Yield Forecasting in Fab-to-Fab Production Migration Based on Bayesian Model Fusion

Ali Ahmadi*, Haralampos-G. Stratigopoulos†, Amit Nahar‡, Bob Orr‡, Michael Pas‡ and Yiorgos Makris*

*Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080
†Sorbonne Universités, UPMC Univ. Paris 06, CNRS, LIP6, 4 place Jussieu, 75005, Paris, France
‡Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, TX 75243

***Abstract*—Yield estimation is an indispensable piece of information at the onset of high-volume production of a device. It can be used to refine the process/design in time so as to guarantee high production yield. In the case of migration of production of a specific device from a source fab to a target fab, yield estimation in the target fab can be accelerated by employing information from the source fab, assuming that the process parameter distributions in the two fabs are similar, but not necessarily the same. In this paper, we employ the Bayesian Model Fusion (BMF) technique for efficient yield prediction of a device in the target fab. BMF adopts prior knowledge from the source fab and combines it intelligently with information from a limited number of early silicon wafers from the target fab. Thus, BMF allows us to obtain quick and accurate yield estimates at the onset of production in the target fab. The proposed methodology is demonstrated on an industrial RF transceiver.**

I. Introduction

The rapidly growing and dynamically changing consumer electronics market introduces interesting challenges to production planning of semiconductor manufacturing companies, calling for agility and flexibility, in order to efficiently respond to fluctuating demand. Contingency plans for dealing with catastrophic events, such as earthquakes and floods, as well as political or sheer financial reasons, also call for flexibility in production planning. In view of the increasing complexities in semiconductor industry, as well as the increasing demand for faster designs with growing quality requirements, a quick and successful migration of production becomes crucial, in order for companies to maintain their profitability. Migrating the production of a device from one fab to another, however, is not a trivial endeavor. Accurate and fast prediction of yield in the target fab is an indispensable piece of information during production migration, in order to identify and quickly resolve any issues that may jeopardize production ramp-up.

Figure 1 shows an overview of the production migration problem. Consider a device currently produced in high volume manufacturing (HVM) in fab A, for which both *e-test* and *probe-test* data is available for a statistically significant number of wafers. E-tests are measurements obtained via simple circuits (i.e. Process Control Monitors) which are typically placed on the scribe lines of the wafer and which reflect how a wafer has been affected by process variations. Probe-tests, on the other hand, are direct measurements of device performances, as obtained at wafer-level. Let us now assume that we want to migrate production of this device to a different fab B of the same technology node. While probe-test data from fab B is not available, since this device has not been produced there before, we assume that e-test data is available for a statistically significant number of wafers from a *previous device* fabricated in fab B. Indeed, since e-tests are technology-specific rather than design-specific, they are typically common across these two devices. Using these three pieces of information, the original incarnation of the yield prognosis problem [1] seeks to predict how well the current device will yield, once its production is migrated from fab A to fab B.

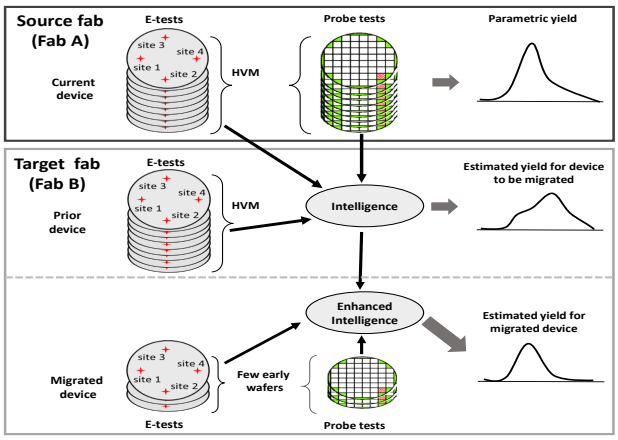


Fig. 1: Yield prediction during fab-to-fab migration.

