Yield Forecasting Across Semiconductor Fabrication Plants and Design Generations

Ali Ahmadi, Haralampos-G. Stratigopoulos, Ke Huang, Amit Nahar, Bob Orr, Michael Pas, John M Carulli, Yiorgos Makris

Yield Forecasting Across Semiconductor Fabrication Plants and Design Generations

Ali Ahmadi, *Student Member, IEEE,* Haralampos-G. Stratigopoulos, *Member, IEEE,* Ke Huang, *Member, IEEE,* Amit Nahar, *Member, IEEE,* Bob Orr, *Member, IEEE,* Michael Pas, *Member, IEEE,* John M. Carulli Jr., *Senior Member, IEEE,* and Yiorgos Makris, *Senior Member, IEEE*

***Abstract*—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. In contrast, herein we introduce yield estimation solutions which rely exclusively on silicon measurements and we apply 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. Effectiveness of the proposed yield forecasting methods is evaluated using actual high-volume production data from two 65nm RF transceiver devices.**

I. Introduction

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 [1], [2]. 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 process design kit (PDK).

In contrast, in this work we focus on yield estimation in two specific scenarios wherein much more 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 [3], [4]. 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 [5]. 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, we consider

Manuscript received October 13, 2016; revised January 18, 2017; accepted January 30, 2017. Date of current version February 12, 2017. This research has been partially supported by the Semiconductor Research Corporation (SRC) Task 2709.001. The authors would also like to thank Texas Instruments Inc. for providing the data on which this study was performed. This paper was recommended by Associate Editor P. Girard.

A. Ahmadi and Y. Makris are with the Department of Electrical Engineering, University of Texas at Dallas, USA (e-mail: {ali.ahmadi, yiorgos.makris}@utdallas.edu).

H.-G. Stratigopoulos is with Sorbonne Universités, UPMC Univ. Paris 6, CNRS, LIP6, France (e-mail: haralampos.stratigopoulos@lip6.fr).

K. Huang is with the Department of Electrical and Computer Engineering, San Diego State University, USA (e-mail: khuang@mail.sdsu.edu).

A. Nahar, B. Orr and M. Pas are with Texas Instruments, USA (e-mail: {a-nahar2, borr, m-pas}@ti.com).

J. M. Carulli, Jr. is with GlobalFoundries, USA (e-mail: John.Carulli@globalfoundries.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier XX.XXXX/TCAD.2017.XXXXXXXX

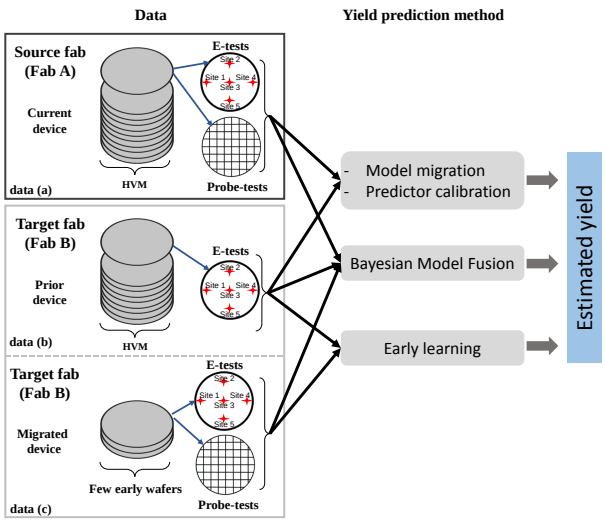


Fig. 1: Yield prediction during fab-to-fab product migration.

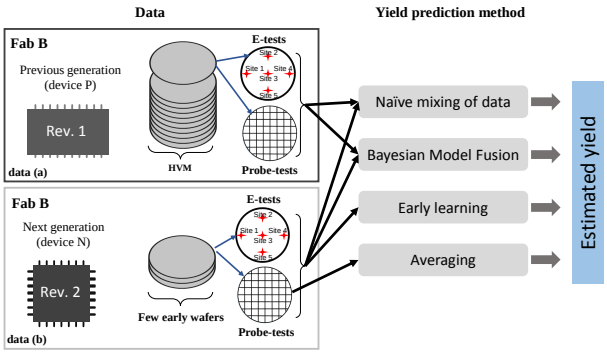


Fig. 2: Yield prediction across design generations.

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, we experiment with various methods 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, we examine four different methods, namely *model migration*, *predictor calibration*, *early learning*, and *Bayesian Model Fusion* (BMF). As illustrated in Fig. 1, 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, we consider 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, we experiment with various methods 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, we consider four different methods, namely *averaging*, *early learning*, *naive mixing of data*, and *Bayesian Model Fusion* (BMF). As shown in Fig. 2, 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).

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) [6]. 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 [7], [8]; 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 [9]–[11]. 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.

