IC LASER TRIMMING SPEED-UP THROUGH

WAFER-LEVEL SPATIAL CORRELATION MODELING

by

Konstantinos Xanthopoulos

APPROVED BY SUPERVISORY COMMITTEE:

Yiorgos Makris, Chair

Mehrdad Nourani

Jeyavijayan Rajendran

IC LASER TRIMMING SPEED-UP THROUGH WAFER-LEVEL SPATIAL CORRELATION MODELING

by

KONSTANTINOS XANTHOPOULOS, BSc

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN

COMPUTER ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2019

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Yiorgos Makris for his guidance and support throughout my research. I would also like to thank Dr. Nourani and Dr. Rajendran, for serving as members of my thesis committee and for providing valuable feedback.

I gratefully acknowledge Dr. Ke Huang, Abbas Poonawala, and Amit Nahar for their essential input that led to the success of this study, as well as Texas Instruments Inc. for providing the data used in this study.

I am grateful to my mother, father, sister, and wife for their love, encouragement, and support.

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

November 2015

IC LASER TRIMMING SPEED-UP THROUGH

WAFER-LEVEL SPATIAL CORRELATION MODELING

Konstantinos Xanthopoulos, MSc
The University of Texas at Dallas, 2019

Supervising Professor: Yiorgos Makris, Chair

Laser trimming is used extensively to ensure accurate values of on-chip precision resistors in the presence of process variations. Such laser resistor trimming is slow and expensive, typically performed in a closed-loop, where the laser is iteratively fired and some circuit parameter (i.e. current) is monitored until a target condition is satisfied. Toward reducing this cost, we introduce a novel methodology for predicting the laser trim length, thereby eliminating the closed-loop control and speeding up the process. Predictions are obtained from wafer-level spatial correlation models, learned from a sparse sample of die on which traditional trimming is performed. Effectiveness is demonstrated on an actual wafer of laser-trimmed ICs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Ensuring performances of high-end integrated circuit (ICs) often relies on precision resistance values or resistance ratios. To this end, IC laser trimming has been extensively used for several decades as a means to controlling the impact of process variation on these sensitive resistors, which, in turn, assist in calibrating shifted electrical parameters of the IC. As reported in (Bloomstein, 1999), over 70% of the world's analog semiconductor companies use laser-based trimming and/or link-blowing in thin-film semiconductor and silicon manufacturing. Several on-chip resistors are placed in ICs to trim the parameters in question and trimming is performed by burning away small portions of these resistors using a laser trimming machine, in order to raise their resistance values until a target is reached. This laser trimming operation is usually conducted while the circuit is being tested by automatic test equipment (ATE), leading to appropriate final values for the resistors in the trimmed circuit. Figure 1.1 illustrates the objective of this trimming procedure, which essentially seeks to center the distribution of IC performances in order to enhance the manufacturing yield of a given design.

While this laser trimming procedure permits test engineers to efficiently center the distribution of IC performances and enhance manufacturing yield, it remains an expensive processing step which significantly increases the IC production costs (Ramirez-Angulo et al., 2011). Indeed, trimming is a lengthy procedure which is typically performed in a closed-loop, where the laser is iteratively fired and the IC parameter in question is monitored until a target condition is satisfied.

In this work[1], we introduce a methodology for reducing the trimming cost by predicting the laser trim length, thereby eliminating the closed-loop control and speeding up the trim-

[1]

Figure 1.1. Trimming procedure to center IC performances that shift due to process variations

ming process. In particular, we employ a spatial correlation modeling methodology based on Gaussian Process (GP), which has been successfully implemented in the context of analog/RF test cost reduction (Kupp et al., 2012b,a; Huang et al., 2013b,a). In this approach, instead of performing the closed-loop trimming procedure for every die on a wafer, we only trim a small sample of devices. The effective trim lengths of the sampled devices are then used to train spatial regression models, which are subsequently used to predict the required trim length for the remaining die on the wafer. The underlying conjecture is that the required trim length is spatially correlated across a wafer, therefore a sample is sufficient for us to accurately predict trim lengths at other die locations. The effectiveness of the proposed approach is demonstrated on an actual wafer of laser-trimmed ICs. Herein, we show that the proposed approach significantly reduces the time and, thereby, the cost of the trimming procedure.

CHAPTER 2

IC LASER TRIMMING

As the semiconductor industry continues scaling devices toward smaller process nodes, maintaining acceptable yield despite process variations has become increasingly challenging. Uncertainty is introduced by various sources during manufacturing and each step, such as lithography, ion implantation, thermal treatments, etc., can be considered as a source of variation. To handle these challenges, designers have traditionally resorted to conservative circuit design approaches, trading off some performance for higher yield and better variation tolerance.

Alternatively, post-silicon calibration methods, such as laser trimming, can be used to adjust various parameters of interest in an IC and to compensate parametric shifts caused by process variation. A number of laser trimmable resistors are often implemented on-die for this purpose and, depending on the application and the required tolerance, various cutting patterns and resistor geometries have been introduced. Accordingly, the IC laser resistor trimming procedure is comprised of three main components (Deluca, 2002): the device under test (DUT), the guided laser trimmer and a measurement tracking system.

During the laser trimming process, the guided laser beam fires pulses on the resistor causing the material on its surface to heat rapidly and vaporize. Consequently, the resistance value is increased after each laser pulse. The amount by which the resistance increases after a single pulse determines the accuracy and speed of the trimming process. Tracking of the desired parameter is achieved by a measurement system, as shown in Figure 2.1. The connection between the DUT and the measurement system is usually done by probes which are connected to the ends of the laser trimmable resistors. As shown in Figure 2.1, the trimmed resistance value is converted to a current value by a current sensor. The converted current value is monitored after each laser pulse and is compared with a user specified target value, as shown on the upper left side of Figure 2.1. The target current value is typically

set according to the specification limits of one or more particular performances of the DUT and is being fed to the comparator by a computer driven control system. When the target is reached, a stop signal is sent to the laser to halt the cutting, as shown on the upper right side of Figure 2.1. The total trimmed length l_{tot} on the resistor during the trimming procedure can then be expressed as:

$$l_{tot} = l_1 + l_2 + \cdots + l_k \tag{2.1}$$

where l_i denotes the trimmed length cut by the i-th laser pulse and k denotes the total number of laser pulses required to reach the target current value. Consequently, the time required for the laser control system to achieve the target current value is expressed as:

$$t_{tot} = \sum_{i=1}^{k} (t_{\text{cut i}} + t_{\text{set i}}) \tag{2.2}$$

where $t_{\text{cut i}}$ denotes the trimming time on the i-th laser pulse and $t_{\text{set i}}$ denotes the settling time that is needed in order for the current measurement that is fed to the comparator to stop rippling. As we can observe in equation (2.2), the total time of the laser trimming process depends on the time required to perform each laser pulse, the total number of laser pulses and the settling time needed to obtain an accurate measurement. In high-accuracy applications, this constraint hinders the laser trimming process as a large number of laser pulses is required to achieve the target, thus increasing the total penalty induced by settling time.