A straightforward approach, called *model migration*, uses the data available in fab A to train statistical models which can predict parametric yield of a wafer as a function of its e-test vector. These models are then "migrated" to fab B and applied on the large-scale e-test data available from a prior device fabricated therein, in order to predict the expected parametric yield of the new device once it is migrated to fab B. Fab-to-fab discrepancies, however, result in rather poor predictions in this case. A more elaborate approach, called *predictor calibration* [1], addresses this limitation by first calibrating the e-test distribution of fab B based on the e-test distribution of fab A. Applying the trained models on the calibrated e-tests of fab B drastically improves prediction accuracy. In either case, the assumption is that no wafers of the device in question have been produced by fab B.

However, prior to migrating a product, a few engineering wafers are typically produced and characterized in the target fab. Therefore, herein we seek to investigate the utility of e-test and probe-test measurements obtained from these few engineering wafers (or from the first few wafers in HVM production), in improving the accuracy of parametric yield prediction in fab B. Once again, a straightforward approach, which we will refer to as *early learning*, is to simply use the limited available data from fab B to train statistical models for predicting parametric yield of a wafer as a function of its e-test vector. These models can, then, be used on the large-scale

978-1-4673-8388-2/15/$31.00 ©2015 IEEE

e-test data available from a prior device fabricated in fab B, in order to predict the expected parametric yield of the new device when it is produced in high-volume. The accuracy of these predictions, however, is expected to be limited, as the number of available wafers is usually small and may not be representative enough to learn a model that accurately predicts the parametric yield for future wafers.

In this work, we address this limitation by intelligently combining data from both fab A and fab B, in order to construct a robust and accurate parametric yield prediction model for fab B. More specifically, we employ the *Bayesian Model Fusion (BMF)* technique to refine the inaccurate prediction model for fab B, which is learned based on the limited data available from fab B, using the prediction model that was learned through the abundance of data available in fab A. Thereby, the fused prediction model becomes far more robust and predicts far more accurately the parametric yield of future wafers that are produced in fab B. BMF is a very powerful technique which has been used successfully for model improvement in various contexts in the past, including pre-silicon validation, yield learning, post-manufacturing tuning, bit error rate estimation, and alternate test [2]–[7].

The remainder of this paper is organized as follows. In Section II, we discuss a regression-based approach for predicting parametric yield based on e-test measurements, both within a fab and during fab-to-fab migration. Then, in Section III, we elaborate on the four yield prediction methods, namely *model migration*, *predictor calibration*, *early learning*, and *Bayesian Model Fusion*. Experimental results comparing the accuracy of these four methods using industrial data are presented in Section IV and conclusions are drawn in Section V.

II. Problem formulation

A. E-test to probe-test correlation

Let us consider a device that is currently being manufactured in fab A. Also assume that we have at hand the e-test measurements from w_A wafers that contain this device and the probe-test measurements from all n_A devices contained in each of these wafers. Let $\boldsymbol{ET}_A^i = [ET_{A,1}^i, \cdots, ET_{A,l}^i]$ denote the l-dimensional e-test measurement pattern of the i-th wafer, where $ET_{A,k}^i$ denotes the k-th e-test measurement. Let $\boldsymbol{PT}_A^{ij} = [PT_{A,1}^{ij}, \cdots, PT_{A,d}^{ij}]^T$ denote the d-dimensional probe-test measurement pattern obtained on the j-th device contained in the i-th wafer, where $PT_{A,k}^{ij}$ denotes the k-th probe-test measurement on the j-th device in the i-th wafer. Let also $\boldsymbol{PT}_A^i = [\boldsymbol{PT}_A^{i1} \cdots \boldsymbol{PT}_A^{in_A}]$ denote the $d \times n_A$ 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 (e.g. wafer-level yield of the k-th probe-test measurement) for the i-th wafer, denoted by $y_{A,k}^i$, as the percentage of devices in the i-th wafer that pass the k-th probe-test specification limits. Let $\boldsymbol{y}_A^i = [y_{A,1}^i, \cdots, y_{A,d}^i]$ denote the d-dimensional parametric yield vector of the probe-test measurements for the i-th wafer. $\boldsymbol{y}_A^i$ is directly computed from $\boldsymbol{PT}_A^i$ given the specifications of the probe-test measurements. Therefore, information from fab A includes

