---|---|---|
| | | P | N |
| Attested | P | TPR = 85.5% | FPR = 14.5% |
| | N | FNR = 15.5% | TNR = 84.5% |

clearly outperforming the one-class SVM reported in Figure 4.2 (d) . This is expected due to the manifold nature of the t-SNE transformed data, which makes it difficult to separate via a single boundary, as the SVM tries to do.

Effectiveness of *AttestUs-I* is assessed by first estimating the performance parameter densities of the ratified fab through the training set. Then, for a batch of ICs originating from the same fab, the performance parameters from all ICs in the batch are measured and the AD and KS membership tests are performed for each of the parameters, as detailed in Section 4.4.1. This experiment, randomly draws batches of sizes in the range [15, 50] from the validation sets of the ratified and the undesired fab; this procedure is repeated 4000 times for each batch size (2000 batches from ratified fab and 2000 batches from undesired fab).

Figure 4.3 (a) shows the *AttestUs-I* results when the Anderson-Darling test is used for distribution test. The horizontal axis denotes the batch size, while the vertical axis is the attestation error rate. As may be observed, this method is very successful in attesting the fab-of-origin of a batch, with accuracy exceeding 96% for batch sizes of as small as 15 ICs. The confusion matrix for this batch size is also provided in the figure. For batches greater than 30 ICs attestation error is quite stable and below 2.5%. Figure 4.3 (b) demonstrates same results for Kolmogorov-Smirnov statistical test. Attestation accuracy for a batch of 15 chips is 93.3% which is lower than that of AD test. This is anticipated, as one of the major




(a) *AttestUs-I: AD statistical test*

(b) *AttestUs-I: KS statistical test*

Figure 4.3. Attestation results for various batch sizes.

advantageous of AD test is dealing with small sample sizes. As can be seen, for a batch size larger than 30 ICs, performance of both tests are very similar and consistent.

To gain further insight, Figure 4.4 shows the distribution of $p$-values for a batch size of 15 ICs for the AD test, where the horizontal axis represents the range of $p$-value and the vertical axis shows the percentage of 2000 randomly selected batches which have a $p$-value within a given range. In the Anderson-Darling distribution test, the null hypothesis is that the 15-dimensional measurement vector of the 15 ICs in the batch comes from a specific population, which is the distribution of the ratified fab. As shown in the top histogram, for the vast majority of the 2000 samples from the ratified fab, the $p$-value is larger than 0.05, hence the null hypothesis is not rejected, i.e., these batches are correctly assumed to have originated from the ratified fab. Conversely, as shown in the bottom histogram, for the vast majority of the 2000 samples from the undesired fab, the $p$-value is smaller than 0.05 and the null hypothesis is rejected, i.e., these batches are correctly assumed to have originated from the undesired fab.

### 4.5.3 Learning from all fabs

In order to quantify the accuracy of the proposed fab-of-origin attestation solutions when test data from both the ratified and the undesired fab is available, the training set is enhanced