In order to speed up the trimming process, a two-resistor approach is commonly employed by designing trimmable resistors into pairs, as shown in Figure 2.1. The first resistor is trimmed to an initial target value with longer pulses and fewer intermediate measurements. This procedure is known as coarse-trim and although it is not as precise as the next step, it

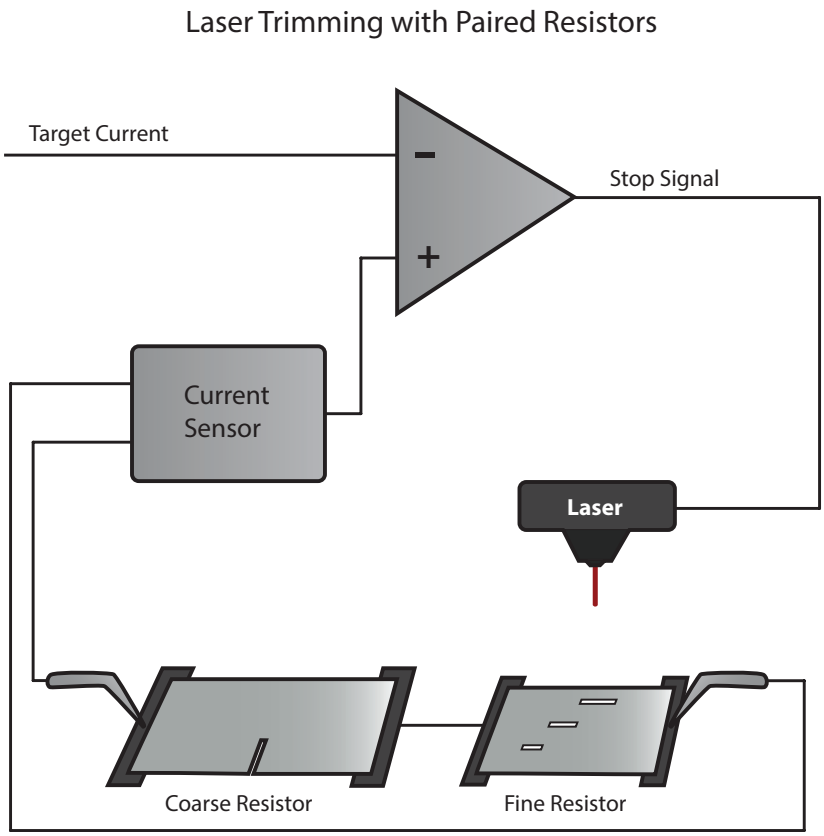

Figure 2.1. Measurement tracking system for laser control

is significantly faster. However, a conservative target is typically set in order to avoid over-trimming beyond the final target, which would result in yield loss. Following the coarse-trim, small cuts are performed on the second resistor until the final target value is reached. This is referred to as fine-trim and is much slower than its coarse counterpart. Figure 2.2 depicts the expected speedup between the single resistor and the paired-resistor laser trimming techniques, with Δt denoting the time saved by the paired-resistor approach.

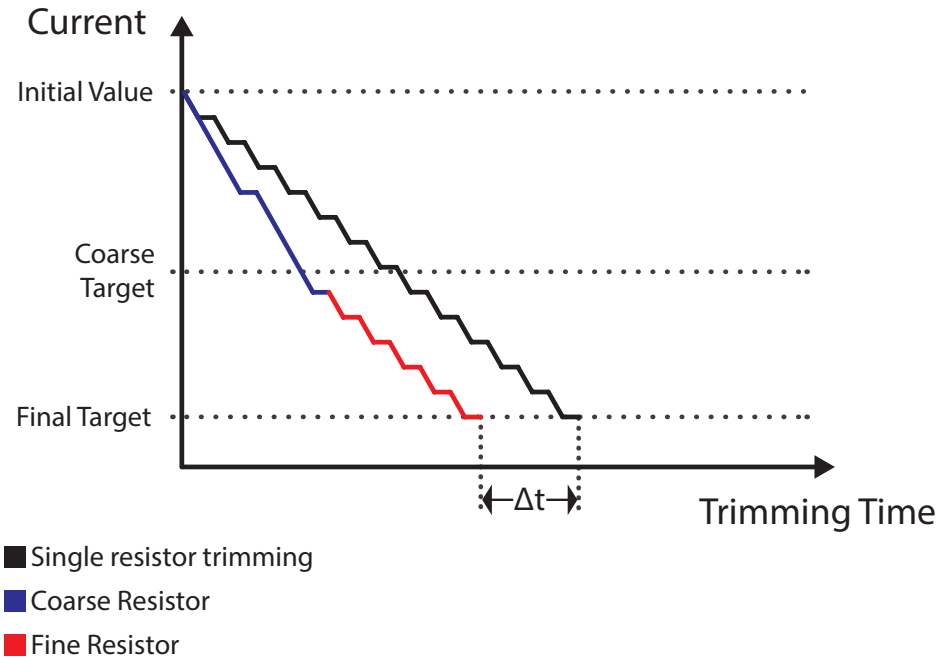

Figure 2.2. Single vs Paired resistor processes

CHAPTER 3

PROPOSED APPROACH

As noted in the previous section, tracking of the current value during trimming is an iterative and time-consuming process. The current value corresponding to the trimmed resistance is measured and compared to the target value after each laser pulse. The number of intermediate measurements delineates the time overhead, as portrayed in Figure 2.2. In this work, we propose a new approach to speed up the trimming procedure by employing a spatial correlation modeling methodology. In the proposed approach, instead of using the time-consuming tracking system to monitor the trimming process for every die location on the wafer, we use it only for a sparse subset of die samples. We, then, use the recorded data to build a model of the trim length as a function of the die coordinates on the wafer, using which we predict the required trim length for the remaining die. Finally, we instruct the laser to cut the resistors based on the predicted length, without taking any intermediate current measurements, i.e. without engaging the closed-loop control. Below, we introduce three methods for predicting the required trim length.

