Hybrid Modeling of Non-Stationary Process Variations

Eva Dyer, Mehrdad Majzoobi, Farinaz Koushanfar
ECE Department, Rice University, Houston, Texas
{ e.dyer, mm7, fk1} @rice.edu

ABSTRACT

Accurate characterization of spatial variation is essential for statistical performance analysis and modeling, post-silicon tuning, and yield analysis. Existing approaches for spatial modeling either assume that: (i) non-stationarities arise due to a smoothly varying trend component or that (ii) the process is stationary within regions associated with a predefined grid. While such assumptions may hold when profiling certain classes of variations, a number of recent modeling studies suggest that non-stationarities arise from both shifts in the process mean as well as fluctuations in the variance of the process. In order to provide a compact model for non-stationary process variations, we introduce a new *hybrid spatial modeling* framework that models the spatially varying random field as a union of non-overlapping rectangular regions where the process is assumed to be locally-stationary within each region. To estimate the parameters in our hybrid spatial model, we develop a host of techniques to both estimate the change-points in the random field and to find an appropriate partitioning of the chip into disjoint regions where the field is locally-stationary. We verify our models and results on measurements collected from 65nm FPGAs.

Categories and Subject Descriptors

J.6 [**Computer-Aided Engineering**]: Computer-aided Design (CAD)

General Terms

Algorithms, Measurement

Keywords

Spatial Correlation, Process Variation Modeling, Non-stationary Variation

1. INTRODUCTION

Over the past few decades, scaling of CMOS to nanometer feature sizes has been the main driving force behind improvements in the performance and functionality of integrated circuits. Despite the obvious advantages of miniaturization of ICs, nanoscale devices exhibit a considerable amount of process variations, causing the device features and performance metrics including power and timing to deviate from their nominal values. Accurate modeling of variations is important for several reasons, including design-time analysis, manufacturing yield maximization, and also for post-silicon optimization. The latter is especially important for reconfigurable devices where variation-aware post-silicon tuning methods have shown to be effective [21].

Recent years have witnessed an explosion of statistical models for CMOS process/performance variations [1, 3, 4, 5, 8, 10, 13, 14, 15, 19, 23]. Whereas early methods characterized process fluctuations as sum of inter-die (global) and independent intra-die (random) variations [4], later models extended this work by assuming the intra-die variation to be spatially correlated [1, 5]. Studies of spatially correlated process variations on a single die can be placed into one of two categories. In grid-based approaches [1, 5, 6], one assumes that the process variations are stationary within regions defined by a pre-specified grid, e.g., uniform, non-uniform, or a quadtree decomposition. The other class of methods assume that the process variations can be modeled as a stationary zero mean random field with an additive baseline component that accounts for any shifts in the mean of the process. In this case, non-stationarities are assumed to arise solely from smooth variations in the baseline component across the extent of the die.

While such assumptions may hold when profiling certain classes of variations, recent studies suggest that spatial variability across the extent of the wafer and within a single die are likely to be highly non-stationary [7]. Furthermore, transitive shifts in the mean of the process variations have been observed in contact plug resistivity [2] and in timing variability on FPGAs [16], invalidating the assumption that non-stationarities arise from a smoothly varying baseline component.

In order to provide a compact spatial model for non-stationary process variations that exhibit such structure, we introduce a new modeling framework that aims to bridge the gap between previous efforts that model process variations as either a continuous random field or as being stationary within a pre-specified grid. In particular, we propose a *hybrid spatial model* for non-stationary process variations that provides a succinct representation of the spatially varying random field by modeling it as a union of non-overlapping regions wherein the process can be assumed to be locally-stationary.

To estimate the parameters in our hybrid spatial model, we begin by developing a set of techniques for *change-point detection* or detecting the points at which the statistics of the random field change. In contrast to a generic change-point detection problem, our aim is to determine an efficient method for change-point detection that exploits the known structure of process variations on CMOS. In particular, we find that there are many settings where layout and

DAC '11 San Diego, California USA
Copyright 2011 ACM 978-1-4503-0636-2/11/06 ...$10.00.

mask-dependent variations will induce a natural partitioning of the space into disjoint regions. For instance, when the spacing between logic blocks exceeds a particular threshold, the process variations on either side of this gap can be treated as independent random fields. This suggests that in many cases, a step-wise transition in the mean of the process may be used to specify regions where the underlying process variations will exhibit altered statistics.

