Probabilistic Analysis of Power and Temperature Under Process Variation for Electronic System Design

Ivan Ukhov, Petru Eles, *Member, IEEE*, and Zebo Peng, *Senior Member, IEEE*

***Abstract*—Electronic system design based on deterministic techniques for power-temperature analysis is, in the context of current and future technologies, both unreliable and inefficient since the presence of uncertainty, in particular, due to process variation, is disregarded. In this work, we propose a flexible probabilistic framework targeted at the quantification of the transient power and temperature variations of an electronic system. The framework is capable of modeling diverse probability laws of the underlying uncertain parameters and arbitrary dependencies of the system on such parameters. For the considered system, under a given workload, our technique delivers analytical representations of the corresponding stochastic power and temperature profiles. These representations allow for a computationally efficient estimation of the probability distributions and accompanying quantities of the power and temperature characteristics of the system. The approximation accuracy and computational time of our approach are assessed by a range of comparisons with Monte Carlo simulations, which confirm the efficiency of the proposed technique.**

***Index Terms*—Power analysis, process variation, system-level design, temperature analysis, uncertainty quantification.**

I. Introduction

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, gate oxide thickness, and threshold voltage, as random 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 stochastic parameters. This concern is especially urgent due to the interdependence between the leakage power and temperature [2], [3]. 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 [4] has evolved into an indispensable asset of electronic system design workflows in order to provide them with guaranties on the efficiency and robustness of products.

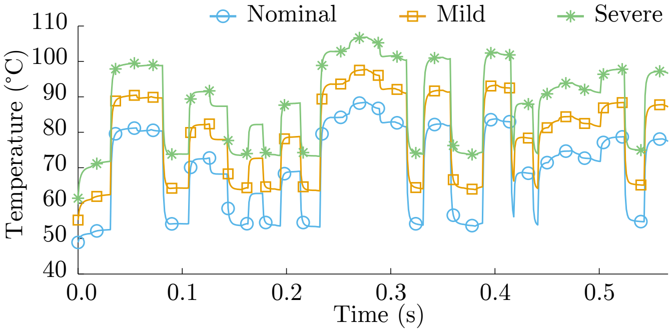

Figure 1. Temperature fluctuations due to process variation.

To illustrate the above concern, consider a quad-core architecture exposed to the uncertainty of the parameters that affect the leakage current. Assume first that these parameters have all nominal values. We can then simulate the system under a certain workload and observe the corresponding temperature profile.[1] The result, labeled as "Nominal," is depicted in Fig. 1 where, for clarity, only one curve, corresponding to one processor, is presented. It can be seen that the temperature is always below 90°C. Now let us assume a mild deviation of the parameters from the nominal values and run the simulation once again. The result is the "Mild" curve in Fig. 1; the maximal temperature is approaching 100°C. Finally, we repeat the experiment considering a severe deviation of the parameters and observe the curve labeled as "Severe" in Fig. 1; the maximal temperature is almost 110°C. Imagine that the designer, when tuning the solution constrained by a maximal temperature of 90°C, was guided exclusively by the nominal parameters. In this case, even with mild deviations, the circuits might be burnt. Another path that the designer could take is to design the system for severe conditions. However, in this scenario, the system might easily end up being too conservative and over-designed. Consequently, such uncertainties have to be addressed in order to pursue efficiency and fail-safeness. Nevertheless, the majority of the literature related to power-temperature analysis of multiprocessor systems ignores this important aspect, e.g., [5], [6], [7], [8].

The remainder of the paper is organized as follows. A summary of the main notations is given in Table I. Sec. II provides an overview of the prior work. In Sec. III, we summarize the contribution of the present paper. The objective of our study is formulated in Sec. IV. The proposed framework is presented in Sec. V. A particular application of our approach is discussed in Sec. VI, and the corresponding results are compared with MC simulations in Sec. VII. Sec. VIII concludes the paper. The work contains a set of supplementary materials with discussions on certain aspects of our framework.

