Statistical Analysis of Process Variation Based on Indirect Measurements for Electronic System Design

Ivan Ukhov
Linköping University
Sweden
ivan.ukhov@liu.se

Mattias Villani
Linköping University
Sweden
mattias.villani@liu.se

Petru Eles
Linköping University
Sweden
petru.eles@liu.se

Zebo Peng
Linköping University
Sweden
zebo.peng@liu.se

***Abstract*—We present a framework for the analysis of process variation across semiconductor wafers. The framework is capable of quantifying the primary parameters affected by process variation, e.g., the effective channel length, which is in contrast with the former techniques wherein only secondary parameters were considered, e.g., the leakage current. Instead of taking direct measurements of the quantity of interest, we employ Bayesian inference to draw conclusions based on indirect observations, e.g., on temperature. The proposed approach has low costs since no deployment of expensive test structures might be needed or only a small subset of the test equipments already deployed for other purposes might need to be activated. The experimental results present an assessment of our framework for a wide range of configurations.**

I. Introduction and Prior Work

Process variation constitutes one of the major concerns of electronic system designs [1, 2]. A crucial implication of process variation is that it renders the key parameters of a technological process, e.g., the effective channel length and gate oxide thickness, as uncertain quantities. Therefore, the same workload applied to two "identical" dies can lead to two different power and, thus, temperature profiles since the dissipation of power and heat essentially depends on the aforementioned quantities. Consequently, process variation leads to performance degradation in the best case and to severe faults or burnt silicon in the worst scenario. Under these circumstances, uncertainty quantification has evolved into an indispensable asset of the fabrication workflows in order to provide guaranties on the efficiency and robustness of products.

An important target of uncertainty quantification is the characterization of the on-wafer distribution of a quantity of interest, deteriorated by process variation, based on measurements. The problem belongs to the class of inverse problems since the analyzed parameter can be seen as an input to the system and the measured data as the corresponding output. Such an inverse problem is addressed in this work: our goal is to characterize arbitrary process parameters with high accuracy and at low costs. The goal is accomplished by tracking supplementary quantities, which are more convenient and less expensive to be measured, and employing Bayesian statistics [3] to infer the needed parameters from the observed data.

Bayesian inference is utilized in [4] to identify the optimal set of locations on a wafer, in which the parameter under consideration should be measured in order to characterize it with the maximal accuracy. The expectation-maximization algorithm is considered in [5] in order to estimate missing test measurements. In [6], the authors consider an inverse problem focused on the inference of the power dissipation based on transient temperature maps using Markov random fields. Another temperature-based characterization of power is developed in [7] wherein a genetic algorithm is employed for the reconstruction of the power model. It should be noted that the procedures in [4, 5] operate on direct measurements, meaning that the output is the same quantity as the one being measured. In particular, [4, 5] rely heavily on the availability of adequate test structures on the dies and are practical only for the secondary quantities affected by process variation, such as delays and currents, but not for the primary ones, such as various geometrical properties. Hence, [4, 5] often lead to excessive costs and have a limited range of application. The approaches [6, 7], on the other hand, concentrating on the power dissipation of a single die, are not concerned with process variation.

Our work makes the following main contribution. We propose a novel approach to the quantification of process variation based on indirect, incomplete, and noisy measurements. Moreover, we develop and implement a solid framework around the proposed idea and perform a thorough study of various aspects of our technique.

II. Motivational Example

Let us consider an important application of the proposed technique: the characterization of the distribution, across a silicon wafer, of the effective channel length, denoted by u. The effective channel length has one of the strongest effects on the subthreshold leakage current and, consequently, on power and temperature [8]; at the same time, u is well known to be severely deteriorated by process variation [1, 2]. Assume the technological process imposes a lower bound u_* on u.[1] This bound separates defective dies ($u < u_*$) from those that are acceptable ($u \geq u_*$). In order to reduce costs, the manufacturer is interested in detecting the faulty dies and taking them out of the production process at early stages. Then the possible actions that they might take with respect to a single die on the wafer are: (a) keep the die if it conforms to the specification; (b) recycle the die otherwise. Let the distribution of u across the wafer be the one depicted on the left side of Fig. 1. The gradient from navy to dark red represents the transition of u from low to high values; hence, the navy regions have a high level of the power and heat dissipation.[2]

In order to quantify the uncertainty due to the variability of the effective channel length u, one can find the above-mentioned distribution by removing the top layer of (thus, destroying) the dies and measuring u directly. Alternatively, despite the fact that the knowledge of u is more preferable, one can step back and decide to characterize process variation using some other parameter that can be measured without the need of damaging the dies, e.g., the

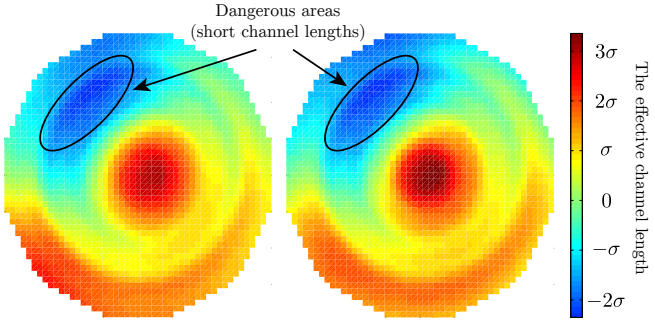

Figure 1. The true (on the left) and inferred (on the right) distributions of the effective channel length across the wafer. The color scheme shows the offset of u from the nominal value where σ stands for the standard deviation of u.

[1]For simplicity, a possible upper bound on the effective channel length is ignored in the motivational example.

[2]The experimental setup will be described in detail in Sec. VI.

leakage current. It should be noted that, in this second case, the chosen surrogate is the final product, and u is left unknown. In either way, adequate test structures have to be present on the dies in order to take the corresponding measurements in sufficiently many points and at the desired level of granularity. Such a sophisticated test structure might not always be readily available, and its deployment might significantly increase production costs. Moreover, the first approach implies that the measured dies have to be recycled afterwards, and the second implies that the further design decisions will be based on a surrogate quantity instead of the primary source of uncertainty, which can compromise the reliability of the decisions. The latter concern is particularly urgent in the situations wherein the production process is not yet completely stable and, hence, the design decisions based on the primary subjects of process variation are desirable.