In order to test this assumption, we begin by introducing a method for detecting change-points in the process mean. We do this by solving a total variation (TV) minimization problem which approximates a set of noisy observations of the random field with a signal containing the minimum number of change-points needed to effectively explain the observations within an appropriate range of fidelity. After obtaining an estimate of the change-points in the process mean, we employ a k-means clustering procedure to partition the chip into a number of non-overlapping regions based upon the assumption that shifts in the mean are correlated with shifts in the statistics of the process. Finally, we present methods for obtaining quantitative measure of stationarity within a random field.

To demonstrate the power of our proposed modeling framework, we study process variations resulting from timing variability across a collection of 65nm FPGAs. We find that detecting transitions in the process mean is sufficient to provide a proxy for detecting change-points in the underlying random field.

Our explicit contributions are as follows: (i) the introduction of a hybrid modeling framework that weds previous continuous and grid-based approaches, (ii) the development of a variational method for detecting sharp transitions in the process mean, (iii) methods for partitioning the chip into non-overlapping regions, and (iv) the development of a quantitative measure of stationarity in a random field.

2. BACKGROUND

Process variations arise from imperfections that lead to both random and spatially uncorrelated variations on wafer as well as systematic and correlated variations. The uncorrelated random variations are typically caused by the fundamental intrinsic atomic-scale randomness of the devices and materials. Systematic correlated components comprise the identifiable portion of the variations that arise from unintentional shifts in processing conditions such as mask errors, lithographic off-axis focusing, and reticle stepper alignment errors. Line edge roughness, variations in channel length and width, variations in gate oxide thickness, energy level quantization belong to this category [17].

There has been a large body of research on modeling process variations. Existing spatial models for process variations can be categorized as either continuous or discrete models. In discrete or grid-based models, the die area is divided into square regions over which the variation is assumed to be constant. Principal component analysis [5], Quad-tree [1], and grid based methods [6] are among the discrete modeling approaches. The work in [6] introduces a grid-based approach where each grid contains a few devices (gates). The parameter variation of devices within the same grids are completely correlated, while those in adjacent grids are highly correlated and those in non-neighboring grids are uncorrelated.

In contrast, continuous models treat the entire die as a continuous random field.These models assume that all non-stationarities in the random field are due to a drift or baseline component that smoothly varies across the extent of the die. This suggests that once the mean or the baseline component is removed, the resulting process is stationary. In [14], the author models the systematic variations as the sum of a first order plane and a spatially correlated 2D gaussian process, and a truly random component. The first order plane accounts for the long range trend component of systematic variation. Any additional trends in the data are later removed with a median polishing algorithm. A Generalized Least Square fitting framework is then used to fit all of the parameters in the model. In a similar attempt in [20], the correlated within-die variation is represented by a two dimensional quadratic polynomial surface, and the model is applied to the timing data collected from an array of ring oscillators on FPGA. The authors in [11] introduce a method for extracting spatial correlation by investigating which spatial correlation functions result in a positive semidefinite correlation matrix. They assume the systematic variation can be predicted with the full knowledge of process steps and instead focus on modeling the random correlated portion of the variation.

In a recent study on measured data from 45nm wafers, the authors demonstrate that different locations on the chip may have very different means and variance, and such variations are more apparent with increasing the chip size [7]. Furthermore, they demonstrate that the correlated variation is mainly due to across-wafer variation and across-field variations on the scale of a single die. To further support claims that process variations exhibit non-stationary behavior, it has been observed that components that lie within the center of the chip exhibit lower variance than components at chip boundaries due to variability in process control.

3. HYBRID MODELS FOR NON-STATIONARY RANDOM FIELDS

In this section, we introduce a novel framework for modeling non-stationary random fields. Let us begin by modeling a continuous random field $Z \in \mathbb{R}^2$ as consisting of two additive components, $Z = A + N$, where A is an additive component containing systematic variations in the mean of the process and N is a spatially correlated zero-mean random field that could exhibit non-stationary behavior. From this point forward, we will refer to A as the *baseline*, *trend*, or *drift* component.

Whereas previous approaches have assumed that after estimation and removal of the baseline component A, the resulting random field is stationary, we assume that the non-stationary behavior of process variations can be attributed to both: (i) systematic shifts in the mean of the process and/or (ii) non-stationarities in the spatially correlated random field. To provide a compact model to account for both of these sources of non-stationarity, we model the continuous random field as a union of non-overlapping regions where the process is locally-stationary within each of the regions. In contrast to grid-based approaches that assume the process is stationary within a square (isotropic) region in the grid, here, we will assume that the chip can be divided into a number of anisotropic regions (with edges at either 0° or 90°). We will refer to this type of model as a *hybrid spatial model*.