3.1 Length-based, original target prediction.

The first method uses directly the trim lengths of the sparse sample of die that are trimmed using the closed-loop approach. The conjecture here is that this physical parameter (i.e. the trim length) is spatially correlated across the wafer. Therefore, a model that predicts the required trim length as a function of the die coordinates can be trained. The underlying idea is shown in Figure 3.1, where the sampled lengths are used to train the statistical model which, in turn, is used to generate the predicted lengths. For this purpose, we use the Gaussian process spatial correlation model, which we explain in detail in Section 4. These lengths will be fed directly to the laser and will be trimmed in a single laser pulse. We refer to

the process of using pre-determined trim lengths as open-loop, as opposed to the closed-loop approach which iteratively trims and measures the current in multiple small increments. In Figure 3.2, we depict the expected time savings of our approach. The black line represents the classic closed-loop trimming method as the baseline. The blue line represents the length-based open-loop method. As may be observed, the open-loop completes the coarse-trim faster and, as a result, the complete trimming (i.e coarse plus fine) is expedited over the closed-loop approach. The saved time is denoted by Δt_1.

3.2 Rate-based, original target prediction.

Alternatively, for reasons that will become apparent in the next subsection, we can predict the required trim length indirectly by first building a GP spatial correlation model that predicts the trim rate of each die as a function of die coordinates. We define the trim rate as the ratio of the difference between the pre-trim and the post-coarse trim current measurements over the trimmed length. Let r_t denote the trim rate for a particular die location, then r_t can be expressed as:

$$r_t = \frac{m_{Pre} - m_{Post}}{L} \tag{3.1}$$

where m_{Pre} denotes the pre-trim current measurement, m_{Post} the post-coarse trim current measurement and L denotes the corresponding trimmed length. In this case, the sampling process depicted in Figure 3.1 would generate the predicted trim rate values $\hat{r_t}$ and pre-trim current measurements $\hat{m_{Pre}}$ for every non-sampled device, which in turn would be used to compute the coarse-trim length $\hat{L}$:

$$\hat{L} = \frac{\hat{m_{Pre}} - m_{target}}{\hat{r_t}} \tag{3.2}$$

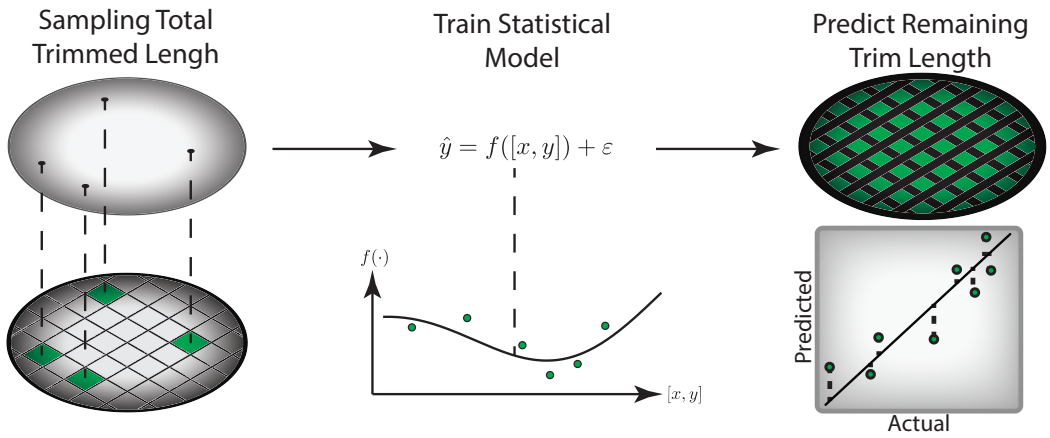

Figure 3.1. Proposed approach for trimming speed-up

where m_{target} denotes the target current which controls the closed-loop and terminates the coarse-trim stage in Figure 2.1.

Since in this method we are predicting the length needed to reach the coarse target, just as we did in Section 3.1, the time savings are the same as before. In other words, the blue line in Figure 3.2 also depicts the expected savings of the rate-based prediction.

3.3 Rate-based, optimized target prediction.

Predicting the required trim length based on the trim rate offers an additional advantage. Specifically, we can now set the target current to any value and predict the corresponding trim length, rather than being constrained by the original target of the closed-loop method. In fact, the long laser beam of the coarse-trim stage forces the test engineer to set a rather pessimistic target value in order to avoid over-trimming. As a result a longer cut, at the much slower fine-stage is required for the final target to be reached. In an open-loop configuration, however, we can be more aggressive at the coarse-trim stage and seek to get closer to the final target value, thereby further improve the overall savings.

Using the value of maximum expected error and the difference between the post-coarse trim target current and the final post-fine target current, we can then determine a new