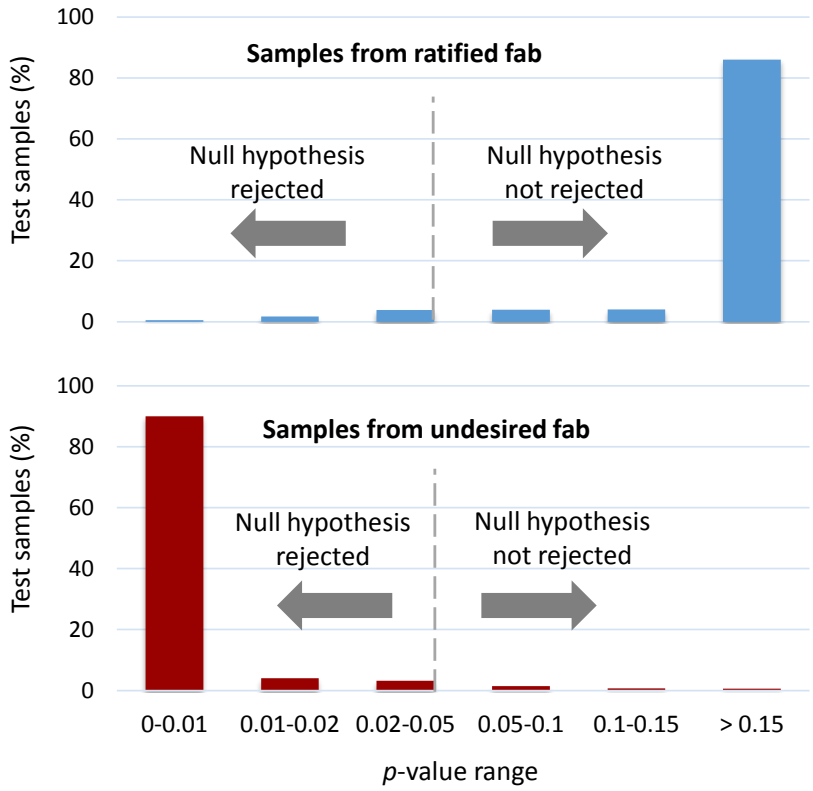



Figure 4.4. Histogram of $p$-values for AD test against the ratified fab distribution for batches of 15 chips (*AttestUs-I*).

so that it contains data from both fabs. Specifically, in addition to all die from 5 randomly selected wafers in each of the 20 lots from the ratified fab, the new training set also includes all die from 5 randomly selected wafers in each of the 20 lots from the undesired fab. The validation set remains unchanged, i.e., it contains all die from a randomly selected wafer from each of the 20 lots of the ratified fab and from each of the 20 lots of the undesired fab (excluding the wafers used for training).

Evaluation of the *AttestMe-II* solution starts with training two-class classifiers which are described earlier in Section 4.4.1, using the training set. The trained classifier is then applied to individually classify each IC in the validation set as originating from the ratified or the undesired fab. *AttestMe-II* results for all five classifiers are summarized in Table

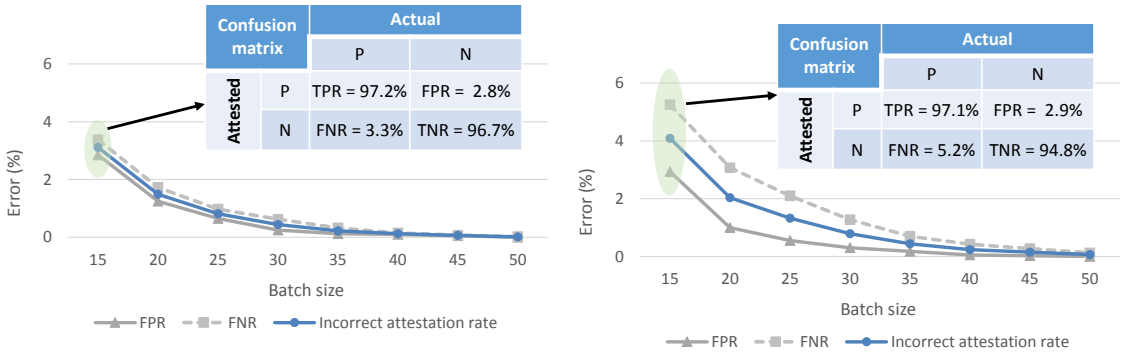
4.3. As may be observed, the Naive Bayes method with 85% attestation accuracy has the lowest performance. The attestation accuracy of other four classifiers is significantly high and superior to the *AttestMe-I* approach. This is expected, because of access to data from both fabs, which simplifies the process of learning the boundary that separates them, as compared to the case where training data is available only from the ratified fab. KNN with 0.9% attestation error outperforms other classification techniques. The conjecture for such low error for KNN classifier is the consistency between training data and validation sets, i.e., having insignificant amount of noise in the validation data. It should be noted that, training and validation wafers are originated from same lots. Therefore, attestation accuracy of KNN for ICs from future lots which will experience performance shift due to process variations, may drop.

Table 4.3. *AttestMe-II* results.

| Classifier | Accuracy | TPR | FPR | TNR | FNR |
|---|---|---|---|---|---|
| Naive Bayes | **85.0%** | 88.0% | 12.0% | 82.0% | 18.0% |
| DNN | **96.5%** | 97.0% | 3.0% | 96.0% | 4.0% |
| SVM | **97.8%** | 97.6% | 2.4% | 98.1% | 1.9% |
| LDA | **98.5%** | 98.5% | 1.5% | 98.4% | 1.6% |
| KNN | **99.1%** | 98.8% | 1.2% | 99.3% | 0.7% |