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

We conjecture that a relationship exists between the parametric yield of the k-th probe-test measurement and the e-test measurement pattern for the i-th wafer, since the purpose of e-test is to reflect process variations that lead to yield loss and to drive yield learning. This relationship, however, is intricate and does not have a known closed-form mathematical expression. For this reason, it is approximated using a regression function $f_{A,k}$. The training data in (1) is used to learn this regression function that predicts the parametric yield of the k-th probe-test measurement for the i-th wafer from its e-test measurement pattern.

$$\hat{y}_{A,k}^i \approx f_{A,k}\left(\boldsymbol{ET}_A^i\right). \tag{2}$$

Once the regression function is learned and its generalization accuracy is validated, we can readily use it to estimate the parametric yield $\hat{\boldsymbol{y}}_A^i$ for future wafers, i.e. $i > w_A$, based solely on their e-test vector. We will show that these estimates approximate accurately the ground truth values $\boldsymbol{y}_A^i$. In this way, in order to compute the parametric yield of a wafer, we only need to obtain the e-test measurements; thereby, we can circumvent the need to obtain the probe-test measurements for all devices in a wafer, thus saving significant cost.

B. Yield prognosis in fab-to-fab migration

Let us now consider that the same device is planned to be manufactured in high-volume in fab B. By employing prior information from fab A, we will show that we can build an accurate parametric yield prediction model for fab B by relying on limited information from the first few wafers manufactured in fab B. For this purpose, we will rely on the BMF technique. In this way, we will show that an accurate parametric yield prediction model for fab B can be generated very quickly, without needing to collect data from a large volume of wafers.

Suppose that we consider the first w_B wafers and that we have at hand e-test measurements from each of these wafers, as well as probe-test measurements from all n_B devices contained in each of these wafers. Following similar notation as in Section II-A, information from fab B includes

$$\text{wafer}_B^i = [\boldsymbol{ET}_B^i, \boldsymbol{y}_B^i], \qquad i = 1, \cdots, w_B. \tag{3}$$

We are interested in learning a regression function that models the relationship between the parametric yield of each k-th probe-test measurement and the e-test measurement pattern for the i-th wafer produced in fab B

$$\hat{y}_{B,k}^i \approx f_{B,k}\left(\boldsymbol{ET}_B^i\right). \tag{4}$$

Let $\boldsymbol{ET'}_B^i = [ET'_{B,1}^i, \cdots, ET'_{B,l}^i]$ denote the l-dimensional e-test measurement pattern of the i-th wafer that contains another device produced in fab B, that is, not the device whose production is planned to be migrated from fab A to fab B.

We will show that the model in (4), which is learned based on the first few wafers using the BMF technique: (a) provides accurate parametric yield predictions that are practically indistinguishable from the predictions of a model that is learned based on a large volume of wafers, as is the model in (2), (b) provides better parametric yield predictions as compared to migrating the models $f_{A,k}$ learned in fab A directly into fab B,

$$\hat{y}_{B,k}^i \approx f_{A,k}\left(\boldsymbol{ET'}_B^i\right), \tag{5}$$

and (c) provides better parametric yield predictions as compared to the predictor calibration method proposed in [1].

III. Yield prediction methods

A. Model migration

As mentioned earlier, 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 signature, $\hat{y}^i_{A,k} \approx f_{A,k}\left(\boldsymbol{ET}^i_A\right)$. Then, the trained regression function is applied directly to the e-tests of fab B, in order to predict parametric yield, $\hat{y}^i_{B,k} \approx f_{A,k}\left(\boldsymbol{ET'}^i_B\right)$. However, since the e-tests of fab A and fab B come from different distributions, the accuracy of this model is expected to be limited.