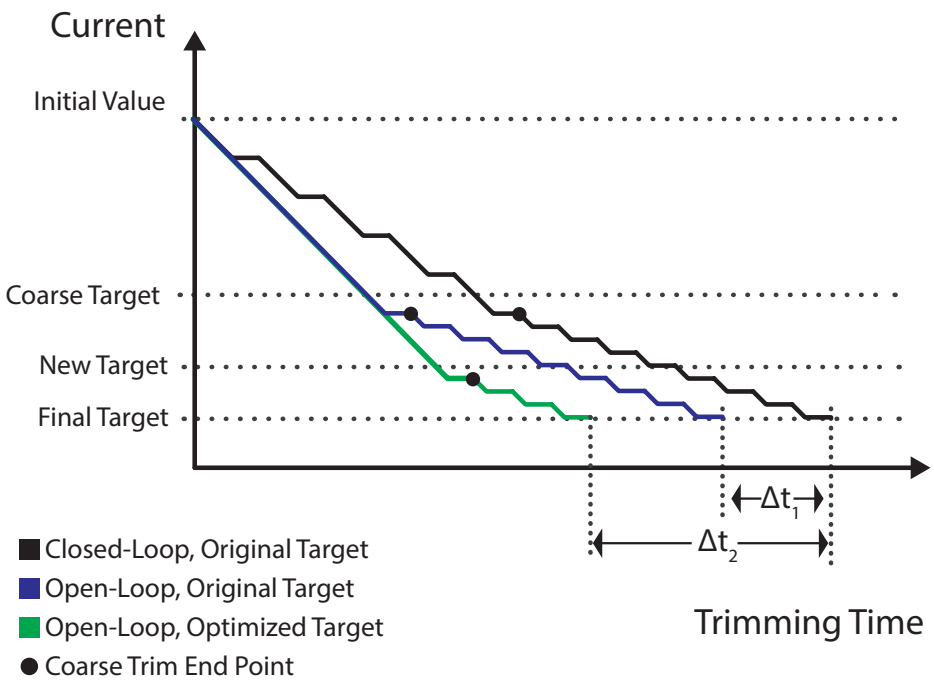

Figure 3.2. Current value vs. trim time in paired resistor trimming approach

safe post-coarse trim target value which is closer to the final post-fine trim target current. Then, we can replace the new post-coarse trim current target in equation 3.2 and get the corresponding coarse-trim length required, as shown below:

$$\hat{L} = \frac{\hat{m_{Pre}} - m_{\text{Optimized target}}}{\hat{r_t}} \tag{3.3}$$

where $m_{\text{Optimized target}}$ is the new target post-coarse trim current value and $\hat{L}$ is the predicted trim length required to reach this new target. Revisiting Figure 3.2, the green line depicts the process when the optimized target has been set closer to the final one. In this case, the open-loop coarse-trim is again faster than the closed-loop version, even though it cuts a longer length, because it requires a single laser pulse. Moreover, additional savings are obtained by the fact that the slow fine-trim stage has a much shorter distance to cover. The overall amount of time that is being saved is reflected by Δt_2.

3.4 Restriction for laser trimming deployment.

As aforementioned, the goal of laser trimming is to control the resistance value with high precision. Thereby, the length of the laser cuts have to be also accurate. As in every machine learning method, in the proposed approach there is a prediction error involved which can be either positive or negative. Without the existence of a secondary resistor (ie. fine-trim) both errors could force the device outside it's specification range. On the other hand, if a fine-trim resistor is used, negative errors are not a threat anymore and it is up to the test engineer to set the appropriate coarse-trim target so that the probability of over-trimming a resistor is low.

While the two-stage resistor configuration is currently being used in the industry, there are cases where a single-stage is preferred. As explained, the proposed method can only be applied to a two-stage setup, which leads to a tread-off between testing time and area overhead.

CHAPTER 4

WAFER-LEVEL SPATIAL CORRELATION MODELING

Recent research on modeling spatial measurement correlation has shown great promise in capturing wafer-level spatial variation and, thereby, reducing test cost of electrical measurements (Liu, 2007; Reda and Nassif, 2010; Zhang et al., 2011; Chang et al., 2011; Kupp et al., 2012b,a; Huang et al., 2013a). The underlying idea is to collect measurements for a sparse subset of die on each wafer and subsequently train statistical spatial models to predict performance outcomes at unobserved die locations. For example, in (Reda and Nassif, 2010), the expectation-maximization (EM) algorithm is used to estimate spatial wafer measurements, assuming that data comes from a multivariate normal distribution and the Box-Cox transformation is used in case data is not normally distributed. The "Virtual Probe" (VP) approach (Zhang et al., 2011) models spatial variation via a Discrete Cosine Transform (DCT) that projects spatial statistics into the frequency domain. The author of (Liu, 2007) laid the groundwork for applying Gaussian Process (GP) models to spatial interpolation of semiconductor data based on Generalized Least Square fitting and a structured correlation function. This fundamental model has been further enhanced using radial feature inclusion, multiple kernel evaluation and introduction of a regularization parameter (Kupp et al., 2012a,b), as well as a clustering approach to handle spatial discontinuous effects (Huang et al., 2013a). The resulting comprehensive GP model has dramatically improved both prediction accuracy and computational time, as compared to the VP model, and is therefore the one that we will use in this work.

4.0.1 Gaussian Process

In this chapter, we briefly articulate the theoretical underpinnings of Gaussian process model.[1] The fundamental concept underlying Gaussian processes is to model function outputs as drawn from a prior distribution with a fixed mean and a kernel-based covariance function. This approach works by extrapolating a function over a Gaussian random field on limited observations (Rasmussen and Williams, 2006). Consider a training set of n_t data points $\{m_1, \ldots, m_{n_t}\}$ located at the Cartesian coordinate denoted by $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_t t}\}$, $\mathbf{x} = [x, y]$. Using the GP approach, we define a Gaussian process as a collection of random variables $f(\mathbf{x}_i)$, $i = 1, \ldots, n_t$, for which any finite set of n_s function evaluations $f(\mathbf{x}_j)$, $j = 1, \ldots, n_s$, $n_s \leq n_t$ over the coordinates is jointly Gaussian-distributed. To derive a GP model for regression, we first consider a noise-free linear model:

$$f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w} \tag{4.1}$$

where $\phi(\mathbf{x})$ is a function of $\mathbf{x}$ mapping the input columns into some high-dimensional feature space, and $\mathbf{w}$ is the coefficient of the linear model which can be assigned a Bayesian prior such that $\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$. By assuming the random variables $f(\mathbf{x}_j)$ have zero mean, we can then specify the GP with mean and covariance functions:

$$\mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0, \tag{4.2}$$

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')] &= \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \phi(\mathbf{x}') \\ &= \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}') \end{aligned} \tag{4.3}$$

[1]Most of this chapter has been adapted from (Kupp et al., 2012b) and is included for the purpose of completeness. Interested readers are also referred to (Rasmussen and Williams, 2006) for further details on Gaussian process modeling.