The remainder of this paper is organized as follows. In Section II, we briefly review state-of-the-art simulation-based yield estimation methods. In Section III, we discuss a regression-based approach for predicting yield based on the e-test profile of a wafer. In Section IV, we briefly present the method used herein for learning regression functions. In Section V, we discuss a feature selection method based on a Genetic Algorithm (GA), which helps in reducing the dimensionality of the problem and improving the overall accuracy of the learned regression models. In Sections VI and VII, we present the proposed yield forecasting methods for the fab-to-fab production migration scenario and the transition to a new design generation scenario, respectively. Experimental results using industrial data are presented in Section VIII and conclusions are drawn in Section IX.

II. Simulation-based yield prediction methods

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

A. Monte Carlo

Monte Carlo (MC) simulation [12], [13] 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.

B. Monte Carlo with speed enhancement

Several methods can be used to speed up MC, including Latin hypercube sampling (LHS) [14], quasi-Monte-Carlo (QMC) [15], and importance sampling [16], [17]. 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.

C. 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 [18]. 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 [18]–[20]. In [21], a recursive strategy is proposed to further accelerate the simulation effort.

D. Response surface and symbolic performance modeling

Another popular method for yield estimation is based on performance modeling [22]–[25]. 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.

E. 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 [26], [27]. 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.

III. 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, we first discuss the use of e-tests as a yield predictor for a specific device fabricated in a specific fab. Given the nature of e-tests, whose role is to reflect process variations that lead to yield loss and to drive yield learning, our 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, we will learn how to approximate them by training regression functions.

Let us consider a device that is currently in production. Assume that we have at hand 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. 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, we can compute the parametric yield of the k-th probe-test measurement for the i-th wafer, 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

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

The training data in (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.

$$y_k^i \approx f_k\left(\boldsymbol{ET}^i\right) \tag{2}$$

$$Y^i \approx f\left(\boldsymbol{ET}^i\right). \tag{3}$$

Once the regression functions are learned and their generalization accuracy is validated, we can readily use them 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. We 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 we only need to obtain the e-test measurements rather than all probe-test measurements for all devices on a wafer.

IV. 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* [28], [29]. In this work, we use MARS [29], which has also been successfully used in several other test cost reduction methods in the past [30], [31].

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.

$$\hat{f}(X) = a_0 + \sum_{m=1}^{M} a_m \cdot B_m(X), \tag{4}$$

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. We build MARS models using e-tests as input variables and yield vectors as the dependent output variables. We use piecewise-cubic basis functions, the maximum number of which is set to half of the number of input variables.

V. 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 [32], 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 [33], [34] and machine learning-based test [35].

In this work, we employ a heuristic-based technique to select a subset of e-test parameters. More specifically, we use

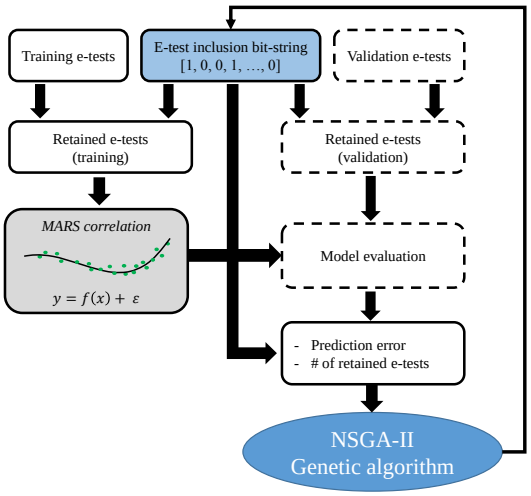


Fig. 3: GA-based feature selection method (NSGA-II).

a multi-objective GA, called NSGA-II [36]. 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 our 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.

Fig. 3 depicts an overview of our 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. We point out that different optimal e-test subsets may be selected for each probe-test. The algorithm stops when there is no significant improvement in the fitness values of a population over a window of the last five generations. We also note that, in each iteration of the GA, the same settings are used in the MARS models.

VI. Yield Prediction During Production Migration

Let us 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. Our 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.

- 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 III, the available information from fab A includes:
$$\text{wafer}_A^i = [\boldsymbol{ET}_A^i, \boldsymbol{y}_A^i, Y_A^i], \qquad i = 1, \cdots, w_A. \tag{5}$$
- 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:
$$\text{wafer}_B^i = [\boldsymbol{ET}_B^i, \boldsymbol{y}_B^i, Y_B^i], \qquad i = 1, \cdots, w_B. \tag{6}$$
- 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. We denote the e-test profile of the i-th fabricated wafer of this prior device as $\boldsymbol{ET'}_B^i$, $i = 1, \cdots, w_0$.

A. 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

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0} \sum_{i=1}^{w_0} f_{A,k}\left(\boldsymbol{ET'}_B^i\right). \tag{7}$$

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.