Effectiveness of *AttestUs-II* requires estimation of the performance parameter densities for both the ratified and the undesired fab using the enhanced training set. Then, for a batch of devices from the same fab, the performance parameters are measured from all ICs in the batch. For each performance parameter, it performs AD and KS membership tests against the densities of both fabs to compute the corresponding $p$-values, and finally decides which

fab the batch originated from, as explained in Section 4.4.1. Once again, this experiment randomly draws batches of sizes in the range [15, 50] from the validation sets of the ratified and the undesired fab, and repeat this procedure 4000 times for every batch size. Figure 4.5 (a) reports the *AttestUs-II* results for AD test, with the horizontal axis denoting the batch size and the vertical axis showing the attestation error. As may be observed, for a batch size of as few as 25 ICs, the accuracy of this solution exceeds 99%, while for a batch size of 40 ICs, it achieves error-free attestation. A comparison to the curves in Figure 4.3 reveals that availability of the additional training information from the undesired fab enhances the accuracy of the membership test and reduces the error. As a point of reference, the confusion matrix for the batch of size 15 is also provided. Figure 4.5 (b) demonstrates the attestation error vs. batch size when KS test is employed for distribution test. As can be observed, the attestation accuracy significantly improved by accessing to the data of undesired fab. As was expected, for small sample sizes (batch size less than 30) its performance is worse than that of AD test.

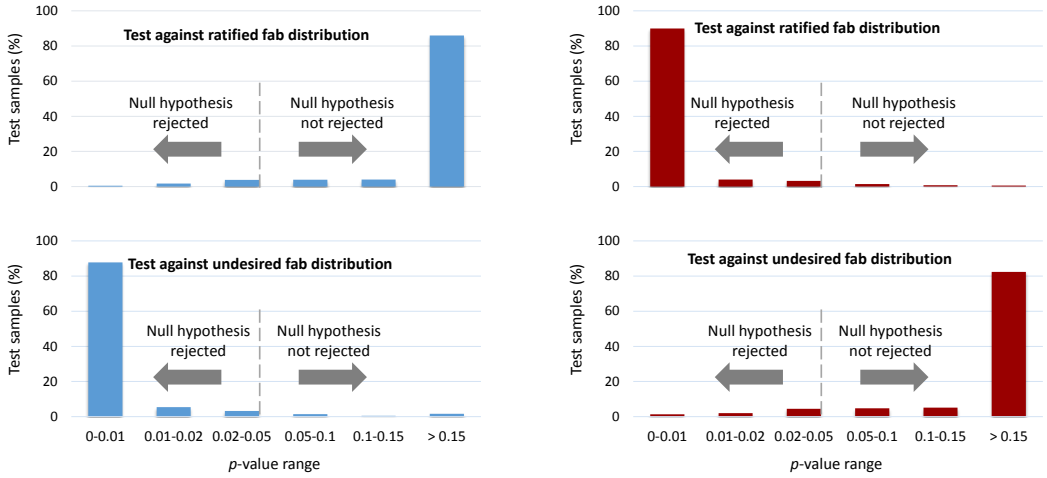



(a) *AttestUs-II: AD statistical test*

(b) *AttestUs-II: KS statistical test*

Figure 4.5. Attestation results for various batch sizes.

Lastly, Figure 4.6 presents the histogram of $p$-values when the AD test is used for batches of 15 ICs. Figure 4.6 (a) shows the $p$-values for 2000 batches originating from the ratified fab, wherein the top and bottom graphs compare these samples against the ratified and the




(a) Samples from ratified fab

(b) Samples from undesired fab

Figure 4.6. Histogram of $p$-values for AD test against the ratified and the undesired fab distributions for batches of 15 chips (*AttestUs-II*).

undesired fab distributions, respectively. Evidently, for the vast majority of samples the null hypothesis is not rejected for the ratified fab but is rejected for the undesired fab, hence these batches are correctly attested as originating from the ratified fab. Conversely, Figure 4.6 (b) demonstrates the same results for 2000 batches originating from the undesired fab, in which case the results are reversed.