Our technique works differently. In order to characterize the effective channel length u, we monitor an auxiliary quantity q that depends on u and is more advantageous from the measurement perspective. The distribution of u across the whole wafer is then obtained via Bayesian inference [3] applied to the collected measurements of q. These measurements are taken only for a small number of locations on the wafer and can potentially be corrupted by the noise due to the imperfection of the measurement equipments.

Let us consider one particular helper q, which can be used to study the effective channel length u; specifically, let q be temperature (we elaborate further on this choice in Sec. VI). We can then apply a fixed workload (e.g., run the same application under the same conditions) to a few dies on the wafer and measure the corresponding temperature profiles. Since temperature does not require extra equipments to be deployed on the wafer and can be tracked using infrared cameras [7] or built-in facilities of the dies, our approach can reduce the costs associated with the analysis of process variation. The results of our framework applied to a set of noisy temperature profiles measured for only 7% of the dies on the wafer are shown on the right side of Fig. 1, and the locations of the selected dies are depicted in Fig. 2. It can be seen that the two maps in Fig. 1 closely match each other implying that our approach is able to reconstruct the distribution of the effective channel length with a high level of accuracy.

Another feature of the proposed framework is that probabilities of various events, e.g., $\mathbb{P}(u \geq u_*)$, can readily be estimated. This is important since, in reality, the true values are unknown for us (otherwise, we would not need to quantify them), and, therefore, we can rely on our decisions only up to a certain probability. We can then reformulate the decision rule defined earlier as follows: (a) keep the die if $\mathbb{P}(u \geq u_*)$ is larger than a certain threshold; (b) recycle the die otherwise. An illustration of this rule is given in Fig. 3 where the lower bound u_* is set to two standard deviations below the mean value of the effective channel length; the probability threshold of the action (a) is set to 0.9; the crosses mark both the true and inferred defective dies (they coincide); and the gradient from light gray to red corresponds to the inferred probability of a die to be defective. It can be seen that the inference accurately detects faulty regions.

In addition, we can introduce a trade-off action: (c) expose the die to a thorough inspection (e.g., via a test structure) if $\mathbb{P}(u \geq u_*)$ is smaller than the threshold of (a) and is larger than some other threshold, e.g., $0.1 < \mathbb{P}(u \geq u_*) < 0.9$. In this case, we can reduce costs by examining only those dies for which there is neither sufficiently strong evidence of their satisfactory nor unsatisfactory condition. Furthermore, one can take into consideration a so-called utility function, which, for each combination of an outcome of u and a taken action, returns the gain that the decision maker obtains. For example, such a function can favor a rare omission of malfunctioning dies to a frequent inspection of correct dies as the latter might involve much more costs. The optimal decision is given by the action that maximizes the expected utility with respect to both the observed

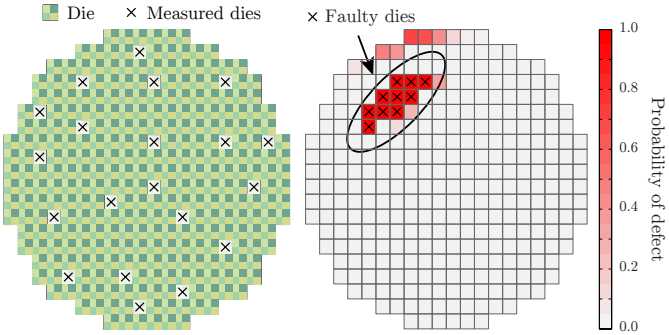

Figure 2. Measurements.

Figure 3. Probability of defect.

data and prior knowledge on u. Thus, all possible u weighted by their probabilities will be taken into account in the final decision, incorporating also the preferences of the user via the utility function.

Finally, we would like to emphasize that temperature is just one option. In certain situations, it might be preferable to perform the above inference based on measurements of some other auxiliary quantity q provided that it depends on the one that we wish to characterize, i.e., on u. For example, q can be the leakage current, which can be readily measured if adequate test structures have already been deployed on the wafer for other purposes.

III. Problem Formulation

Consider a generic electronic system, which is fabricated on a silicon wafer hosting n_d dies. The system depends on a process parameter u, which we are interested in studying and shall refer to as the quantity of interest (QOI). Due to the presence of process variation, the value of u deviates from the nominal one, and this deviation can be different at different locations on the wafer. The QOI is assumed to be expensive/impractical for direct measurements.

The goal of this work is to develop a statistical framework targeted at the identification of the on-wafer distribution of u with the following properties: (a) low measurement costs; (b) high computational speed; (c) robustness to the measurement noise; (d) ability to accommodate prior knowledge on u; and (e) ability to assess the trustworthiness of the collected data and corresponding predictions.

In order to achieve the established goal, we propose the use of indirect measurements. Specifically, instead of u, we measure an auxiliary parameter q, which we shall refer to as the quantity of measurement (QOM). The observations of q are then processed via Bayesian inference in order to derive the distribution of the QOI, u. The QOM is chosen such that: (a) q is convenient and cheap to be tracked; (b) q depends on u, which is signified by $q = f(u)$; and (c) there is a way to compute q for a given u. The last means that f should be known; however, it does not have to be explicitly given: our framework treats f as a "black box." For example, f can be a piece of code or an output of an adequate simulator.

As the first step, the user of the proposed framework is supposed to harvest a set of observations of q at several locations on the wafer (recall Sec. II). Without loss of generality, we shall adhere to the following convention. One die corresponds to one potential measurement site, and $n'_\mathrm{d} \ll n_\mathrm{d}$ denotes the number of those sites that have been selected for measurements. Each site comprises n_p measurement points, and each point contains n_t data instances. For example, in Sec. II, each observation was an $n_\mathrm{p} \times n_\mathrm{t}$ matrix capturing temperature of n_p processing elements for n_t moments of time. Denote by $\mathcal{Q} = \{q_i^\mathrm{msr}\}_{i=1}^{n'_\mathrm{d}}$ the collected data set where $q_i^\mathrm{msr} \in \mathbb{R}^{n_\mathrm{p} \times n_\mathrm{t}}$ stands for one observation (one site) of the QOM. It is implied that the placement of each selected site is recorded along with $\mathcal{Q}$.