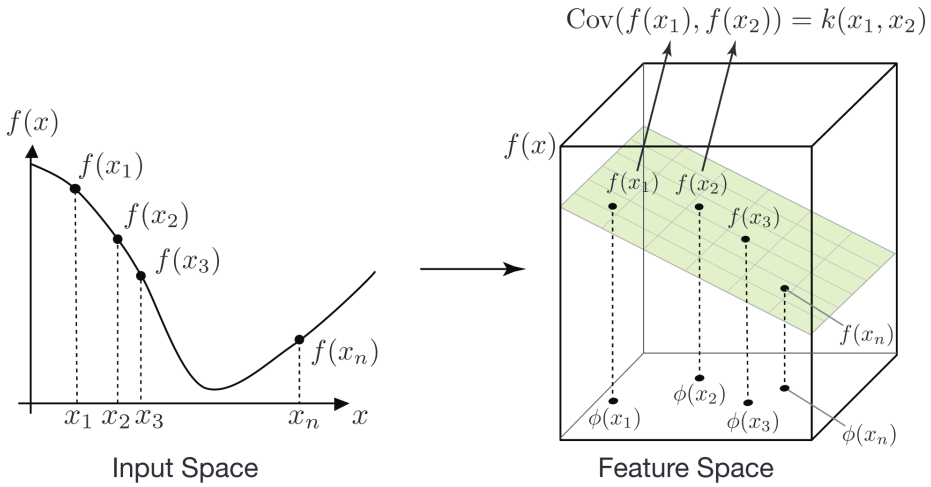

Figure 4.1. Overview of Gaussian Process Regression.

Figure 4.1 depicts exactly how a non linear input space shown at the left side can be mapped to a higher dimensional feature space, represented by the bottom plane at the right side of the same figure.

Recall that our ultimate goal of building a Gaussian process-based regression model is to somehow capture spatial variation in $f(\mathbf{x})$ as a function of the coordinates $\mathbf{x}$. The following discussion demonstrates how we can accomplish this task by modeling our data as drawn from a process with a covariance function that depends on spatial location. By taking this approach, proximal data points are modeled as being highly covariant, and distant points are modeled with low covariance. This codifies our intuition and *a priori* knowledge of the domain; we expect the variation of wafer-level measurement data to strongly correlate to spatial coordinates.

Consider the covariance function specified in equation (4.3). We can redefine the covariance matrices Σ_p as $(\Sigma_p^{1/2})^2$, and rewrite Equation 4.3 as:

$$\mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}') \tag{4.4}$$

$$= \phi(\mathbf{x})^\top (\Sigma_p^{1/2})^\top \Sigma_p^{1/2} \phi(\mathbf{x}') \tag{4.5}$$

We now introduce the parameter $\psi(\mathbf{x})$ by defining $\psi(\mathbf{x}) = \Sigma_p^{1/2}\phi(\mathbf{x})$, and subsequently rewrite the covariance of Equation 4.3 as:

$$\mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \phi(\mathbf{x})^\top(\Sigma_p^{1/2})^\top\Sigma_p^{1/2}\phi(\mathbf{x}') \tag{4.6}$$

$$= \langle\psi(\mathbf{x}), \psi(\mathbf{x}')\rangle \tag{4.7}$$

Crucially, this covariance function is formed as an inner product, permitting us to leverage the kernel trick(Aizerman et al., 1964) and express equation (4.7) as a kernel function $k(\mathbf{x}, \mathbf{x}')$. In other words, the covariance between any outputs can be written as a function of the inputs using a kernel function, without needing to explicitly computing $\phi(\mathbf{x})$ as shown in the right side of Figure 4.1. Many kernel functions exist, and any function $k(\cdot, \cdot)$ that satisfies Mercer's condition(Vapnik, 1995) is a valid kernel function. However, only a handful of kernels are commonly used. Among these common kernels, the most prevalent is the squared exponential, also known as the radial basis function kernel. In this work, we employed a squared exponential kernel of the form:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2l^2}|\mathbf{x} - \mathbf{x}'|^2\right) \tag{4.8}$$

where l is some characteristic length-scale of the squared exponential kernel. Employing this kernel is equivalent to training a linear regression model with an infinite-dimensional feature space. Substituting our squared-exponential covariance function into the definition of the Gaussian process, we arrive at a Gaussian process formulation as:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{4.9}$$

Once the covariance function is specified, for new input $\mathbf{x}_*$, we can readily predict $f_*(\mathbf{x}_*)$ by computing the conditional distributions of the joint Gaussian distribution.

$$f_*|X, \mathbf{t}, \mathbf{x}_* \sim \mathcal{N}(\mathbf{k}_*^\top K^{-1}\mathbf{t}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top K^{-1}\mathbf{k}_* \tag{4.10}$$

where X is a matrix denoting observed die locations, $\mathbf{t}$ denotes measurement values at observed die locations, $\mathbf{x}_*$ is a location we wish to extrapolate to, and where we have defined $K = K(X, X')$ as the matrix of the kernel function $k(\mathbf{x}, \mathbf{x}')$ evaluated at all pairs of observed training die locations. We have also defined $\mathbf{k}_* = K(X, \mathbf{x}_*)$ as the column vector of kernel evaluations between the unobserved target point and the entire set of observed training points, and lastly, $k(\mathbf{x}_*, \mathbf{x}_*)$ as the variance of the test function value at unobserved point $\mathbf{x}_*$.

In this work, we primarily concern ourselves with point predictions, and so we simply use the distribution mean $\bar{f}_* = \mathbf{k}_*^\top K^{-1}\mathbf{t}$ to generate a point prediction from the predictive distribution. This corresponds to decision-theoretic risk minimization (Vapnik, 1998) using a squared-loss function.

CHAPTER 5

EXPERIMENTAL RESULTS

We now evaluate the effectiveness of the proposed method on two dual-stage laser-trimmed precision resistors on a current transmitter device. The resistors have a top hat geometry (Deluca, 2002) and they are coarse-trimmed with a single plunge type cut (Deluca, 2002). Our dataset consists of one wafer with 1,924 devices, on which both coarse and fine laser trim has been performed for each of the two resistors of interest. Figure 5.1 shows the data that was collected during each stage of the process. Specifically, prior to trimming, the pre-trim current value is measured and logged. Coarse trim is then performed in a closed-loop, until the monitored current reaches the target value. At that point, the post-coarse trim current value, the coarse-trim length, and the coarse-trim time are logged. Subsequently, fine-trim is performed in a closed-loop, until the final current target is reached. The fine-trim time is then logged. In our analysis, current and length measurements are used for building the spatial correlation models, while time measurements are used for predicting the speed-up achieved by our method.