B. Predictor calibration

This technique was proposed in [1] as a solution for parametric yield prediction during fab-to-fab migration. The proposed method is based on e-test and probe-test measurements of fab A, as well as the e-test profile of fab B, which can be obtained using another device that is fabricated in the same technology node in fab B. Therein, the authors proposed an algorithm to calibrate the e-test distribution of fab B based on the e-test distribution of fab A. Then, the calibrated e-tests are utilized for parametric yield prediction in fab B. In summary, this approach comprises the following steps:

- A regression function is first trained to express parametric yield in fab A as a function of the e-test signature, e.g. $\hat{y}^i_{A,k} \approx f_{A,k}\left(\boldsymbol{ET}^i_A\right)$.
- Then, the calibration algorithm maps the distribution of e-test measurements in fab B into the distribution of e-test measurements in fab A, e.g. $\widehat{\boldsymbol{ET'}}^i_B = F_A^{-1}\left(F_B(\boldsymbol{ET'}^i_B)\right)$, where F_B is the Cumulative Distribution Function of the e-test profile in fab B.
- Finally, in order to predict parametric yield in fab B, the trained regression model is applied to the calibrated e-test measurements, e.g. $\hat{y}^i_{B,k} \approx f_{A,k}\left(\widehat{\boldsymbol{ET'}}^i_B\right)$.

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-test measurements are available for this device from fab B. We now consider the scenario where we have access to probe-test measurements from w_B early silicon wafers from fab B during production migration. This enables us to train a regression model to express parametric yield as a function of the e-test signature, i.e. $\hat{y}^i_{B,k} \approx f_{B,k}\left(\boldsymbol{ET}^i_B\right)$. Subsequently, this model can be applied to the available e-test profile from fab B, which can be obtained from a prior device fabricated therein, in order to predict parametric yield, i.e. $\hat{y}^i_{B,k} \approx f_{B,k}\left(\boldsymbol{ET'}^i_B\right)$. The accuracy of this method is very limited, however, because the regression model is trained on a small, possibly not representative, training set.

D. Bayesian Model Fusion (BMF)

Early learning solely uses data from a few initial wafers to build a regression model for fab B. However, a more elaborate technique can utilize rich measurements from fab A to enhance the prediction accuracy. In this work, we employ BMF to intelligently fuse data from both fabs, in order to provide an accurate yield prediction in fab B.

Suppose that we have w_A wafers from fab A. The training data

$$\text{wafer}^i_A = [\boldsymbol{ET}^i_A, \boldsymbol{y}^i_A], \qquad i = 1, \cdots, w_A \tag{6}$$

allows us to learn an accurate regression function for predicting yield of the k-th probe-test

$$\hat{y}^i_{A,k} \approx f_{A,k}\left(\boldsymbol{ET}^i_A\right) = \sum_{m=1}^{M} a_{A,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}^i_A\right). \tag{7}$$

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, such as polynomial, Multi Adaptive Regression Splines (MARS), etc. [8], [9].

For small w_B, given the limited training data

$$\text{wafer}^i_B = [\boldsymbol{ET}^i_B, \boldsymbol{y}^i_B], \qquad i = 1, \cdots, w_B, \tag{8}$$

our objective is to learn an accurate regression function

$$\hat{y}^i_{B,k} \approx f_{B,k}\left(\boldsymbol{ET}^i_B\right) = \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}^i_B\right), \tag{9}$$

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 (8) 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 (8) 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 (9) by leveraging information from the data in (6), 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,

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}|\mathbf{wafer}_B), \tag{10}$$

where $\mathbf{wafer}_B = [\text{wafer}^1_B, \cdots, \text{wafer}^{w_B}_B]$. 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

$$\text{pdf}(\boldsymbol{a}_{B,k}|\mathbf{wafer}_B) \propto \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\mathbf{wafer}_B|\boldsymbol{a}_{B,k}). \tag{11}$$