Note that, if f is the identity function, i.e., $q \equiv u$, the proposed technique will primarily focus on the reconstruction of any missing

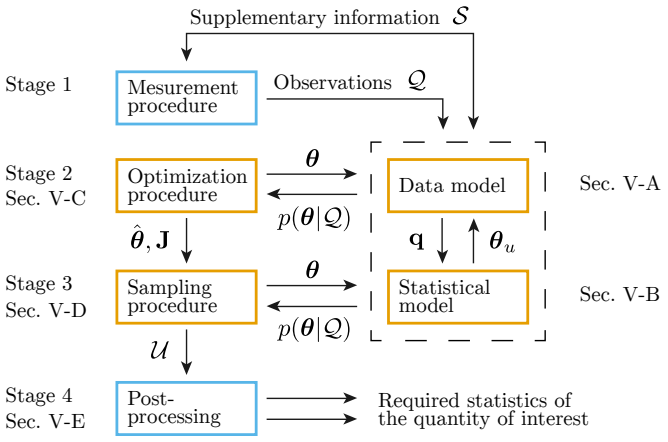

Figure 4. The proposed framework.

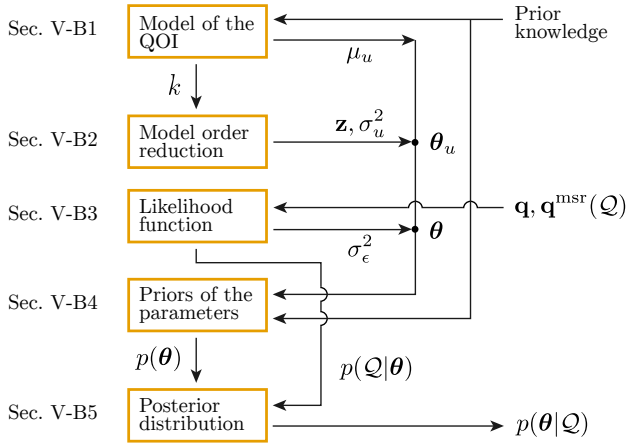

Figure 5. The statistical model.

observations (defined in Sec. V-B2) in $\mathcal{Q}$. From this standpoint, our approach is a generalization of those developed in [4, 5].

For convenience, we denote by $\mathcal{S}$ all the information relevant to the production and measurement processes including: (a) the layout of the wafer and (b) the floorplan of a die on the wafer.

IV. Preliminaries

In order to give a clear presentation of the proposed technique, we first overview the basics of Bayesian inference [3]. Let $\boldsymbol{\theta}$ be a set of unknown parameters (in our case, related to, e.g., the effective channel length), which we would like to characterize. Our arsenal to solve the problem includes: (a) a set of observations $\mathcal{Q}$ (in our case, e.g., temperature or current); (b) a data model connecting $\boldsymbol{\theta}$ with $\mathcal{Q}$; and (c) prior beliefs on $\boldsymbol{\theta}$. A natural solution is Bayes' rule:

$$p(\boldsymbol{\theta}|\mathcal{Q}) \propto p(\mathcal{Q}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{1}$$

where $p(\cdot)$ denotes a probability density function. $p(\mathcal{Q}|\boldsymbol{\theta})$ is known as the likelihood function, which accommodates the data model and yields the probability of observing the data $\mathcal{Q}$ given the parameters $\boldsymbol{\theta}$. $p(\boldsymbol{\theta})$ is called the prior of $\boldsymbol{\theta}$, which represents our knowledge on $\boldsymbol{\theta}$ prior to any observations. $p(\boldsymbol{\theta}|\mathcal{Q})$ reads as the posterior of $\boldsymbol{\theta}$ given $\mathcal{Q}$. Such a posterior is an exhaustive solution to our problem: having constructed $p(\boldsymbol{\theta}|\mathcal{Q})$, all the needed characteristics of $\boldsymbol{\theta}$ can be trivially estimated by drawing samples from this posterior.

Unfortunately, the posterior distribution often does not belong to any of the common families of probability distributions, which is primarily due to the data model involved in the likelihood function, and, therefore, the sampling procedure is not straightforward. To tackle the difficulty, one usually relies on such techniques as Markov Chain Monte Carlo (MCMC) sampling [3]. In this case, an ergodic Markov chain with the stationary distribution equal to the target posterior distribution is constructed and then utilized for the probability space exploration. A popular instantiation of MCMC sampling is the Metropolis-Hastings (MH) algorithm wherein such a Markov chain is attained via sampling from an auxiliary, computationally convenient distribution known as the proposal distribution. We shall further elaborate on this algorithm in Sec. V-B–Sec. V-D.

V. Proposed Framework

In this section, we present our statistical framework for the characterization of process variation. The technique is divided into four major stages depicted in Fig. 4. Stage 1 is the data-harvesting stage wherein the user collects a set of observations of the QOM, q, forming the input set $\mathcal{Q}$. At Stage 2, we undertake an optimization procedure, which assists MCMC sampling at Stage 3 in the construction of an efficient proposal distribution. Stage 3 produces a collection of samples of the QOI, u, such as the effective channel length, which is then processed at Stage 4 in order to estimate all the needed characteristics with respect to this QOI, e.g., the probability of the effective channel length to be smaller than a certain threshold as motivated in Sec. II. As it can be seen in Fig. 4, Stage 2 and Stage 3 actively communicate with the two models on the right, called the data and statistical models, which we discuss next.

A. Data Model

The data model is essentially a directed relation between the QOI, u, and the QOM, q, which we denote by the "black-box" transformation $q = f(u)$. f depends on the choice of q and is specified by the user according to the guidelines in Sec. III.

The data model is utilized to predict the values of the QOM at the same sites, at the same inner points, and with the same amount as the ones in $\mathcal{Q}$. The resulting data are then stacked into one vector with $n_{\text{d}}' n_{\text{p}} n_{\text{t}}$ elements (see Sec. III), which is denoted by $\mathbf{q}$. We also let $\mathbf{q}^{\text{msr}} \in \mathbb{R}^{n_{\text{d}}' n_{\text{p}} n_{\text{t}}}$ be a stacked version of the data in $\mathcal{Q}$ such that the respective elements of $\mathbf{q}$ and $\mathbf{q}^{\text{msr}}$ correspond to the same locations.