B. 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 $\boldsymbol{ET}_{A,j} =$

$[ET^1_{A,j}, \cdots, ET^{w_A}_{A,j}]$ and $\boldsymbol{ET'}_{B,j} = [ET^1_{B,j}, \cdots, ET^{'w_0}_{B,j}]$ 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 $\Delta(\mu_j)$

$$\hat{\boldsymbol{ET'}}_{B,j} = \boldsymbol{ET'}_{B,j} - \Delta(\mu_j), \tag{8}$$

$$\Delta(\mu_j) = \mu(\boldsymbol{ET'}_{B,j}) - \mu(\boldsymbol{ET}_{A,j}). \tag{9}$$

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, we employ a two-step procedure. First, using the cumulative distribution function (CDF) of the j-th e-test in fab B, $F_{B,j}$, we find the cumulative probability associated with each sample, $x^i_j = F_{B,j}(ET'^i_{B,j})$. Then, using the inverse CDF of fab A, we determine the e-test value associated with cumulative probability x^i_j, $\hat{ET'}^i_{B,j} = F^{-1}_{A,j}(x^i_j)$, where $F^{-1}_{A,j}$ is the inverse CDF of the j-th e-test for fab A. We employ the kernel density estimation (KDE) [37] 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., $y^i_{A,k} \approx f_{A,k}\left(\boldsymbol{ET}^i_A\right)$. 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

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0}\sum_{i=1}^{w_0} f_{A,k}\left(\hat{\boldsymbol{ET'}}^i_B\right). \tag{10}$$

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.

C. 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. We now consider the scenario where we have access to probe-tests from a small number w_B of early silicon wafers from fab B, containing this device. This enables us 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. $y^i_{B,k} \approx f_{B,k}\left(\boldsymbol{ET}^i_B\right)$. 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

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0}\sum_{i=1}^{w_0} f_{B,k}\left(\boldsymbol{ET'}^i_B\right). \tag{11}$$

D. 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 [38]–[43].

The training data in (5) allow us to learn an accurate regression function for predicting parametric yield of the k-th probe-test in fab A

$$\hat{y}^i_{A,k} \approx f_{A,k}\left(\boldsymbol{ET}^i_A\right) = \sum_{m=1}^{M} a_{A,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}^i_A\right). \tag{12}$$

We have 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 IV.

For small w_B, given the limited training data in (6), our objective is to learn an accurate regression function for fab B

$$\hat{y}^i_{B,k} \approx f^{'}_{B,k}\left(\boldsymbol{ET}^i_B\right) = \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}^i_B\right), \tag{13}$$

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 (6) for training and the rest for assessing the generalization ability of the regression function on previously unseen wafers. However, since we are interested in learning the regression function based on the very first few wafers, the data in (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 (13) by leveraging information from the data in (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 $\text{pdf}(\boldsymbol{a}_{B,k}|\textbf{wafer}_B)$, that is,

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}|\textbf{wafer}_B), \tag{14}$$

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

By applying Bayes' theorem, we can write

$$\text{pdf}(\boldsymbol{a}_{B,k}|\textbf{wafer}_B) \propto \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k}). \tag{15}$$

Thus, the problem boils down to

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k}). \tag{16}$$

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

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

$$\text{pdf}(\boldsymbol{a}_{B,k}) = \prod_{m=1}^{M} \text{pdf}(a_{B,k,m}). \tag{17}$$

We 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 $\lambda|a_{A,k,m}|$

$$\text{pdf}(a_{B,k,m}) = \frac{1}{\sqrt{2\pi}\lambda|a_{A,k,m}|} \cdot \exp\left[-\frac{(a_{B,k,m} - a_{A,k,m})^2}{2\lambda^2 a_{A,k,m}^2}\right]. \tag{18}$$

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* $\text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k})$ is expressed in terms of the data in (6). Specifically, since the data from each wafer is independent, we can write

$$\text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k}) = \prod_{i=1}^{w_B} \text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right). \tag{19}$$

Furthermore,

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \text{pdf}(\varepsilon^i), \tag{20}$$

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

$$\varepsilon^i = y_{B,k}^i - f_{B,k}\left(\boldsymbol{ET}_B^i\right). \tag{21}$$

This error is a random variable that is assumed to follow a zero-mean Gaussian distribution with some standard deviation σ_0

$$\text{pdf}(\varepsilon^i) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left(-\frac{\left(\varepsilon^i\right)^2}{2\sigma_0^2}\right). \tag{22}$$

Therefore, combining (20), (21), (22), and (13), we can write

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \cdot \exp\left\{-\frac{1}{2\sigma_0^2} \cdot \left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2\right\}. \tag{23}$$

By combining (17), (18), (19), and (23), we obtain an expression of $\text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k})$. By taking the natural logarithm of this expression, the maximization problem in (16), after eliminating constant terms, becomes

$$\max_{\boldsymbol{a}_{B,k}} -\left(\frac{\sigma_0}{\lambda}\right)^2 \sum_{m=1}^{M} \frac{(a_{B,k,m} - a_{A,k,m})^2}{a_{A,k,m}^2} - \sum_{i=1}^{w_B}\left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2. \tag{24}$$

The optimal values of σ_0 and λ are determined by k-fold cross-validation [28], [29].