In order to model the continuous random field Z, we first collect a set of samples from the field at discrete points in space. For a particular point in space with coordinates (x_i, y_i), we obtain a single measurement $z_i = Z(x_i, y_i) \in \mathbb{R}$. If we sample the field on a grid of $n_1 \times n_2$ points in space, this collection of measurements can either be written as a matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ or we can reorder the entries and stack them into a vector, which we will denote as $\mathbf{z} \in \mathbb{R}^n$, where $n = n_1 n_2$. We will use this vector and matrix notation interchangeably.

To estimate the parameters that specify our hybrid model, we must first estimate (i) the minimum number of regions required to provide an accurate characterization of the field, (ii) the boundaries for all k regions in our model $\{\mathcal{R}_i\}_{i=1}^k$, and (iii) the mean and autocorrelation of the random field within each of the specified regions. Upon estimating these parameters our final hybrid spatial model can be expressed as, $\mathcal{H}(Z) = \{Y_i, \mathcal{R}_i\}_{i=1}^k$, where each local process is defined according to its mean and autocorrelation, $Y_i = (\mu_i, \Sigma_i)$, and μ_i and Σ_i are the mean and autocorrelation for the local process Y_i defined over the region $\mathcal{R}_i$. Due to our assumption that all of the regions are non-overlapping, $\mathcal{R}_i \cap \mathcal{R}_j = \{\emptyset\}$, $\forall i \neq j$.

4. VARIATIONAL METHODS FOR DRIFT ESTIMATION

In the previous section, we introduced our hybrid model for non-stationary process variations. The first step in specifying this model is to detect sharp transitions in the process mean that could be due to layout-dependent effects as well as variability introduced by the mask structure. In this section, we will introduce a method for estimating these change points in the mean by assuming that the baseline of our variations exhibit 2D piecewise constant structure.

To extract the 2D baseline component from a set of measurements $\mathbf{Z}$, we propose the use of regularized TV norm minimization. This method has been used in a number of settings where piecewise smooth signal components must be separated from noise [18] and also in geometric separation tasks where natural image content must be extracted from images consisting of a combination of image content and periodic texture [12]. More recently, TV minimization has been applied to the recovery of images from compressive measurements [22]. In all of these settings, knowledge that the signal of interest has a sparse gradient field can be leveraged to recover an approximation to the signal from incomplete or noisy measurements.

To motivate the use of total variation for drift estimation, we point the reader to Figure 1 where we show a piecewise constant signal in black and the same signal with additive noise in red. Above this plot, we also show the result of a differencing operation over each of these signals, where the difference signal $\mathbf{g}(\mathbf{y})$ can be written as:

$$\mathbf{g}(\mathbf{y}) = \mathbf{y}(i+1) - \mathbf{y}(i), \quad \forall i = \{1, \ldots, n-1\}, \qquad (1)$$

for a vector $\mathbf{y} \in \mathbb{R}^n$. We see that the non-zero elements of the difference signal $\mathbf{g}(\mathbf{y})$ are sparse, with non-zeros only occurring at the step-like transition points. Thus, we say that the original piecewise constant signal $\mathbf{x}$ exhibits a 2-sparse difference vector $\mathbf{g}(\mathbf{x})$, where the non-zero entries correspond to the locations where transitions occur. When no noise is present, finding the change points from the difference signal is easy. However, when noise is introduced into the piecewise constant signal, the resulting difference vector is no longer 2-sparse but has two large non-zero components and a number of smaller non-zero components.

In order to find an approximation to a noisy signal $\mathbf{y}$ that also exhibits a sparse difference vector (small total variation), we aim to tradeoff the error in our final representation with the sparsity of the difference vector for our estimate. Finding the sparsest approximation of our signal (as measured by the ℓ_0-norm or number of non-zero coefficients in the signal) in NP hard, however, we may replace the ℓ_0 penalty with an ℓ_1 norm, resulting in a convex problem that has a unique and global minimizer.

By adding a sparsifying penalty to our objective function, we aim to minimize a combination of the ℓ_2 error and the total variation of the signal,

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{g}(\mathbf{x})\|_1, \qquad (2)$$

Here, a standard least-squares term is added to a scaled ℓ_1 penalty on the absolute differences between successive coefficients, where λ is known as the *regularization parameter*.