In order to acquire a better understanding of the data model, let us return to the setup considered in Sec. II. In this case, u stands for the effective channel length, and q stands for the temperature profile corresponding to a fixed workload. The data model $q = f(u)$ can be roughly divided into two transitions: (a) the effective channel length u to the leakage power p_{leak} and (b) the leakage power p_{leak} to the corresponding temperature profile q. The first transition is accomplished using one of the leakage models broadly available in the contemporary literature; see, e.g., [1, 2, 8]. In particular, a leakage model can be constructed via a fitting procedure applied to a data set of SPICE simulations of reference electrical circuits. The only requirement to such a model is that it should be parametrized by u. In addition, it can also be parametrized by temperature in order to account for the well-known interdependency between leakage and temperature. The second transition is undertaken by combining the leakage power p_{leak} with the dynamic power p_{dyn} that corresponds to the considered workload. The obtained total power along with the temperature-related information contained in $\mathcal{S}$ (mainly, the floorplan and thermal parameters of the die) are fed to a thermal simulator (see Sec. VI) in order to acquire the corresponding temperature q.

B. Statistical Model

Once the wafer has been fabricated, the values of u are fixed for all locations on the wafer; however, they remain unknown for us. In order to infer them, we employ the procedure, called the statistical model, developed in the current subsection and displayed in Fig. 5. The development consists of the five components described below.

1) Model of the QOI: The first step is to assign an adequate model to the unknown u. We model u as a Gaussian process [9] since: (a) it is flexible in capturing the correlation patterns induced by the manufacturing process; (b) it is computationally efficient; and (c) Gaussian distributions are often natural and accurate models for uncertainties due to process variation [2, 5, 8]. Thus, we have

$$u|\boldsymbol{\theta}_u \sim \mathcal{GP}\left(\mu, k\right) \tag{2}$$

where $\mu(r)$ and $k(r, r')$ are the mean and covariance functions of u, respectively, and $r, r' \in \mathbb{R}^2$ denote coordinates on the wafer. Hereafter, the vertical bar, pronounced as "given," is used to mark the parameters that the probability distribution on the right-hand side depends on. In this case, such parameters are $\boldsymbol{\theta}_u$, which we shall identify later on. Prior to taking any measurements, u is assumed to be spatially unbiased; therefore, we let μ be a single location-independent parameter μ_u, i.e., $\mu(r) = \mu_u, \forall r \in \mathbb{R}^2$. The covariance function k is chosen to be the following composition:

$$k(r, r') = \sigma_u^2 \left(\eta\, k_{\text{SE}}(r, r') + (1-\eta) k_{\text{OU}}(r, r')\right) \tag{3}$$

where

$$k_{\text{SE}}(r, r') = \exp\left(-\frac{\|r - r'\|^2}{\ell_{\text{SE}}^2}\right) \text{ and}$$
$$k_{\text{OU}}(r, r') = \exp\left(-\frac{|\,\|r\| - \|r'\|\,|}{\ell_{\text{OU}}}\right)$$

are the squared exponential and Ornstein-Uhlenbeck correlation functions [9], respectively; σ_u^2 represents the variance of u; $\eta \in [0, 1]$ is a weighting coefficient; ℓ_{SE} and $\ell_{\text{OU}} > 0$ are so-called length-scale parameters; and $\|\cdot\|$ stands for the Euclidean distance. The choice of the covariance function k is guided by the observations of the correlation structures induced by the fabrication process [1, 10]: k_{SE} imposes similarities between the points on the wafer that are close to each other, and k_{OU} imposes similarities between the points that are at the same distance from the center of the wafer. ℓ_{SE} and ℓ_{OU} control the extend of these similarities, i.e., the range wherein the influence of one point on another is significant. Although all the above parameters of the model of u can be inferred from the data, for simplicity, we shall focus on μ_u and σ_u^2. The rest of the parameters, namely, η, ℓ_{SE}, and ℓ_{OU}, are assumed to be determined prior to our analysis based on the knowledge of the correlation patterns typical for the production process utilized (see [11] and references therein).

We have established a model for u given as a stochastic process. Now the model requires one additional treatment in order to make it computationally tractable, which we shall discuss next.

2) Model order reduction: The model of the QOI is an infinite-dimensional object as it characterizes a continuum of locations. For practical computations, however, it should be reduced to a finite-dimensional one. First, u is discretized with respect to the union of two sets of points: the first one is composed of the $n'_{\text{d}} n_{\text{p}}$ points where the observations in $\mathcal{Q}$ were made (n'_{d} selected sites with n_{p} inner locations each), and the other of the points where the user wishes to characterize u. For simplicity, we assume that the user is interested in all the sites, which is $n_{\text{d}} n_{\text{p}}$ points in total. Thus, we obtain an $n_{\text{d}} n_{\text{p}}$-dimensional representation of u denoted by $\mathbf{u} \in \mathbb{R}^{n_{\text{d}} n_{\text{p}}}$. Second, the dimensionality is reduced even further by applying the well-known principal component analysis to the covariance matrix of $\mathbf{u}$ computed via Eq. (3). More precisely, we factorize this matrix using the eigenvalue decomposition [12] and discard those eigenvalues (and their eigenvectors) whose contribution to the total sum of the eigenvalues is below a certain threshold. The result is

$$\mathbf{u} = \mu_u \mathbf{e} + \sigma_u \mathbf{L}\mathbf{z} \tag{4}$$

where $\mathbf{e} = (e_i = 1) \in \mathbb{R}^{n_{\text{d}} n_{\text{p}}}$, $\mathbf{L} \in \mathbb{R}^{n_{\text{d}} n_{\text{p}} \times n_{\text{v}}}$, and $\mathbf{z} = (z_i) \in \mathbb{R}^{n_{\text{v}}}$ obey the standard Gaussian distribution. n_{v} is the final dimensionality of the model of u; typically, $n_{\text{v}} \ll n_{\text{d}} n_{\text{p}}$. Consequently, the QOI is now ready for practical computations. In what follows, the parameters of Eq. (2) are defined by $\boldsymbol{\theta}_u = \{\mathbf{z}, \mu_u, \sigma_u^2\}$ (see Fig. 5).