Thus, the problem boils down to

$$\max_{\boldsymbol{a}_{B,k}} \text{pdf}(\boldsymbol{a}_{B,k}) \cdot \text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k}). \quad (12)$$

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

$$\text{pdf}(\boldsymbol{a}_{B,k}) = \prod_{m=1}^{M} \text{pdf}(a_{B,k,m}). \quad (13)$$

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

$$\text{pdf}(a_{B,k,m}) = \frac{1}{\sqrt{2\pi}\lambda|a_{A,k,m}|} \cdot \exp\left[-\frac{(a_{B,k,m} - a_{A,k,m})^2}{2\lambda^2 a_{A,k,m}^2}\right]. \quad (14)$$

This approach accounts for the fact that $a_{B,k,m}$ is expected to be similar to $a_{A,k,m}$ and deviate from $a_{A,k,m}$ 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 (8). Specifically, since the data from each wafer are independent, we can write

$$\text{pdf}(\textbf{wafer}_B|\boldsymbol{a}_{B,k}) = \prod_{i=1}^{w_B} \text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right). \quad (15)$$

Furthermore,

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \text{pdf}(\varepsilon^i), \quad (16)$$

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

$$\varepsilon^i = y_{B,k}^i - f_{B,k}\left(\boldsymbol{ET}_B^i\right). \quad (17)$$

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

$$\text{pdf}(\varepsilon^i) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left(-\frac{\left(\varepsilon^i\right)^2}{2\sigma_0^2}\right). \quad (18)$$

Therefore, combining (16), (17), (18), and (9), we can write

$$\text{pdf}\left(\text{wafer}_B^i|\boldsymbol{a}_{B,k}\right) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \cdot \exp\left\{-\frac{1}{2\sigma_0^2} \cdot \left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2\right\}. \quad (19)$$

By combining (13), (14), (15), and (19), 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 (12), after eliminating constant terms, becomes

$$\max_{\boldsymbol{a}_{B,k}} -\left(\frac{\sigma_0}{\lambda}\right)^2 \sum_{m=1}^{M} \frac{(a_{B,k,m} - a_{A,k,m})^2}{a_{A,k,m}^2} - \sum_{i=1}^{w_B}\left[y_{B,k}^i - \sum_{m=1}^{M} a_{B,k,m} \cdot b_{k,m}\left(\boldsymbol{ET}_B^i\right)\right]^2. \quad (20)$$

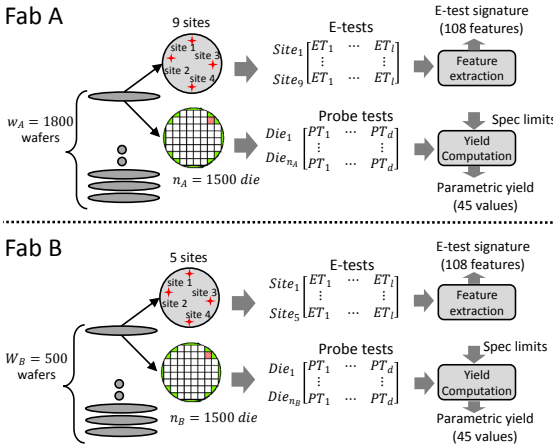


Fig. 2: Datasets from fab A and fab B.

The optimal values of σ_0 and λ are determined by v-fold cross-validation [8], [9].