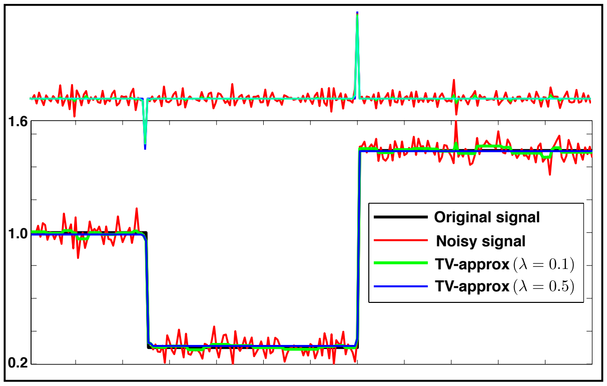


Figure 1: 1D piecewise-constant signal (black) with additive Gaussian noise (red). The approximations obtained by TV-minimization for different choices of regularization parameter λ, where $\lambda = 0.1, 0.5$ for the green and blue curves respectively. On the top of the figure, we show the resulting gradient obtained from each the signals in the bottom figure, where it is easy to see that for piecewise constant signals, the gradient vector is extremely sparse.

More explicitly, the ℓ_2 term measures the total error incurred by approximating the signal $\mathbf{y} \in \mathbb{R}^n$ by $\mathbf{x}$ and the term $\|\mathbf{g}(\mathbf{x})\|_1$ encourages solutions that exhibit sparse absolute differences, where $\|\mathbf{g}(\mathbf{x})\|_1 = \sum_{i=1}^{n-1} |\mathbf{x}(i+1) - \mathbf{x}(i)|$ is referred to as the total variation (TV) norm of $\mathbf{x}$ because it provides a measure of the total variation within a particular signal. One great advantage of the optimization above is that it has a convex form for a particular value of the regularization parameter λ, where λ can be chosen by cross-validation or by employing any number of information criterion for model selection. We note that in many cases, choosing this parameter can be challenging; however, in practice, the sparse gradient fields recovered by this algorithm tend to be quite robust across different choices of λ.

Upon solving problem 2 for the noisy signal in Figure 1 for different values of λ, we obtain approximations to the noisy signal that have small TV norm. These approximations are shown in blue and green for higher and lower regularization parameters respectively. We find that for sufficiently large λ we obtain an approximation of the underlying piecewise constant signal that corresponds very closely to the original signal.

Now that we have motivated the use of TV minimization for recovering piecewise-constant signals, we now extend the TV minimization problem defined for 1D signals to the two-dimensional setting. In this case, we can define the anisotropic TV norm of an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ as:

$$TV(\mathbf{X}) = \sum_{j=1}^{n_2-1} \sum_{i=1}^{n_1-1} |\mathbf{X}_{i,j} - \mathbf{X}_{i+1,j}| + |\mathbf{X}_{i,j} - \mathbf{X}_{i,j+1}|, \quad (3)$$

where $\mathbf{X}_{ij}$ is the $(i,j)^{th}$ entry of the matrix $\mathbf{X}$.

In words, the anisotropic total variation of a 2D function measures the amount of absolute variation between the intensity at single point in space and points directly below and to the right. By treating the variation in each direction independently, we obtain a measure of the anisotropic total variation of the images. We note that due to the rectangular packing of chips, the anisotropic total variation produces a more accurate estimate of the baseline signal, however, there is an equivalent definition for the isotropic TV in 2D

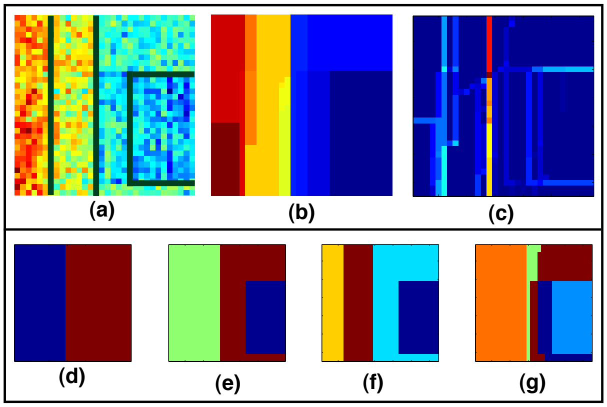


Figure 2: Optimal Grid Detection: (a) Raw FPGA data with optimal grid overlaid in black, (b) TV signature for $\lambda = 0.1$, (c) Absolute differences in x and y (showing transitions in TV signature), and in (d), (e), (f), (g) we show the partitions obtained with k-means clustering for $k = \{2, 3, 4, 5\}$.

that can be used to recover smoothly varying trend components.