3) Likelihood function: In a Bayesian context, the observed information is taken into account via a likelihood function (see Sec. IV). In our case, the observed information is the measurements $\mathcal{Q}$ stacked into $\mathbf{q}^{\text{msr}}$ as described in Sec. V-A. Since the measurement process is not perfect, we should also take into consideration the measurement noise. To this end, for a given u, the observed $\mathbf{q}^{\text{msr}}$ is assumed to deviate from the data model prediction $\mathbf{q}$ as follows:

$$\mathbf{q}^{\text{msr}} = \mathbf{q} + \boldsymbol{\epsilon} \tag{5}$$

where $\boldsymbol{\epsilon}$ is an $n'_{\text{d}} n_{\text{p}} n_{\text{t}}$-dimensional vector of noise, which is typically assumed to be a white Gaussian noise [9, 11]. Without loss of generality, the noise is assumed to be independent of u and to have the same magnitude for all measurements (characterized by the utilized instruments). Hence, the model of the noise is

$$\boldsymbol{\epsilon}|\sigma_\epsilon^2 \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right) \tag{6}$$

where σ_ϵ^2 is the variance of the noise, $\mathbf{0}$ is a vector of zeros, and $\mathbf{I}$ is the identity matrix. Let us denote the parameters of the inference by $\boldsymbol{\theta} = \boldsymbol{\theta}_u \cup \{\sigma_\epsilon^2\} = \{\mathbf{z}, \mu_u, \sigma_u^2, \sigma_\epsilon^2\}$ (observe this union in Fig. 5). Finally, combining Eq. (5) and Eq. (6), we obtain

$$\mathbf{q}^{\text{msr}}|\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{q}, \sigma_\epsilon^2 \mathbf{I}\right). \tag{7}$$

The probability density function of this distribution is the likelihood function $p(\mathcal{Q}|\boldsymbol{\theta})$ of our statistical model, which is the first of the two components needed for the posterior given in Eq. (1).

4) Priors of the parameters: The second component of the posterior in Eq. (1) is the prior $p(\boldsymbol{\theta})$, which we now need to decide on. In this paper, we put the following priors on $\boldsymbol{\theta}$:

$$\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{8}$$
$$\mu_u \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), \tag{9}$$
$$\sigma_u^2 \sim \text{Scale-inv-}\chi^2\left(\nu_u, \tau_u^2\right), \text{ and} \tag{10}$$
$$\sigma_\epsilon^2 \sim \text{Scale-inv-}\chi^2\left(\nu_\epsilon, \tau_\epsilon^2\right). \tag{11}$$

The prior for $\mathbf{z}$ is due to the properties of the decomposition in Eq. (4). The next three priors, i.e., a Gaussian and two scaled inverse chi-squared distributions, are a common choice for a Gaussian model with the mean and variance being unknown. The parameters μ_0, τ_u^2, and τ_ϵ^2 represent the presumable values of μ_u, σ_u^2, and σ_ϵ^2, respectively, and are set by the user based on the prior knowledge of the technological process and measurement instruments employed. The parameters σ_0, ν_u, and ν_ϵ reflect the precision of this prior information. When the prior knowledge is weak, non-informative priors can be utilized [3]. Taking the product of the densities in Eq. (8)–Eq. (11), we obtain the prior $p(\boldsymbol{\theta})$ completing Eq. (1).

5) Posterior: At this point, we have obtained the two pieces of the posterior shown in Eq. (1): the likelihood function, which is the density in Eq. (7), and the prior, which is the product of the four densities in Eq. (8)–Eq. (11). Thus, the posterior is

$$p(\boldsymbol{\theta}|\mathcal{Q}) \propto p(\mathbf{q}^{\text{msr}}|\mathbf{z}, \mu_u, \sigma_u^2, \sigma_\epsilon^2)\; p(\mathbf{z})\; p(\mu_u)\; p(\sigma_u^2)\; p(\sigma_\epsilon^2). \tag{12}$$

Provided that we have a way of drawing samples from Eq. (12), the QOI can be readily analyzed as we shall see in Sec. V-E. The problem, however, is that the direct sampling of the posterior is not possible due to the data model involved in the likelihood function via $\mathbf{q}$ (see Eq. (7) and Sec. V-A). In order to circumvent this problem, we utilize the Metropolis-Hastings (MH) algorithm [3] mentioned in Sec. IV. The algorithm operates on an auxiliary distribution called the proposal distribution, which is chosen to be convenient for sampling. Each sample, drawn from this proposal, is then used in Eq. (12) to evaluate the posterior probability of this sample and decide

whether it should be accepted or rejected.[3] The acceptance/rejection strategy of the MH algorithm pushes the produced chain of samples towards regions of high posterior probability, which, after a sufficient number of steps, depending on the starting point of the chain and the efficiency of the moves, results in a good approximation of the target posterior distribution in Eq. (12). The preliminary computations needed for the proposal construction are discussed next, and the subsequent sampling procedure in Sec. V-D.

C. Optimization of the Proposal Distribution

In this section, we describe the objective of Stage 2 in Fig. 4. Although the requirements to the proposal distribution mentioned earlier are rather weak, it is often difficult to pick an efficient proposal, which would yield a good approximation with as few evaluations of the posterior in Eq. (12) and, thus, of the data model in Sec. V-A as possible. This choice is especially severe for high-dimensional problems, and our problem, involving around 30 parameters as we shall see in Sec. VI, is one them. Therefore, a careful construction of the proposal distribution is an essential component of our framework.[4] A common technique to construct a high-quality proposal is to perform an optimization of the posterior given by Eq. (12). More specifically, we seek for such a value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ that maximizes Eq. (12) and, hence, has the maximal posterior probability. We also compute the negative of the Hessian matrix at $\hat{\boldsymbol{\theta}}$, which is called the observed information matrix and denoted by $\mathbf{J}$ (see the output of Stage 2 in Fig. 4). Using $\hat{\boldsymbol{\theta}}$ and $\mathbf{J}$, we can now construct such a proposal, which will allow the MH algorithm (a) to start producing samples directly from the desired regions of high probability and (b) to explore those regions more rapidly.