For the purpose of our experiment, we randomly sample 10% of the available die in order to build wafer-level spatial correlation models. Figures 5.2(2) and 5.3(2) show the sampled die locations for each resistor. We note that the same 10% sample is used for both resistors, emulating what would happen in manufacturing (i.e. the probe would collect all data from a limited sample of die). Using the spatial correlation models, we then predict the required coarse-trim length for the remaining 90% of the die on the wafer. As explained in Section 3, we predict this coarse-trim length in three ways.

5.1 Length-based, original target prediction.

The first prediction method is based on the actual length that was coarse trimmed for the 10% sample of resistors. The conjecture is that the coarse-trim length required is spatially

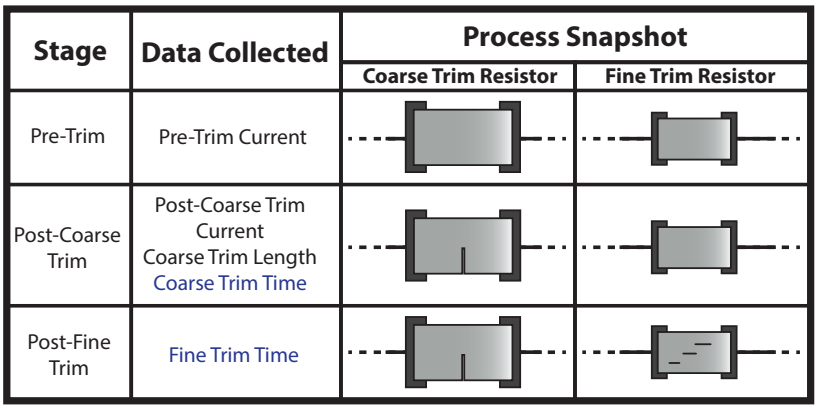

Figure 5.1. Per stage data collection of the laser trimming process

correlated. Hence, we can use this sample to build a GP wafer-level spatial correlation model that will predict the required length for the remaining 90% of the die on the wafer as a function of their die coordinate. Thereby, we can blindly perform coarse-trim in an open-loop by simply providing the length, which is faster than the closed-loop iterative laser firing and current monitoring method.

Figures 5.2(1) and 5.3(1) show the actual coarse-trim length across the entire wafer. As can be observed, spatial correlation indeed exists on the wafer, with a radial component being prominent. GP is very effective in modeling this spatial correlation based on the 10% sample. Indeed, this can be observed in Figures 5.2(3) and 5.3(3), where the predicted coarse-trim length wafer maps are shown for the two resistors. A visual comparison with the actual values confirms that the coarse-trim length can be accurately predicted by our method. For further visualization, the prediction error, which is the difference between the actual wafer maps (Figures 5.2(1) and 5.3(1)) and predicted wafer maps (Figures 5.2(3) and 5.3(3)), is also shown in Figures 5.2(5) and 5.3(5) for the two resistors, respectively. Evidently, the error variance is very low and balanced across the wafer. The maximum prediction error for Resistors A and B (corresponding to the yellow colored die on the error maps) is less than 3%, while the overall mean percent prediction error is 0.68% for Resistor A and 0.55% for Resistor B, as shown in the second column of Table 5.1.

5.2 Rate-based, original target prediction.

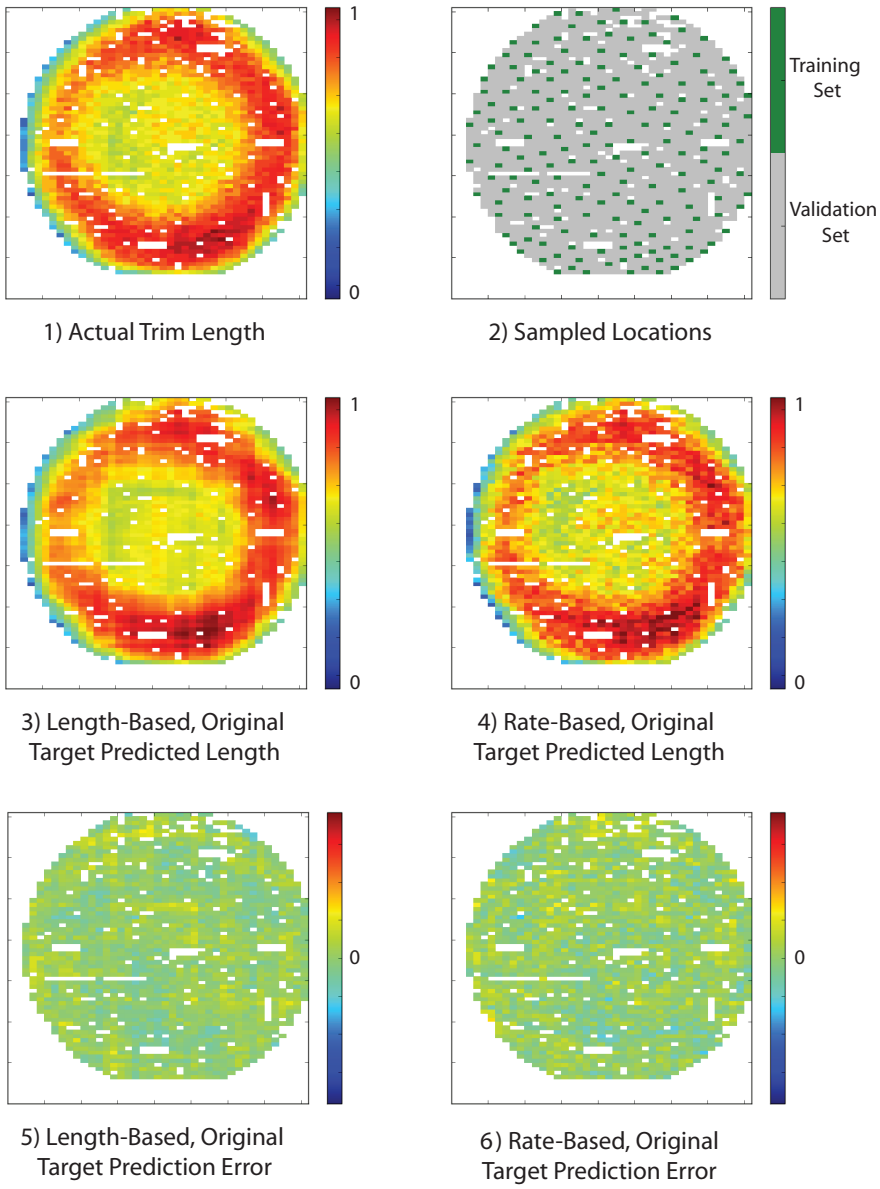

Figure 5.2. Resistor A: Actual and predicted Lengths for each proposed method

The second method also seeks to predict the coarse-trim length but in an indirect way. Specifically, instead of building a wafer-level spatial correlation model for the coarse-trim length itself, we build a spatial correlation model for the coarse-trim rate. The trim rate is