To find an approximation to a piecewise constant 2D signal $\mathbf{A}$ from a collection of observations $\mathbf{Z} = \mathbf{N} + \mathbf{A}$, we can regularize the growth of the TV norm—to encourage the selection of signals with sparse gradient fields—while trading off the total ℓ_2 error in the approximation by varying λ. To do this, the following convex optimization problem can be solved for a particular value of the regularization parameter λ:

$$\widehat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{z} - \mathbf{a}\|_2^2 + \lambda\, TV(\mathbf{A}), \tag{4}$$

where the ℓ_2 term measures the total point-wise error incurred by approximating the measured signal $\mathbf{z} \in \mathbb{R}^n$ by $\widehat{\mathbf{a}}$ and the TV term regularizes (penalizes) growth in the anisotropic TV of the estimated 2D baseline component $\widehat{\mathbf{A}}$.

5. METHODS FOR GRID DETECTION

In the previous section, we demonstrated that TV minimization can be used to obtain robust estimates of the change-points in the process mean. The next step is to use the baseline signature extracted by our TV procedure to provide a partitioning of the chip into non-overlapping regions. To do this, we employ a standard clustering procedure called k-means. Following this, we introduce a method for obtaining a quantitative measure of stationarity within each region obtained from the clustering procedure.

To determine a partitioning of the chip, we cluster the TV signature for different values of k, where k corresponds to the number of regions the chip is divided into. An upper limit to the number of regions can be set by simply rejecting a clustering when the number of pixels per region is very small because in this case, the autocorrelation function can not be computed accurately. For a reasonable range of number of regions, for $k = 1, \ldots, k_{max}$ we can determine how good the partitioning is by measuring the stationarity within each region. In Figure 2, we show raw data collected from a FPGA, its TV signature, and the clusterings obtained for $k = \{2, 3, 4, 5\}$. After obtaining these possible partitions for different values of k, we select the partition that produces the highest measure of stationarity across all blocks. Now, we will describe a method for computing the stationarity of the field a particular region.

Although tests for stationarity are very common in 1D time-series data, deriving an optimal test for the stationarity of a random field is more challenging. We propose an approach that leverages the fact that for a stationary process, the autocorrelation function should remain constant over shifts in the origin.

To compute the 2D spatial autocorrelation function, first the 2D power spectral density (PSD) must be computed. If we assume that the field is stationary over the entire region, then to obtain an unbiased estimate of the PSD we employ the Welch-Bartlett procedure. The procedure is as follows: (1) the region of interest is either divided into overlapping blocks of equal size, (2) the PSD is computed for each window, and then (3) the local estimates are averaged according to the number of sub-blocks that are selected. After averaging the PSD estimates, the inverse FFT is taken to produce the final autocorrelation estimate.

Here, we wish to use the intuition behind this procedure to obtain a quantitative measure of stationarity. In particular, for non-stationary random fields, local PSD estimates obtained in each sub-block will vary as we sweep across the region of interest. Thus, we can study the diversity of autocorrelation estimates obtained at different sub-blocks to understand the degree of stationarity within the region. To do this, we compute local PSD estimates by sweeping a window over the region of interest as in the Welch-Bartlett procedure. However, instead of averaging each of these estimates and then taking the inverse FFT, we take the inverse FFT over each sub-block and measure the similarity between the measured autocorrelation matrices obtained within each block in the region.

To make this precise, for each region, we sweep a $m \times m$ rectangular window across the region and collect p autocorrelation matrices each of dimension $(2m - 1) \times (2m - 1)$. We can then reorder the resulting autocorrelation matrices and stack their entries into vectors, $\mathbf{c}_i \in \mathbb{R}^N$, where $N = (2m-1)(2m-1)$, for all $i = \{1, \ldots, p\}$. To measure the similarity between the autocorrelation vectors computed for all sub-blocks within the region, we take normalized inner-products and stack them into a similarity matrix $\mathbf{G}$, where the $(i, j)^{th}$ entry of the similarity matrix is given by:

$$\mathbf{G}_{ij} = \frac{|\mathbf{c}_i^T \mathbf{c}_j|}{\|\mathbf{c}_i\|_2 \|\mathbf{c}_j\|_2}. \tag{5}$$

Finally, a quantitative measure of the stationarity for a region $\mathcal{R}$ with similarity matrix $\mathbf{G} \in \mathbb{R}^{p \times p}$, is defined as:

$$s(\mathcal{R}) = \min_i \quad \frac{1}{p}\left(\sum_{j=1}^{p} \mathbf{G}_{ij}\right). \tag{6}$$

In other words, a measure of the stationarity of the field is defined with respect to the sub-block that has minimal correlation with all other $p - 1$ sub-blocks in the region.