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

$$\hat{\bar{y}}_{B,k} = \frac{1}{w_0}\sum_{i=1}^{w_0} f'_{B,k}\left(\boldsymbol{ET'}_B^i\right). \tag{25}$$

VII. 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 us 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 we have access to the e-test and probe-test data from w_P wafers. Using similar notation as in Section III, information from device P includes

$$\text{wafer}_P^i = [\boldsymbol{ET}_P^i, \boldsymbol{y}_P^i, Y_P^i], \qquad i = 1, \cdots, w_P. \tag{26}$$

Let us also assume that we have at hand 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

$$\text{wafer}_N^i = [\boldsymbol{ET}_N^i, \boldsymbol{y}_N^i, Y_N^i], \qquad i = 1, \cdots, w_n. \tag{27}$$

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

A. 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

$$\hat{\bar{Y}}_N = \frac{1}{w_n}\sum_{i=1}^{w_n} Y_N^i. \tag{28}$$

B. Early learning

Another approach is to use the data in (27) as a training set and learn a regression model to express wafer yield as a function of the e-tests for device N

$$Y^i \approx f_N\left(\boldsymbol{ET}_N^i\right). \tag{29}$$

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

$$\hat{\bar{Y}}_N = \frac{1}{w_P}\sum_{i=1}^{w_P} f_N\left(\boldsymbol{ET}_P^i\right). \tag{30}$$

C. Naive mixing of data

A third approach is to naively mix data in (26) and (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

$$Y^i \approx f_{PN}\left(\boldsymbol{ET}^i\right). \tag{31}$$

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

$$\hat{\bar{Y}}_N = \frac{1}{w_P}\sum_{i=1}^{w_P} f_{PN}\left(\boldsymbol{ET}_P^i\right). \tag{32}$$

D. Bayesian Model Fusion

Finally, similar to Section VI-D, we 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 we can learn regression models

$$\hat{Y}_P^i \approx f_P\left(\boldsymbol{ET}_P^i\right) = \sum_{m=1}^{M} a_{P,m}\cdot b_m\left(\boldsymbol{ET}_P^i\right) \tag{33}$$

and

$$\hat{Y}_N^i \approx f^{'}{}_N\left(\boldsymbol{ET}_N^i\right) = \sum_{m=1}^{M} a_{N,m}\cdot b_m\left(\boldsymbol{ET}_N^i\right), \tag{34}$$

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 (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

$$\max_{\boldsymbol{a}_N} \text{pdf}(\boldsymbol{a}_N|\textbf{wafer}_N), \tag{35}$$

where $\text{pdf}(\boldsymbol{a}_N|\textbf{wafer}_N) \propto \text{pdf}(\boldsymbol{a}_N)\text{pdf}(\textbf{wafer}_N|\boldsymbol{a}_N)$, $\text{pdf}(\boldsymbol{a}_N)$ is the *prior* distribution, $\text{pdf}(\textbf{wafer}_N|\boldsymbol{a}_N)$ is the *likelihood function*, and $\textbf{wafer}_N = [\text{wafer}_N^1, \cdots, \text{wafer}_N^{w_n}]$. Similar steps as in Section VI-D 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

$$\hat{\bar{Y}}_N \approx= \frac{1}{w_P}\sum_{i=1}^{w_P} f_N^{'}\left(\boldsymbol{ET}_P^i\right). \tag{36}$$

VIII. Experimental Results

A. 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, we use actual HVM production datasets from two consecutive design generations of a Texas Instruments 65nm RF transceiver[1]. We will refer to these two design generations as device P and device N, respectively, emphasizing that device N is the new-generation of device P with slight enhancements. Our datasets originate from two geographically dispersed fabs, which we will refer 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 Fig. 4, 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), we use as its e-test signature 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 we can compute the parametric yield of each probe-test on every wafer, as well as the overall yield of each wafer.

Using these datasets, we seek 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.

B. 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, we use the entire datasets of device N from both fab A and fab B to perform two independent experiments, one for each fab. The regression models are trained using MARS and we use 5-fold cross validation to report robust prediction error values. Specifically, for a given fab, the dataset is divided into

[1]Details regarding the devices may not be released due to a binding NDA.