### 4.5.4 Future production attestation accuracy

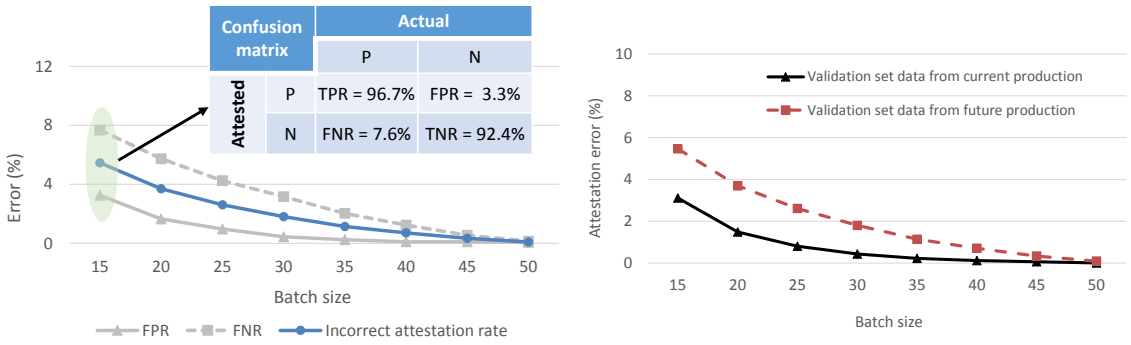
As a final experiment, the goal is to evaluate the robustness of the proposed solutions against fabrication process shifts. To do so, probe-test data from a new set of wafers from 10 lots is used, which were fabricated in each of the two fabs a few months after the wafers of the original dataset. "Future wafers" refers to these new wafers. The training set remains the same, but the new validation set now comprises all die from 20 randomly selected future wafers, equally distributed across the 10 new lots from each of the two fabs. Table 4.4 includes the effectiveness of the five classifiers as explained in Section 4.4.1 for *AttestMe-II* solution on the new validation set, which comprises future wafers. As may be observed, SVM, DNN

and LDA offer robust and accurate attestation in comparison with Table 4.3. Among them, LDA has a slightly better attestation accuracy, and outperforms other classification methods. KNN with 81.3% attestation accuracy shows a significant performance reduction, compared to 99.1% accuracy when the validation set was from training lots. As mentioned earlier, due to process variations, the performance parameters of devices from new lots shifts slightly which translates to noise in the KNN method.

Table 4.4. *AttestMe-II* results for chips from future production.

| Classifier | Accuracy | TPR | FPR | TNR | FNR |
|---|---|---|---|---|---|
| Naive Bayes | **80.8%** | 87.7% | 12.3% | 73.9% | 26.1% |
| KNN | **81.3%** | 77.9% | 22.1% | 84.7% | 15.3% |
| SVM | **93.6%** | 90.3% | 9.7% | 96.9% | 3.1% |
| DNN | **94.0%** | 96.0% | 4.0% | 92.0% | 8.0% |
| LDA | **94.6%** | 91.9% | 8.1% | 97.2% | 2.8% |

Similarly, Figure 4.7 (a) demonstrates the effectiveness of the *AttestUs-II* solution (AD test) for batch of ICs from future wafers, where the horizontal axis is the batch size and the vertical axis denotes the attestation error. As it can be seen, AD distribution test successfully attests batch of chips from future production even for small sample size. Figure 4.7 (b) shows the attestation error of AD test method for chips from current and future production. For small sample sizes the attestation accuracy drops slightly, however for batch sizes larger than 35 ICs, the difference in the two scenarios is negligible, confirming the fact that the *AttestUs-II* solution is robust to process variations. Figure 4.8 presents similar results for KS test method. As may be observed, the sensitivity of this test to the process variations is higher

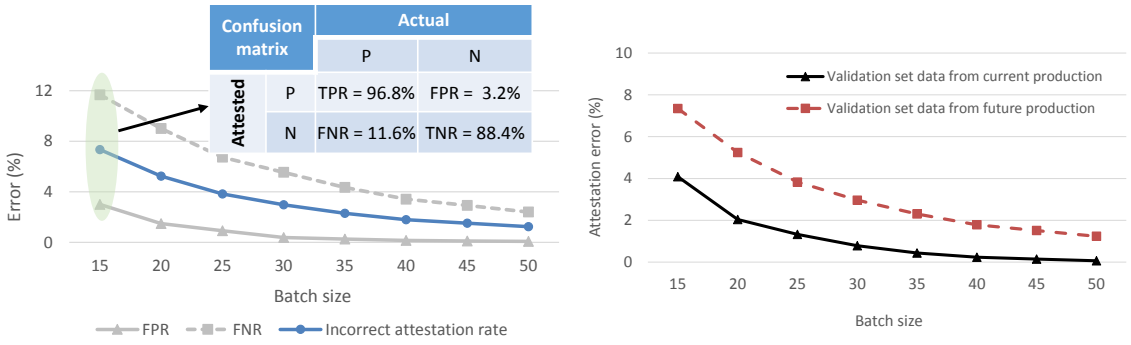



(a) Attestation results for chips from future production

(b) Attestation results of chips from current production vs. chips from future production

Figure 4.7. *AttestUs-II*: AD test, (a) results for chips from future production, (b) comparison of results for chips from current and future production.

than AD test, specifically for small batch sizes. As an ancillary measure for maintaining robustness, the underlying trained models can be periodically updated.