6. EXPERIMENTAL RESULTS

For our evaluations, we measured the timing variability of twelve different FPGA Virtex 5 devices. To extract the delay measurements, we use the delay characterization system presented in [16]. In our implementation, we place a 32×32 array of at-speed delay test circuits on 12 Virtex 5 FPGA chips. Each test circuit is inserted inside two slices with a single CLB. The test circuit measures the effective delay of circuit under test which consists of 4 cascaded inverters. To measure the delay of the circuit under test, the clock frequency is continuously increased until the test circuit captures timing failure at a rate of 50%. We use an ordinary desktop function generator to sweep the clock frequency from $10 - 15$MHz and shift the base frequency 34 times using the PLL inside the FPGA. The measured effective delays have a accuracy of ± 3 pico seconds.

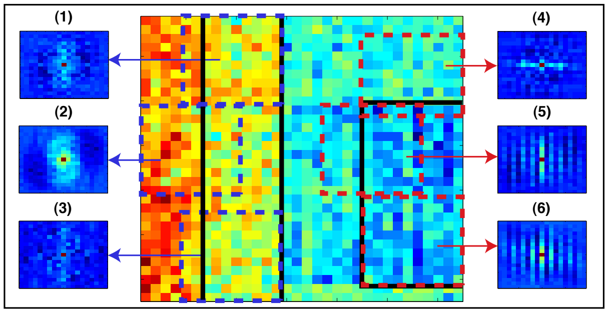


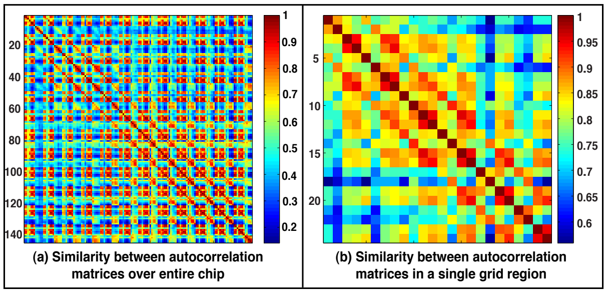
Figure 3: Diversity of local autocorrelation structures for different sub-blocks within an array of CLBs on Virtex 5 FPGA. On the left, we show the autocorrelation computed for each circled sub-block. On the right, we demonstrate that within regions defined by our grid detection algorithm, the autocorrelation functions are very similar within a block.

After scanning across the entire array of CLBs, we obtain a 32×32 array of delay measurements, which we denote as $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$, where $\mathbf{Z}_{i,j}$ corresponds to the mean delay of the circuit inside the $(i,j)^{th}$ CLB. This array of non-zero real numbers can also be written as a vector, $\mathbf{z} \in \mathbb{R}^{1024}$, by simply reordering the rows or columns of the matrix. In Figure 5, we show delay measurements obtained for six different Virtex 5 FPGAs.

After collecting the raw timing measurements, to solve problem (2) we used CVX, a package for specifying and solving convex programs [9] for $\lambda = \{0.02, 0.03, \ldots, 0.1\}$. We selected the TV signature that had the smallest TV norm but still contained enough structural information about regions on the chip, in this case $\lambda = 0.1$. Finally, we performed k-means clustering on the TV signature for $k = \{2, 3, 4, 5\}$. The results of this clustering procedure is shown in Figure 2. We note that a number of different regularization parameters yield very similar TV signatures and hence will produce very similar partitions. A more systematic selection of the regularization parameter should be studied further.

If we look at the physical layout of the chip, a number of the transitions that we detect with our algorithm actually correspond to locations where CLBs are spaced farther apart on the chip. This suggests that when the distance between logic blocks exceeds a certain threshold, this results in a sharp transition in the mean and variance of process variations as you traverse the boundary. This observation is in stark contrast to previous assumptions that the correlation between any two points in the random field is only dependent upon their distance. In addition to layout-dependent effects, we also observe transitions at points in the array that cannot be predicted by the layout. For the example in Figure 2, when the number of regions is set to 2, the resulting boundary exactly corresponds to a point where the spacing between CLBs is higher. When we consider 3 regions, we obtain a boundary that is not layout-dependent. Finally for 4 regions, we obtain another boundary that is layout-dependent. We note that although it may seem obvious that layout-dependent effects could induce sharp transitions in the variations within the circuit, we are not aware of any spatial models that exploit this fact. Furthermore, we hypothesize that for application-specific ICs that exhibit less regularity in their layout than FPGAs, one would observe even more transitions that are independent of the layout of the chip.