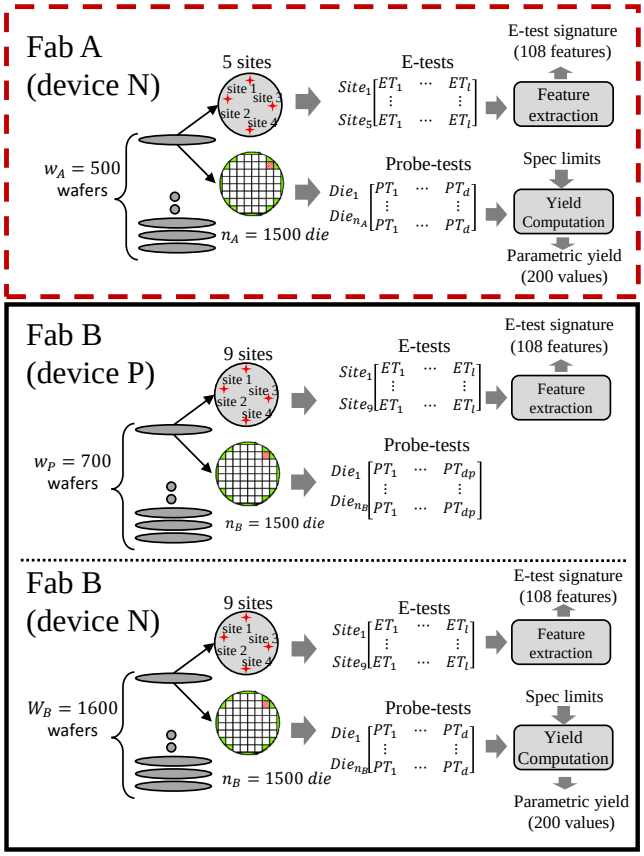


Fig. 4: Datasets from fab A and fab B.

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, we report the average prediction error across the 5 iterations.

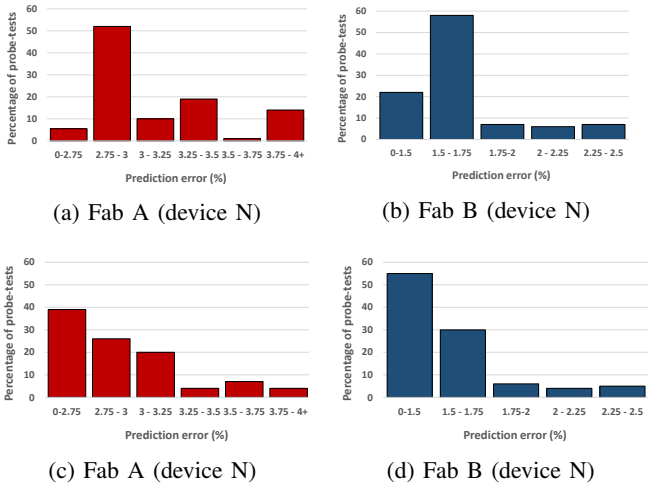
We use the following expression for calculating the error in predicting the parametric yield of the k-th probe test

$$\delta_k = 100 \cdot \frac{1}{w} \sum_{i=1}^{w} \frac{|\hat{y}_k^i - y_k^i|}{y_k^i}, \tag{37}$$

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.

Figs. 5(a)-(b) present the parametric yield prediction results for the datasets of device N from fab A and fab B, respectively. In this experiment, we consider all 108 e-test features. 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 Fig. 5(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.

Figs. 5(c)-(d) present the same results as in Figs. 5(a)-(b), but this time using only the subset of e-test features that are



(a) Fab A (device N) (b) Fab B (device N)

(c) Fab A (device N) (d) Fab B (device N)

Fig. 5: 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.

selected by the GA-based feature selection method of Section V. 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. Figs. 5(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 Figs. 5(a)-(b). These results corroborate that, by reducing the dimensionality of the e-test signature, feature selection improves significantly the quality of predictions. We note 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, we examine the use of e-tests for predicting wafer yield. As before, the regression models are trained using MARS, we employ 5-fold cross validation to report robust prediction errors values, and we use a similar expression for evaluating the prediction error of the overall wafer yield

$$\delta = 100 \cdot \frac{1}{w} \sum_{i=1}^{w} \frac{|\hat{Y}^i - Y^i|}{Y^i} \tag{38}$$

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 I 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

TABLE I: 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%

in lower prediction error. In order to quantitatively demonstrate this improvement, we use the metric $\Delta\epsilon$, defined as

$$\boldsymbol{\Delta}\epsilon = |\frac{\text{All e-tests error - Subset of e-tests error}}{\text{All e-tests error}} \times 100|. \tag{39}$$

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 Fig. 5 and Table I, for the rest of experiments all regression models are trained with the subset of e-tests selected by this method.