(a) Attestation results for chips from future production

(b) Attestation results of chips from current production vs. chips from future production

Figure 4.8. *AttestUs-II*: KS test, (a) results for chips from future production, (b) comparison of results for chips from current and future production

## 4.6 Conclusion

Parametric measurements, such as the ones taken during manufacturing testing, comprise valuable information which reflects the interaction between the design of an IC and the fabrication process through which it was produced. In conjunction with machine learning methods, this information may be harnessed to provide effective solutions to numerous variants of the fab-of-origin attestation problem, without requiring design modifications, custom processing steps, or specialized characterization equipment. Four such solutions were developed and evaluated using actual test data from a large number of ICs implementing an RF transceiver design, which were fabricated in two geographically dispersed foundries. Results indicate that the accuracy of these fab-of-origin attestation solutions reaches 99.1% when deciding whether a single IC originated from a ratified fab or an unknown/undesired facility and 100% when collectively making the same decision for a batch of as few as 40 ICs.

It is worth noting that while precise cloning of an IC could evade the proposed methods, this study was performed on two fabs of the same manufacturer so it resembles the best cloned devices one can build. Thus, it is expected even higher attestation accuracy when the fabrication facilities are independent. Also, it is possible that changes in fabrication process, such as machine part replacements, software updates or new material suppliers, may shift the process parameters and affect the accuracy of proposed models over time. Nevertheless, these methods were able to attest future productions with only minor accuracy reduction, demonstrating robustness of the models to such changes.

# CHAPTER 5

# CONCLUSION AND FUTURE DIRECTION

The semiconductor industry is rapidly growing and dynamically changing in order to meet the consumer market requirements. This has brought complex challenges into the manufacturing process of integrated circuits. Test cost need to be reduced without jeopardizing product quality. Fast and accurate prediction of high volume manufacturing yield is required to identify process problems and ramp-up production in a short time. Security-related challenges need to be properly addressed in order to guarantee the required performance and reliability for fab-less customers and end-user applications. This dissertation presented three machine learning based methodologies to address these challenges. To reduce the probe-test time, an adaptive method was proposed to optimize probe-test flow using process variations captured by e-test measurements. To accomplish this, process signature of each wafer was extracted at an early stage before the wafer reaches the probe station and this drives a selection engine to select the optimized test flow. The third Chapter of this work introduced a fast and accurate yield estimation methodology for fab-to-fab production migration and during transition to a new design generation. The proposed methodology is based on the correlation between e-test measurements and yield, and utilized silicon data from early engineering wafers. Finally, a machine learning approach was presented in Chapter four as an attestation tool to verify the fabrication facility that manufactured an integrated circuit. Experimental results using multiple large datasets of actual test measurements confirmed the aptitude of the proposed methods in effectively reducing test cost, efficiently estimate high volume manufacturing yield and precise attestation of fab-of-origin of integrated circuits.

This work introduces three machine learning based solutions to address challenges in the semiconductor manufacturing. It provides a platform, that can be further used to enhance the manufacturing process of ICs. The future directions for this work are as follows:

1. The core of test cost reduction method was to extract the process variation signature of a wafer from e-test measurements. By identifying the process signatures in which some test groups can be eliminated from the original test flow, one can determine the sweet spot in the process space. In other words, the desired process point for the operation can be identified in order to reduce the test cost significantly while having lower failure rate. Therefore, process engineer can tune the process table accordingly such that manufactured wafers stay in the process region of interest.

2. The proposed yield estimation methodology models the yield as a function of e-test measurements. The regression models can be used to determine the corresponding values for e-test parameters in order to meet a target yield. Similarly, process engineer tunes the process table such that those process parameters stay in the target range.

## REFERENCES


[1] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Design Automation Conference*, 2007, pp. 817–822.

[2] S. Reda and S. R. Nassif, "Accurate spatial estimation and decomposition techniques for variability characterization," *IEEE Transactions on Semiconductor Manufacturing*, vol. 23, no. 3, pp. 345–357, 2010.

[3] W. Zhang, X. Li, F. Liu, E. Acar, R.A. Rutenbar, and R.D. Blanton, "Virtual probe: a statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 12, pp. 1814–1827, 2011.

[4] H.-M. Chang, K.-T. Cheng, W. Zhang, X. Li, and K.M. Butler, "Test cost reduction through performance prediction using virtual probe," in *IEEE International Test Conference*, 2011, pp. 1–9.

[5] N. Kupp, K. Huang, J.M. Carulli, and Y. Makris, "Spatial estimation of wafer measurement parameters using Gaussian process models," in *IEEE International Test Conference*, 2012, pp. 1 – 8.