D. Sampling via the Metropolis-Hastings Algorithm

Let us turn to Stage 3 in Fig. 4. We have at our disposal $\hat{\boldsymbol{\theta}}$ and $\mathbf{J}$ from Stage 2 in order to construct an adequate proposal and utilize it for sampling. A commonly used proposal is a multivariate Gaussian distribution wherein the mean is the current location of the chain of samples started at $\hat{\boldsymbol{\theta}}$, and the covariance matrix is the inverse of $\mathbf{J}$ [3]. In order to speed up the sampling process, we would like to make use of the potential of multicore parallelization. The above proposal, however, is purely sequential as the mean for the next sample draw is dependent on the previous sample. Therefore, we appeal to a variation of the MH algorithm known as the independence sampler [3]. In this case, a typical choice of the proposal is a multivariate t-distribution, independent of the current position of the chain:

$$\boldsymbol{\theta} \sim t_\nu \left(\hat{\boldsymbol{\theta}}, \alpha^2 \mathbf{J}^{-1} \right) \tag{13}$$

where $\hat{\boldsymbol{\theta}}$ and $\mathbf{J}$ are as in Sec. V-C, ν is the number of degrees of freedom, and α is a tuning constant controlling the standard deviation of the proposal. Now the proposal samples and the time-consuming evaluation of their posterior in Eq. (12) can be computed for all samples in parallel. Then the precomputed samples can subsequently be accepted or rejected as in the usual MH algorithm.

Having completed the sampling procedure, we obtain a collection of samples of $\boldsymbol{\theta}$. The first portion of the drawn samples is typically discarded before the final computations as being unrepresentative; this portion is also known as the burn-in period. Each of the preserved samples of $\boldsymbol{\theta}$, comprising $\mathbf{z}$, μ_u, and σ_u^2, is then used in Eq. (4) to compute a sample of u, $\mathbf{u}_i \in \mathbb{R}^{n_\mathrm{d} n_\mathrm{p}}$. Denote such a data set with n_mc samples of the QOI by $\mathcal{U} = \{\mathbf{u}_i\}_{i=1}^{n_\mathrm{mc}}$.

[3] A reject means that the sequence of samples advances using the last accepted sample; therefore, the chain of samples is never interrupted.

[4] This has been also confirmed by our experiments. Without optimization, even for small examples, no adequate results were obtained in an affordable time. Therefore, all the experiments in Sec. VI include the optimization step.

E. Post-processing

At Stage 4 in Fig. 4, using the set of samples $\mathcal{U}$, the user computes the desired statistics of the QOI such as the most probable value of the effective channel length at some location of interest, the probability of a certain area on the wafer to be defective, etc. The computations boil down to the estimation of expected values with respect to the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{Q})$. This estimation is done in the standard sample-based fashion, that is, in order to compute some arbitrary quantity dependent on u, one needs to evaluate this quantity for each $\mathbf{u}_i$ in $\mathcal{U}$ and then take the average.

The strength of the Bayesian approach to inference really starts to shine when we are also interested in assessing the trustworthiness of the measured data and, therefore, the reliability of the estimates/decisions based on these data. Such an assessment can readily be undertaken using our framework since the delivered posterior distribution contains all the needed information about the QOI. This is especially helpful in decision making as exemplified in Sec. II.

VI. Experimental Results

In this section, we assess our framework using the inference of the effective channel length u based on temperature q. This choice for illustration is dictated by the fact that such a high-level parameter as temperature constitutes a challenging task for the inference of such a low-level parameter as the effective channel length, which implies a strong assessment of the proposed technique. On the other hand, the effective channel length is an important target *per se* as it is strongly affected by process variation and considerably impacts the power/heat dissipation [1, 2, 8]; in particular, it also influences other process-related characteristics such as the threshold voltage. The performance of our approach is expected to only increase when the auxiliary parameter q resides "closer" to the target parameter u with respect to the transformation $q = f(u)$. For instance, such a "closer" quantity q can be the leakage current, which, however, might not always be the most preferable parameter to measure.

Now we shall describe the default configuration of our setup, which will be later adjusted according to the purpose of each particular experiment. We consider a 45-nanometer technological process. The diameter of the wafer is 20 dies, and the total number of dies n_d is 316. The number of measured dies n'_d is 20, and these dies are chosen by an algorithm, which pursues an even coverage of the wafer. The number of processing elements in each die is four, and they are the points of taking measurements, i.e., $n_\mathrm{p} = 4$. The floorplans of the multiprocessor platforms are constructed in such a way that the processing elements form regular grids. The dynamic power profiles involved in the experiments are based on simulations of randomly generated task graphs via TGFF v3.5 [13]. The sampling interval of these profiles is 1 ms. The leakage model, parametrized by temperature and the effective channel length, is constructed by fitting to SPICE simulations of reference electrical circuits composed of BSIM4 v4.7 devices [14] configured according to the 45-nm PTM HP model [15]. The temperature calculations are undertaken using the approach described in [16], based on HotSpot v5.02 [17].[5] The input data set $\mathcal{Q}$ is obtained as follows: (a) draw a sample of u from a Gaussian distribution with the mean value equal to 17.5 nm, according to the considered technological process [15], and the covariance function given by Eq. (3) wherein the standard deviation is 2.25 nm; (b) perform one fine-grained temperature simulation per each of the n'_d selected dies under the corresponding dynamic power profile; (c) shrink the temperature profiles to keep only n_t, which is equal to 20 by default, evenly spaced moments of time; and (d) perturb the obtained data set using a white Gaussian noise with the standard deviation of 1 K (Kelvin).

[5] The floorplans of the platforms, task graphs of the applications, thermal configuration of HotSpot, etc. are available online at [18].