C. Yield prediction during migration from fab A to fab B

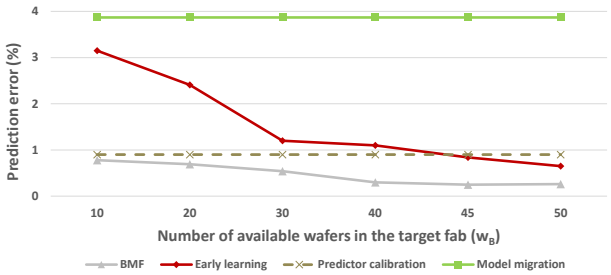
In order to quantify yield prediction accuracy during fab-to-fab production migration using the methods discussed in Section VI, we performed the following experiment using fab A as the source fab and fab B as the target fab. The model migration and predictor calibration methods assume access to both 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 $w_B \ll W_B$. We vary w_B 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, we use bootstrapping to report robust prediction errors, smoothen them, and assist with the interpretation of the overall results. In total, we perform 10 bootstrap iterations and, in each iteration, we sample w_B wafers uniformly at random from the W_B wafers and we perform 5-fold cross validation. The reported prediction errors are averaged over these 50 iterations. In each iteration, we use the following expressions for evaluating the prediction error of the HVM parametric yield of the k-th probe-test and the HVM wafer yield

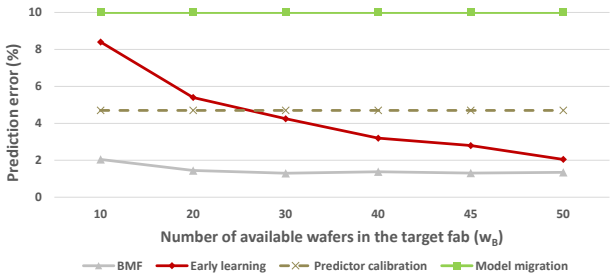
$$\delta_k = 100 \cdot \frac{|\hat{\bar{y}}_{B,k} - \bar{y}_{B,k}|}{\bar{y}_{B,k}}, \tag{40}$$

$$\delta = 100 \cdot \frac{|\hat{\bar{Y}}_B - \bar{Y}_B|}{\bar{Y}_B}, \tag{41}$$

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,



(a) Results for a randomly-selected probe-test



(b) Results for overall wafer yield

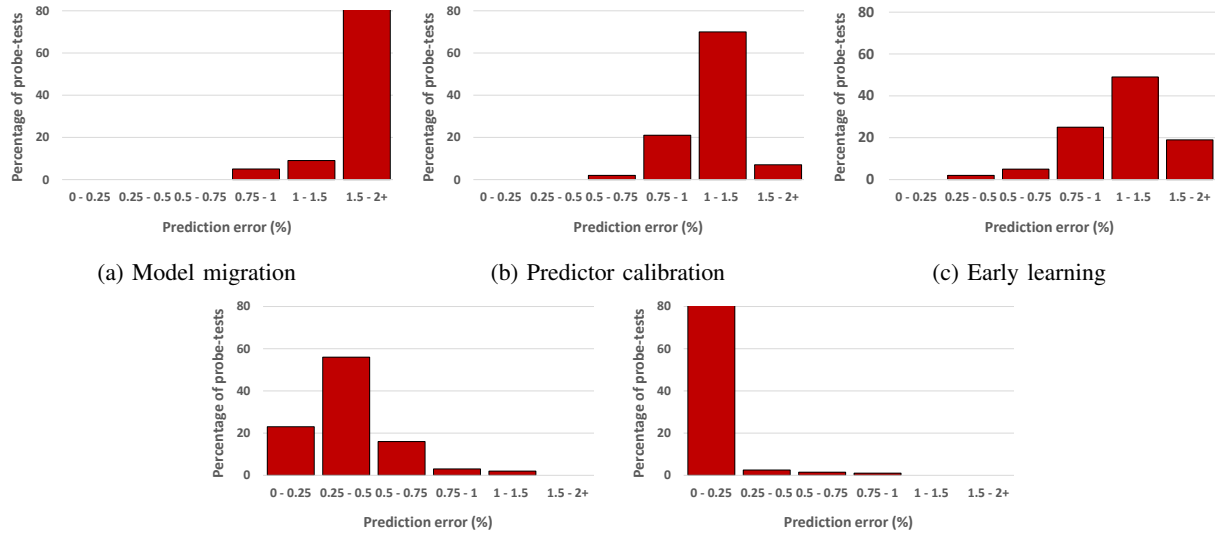
Fig. 6: Yield prediction error during production migration.

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 VI is demonstrated in Figs. 6(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 Fig. 6, 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 our 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



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

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

Fig. 7: Yield prediction error across all 200 probe measurements during fab A to fab B production migration with $w_B = 30$.

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 [38]–[43].