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

[7] K. Huang, N. Kupp, J.M. Carulli, and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," in *Design, Automation & Test in Europe Conference*, 2013, pp. 553 – 558.

[8] M. Heydarzadeh and M. Nourani, "A two-stage fault detection and isolation platform for industrial systems using residual evaluation," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 10, pp. 2424–2432, 2016.

[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] M. Heydarzadeh, H. Luo, and M. Nourani, "Model-free testing of analog circuits," in *IEEE Asian Test Symposium*, 2016, pp. 102–106.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.

[23] 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.

[24] A. Ahmadi, C Xanthopoulos, A. Nahar, B. Orr, M. Pas, and Y. Makris, "Harnessing process variations for optimizing wafer-level probe-test flow," in *IEEE International Test Conference*, 2016, pp. 1–8.

[25] B. Liu, F. V. Fernández, and G. G. E. Gielen, "Efficient and accurate statistical Analog yield optimization and variation-aware circuit sizing based on computational intelligence techniques," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 6, pp. 793–805, 2011.

[26] F. Gong, H. Yu, Y. Shi, and L. He, "Variability-aware parametric yield estimation for Analog/Mixed-signal circuits: Concepts, algorithms, and challenges," *IEEE Design & Test*, vol. 31, no. 4, pp. 6–15, 2014.

[27] J. Swidzinski and K. Chang, "Nonlinear statistical modeling and yield estimation technique for use in Monte Carlo simulations," *IEEE Transactions on Microwave Theory and Techniques*, vol. 48, no. 12, pp. 2316–2324, 2000.

[28] A. Dharchoudhury and S.-M. Kang, "Worst-case analysis and optimization of VLSI circuit performances," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 4, pp. 481–492, 1995.

[29] M. Stein, "Large sample properties of simulations using latin hypercube sampling," *Technometrics*, vol. 29, no. 2, pp. 143–151, 1987.

[30] A. Singhee and R. A. Rutenbar, "From finance to flip flops: A study of fast Quasi-Monte Carlo methods from computational finance applied to statistical circuit analysis," in *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 685–692.

[31] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," in *Proc. IEEE/ACM Design Automation Conference*, 2006, pp. 69–72.

[32] T. Doorn, E. Ter Maten, J. Croon, A. D. Bucchianico, and O. Wittich, "Importance sampling Monte Carlo simulations for accurate estimation of SRAM yield," in *Proc. IEEE Solid-State Circuits Conference*, 2008, pp. 230–233.

[33] A. Singhee and R. A. Rutenbar, "Statistical blockade: very fast statistical simulation and modeling of rare circuit events and its application to memory design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 8, pp. 1176–1189, 2009.

[34] N. E. Evmorfopoulos, G. I. Stamoulis, and J. N. Avaritsiotis, "A Monte Carlo approach for maximum power estimation based on extreme value theory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 4, pp. 415–432, 2002.

[35] H.-G. Stratigopoulos, P. Faubet, Y. Courant, and F. Mohamed, "Multidimensional Analog test metrics estimation using extreme value theory and statistical blockade," in *Proc. IEEE/ACM Design Automation Conference*, 2013, pp. 1–7.

[36] A. Singhee, J. Wang, B. H. Calhoun, and R. A. Rutenbar, "Recursive statistical blockade: An enhanced technique for rare event simulation with application to SRAM circuit design," in *Proc. IEEE International Conference on VLSI Design*, 2008, pp. 131–136.

[37] X. Li, J. Le, P. Gopalakrishnan, and L. T. Pileggi, "Asymptotic probability extraction for nonnormal performance distributions," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 1, pp. 16–37, 2007.

[38] H. Liu, A. Singhee, R. A. Rutenbar, and L. R. Carley, "Remembrance of circuits past: macromodeling by data mining in large Analog design spaces," in *Proc. IEEE/ACM Design Automation Conference*, 2002, pp. 437–442.

[39] X. Li, Y. Zhan, and L. T. Pileggi, "Quadratic statistical approximation for parametric yield estimation of Analog/RF integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 5, pp. 831–843, 2008.

[40] L. Milor and A. S. Vincentelli, "Computing parametric yield accurately and efficiently," in *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 1990, pp. 116–119.

[41] C. M. Kurker, J. J. Paulos, R. S. Gyurcsik, and J.-C. Lu, "Hierarchical yield estimation of large Analog integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 3, pp. 203–209, 1993.