Table I
MEASURED SITES n'_{d}

*	1	10	20	40	80	160
OT, m	0.41	2.49	3.34	4.59	7.33	10.29
ST, m	**2.40**	**3.99**	**4.60**	**5.79**	**8.49**	**12.96**
TT, m	2.81	6.47	7.94	10.38	15.81	23.25
PT, m	0.61	1.02	1.18	1.51	2.16	3.62
TT, m	1.02	3.50	4.52	6.10	9.49	13.91
E, %	30.49	4.40	3.42	1.09	0.85	0.67

Table II
MEASURED POINTS PER SITE n_{p}

2	4	8	16	32
2.67	3.34	5.20	7.37	13.85
3.71	**4.60**	**6.03**	**8.92**	**14.77**
6.38	7.94	11.23	16.29	28.62
0.98	1.18	1.58	2.51	5.30
3.65	4.52	6.78	9.88	19.15
4.71	3.42	3.68	2.73	1.94

Table III
DATA AMOUNT PER POINT n_{t}

1	10	20	40	80	160
1.12	3.02	3.34	3.62	3.64	4.20
2.40	**4.38**	**4.60**	**4.67**	**4.80**	**4.97**
3.52	7.40	7.94	8.29	8.44	9.16
0.62	1.13	1.18	1.22	1.25	1.30
1.74	4.16	4.52	4.84	4.89	5.50
7.48	2.72	3.42	1.83	2.34	1.32

Table IV
NOISE DEVIATION σ_ϵ

0 K	0.5 K	1 K	2 K
5.08	3.73	3.34	3.19
4.76	**4.70**	**4.60**	**4.71**
9.84	8.43	7.94	7.90
1.19	1.17	1.18	1.18
6.27	4.91	4.52	4.37
0.02	2.71	3.42	4.05

* OT — optimization time, ST — sequential sampling time, PT — parallel sampling time, TT — total time (optimization plus sampling), and E — NRMSE.

Let us turn to the statistical model in Sec. V-B and summarize the intuition and our assignment for each parameter of this model. In the covariance function given by Eq. (3), the weight parameter η and the two length-scale parameters ℓ_{SE} and ℓ_{OU} should be set according to the correlation patterns typical for the production process at hand [1, 10]; we set η to 0.7 and ℓ_{SE} and ℓ_{OU} to half the radius of the wafer. The threshold parameter of the model order reduction procedure described in Sec. V-B2 and utilized in Eq. (4) should be set high enough to preserve a sufficiently large portion of the variance of the data and, thus, to keep the corresponding results accurate; we set it to 0.99 preserving 99% of this variance. The resulting dimensionality n_{v} of $\mathbf{z}$ in Eq. (4) was found to be 27–28. The parameters μ_0 and τ_u of the priors in Eq. (9) and Eq. (10), respectively, are specific to the considered technological process; we set μ_0 to 17.5 nm and τ_u to 2.25 nm. The parameters σ_0 and ν_u in Eq. (9) and Eq. (10), respectively, determine the precision of the information on μ_0 and τ_u and are set according to the beliefs of the user; we set σ_0 to 0.45 nm and ν_u to 10. The latter can be thought of as the number of imaginary observations that the choice of τ_u is based on. The parameter τ_ϵ in Eq. (11) represents the precision (deviation) of the equipments utilized to collect the data set $\mathcal{Q}$ and can be found in the technical specification of these equipments; we set τ_ϵ to 1 K. The parameter ν_ϵ in Eq. (11) has the same interpretation as ν_u in Eq. (10); we set it to 10 as well. In Eq. (13), ν and α are tuning parameters, which can be configured based on experiments; we set ν to eight and α to 0.5. The number of sample draws is another tuning parameter, which we set to 10^4; the first half of these samples is ascribed to the burn-in period leaving $5 \cdot 10^3$ effective samples n_{mc}. For the optimization in Sec. V-C, we use the Quasi-Newton algorithm [12]. For parallel computations, we utilize four processors. All the experiments are conducted on a GNU/Linux machine with Intel Core i7 2.66 GHz and 8 GB of RAM.

To ensure that the experimental setup is adequate, we first perform a detailed inspection of the results obtained for one particular example with the default configuration. The true and inferred distributions of the QOI are shown in Fig. 1 where the normalized root-mean-square error (NRMSE) is below 2.8%, and the absolute error is bounded by 1.4 nm, which suggests that the framework produces a close match to the true value of the QOI. We have also looked at the behavior of the constructed Markov chains and the quality of the proposal distribution; however, due to the shortage of space, these results are not presented here. All the observations suggest that the optimization and sampling procedures are properly configured.

Next we use the assessed configuration and alter only one parameter at a time: the number of measured sites/dies n'_{d}; the number of processing elements/measured points n_{p} on a site; the amount of data per measurement point n_{t}; and the noise deviation σ_ϵ.

A. Number of Measured Sites

Let us change the number of dies n'_{d} that have been measured. The considered scenarios are 1, 10, 20, 40, 80, and 160 measured dies, respectively. The results are reported in Tab. I. In this and the following tables, we report the optimization (Stage 2 in Fig. 4) and sampling (Stage 3 in Fig. 4) times separately (given in minutes). In addition, the sampling time is given for two cases: sequential and parallel computing, which is followed by the total time and error (NRMSE). The computational time of the post-processing phase (Stage 4 in Fig. 4) is not given as it is negligibly small. The sequential sampling time is the most representative indicator of the computational complexity scaling as the number of samples is always fixed, and there is no parallelization; thus, we shall watch this value in most of the discussions below (highlighted in bold).

We see in Tab. I that the more data the proposed framework needs to process, the longer the execution times, which is reasonable. The trend, however, is rather modest: with the doubling of n'_{d}, all the computational times increase less than two times. The error firmly decreases and drops below 4% with around 20 sites measured, which is only 6.3% of the total number of dies on the wafer.

B. Number of Measured Points Per Site

Here we consider five platforms with the number of processing elements/measurement points n_{p} on each die equal to 2, 4, 8, 16, and 32, respectively. The results are summarized in Tab. II. All the computational times grow with n_{p}. This behavior is expected as the granularity of the utilized thermal model (see Sec. V-A and [16]) is bound to the number of processing elements; therefore, the temperature simulations become more intensive. Nevertheless, even for large examples, the timing is readily acceptable, taking into account the complexity of the inference procedure behind and the yielded accuracy. An interesting observation can be made from the NRMSE: the error tends to decrease as n_{p} grows. The explanation is that, with each processing element, $\mathcal{Q}$ delivers more information to the inference to work with since the temperature profiles are collected for all the processing elements simultaneously.