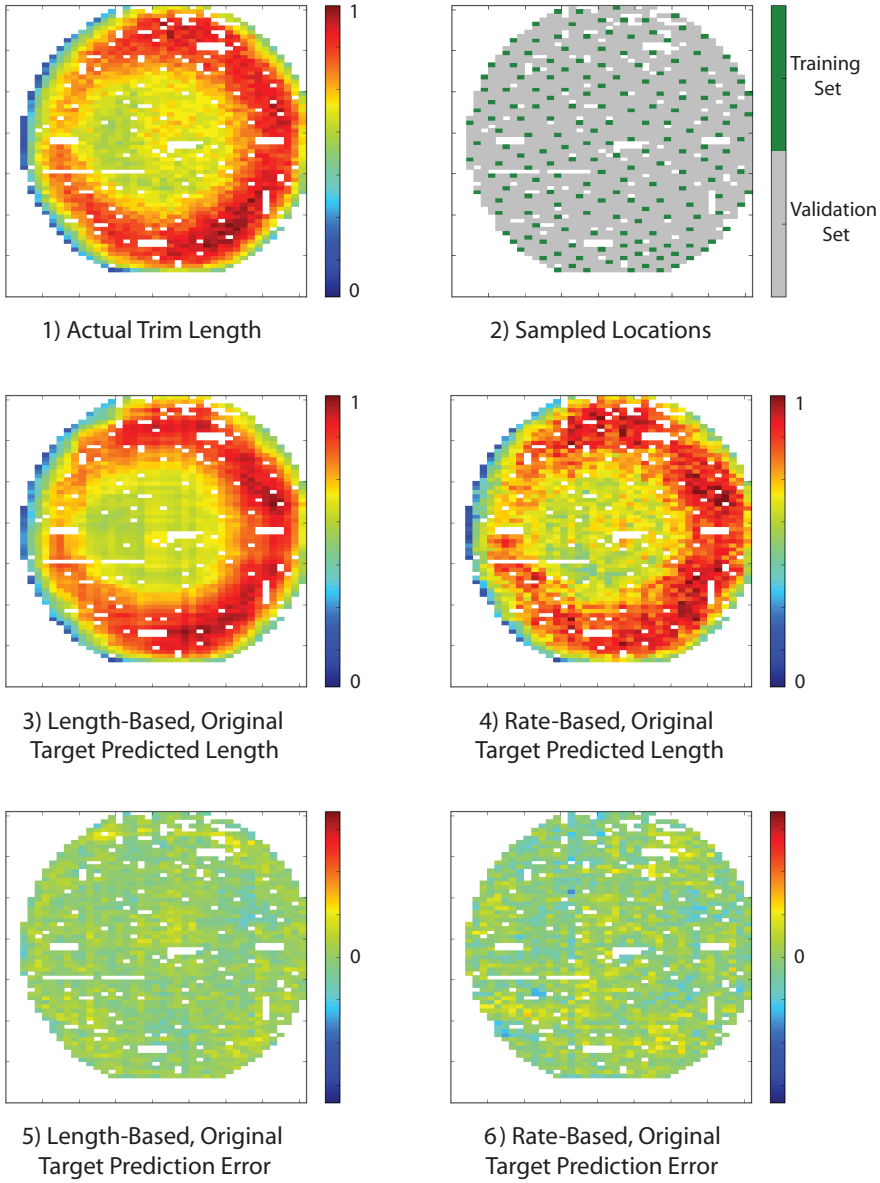

Figure 5.3. Resistor B: Actual and predicted Lengths for each proposed method

computed through Equation 3.1 using the collected pre-coarse trim current, post-coarse trim current, and coarse-trim length measurements on the 10% die sample. Then, using the trim rate predicted for each die as a function of its spatial coordinates, as well as the pre-coarse

trim current measurement and the targeted post-coarse trim current measurement[1], we can predict the required coarse-trim length for each of the remaining 90% die locations.

The predicted wafer maps for the two resistors are presented in Figures 5.2(4) 5.3(4), respectively, and corresponding error maps in Figures 5.2(6) and 5.3(6), respectively. As can be observed, the trim rate-based method is also very effective in correctly predicting the required coarse-trim lengths. The overall mean percent prediction error is 0.78% for Resistor A and 0.76% for Resistor B, as shown in the third column of Table 5.1. While this is slightly higher than the corresponding error of the length-based method, the maximum prediction error remains below 3%. This can also be verified by visual inspection of the Figures 5.2(6) and 5.3(6), where the highest colored value is similar to that of the length-based prediction error maps, shown in Figures 5.2(5) and 5.3(5).

Table 5.1. Overall Mean Percent Prediction Error

Resistor	**Length-based, Original Target**	**Rate-based, Original Target**
A	0.68%	0.78%
B	0.55%	0.76%

In both of the above methods, the error maybe either in the positive or in the negative direction. We note that if this error results in a shorter laser cut in the coarse-trim stage, the fine-trim stage will have to compensate with longer cuts (i.e. extra fine-trim time), while if the error results in a longer laser cut, the fine-trim will have to compensate with shorter cuts (i.e. less fine-trim time). In either case, the time savings from performing open-loop coarse-trim outweigh any additional fine-trim time. To evaluate the obtained speedup of open-loop over closed-loop coarse-trim, we measured actual trim times on 200 die locations on a new, untrimmed wafer, on which we performed open-loop coarse-trim.

[1]In order to accurately calculate the coarse-trim length prediction error of the rate-based method, the post-coarse trim current target value was set to the observed post-coarse trim current measurement for each die.