[42] H.-G. Stratigopoulos, M. J. Barragan, S. Mir, H. L. Gall, N. Bhargava, and A. Bal, "Evaluation of low-cost mixed-signal test techniques for circuits with long simulation times," in *Proc. IEEE International Test Conference*, 2015, pp. 1–7.

[43] A. Ahmadi, K. Huang, A. Nahar, B. Orr, M. Pas, J. Carulli, and Y. Makris, "Yield prognosis for Fab-to-Fab product migration," in *Proc. IEEE VLSI Test Symposium*, 2015, pp. 1–6.

[44] 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 *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2015, pp. 9–14.

[45] A. Ahmadi, H.-G. Stratigopoulos, A. Nahar, B. Orr, M. Pas, and Y. Makris, "Harnessing fabrication process signature for predicting yield across designs," in *Proc. IEEE International Symposium on Circuits and Systems*, 2016, pp. 898–901.

[46] S. S. Sapatnekar, "Overcoming variations in nanometer-scale technologies," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 1, pp. 5–18, 2011.

[47] L. Abdallah, H.-G. Stratigopoulos, S. Mir, and J. Altet, "Defect-oriented non-intrusive RF test using on-chip temperature sensors," in *Proc. IEEE VLSI Test Symposium*, 2013, pp. 1–6.

[48] P. Ituero, M. López-Vallejo, and C. López-Barrio, "A 0.0016 $mm^2$ 0.64 nj leakage-based CMOS temperature sensor," *Sensors*, vol. 13, no. 9, pp. 12648–12662, 2013.

[49] B. Razavi, "CMOS technology characterization for Analog and RF design," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 3, pp. 268–276, 1999.

[50] M. Bhushan, A. Gattiker, M. B. Ketchen, and K. K. Das, "Ring oscillators for CMOS process tuning and variability control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10–18, 2006.

[51] L.-T. Pang and B. Nikolic, "Measurements and analysis of process variability in 90 nm CMOS," *IEEE Journal on Solid-State Circuits*, vol. 44, no. 5, pp. 1655–1663, 2009.

[52] V. Cherkassky and F. Mulier, *Learning from data: concepts, theory, and methods*, John Wiley & Sons, 2007.

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

[54] P. 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.

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

[56] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[57] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[58] A. Ahmadi, H.-G. Stratigopoulos, K. Huang, A. Nahar, B. Orr, M. Pas, J. M. Carulli, and Y. Makris, "Yield forecasting across semiconductor fabrication plants and design generations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017.

[59] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26, CRC press, 1986.

[60] X. Li, W. Zhang, F. Wang, S. Sun, and C. Gu, "Efficient parametric yield estimation of Analog/Mixed-signal circuits via Bayesian Model Fusion," in *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2012, pp. 627–634.

[61] F. Wang, W. Zhang, S. Sun, X. Li, and C. Gu, "Bayesian Model Fusion: large-scale performance modeling of Analog and mixed-signal circuits by reusing early-stage data," in *Proc. IEEE/ACM Design Automation Conference*, 2013, pp. 59–64.

[62] C. Gu, E. Chiprout, and X. Li, "Efficient moment estimation with extremely small sample size via Bayesian inference for Analog/Mixed-signal validation," in *Proc. IEEE/ACM Design Automation Conference*, 2013, pp. 1–7.

[63] S. Sun, F. Wang, S. Yaldiz, X. Li, L. Pileggi, A. Natarajan, M. Ferriss, J. Plouchart, B. Sadhu, B. Parker, et al., "Indirect performance sensing for on-chip Analog self-healing via Bayesian Model Fusion," in *Proc. IEEE Custom Integrated Circuits Conference*, 2013, pp. 1–4.

[64] J. Liaperdos, H.-G. Stratigopoulos, L. Abdallah, Y. Tsiatouhas, A. Arapoyanni, and X. Li, "Fast deployment of alternate Analog test using Bayesian Model Fusion," in *Proc. IEEE Design, Automation & Test in Europe Conference*, 2015, pp. 1030–1035.

[65] C. Fang, Q. Huang, F. Yang, X. Zeng, X. Li, and C. Gu, "Efficient bit error rate estimation for high-speed link by Bayesian Model Fusion," in *Proc. IEEE Design, Automation & Test in Europe Conference*, 2015, pp. 1024–1029.

[66] DARPA, *SB133-003: Electronic Component Fingerprinting to Determine Manufacturing Origin.*, Online. Available: http://www.acq.osd.mil/osbp/sbir/solicitations/sbir20133/darpa133.htm.

[67] A. Ahmadi, M.-M. Bidmeshki, A. Nahar, B. Orr, M. Pas, and Y. Makris, "A machine learning approach to fab-of-origin attestation," in *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2016, p. 92.