C. Amount of Data Per Measured Point

In this subsection, we sweep the number of moments of time n_{t} captured by the measured temperature profiles. The scenarios are 1, 10, 20, 40, 80, and 160 time moments, respectively. The results are aggregated in Tab. III. As we see, the growth of the computational time is relatively small. One might have expected this growth due to n_{t} to be the same as the one due to n_{p} since, formally, the influence of n_{p} and n_{t} on the dimensionality of $\mathcal{Q}$ is identical (recall $\mathbf{q}^{\mathrm{msr}} \in \mathbb{R}^{n'_{\mathrm{d}} n_{\mathrm{p}} n_{\mathrm{t}}}$). However, the meaning of the two numbers, n_{p} and n_{t}, is completely different, and, therefore, the way they manifest themselves in the algorithm is also different. Therefore, the corresponding amounts of extra data are being treated differently leading to the discordant timing shown in Tab. II and Tab. III. The NRMSE in Tab. III has a decreasing trend; however, this trend is less steady than the ones discovered before. The finding can be explained as follows. The distribution of the time moments in $\mathcal{Q}$ changes since these moments are kept evenly spaced across the corresponding time spans of the input power profiles. Some moments of time can be more informative than the other. Hence, more or less representative

samples can end up in $\mathcal{Q}$ helping or misleading the inference. We can also conclude that a larger number of spatial measurements is more advantageous than a larger number of temporal measurements.

D. Deviation of the Measurement Noise

Next we vary the standard deviation of the noise (in Kelvins), affecting the data $\mathcal{Q}$, within the set $\{0, 0.5, 1, 2\}$ coherent with the literature [7]. Note that the corresponding prior distribution in Eq. (11) is kept unchanged. The results are given in Tab. IV. The sampling time is approximately constant. However, we observe an increase of the optimization time with the decrease of the noise level, which can be ascribed to wider possibilities of perfection for the optimization procedure. A more important observation, revealed by this experiment, is that, in spite of the fact that the inference operates on indirect and drastically incomplete data, a thoroughly calibrated equipment can considerably improve the quality of predictions. However, even with a high level of noise of two degrees—meaning that measurements are dispersed over a wide band of 8 K with a large probability of more than 0.95—the NRMSE is still only 4%.

E. Sequential vs. Parallel Sampling

Let us summarize the results of the sequential and parallel sampling strategies. In the sequential MH algorithm, the optimization time is typically smaller than the time needed for drawing posterior samples. The situation changes when parallel computing is utilized. With four parallel processors, the sampling time decreases 3.81 times on average, which indicates good parallelization properties of the chosen sampling strategy. The overall speedup ranges from 1.49 to 2.75 with the average value of 1.77 times, which can be pushed even further employing more parallel processors.

VII. Conclusion

We proposed a framework for the analysis of process variation across semiconductor wafers based on cost-efficient, indirect measurements. The technique was exposed to an extensive study of various aspects concerning its implementation. The obtained results support the computational efficiency and accuracy of our approach.

We would like to note that, although the framework was demonstrated on the effective channel length and temperature, it can be readily utilized to analyze any other QOIs based on any other QOMs.

References

[1] A. Chandrakasan, F. Fox, W. Bowhill, and W. Bowhill. *Design of High-performance Microprocessor Circuits*. IEEE Press, 2001.

[2] A. Srivastava, D. Sylvester, and D. Blaauw. *Statistical Analysis and Optimization for VLSI: Timing and Power*. Springer, 2010.

[3] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman&Hall/CRC, 2004.

[4] W. Zhang, X. Li, and R. Rutenbar. "Bayesian Virtual Probe: Minimizing Variation Characterization Cost for Nanoscale IC Technologies via Bayesian Inference". In: *DAC*. 2010, pp. 262–267.

[5] S. Reda and S. R. Nassif. "Analyzing the impact of process variations on parametric measurements: novel models and applications". In: *DATE*. 2009, pp. 375–380.

[6] S. Paek, S.-H. Moon, W. Shin, J. Sim, and L.-S. Kim. "PowerField: A Transient Temperature-to-power Technique Based on MRF Theory". In: *DAC*. 2012, pp. 630–635.

[7] F. Mesa-Martinez, J. Nayfach-Battilana, and J. Renau. "Power Model Validation Through Thermal Measurements". In: *ISCA* (2007), pp. 302–311.

[8] D.-C. Juan, Y.-L. Chuang, D. Marculescu, and Y.-W. Chang. "Statistical Thermal Modeling and Optimization Considering Leakage Power Variations". In: *DATE*. 2012, pp. 605–610.

[9] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[10] 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: *IEEE Transactions on CAD of ICs and Systems* 30.3 (2011), pp. 388–401.

[11] Y. Marzouk and H. Najm. "Dimensionality Reduction and Polynomial Chaos Acceleration of Bayesian Inference in Inverse Problems". In: *Journal of Computational Physics* 228.6 (2009), pp. 1862–1902.

[12] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 2007.

[13] R. Dick, D. Rhodes, and W. Wolf. "TGFF: Task Graphs for Free". In: *CODES/CASHE*. 1998, pp. 97–101.

[14] *BSIM4*. Berkeley Short-channel IGFET Model Group at the University of California, Berkeley. URL: http://www-device.eecs.berkeley.edu/bsim/.

[15] *PTM*. Nanoscale Integration and Modeling Group at Arizona State University. URL: http://ptm.asu.edu/.

[16] I. Ukhov, M. Bao, P. Eles, and Z. Peng. "Steady-state Dynamic Temperature Analysis and Reliability Optimization for Embedded Multiprocessor Systems". In: *DAC*. 2012, pp. 197–204.

[17] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan. "Temperature-aware Microarchitecture: Modeling and Implementation". In: *ACM Transactions on Architecture and Code Optimization* 1.1 (Mar. 2004), pp. 94–125.

[18] Embedded Systems Laboratory, Linköping University. URL: http://www.ida.liu.se/~ivauk83/research/SAPV.