IV. Experimental Results

A. Case study and datasets

In order to experimentally evaluate the effectiveness of the proposed yield prediction method for the fab-to-fab production migration problem, we use actual production data from a 65nm RF tranceiver currently in HVM production by Texas Instruments[1]. These data, which are depicted in Figure 2, correspond to devices from two geographically dispersed fabs wherein this device is fabricated, which we will refer to as fab A and fab B. The dataset for fab A includes l=54 e-test and d=45 probe-test measurements from a total of w_A=1800 wafers, each of which has 9 e-test measurement sites and approximately 1500 die per wafer. The dataset for fab B includes the same e-test and probe-test measurements from a total of W_B=500 wafers, with the only difference being that e-test measurements are obtained on only 5 instead of 9 sites. These two datasets were obtained from the two fabs at approximately the same time period. Along with the data, we are also provided with the specification limits for each of the 45 probe-test measurements, hence for each of the two fabs we can compute the parametric yield of each probe-test measurement on every wafer and, thereby, the parametric yield of each wafer. Additionally, for each of the 54 e-test measurements, we compute the mean and the standard deviation across the 9 sites on wafers produced in fab A (5 sites on wafers produced in fab B), hence the e-test signature of each wafer consists of 108 features.

B. Experiment design

We consider the four prediction techniques described in Section III. These are summarized in Table I, along with their corresponding training and validation set.

In our experiment, we vary w_B in the range [10, 50], in order to study the influence of the size of the training set on the early learning and BMF methods. We use MARS to learn the regression functions, and we train a separate regression model for each of the 45 probe-tests. As error prediction metric

[1]Details regarding the device cannot be released due to an NDA under which these data have been provided to us.

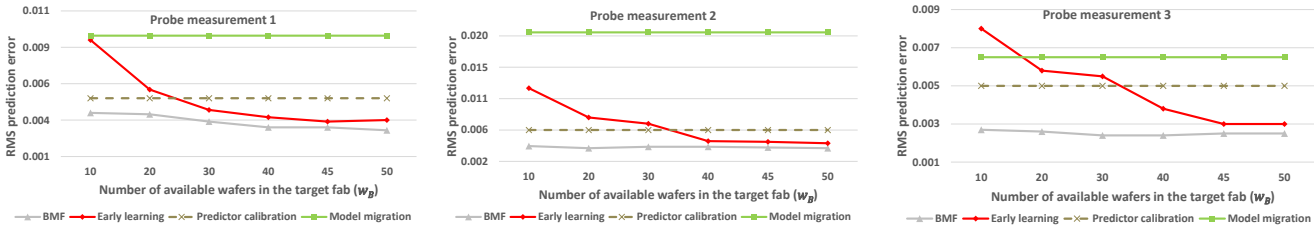


Fig. 3: RMS parametric yield prediction error during migration from fab A to fab B.

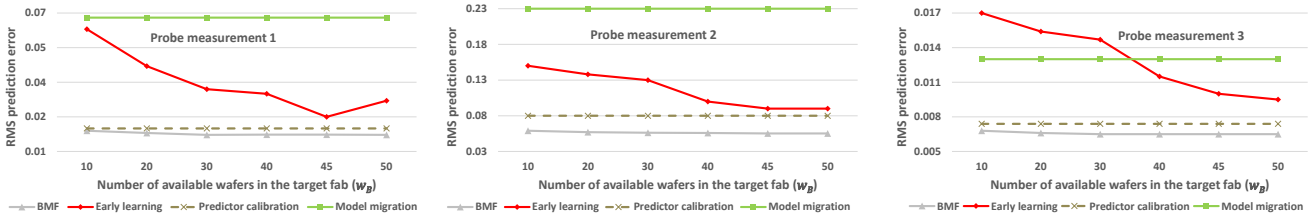


Fig. 4: RMS parametric yield prediction error during migration from fab B to fab A.

TABLE I: Prediction techniques

Learning method	Training set	Validation set
BMF	Intelligent mixture of data from fab A and early wafers from fab B: $\text{wafer}_A^i = [\boldsymbol{ET}_A^i, \boldsymbol{y}_A^i], i = 1, \cdots, w_A$ and $\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = 1, \cdots, w_B$	$\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = w_B + 1, \cdots, W_B$
Early learning	Early wafers from fab B: $\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = 1, \cdots, w_B$	$\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = w_B + 1, \cdots, W_B$
Model migration	Data from fab A: $\text{wafer}_A^i = [\boldsymbol{ET}_A^i, \boldsymbol{y}_A^i], i = 1, \cdots, w_A$	$\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = 1, \cdots, W_B$
Predictor calibration	Data from fab A and fab B: $\text{wafer}_A^i = [\boldsymbol{ET}_A^i, \boldsymbol{y}_A^i], i = 1, \cdots, w_A$ and $\text{wafer}_B^j = [\boldsymbol{ET}_B^j], j = 1, \cdots, W_B$	$\text{wafer}_B^j = [\boldsymbol{ET}_B^j, \boldsymbol{y}_B^j], j = 1, \cdots, W_B$