Accordingly, in the second and fourth column of Table 5.2 we report the expected average trim time per die for the closed-loop and open-loop case respectively, for each of the two resistors. As expected, there is a significant improvement in coarse-trim time, as a result of eliminating the need for the time-consuming closed-loop current monitoring. Note that the fine-trim time, reported in the third and fifth columns of this table, remains the same, as we are essentially trimming the same length in the coarse stage, whether in close-loop or in open-loop configuration. The overall (i.e. coarse and fine-trim) average time for the two configurations and the speedup are shown in the second through fourth columns of Table 5.3. As may be observed, the speedup is is 1.25 for Resistor A and 1.32 for Resistor B, indicating that significant savings can be obtained by applying the proposed method, even if our coarse-trim target of the open-loop option remains the same as the original target of the closed-loop configuration.

Table 5.2. Average Trimming Times Per Die

Resistor	**Closed-Loop**		**Open-Loop, Original Target**		**Open-Loop, Optimized Target**	
	Coarse	Fine	Coarse	Fine	Coarse	Fine
A	312 ms	281 ms	190 ms	281 ms	233 ms	66 ms
B	338 ms	231 ms	200 ms	231 ms	244 ms	57 ms

Table 5.3. Total Average Trimming Times And Speedup Per Die

Resistor	**Closed-Loop**	**Open-Loop Original Target**		**Open-Loop Optimized Target**	
	Total Time	Total Time	Speedup	Total Time	Speedup
A	593 ms	471 ms	1.25	299 ms	1.98
B	569 ms	431 ms	1.32	301 ms	1.89

5.3 Rate-based, optimized target prediction.

The third method takes advantage of the fact that the required coarse-trim length can be accurately predicted through the trim-rate estimated for each die location using a GP wafer-level spatial correlation model. Since the maximum prediction error remains very low and the range of pre-trim to post-fine trim current values is known, we can adaptively choose a new, optimized coarse-trim target. Thereby, we can use the fast, open-loop, coarse-trim option to get closer to the final target, leaving less work for the slower fine-trim stage. To this end, we can use equation (3.3) to predict the trim lengths for the new optimized post-coarse trim current target. In our analysis, we set this target such that 75% of the fine-trim effort could be replaced by a more aggressive coarse-trim stage[2]. We note that this optimized target leaves plenty of margin so that even the maximal prediction error observed in our experiments would not overshoot the final post-trim target, therefore no yield loss is incurred by this method.

The sixth and seventh columns of Table 5.2 show the extrapolated trim times for this rate-based optimized-target coarse-trim length prediction method, while the fifth and sixth columns of Table 5.3 show the anticipated speedup over the baseline closed-loop method, which is estimated at 1.98 for Resistor A and 1.89 for Resistor B, respectively. Based on these promising results, we plan to deploy and evaluate the rate-based optimized target method on a larger-scale experiment as our future work.

[2]An NDA under which this data was provided to us prevents us from disclosing the actual current values.

CHAPTER 6

CONCLUSION

The key conjecture corroborated by the research described in this paper is that the physical parameters which are typically calibrated through IC laser trimming are spatially correlated. Therefore, wafer-level spatial correlation modeling methods, which have previously been introduced and leveraged for electrical test cost reduction, may also be used to reduce the cost of the expensive and time-consuming IC laser-trimming process. Experimental results with two laser-trimmed precision resistors on a wafer of $\sim$2K devices indicate that almost half of the time needed for laser trimming can be eliminated without impacting yield, thereby offering substantial savings in high volume production of such devices.

REFERENCES

Aizerman, M., E. Braverman, and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*.

Bloomstein, W. W. (1999). Marketwatch: Laser trimming meets IC manufacturing demands. *Laser Focus World (LFW)*.

Chang, H.-M., K.-T. Cheng, W. Zhang, X. Li, and K. Butler (2011). Test cost reduction through performance prediction using virtual probe. In *IEEE International Test Conference*.

Deluca, P. (2002). A review of thirty-five years of laser trimming with a look to the future. *Proceedings of the IEEE*.

Huang, K., N. Kupp, J. Carulli, and Y. Makris (2013a). Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests. In *Design, Automation & Test in Europe Conference (DATE)*.

Huang, K., N. Kupp, J. Carulli, and Y. Makris (2013b). On combining alternate test with spatial correlation modeling in analog/RF ICs. In *IEEE European Test Symposium (ETS)*.

Kupp, N., K. Huang, J. Carulli, and Y. Makris (2012a). Spatial correlation modeling for probe test cost reduction in RF devices. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*.

Kupp, N., K. Huang, J. Carulli, and Y. Makris (2012b). Spatial estimation of wafer measurement parameters using gaussian process models. In *IEEE International Test Conference (ITC)*.

Liu, F. (2007). A general framework for spatial correlation modeling in VLSI design. In *Design Automation Conference*.

Ramirez-Angulo, J., R. Geiger, and E. Sanchez-Sinencio (2011). Characterization, evaluation, and comparison of laser-trimmed film resistors. *IEEE Journal of Solid-State Circuits*.

Rasmussen, C. and C. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Reda, S. and S. R. Nassif (2010). Accurate spatial estimation and decomposition techniques for variability characterization. *IEEE Transactions on Semiconductor Manufacturing*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory.* Adaptive and learning systems for signal processing, communications, and control.

Zhang, W., X. Li, F. Liu, E. Acar, R. Rutenbar, and R. Blanton (2011). Virtual probe: A statistical framework for low-cost silicon characterization of nanoscale integrated circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

BIOGRAPHICAL SKETCH

Konstantinos Xanthopoulos graduated with a BSc in Computer Science from the University of Piraeus in 2012. In the Spring of 2013 he started his master's degree in Computer Engineering at The University of Texas at Dallas. At the same time, he joined the Trusted and RELiable Architectures (TRELA) lab under the supervision of Dr. Yiorgos Makris, to pursue his doctorate degree with focus on machine learning applications and statistical analysis on semiconductor manufacturing data.

During the summer semesters of 2013 and 2014, Konstantinos worked at Texas Instruments Inc. with the goal to develop and deploy adaptive testing solutions for test cost reduction. He is a student member of the IEEE.

CURRICULUM VITAE

Constantinos (Konstantinos) Xanthopoulos

November 18th, 2015

Contact Information:

Department of Computer Engineering
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.
Email: Konstantinos.xanthopoulos@utdallas.edu

Educational History:

BSc, Computer Science, University of Piraeus

Employment History:

Product Engineer (Co-Op), Texas Instruments Inc., Summer 2014
Product Engineer (Co-Op), Texas Instruments Inc., Summer 2013