To demonstrate the fact that process variations can exhibit diverse autocorrelation structures, in Figure 3 we show the autocorrelation computed for six different sub-blocks of size 10×10 across different points in space. In the labeled regions (1) and (3) on the left, we see that each of these sub-blocks contain very similar autocorrelation structure (nearly white noise) and lie in the same region specified by our proposed grid detection methods. In region (2), we see that as the block is shifted to the left and is placed in between two distinct regions with different mean and variance, the autocorrelation exhibits very different structure. This last example demonstrates that when using a grid-based approach on a non-optimal grid, any estimates of the mean or variance within a suboptimal block will be biased. On the right, in regions (5) and (6), we show two sub-blocks with similar autocorrelation, even though block (5) has been slightly shifted outside of the region detected by our algorithm. However, when we look at region (4), we see that even for a small shift (2 CLBs) out of a region detected by our algorithm, we obtain a very different estimate of the autocorrelation.



(a) Similarity between autocorrelation matrices over entire chip

(b) Similarity between autocorrelation matrices in a single grid region

Figure 4: Autocorrelation matrices for 10×10 sub-blocks (left) across the entire chip, (right) within a region (highlighted in Figure 3) identified by our grid detection algorithm.

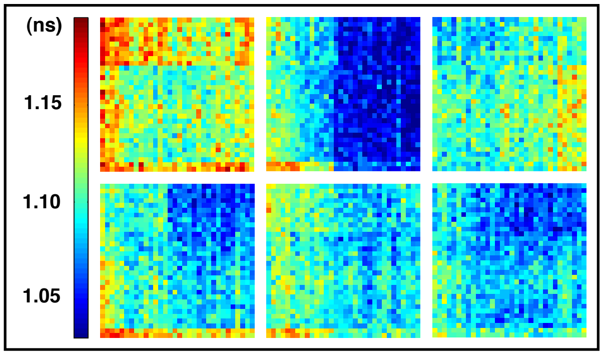


Figure 5: Timing measurements collected from a 32×32 array of CLBs on six different Virtex 5 FPGAs.

To obtain a measure of the stationarity of the process within each of the $m_i \times m_j$ regions specified by our clustering procedure, we swept a $q \times q$ pixel window at 2 pixel shifts over each region of interest, where $q = \min(m_i, m_j, 10)$, and computed the autocorrelation within each sub-block. For all p sub-blocks, we reordered each of the autocorrelation matrices to obtain p autocorrelation vectors each of dimension N, where $N = (2q-1)(2q-1)$. Following this, we computed the normalized inner products between the correlation vectors in accordance with Equation 5.

In Figure 4, on the left, we show the similarity matrices computed for 10×10 blocks shifted across the entire chip and on the right, we show the similarity between the autocorrelation functions computed within one of the larger blocks detected by our algorithm (the region containing sub-block (6) in Figure 3). We note that for autocorrelations computed over the entire chip, many of the correlation functions have very low correlation (i.e., 0.2) but for the autocorrelations computed within a single region, the correlations are much higher. When we compute the score as in 6 for each of these similarity matrices, we obtained scores of $s(\mathcal{R}) = 0.7059$ and $s(\mathcal{C}) = 0.3703$, where $s(\mathcal{R})$ is the stationarity measure for the region outlined in black and $s(\mathcal{C})$ is the stationarity measure computed for the entire chip.

7. CONCLUSIONS

In this paper, we presented a set of novel methods for spatial modeling of non-stationary process variations. To do this, we developed a hybrid spatial model that provides a compact representation for random fields that exhibit non-stationarities arising from both fluctuations in the process mean as well as non-stationarities in the residual spatially correlated random field.

To estimate the parameters in our hybrid model, we developed a TV denoising method for baseline estimation that exploits the transitive nature of process variations to detect change-points in the process mean. Following this, we presented a method for partitioning the chip into a collection of non-overlapping regions that leverages the fact that sharp transitions in the process mean tend to be correlated with change-points in the variance of the process. Finally, we introduced a method for testing whether a process is stationary over a particular region in space.