Finally, in Fig. 7, we compare 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, we also include a "lower bound" result where we apply the early learning method 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.

D. Yield prediction across design generations

In order to quantify yield prediction accuracy across design generations using the methods discussed in Section VII, we performed the following experiment 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 our case, we used wafers from the first two lots in our dataset. In addition, the rest of the methods assume access to the entire dataset of the previous-generation device P. We perform 10 bootstrap iterations and, in each iteration, we sample w_B wafers uniformly at random from the available w_n wafers and we perform 5-fold cross validation. The reported prediction errors are averaged over these 50 iterations. We repeat the experiment by varying w_B in the range [10, 50]. We use the following expression for evaluating prediction error of the HVM overall wafer yield of device N

$$\delta = 100 \cdot \frac{|\hat{\bar{Y}}_N - \bar{Y}_N|}{\bar{Y}_N}, \tag{42}$$

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

Fig. 8 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-

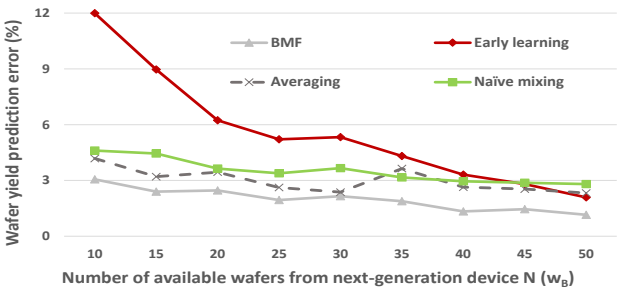


Fig. 8: Error in predicting device N yield from early wafers.

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.

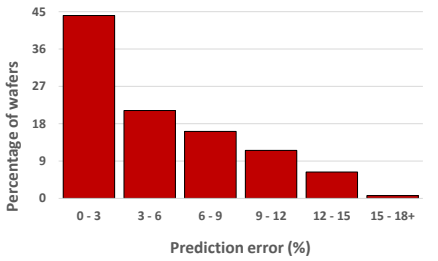
To gain better insight, we consider $w_B = 20$ and we illustrate, in Fig. 9, the distribution of wafer level prediction error for all wafers in the validation set for the BMF and early learning methods. The prediction error is expressed as

$$\delta_i = 100 \cdot \frac{|\hat{Y}_N^i - Y_N^i|}{Y_N^i}, \tag{43}$$

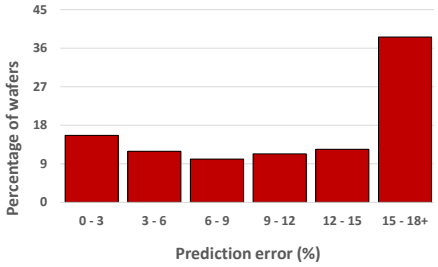
where $\hat{Y}_N^i$ and Y_N^i are the predicted and actual wafer yield values for the i-th wafer, respectively. In each histogram, the horizontal axis represents the prediction error range and the vertical axis represents the percentage of the wafers in the validation set whose wafer yield is predicted 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%.

IX. Conclusion

We 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 at our disposal, 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



(a) BMF



(b) Early learning

Fig. 9: Wafer yield prediction error of device N with $w_B = 20$.

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

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

[33] H.-G. 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.

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

[35] H.-G. 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.

[36] 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.

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

[38] 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.

[39] 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.

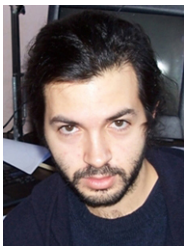
[40] C. Gu, E. Chiprout, and X. Li, "Efficient moment estimation with extremely small sample size via Bayesian inference for Analog/Mixed-signal validation," in *Proc. IEEE/ACM Design Automation Conference*, 2013, pp. 1–7.

[41] 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.

[42] 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.

[43] 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.

Ali Ahmadi (S'11) received the B.S. degree in computer engineering from University of Isfahan, Iran, in 2006 and the M.S. degree in computer engineering from University of Tehran, Iran, in 2009. He is currently a Ph.D. student in electrical engineering at the University of Texas at Dallas, USA. His research interests are applications of machine learning and data mining in semiconductor manufacturing for test cost reduction, yield improvement and defect modeling. He was a recipient of the Best Paper Award from the 2015 IEEE VLSI Test Symposium.