we use the root mean square (RMS) error computed on the validation set. Finally, we use bootstrapping in all learning procedures so as to (a) report a faithful prediction error, and (b) smoothen the prediction errors and ease the interpretation of the results, which is especially important for the runs where w_B is small. In each bootstrap iteration, we sample w_B wafers uniformly at random from the W_B wafers. In total, we perform 10 iterations.

C. Results

The accuracy of the four yield prediction methods summarized in Table I is demonstrated in Figure 3 for three different, randomly-chosen, probe-test measurements. The plots show the RMS prediction error on the validation set as a function of the training set size w_B for the BMF and early learning methods. We recall that w_B is a set of early wafers produced in fab B containing the device under consideration. The model migration and predictor calibration methods do not utilize any device-specific information from fab B for training purposes. They only rely on e-test measurements from produced wafers in fab B which, in theory, could contain any other device, although in our datasets they actually contain the device under consideration. Therefore, the corresponding curves for these two methods are flat and independent of w_B. As it can be seen from Figure 3, the model migration method shows the worst performance, which is expected since it naively uses the model that is learned on data from fab A for predicting the yield in fab B. The early learning method 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. In some cases, for very small training set sizes, it turns out that the early learning method presents an even worse performance, as compared to the model migration method. The predictor calibration method outperforms the model migration method and, in the case of small w_B, it also outperforms the early learning method, despite the fact that it does not use any device-specific information.

The BMF method outperforms all other methods regardless of the size of the training set. It shows a remarkably stable behavior, maintaining nearly constant prediction error even when the training set size is very small. This implies that the BMF method, by statistically extracting prior knowledge from fab A, is capable of generating accurate prediction models for fab B based only on few early wafers from fab B. Thus, the BMF learning procedure can be used to quickly estimate parametric yield from 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 [2]–[7].

Figure 4 shows the prediction error of the different methods when we reverse the roles of fab A and fab B, that is when the production of the device is migrated from fab B to fab A. Once again, we observe similar trends as in Figure 3. Next, we investigate whether the BMF method could perform equally

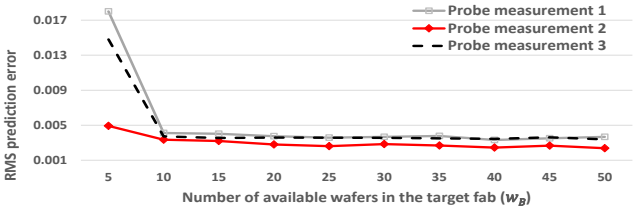


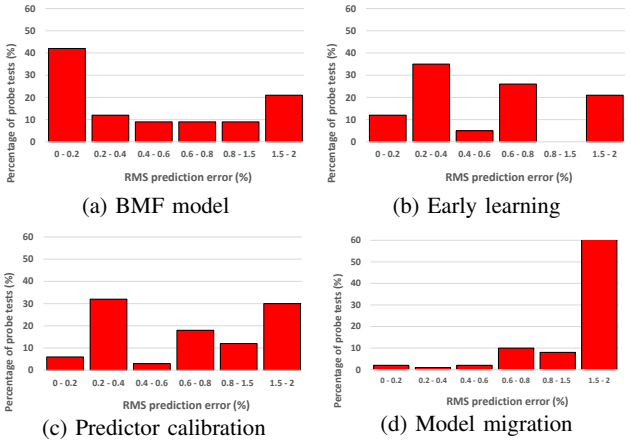
Fig. 5: RMS parametric yield prediction error of BMF technique during migration from fab A to fab B for different values for w_B.