To validate our hybrid modeling approach, we applied our methods to timing measurements collected from a family of Virtex 5 FPGAs. In line with our previous studies of timing variability on FPGAs where we showed that timing measurements exhibit low-rank (block) structure [16], we find that timing variability is far from smooth. Furthermore, we demonstrated that our measured data supports our conjecture that change-points in the process mean are correlated with change-points in the statistics of the field. This fact enabled us to partition each FPGA into a set of non-overlapping anisotropic regions where the variations exhibit stationary behavior within each region. Although we have only observed these trends in timing data collected from FPGAs, we expect that block structure would also be observed in other types of process variations and for other CMOS technologies due to both the rectangular layout of chips and mask-dependent effects. If step-wise structure is a ubiquitous feature of process variations on ICs, then the development of methods for automatic extraction of this type of structure is essential for the computation of accurate and robust spatial models for process variations. To the best of our knowledge, this work is the first to exploit block structure in process variations to provide a compact model for non-stationary random fields.

8. REFERENCES

[1] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaranand, M. Zhao, K. Gala, and R. Panda. Statistical delay computation considering spatial correlations. In *ASPDAC*, pages 271–276, 2003.

[2] K. Balakrishnan and D. Boning. Measurement and analysis of contact plug resistance variability. In *Custom Integrated Circuits Conference (CICC)*, pages 415–422, 2009.

[3] D. Boning, J. Chung, D. Ouma, and R. Divecha. Spatial variation in semiconductor processes. *Process Control Diagnostics and Modeling in Semiconductor Manufacturing*, 97(9):92–83, 1997.

[4] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameter variations and impact on circuits and microarchitecture. In *DAC*, pages 338–342, 2003.

[5] H. Chang and S. S. Sapatnekar. Statistical timing analysis considering spatial correlations using a single pert-like traversal. In *ICCAD*, page 621, 2003.

[6] H. Chang and S. S. Sapatnekar. Prediction of leakage power under process uncertainties. *ACM Trans. Des. Autom. Electron. Syst.*, 12(2):12, 2007.

[7] L. Cheng, P. Gupta, C. Spanos, K. Qian, and L. He. Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability. In *DAC*, pages 104–109, 2009.

[8] Q. Fu, W.-S. Luk, J. Tao, and C. Y. and X. Zeng. Characterizing intradie spatial correlation using spectral density method. In *Symposium on Quality of Electronic Design*, pages 718–723, 2008.

[9] M. Grant and S. Boyd. *CVX: Matlab software for disciplined convex programming*. 2009.

[10] S. i. Ohkawa, H. Masuda, and Y. Inoue. A novel expression of spatial correlation by a random curved surface model and its application to LSI design. E91-A:1062–1070, 2008.

[11] J. Jinjun, V. Zolotov, and H. Lei. Robust extraction of spatial correlation. *TCAD*, 26(4):619 –631, 2007.

[12] J. l. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14:1570–1582, 2004.

[13] B. Liu. Spatial correlation extraction via random field simulation and production chip performance regression. In *DATE*, pages 527–532, 2008.

[14] F. Liu. A general framework for spatial correlation modeling in VLSI design. In *DAC*, pages 817–822, 2007.

[15] J.-H. Liu, M.-F. Tsai, L. Chen, and C. C.-P. Chen. Accurate and analytical statistical spatial correlation modeling for VLSI DFM applications. In *DAC*, pages 694–697, 2008.

[16] M. Majzoobi, E. Dyer, A. Elnably, and F. Koushanfar. Rapid FPGA characterization using clock synthesis and signal sparsity. In *ITC*, 2010.

[17] M. Orshansky, S. R. Nassif, and D. Boning. *Design for Manufacturability and Statistical Design: A Constructive Approach*. Springer, 2007.

[18] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[19] T. Sato, H. Ueyama, N. Nakayama, and K. Masu. Determination of optimal polynomial regression function to decompose on-die systematic and random variations. In *ASPDAC*, pages 518–523, 2008.

[20] P. Sedcole and P. Y. K. Cheung. Within-die delay variability in 90nm FPGAs and beyond. In *FPT*, pages 97–104, 2006.

[21] P. Sedcole, S. P. Wong, and P. Y. K. Cheung. Compensating for variability in FPGAs by re-mapping and re-placement. In *FPL*, pages 613–616, 2009.

[22] M. Shiqian, Y. Wotao, Z. Yin, and A. Chakraborty. An efficient algorithm for compressed mr imaging using total variation and wavelets. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1 –8, 2008.

[23] W. Zhang, W. Yu, Z. Wang, Z. Yu, R. Jiang, and J. Xiong. An efficient method for chip-level statistical capacitance extraction considering process variations with spatial correlation. In *Proceedings of the conference on Design, automation and test in Europe (DATE)*, pages 580–585, 2008.