Haralampos-G. Stratigopoulos (S'02-M'07) received the Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2001 and the Ph.D. in electrical engineering from Yale University, USA, in 2006. From October 2007 to May 2015 he was a Researcher with the French National Center for Scientific Research (CNRS) at TIMA Laboratory, Grenoble, France. Currently he is a researcher with the CNRS at LIP6 Laboratory, Paris, France. His main research interests are in the areas of design-for-test and built-in test for analog, mixed-signal, RF circuits and systems, computer-aided design, and machine learning. He was the General Chair of the 2015 IEEE International Mixed-Signal Testing Workshop (IMSTW) and the Program Chair of the 2017 IEEE European Test Symposium (ETS). He has served on the Technical Program Committees of Design, Automation, and Test in Europe Conference (DATE), IEEE International Conference on Computer-Aided Design (ICCAD), IEEE VLSI Test Symposium (VTS), IEEE European Test Symposium (ETS), IEEE International Test Conference (ITC) and several others international conferences. He is an Associate Editor of Springer Journal of Electronic Testing: Theory & Applications, IEEE Design & Test Magazine, and IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. He received the Best Paper Award in the 2009, 2012, and 2015 IEEE European Test Symposium (ETS).

Ke Huang (S'10, M'12) received the B.S. and M.S. degrees in electrical engineering from Joseph Fourier University (Grenoble I University), Grenoble, France, in 2006 and 2008, respectively, and the Ph.D. degree in electrical engineering from the University of Grenoble, Grenoble, France, in 2011. He was a Postdoctoral Research Associate at the University of Texas at Dallas, USA, from 2012 to 2014. In 2014, he joined the Department of Electrical and Computer Engineering at San Diego State University, San Diego, CA, USA, where he is now an Assistant Professor. His research focuses on very-large-scale integration (VLSI) testing and security, computer-aided design of integrated circuits (ICs), and intelligent vehicles. Prof. Huang was awarded a Ph.D. Fellowship from the French Ministry of National Education from 2008 to 2011. He was a recipient of the Best Paper Award from the 2013 Design Automation and Test in Europe (DATE13) conference and a recipient of the Best Paper Award from the 2015 IEEE VLSI Test Symposium (VTS15).

Amit Nahar received his M.S. degree in Electrical and Computer Engineering from Portland State University in 2005. He joined Texas Instruments Inc. in 2005 and currently is a Test Manager working on defining, architecting and developing methodologies for adaptive test and volume test, manufacturing and design data analytics. He has published over 18 papers and holds 3 US patents.

Bob Orr received a Bachelor of Science in Electrical Engineering from the Rose-Hulman Institute of Technology in 1980. Currently he is an Engineering Manager in the Test Technology and Product Engineering group at Texas Instruments Incorporated defining standards, methods and tools to support data acquisition and integration infrastructure requirements for TI Manufacturing and business group engineers, including advanced statistical applications for production test. Roles at TI have included Factory and Business Product and Test engineering for standard, custom VLSI and custom memory products.

Michael Pas received his Ph.D. degree in 1989 from Texas A&M University in Physical Chemistry. He completed a Post Doc at the Center for Numerical Intensive Computing at IBM Kingston, NY in 1991. He joined Texas Instruments Inc. in 1991. He has held a variety of engineering and management positions at TI. He has published over 35 papers in MRS, ECS and IEEE journals. He holds 19 US patents.

John M. Carulli, Jr. (SM'12) received the M.S.E.E. degree from the University of Vermont, Burlington, VT, USA, in 1990. He leads the Test organization at GLOBALFOUNDRIES Fab8 in Malta, NY working on leading edge CMOS technologies. He had 21 years at Texas Instruments where he was a Distinguished Member of the Technical Staff. While in the Analog Engineering Operations organization he led test and design data mining methods targeted at test cost reduction. While in the Silicon Technology Development organization, he was the Manager of the Product Reliability group responsible for product and design reliability activities for new technology development. John holds 7 US Patents and has over 50 publications in the areas of reliability, test, and process development. He is co-recipient of two Best Paper Awards and two Best Paper Nominations working in close collaboration with university partners. John serves on the organizing or program committees of several conferences including the International Test Conference, VLSI Test Symposium, and European Test Symposium.

Yiorgos Makris (SM'08) received the Diploma of Computer Engineering from the University of Patras, Greece, in 1995 and the M.S. and Ph.D. degrees in Computer Engineering from the University of California, San Diego, in 1998 and 2001, respectively. After spending a decade on the faculty of Yale University, he joined UT Dallas where he is now a Professor of Electrical Engineering, leading the Trusted and RELiable Architectures (TRELA) Research Laboratory. His research focuses on applications of machine learning and statistical analysis in the development of trusted and reliable integrated circuits and systems, with particular emphasis in the analog/RF domain. Prof. Makris serves as an Associate Editor of the IEEE Transactions on Information Forensics and Security and the IEEE Design & Test of Computers Periodical and he has also served as a guest editor for the IEEE Transactions on Computers and the IEEE Transactions on Computer-Aided Design of VLSI Circuits and Systems. He is a recipient of the 2006 Sheffield Distinguished Teaching Award, as well as Best Paper Awards from the 2013 Design Automation and Test in Europe (DATE'13) conference and the 2015 VLSI Test Symposium (VTS'15).