[68] K. Huang, N. Kupp, J. M. Carulli, and Y. Makris, "Process monitoring through wafer-level spatial variation decomposition," in *2013 IEEE International Test Conference*, 2013, pp. 1–10.

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

[70] W. R. Daasch, J. McNames, R. Madge, and K. Cota, "Neighborhood selection for IDDQ outlier screening at wafer sort," *IEEE Design & Test*, vol. 19, no. 5, pp. 74–81, 2002.

[71] K. Huang, J. M. Carulli, and Y. Makris, "Parametric counterfeit ic detection via support vector machines," in *Proc. IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*, 2012, pp. 7–12.

[72] D Chang, S Ozev, O Sinanoglu, and R Karri, "Approximating the age of rf/analog circuits through re-characterization and statistical estimation," in *Proc. IEEE Design, Automation & Test in Europe Conference & Exhibition*, 2014, pp. 1–4.

[73] J. B. Wendt, F. Koushanfar, and M. Potkonjak, "Techniques for foundry identification," in *Proc. ACM Design Automation Conferenc*, 2014, pp. 1–6.

[74] R. L. Helinski, E. I. Cole, G. Robertson, J. Woodbridge, and L. G. Pierson, "Electronic forensic techniques for manufacturer attribution," in *Proc. IEEE International Symposium on Hardware Oriented Security and Trust*, 2016, pp. 139–144.

[75] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *the Journal of machine Learning research*, vol. 2, pp. 139–154, 2002.

[76] 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.

[77] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[78] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *The annals of mathematical statistics*, pp. 193–212, 1952.

[79] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[80] H.-G. Stratigopoulos, S. Mir, and Y. Makris, "Enrichment of limited training sets in machine-learning-based analog/rf test," in *Proc. IEEE Design, Automation & Test in Europe Conference & Exhibition*, 2009, pp. 1668–1673.

[81] A. N. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*, na, 1933.

[82] N. V. Smirnov, "Estimate of deviation between empirical distribution functions in two independent samples," *Bulletin Moscow University*, vol. 2, no. 2, pp. 3–16, 1939.

[83] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*. IBM New York, 2001, vol. 3, pp. 41–46.

[84] I. Rish, J. Hellerstein, and J. Thathachar, "An analysis of data characteristics that affect naive bayes performance," *IBM TJ Watson Research Center*, vol. 30, 2001.

[85] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transaction on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[86] V. Vapnik, *The nature of statistical learning theory*, Springer Science & Business Media, 2013.

[87] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transaction on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[88] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.

[89] A. R. Webb, *Statistical pattern recognition*, John Wiley & Sons, 2003.

[90] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

## BIOGRAPHICAL SKETCH

Ali Ahmadi was born in Iran. Ali received his B.S. degree from the University of Isfahan, Iran in 2006, and his M.S. degree from University of Tehran, Iran, in 2009, both in computer engineering. He started his Ph.D degree in electrical engineering at The University of Texas at Dallas in August 2010. His research interests include applications of machine learning in semiconductor manufacturing for test cost reduction, yield estimation, foundry identification, and defect analysis.

**CURRICULUM VITAE**

# Ali Ahmadi

April 15, 2017

## Contact Information:

Department of Electrical Engineering
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.

Email: ali.ahmadi@utdallas.edu

## Educational History:

B.S., Computer Engineering, University of Isfahan, IRAN, 2006
M.S., Computer Engineering, University of Tehran, IRAN, 2009
Ph.D., Electrical Engineering, The University of Texas at Dallas, 2017

*Applications of Machine Learning in Test Cost Reduction, Yield Estimation and Fab-of-Origin Attestation of Integrated Circuits*
Ph.D. Dissertation
Electrical Engineering Department, The University of Texas at Dallas
Advisor: Dr. Yiorgos Makris

## Employment History:

Sr. Diagnostics Engineer, GLOBALFOUNDRIES, January 2017 – present
Research Assistant, The University of Texas at Dallas, August 2010 – December 2016

## Professional Recognitions and Honors:

Nominated for Best Paper award for VLSI Test Symposium, 2016
Received Best Paper award for VLSI Test Symposium, 2015
Received Ericsson Graduate Fellowship award at UT Dallas, 2015
Received Best Poster award from Semiconductor Research Corporation and Texas Analog Center of Excellence, 2015

## Professional Memberships:

Institute of Electrical and Electronics Engineers (IEEE), 2011–present
Association of Computing Machinery (ACM), 2015–present