for even smaller training sets than the ones shown in Figures 3 and 4. Figure 5 plots the prediction results for the same three probe-tests. As it can be seen, the prediction error increases drastically when the training set size is reduced from $w_B = 10$ to $w_B = 5$, showing that for $w_B = 10$ the BMF method reaches its limits. However, the prediction error for $w_B = 5$ remains significantly smaller than that of the other three methods and for devices that do not have a very high yield may even be considered acceptable. In any case, $w_B = 10$ is a very small number of wafers that should be quickly become available at the onset of production.

Finally, in Figure 6 we illustrate the cumulative comparative results for all 45 probe-tests for the fab A to fab B production migration scenario in the case where $w_B = 40$. For each prediction method we present a histogram where each bar shows the percentage of probe-tests that have a prediction error within a specific range. As it can be seen, the histogram of the BMF method has most of its weight more on the left side, i.e. towards smaller prediction errors, as compared to the histograms of the other three methods. The advantage of the BMF method becomes even more evident in Figure 7, which illustrates the same result in the case where $w_B = 10$.

V. Conclusion

We presented the use of the BMF learning technique in the context of predicting parametric yield while migrating production of a device from a source to a target fab. More specifically, we discussed how HVM e-test and probe-test data from the source fab can be intelligently combined with e-test and probe-test data from a very small number of early silicon wafers pro-

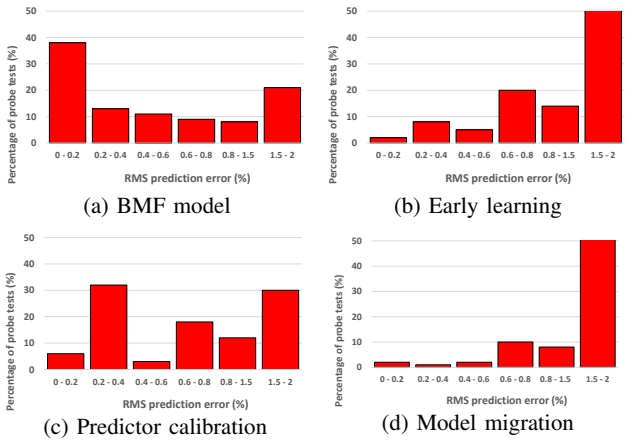


(a) BMF model (b) Early learning

(c) Predictor calibration (d) Model migration

Fig. 6: RMS parametric yield prediction error during migration from fab A to fab B across all 45 probe measurements when $w_B = 40$.

duced in the target fab, in order to develop accurate and robust



(a) BMF model (b) Early learning

(c) Predictor calibration (d) Model migration

Fig. 7: RMS parametric yield prediction error during migration from fab A to fab B across all 45 probe measurements when $w_B = 10$.

models for forecasting parametric yield when production of the device is ramped up. As we demonstrated using a large dataset from a 65nm Texas Instruments RF transceiver produced in two different fabs, the proposed approach outperforms earlier methods which forecast parametric yield using HVM e-test and probe-test data from the source fab along with HVM data from a different device in the target fab, or which rely only on the limited statistics available in the e-test and probe-test data of the few early silicon wafers. Indeed, information obtained from as few as 10 wafers in the target fab, suffices to reduce parametric yield prediction error to levels comparable to those achievable only when a large HVM population of wafers is available.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] S. Sun *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.

[6] J. Liaperdos, H. Stratigopoulos, L. Abdallah, Y. Tsiatouhas, A. Arapoyanni, and X. Li, "Fast deployment of alternate analog test using Bayesian model fusion," in *Proc. Design, Automation & Test in Europe Conference*, 2015, pp. 1030–1035.

[7] 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. Design, Automation & Test in Europe Conference*, 2015, pp. 1024–1029.

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

[9] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.