II. Prior Work

Since the appearance of the first digital computers in 1940s, Monte Carlo (MC) sampling remains one of the most well-known and widely used methods for the analysis of stochastic systems. The reason for this popularity lies in the ease of

Manuscript received June 13, 2013; revised September 3, 2013, and November 20, 2013; accepted December 24, 2013.

The authors are with the Department of Computer and Information Science, Linköping University, Linköping 581 83, Sweden (emails: ivan.ukhov@liu.se, petru.eles@liu.se, and zebo.peng@liu.se).

Digital Object Identifier 10.1109/TCAD.2014.2301672

[1] The experimental setup will be detailed in Sec. VI and Sec. VII.

implementation, in the independence of the stochastic dimensionality of the considered problems, and in the fact that the quantities estimated using MC simulations asymptotically approach the true values (the law of large numbers). The crucial problem with MC sampling, however, is the low rate of convergence: the error decreases at the order of ${n_\mathrm{mc}}^{-1/2}$ where n_mc is the number of samples.[2] This means that, in order to get an additional decimal point of accuracy, one has to obtain hundred times more samples. Each such sample implies a complete realization of the whole system, which renders MC-based methods slow and often infeasible since the needed number of simulations can be extremely large [9].

In order to overcome the limitations of deterministic power-temperature analysis (PTA) and, at the same time, to completely eliminate or, at least, mitigate the costs associated with MC sampling, a number of alternative stochastic PTA techniques have been recently introduced. Due to the fact that the leakage component of the power dissipation is influenced by process variation the most [1], [2], [10], [11], the techniques discussed below primarily focus on the variability of leakage.

A solely power-targeted but temperature-aware solution is proposed in [12] wherein the driving force of the analysis is MC sampling with partially precomputed data. A learning-based approach is presented in [10] to estimate the maximal temperature under the steady-state condition. Temperature-related issues originating from process variation are also considered in [11] where a statistical model of the steady-state temperature based on Gaussian distributions is derived. A statistical steady-state temperature simulator is developed in [13] using polynomial chaos (PC) expansions and the Karhunen-Loève (KL) decomposition [4], [14]. A KL-aided stochastic collocation [4] approach to steady-state temperature analysis is presented in [15]. In [16], PC expansions are employed to estimate the full-chip leakage power. The KL decomposition is utilized in [17] for leakage calculations. In [18], the total leakage is quantified using the PC and KL methods. The same combination of tools is employed in [19] and [20] to analyze the response of interconnect networks and power grids, respectively, under process variation.

Table I
MAIN NOTATIONS

Notation	Meaning	Notation	Meaning
$\mathbf{p}$	Power	n_p	# of processing elements
$\mathbf{q}$	Temperature	n_n	# of thermal nodes
$\mathbb{T}[\cdot]$	Probability transform	n_t	# of time moments
$\mathbf{u}$	Uncertain parameters	n_u	# of elements in $\mathbf{u}$
$\boldsymbol{\xi}$	Independent variables	n_ξ	# of elements in $\boldsymbol{\xi}$
$\mathcal{C}^{n_\xi}_{n_\mathrm{po}}[\,\cdot\,]$	PC expansion, (4)	n_po	PC order
ψ_i	Basis polynomials	n_pc	# of PC coefficients, (5)
$\langle\cdot,\cdot\rangle$	Inner product, (19)	n_qp	# of quadrature points

The last five of the aforementioned techniques, i.e., [16], [17], [18], [19], [20], perform only stochastic power analysis and ignore the interdependence between leakage and temperature. The others are temperature-related approaches, but none of them attempts to tackle stochastic transient PTA and to compute the evolving-in-time probability distributions of temperature. However, such transient curves are of practical importance. First of all, certain procedures cannot be undertaken without the knowledge of the time-dependent temperature variations, e.g., reliability optimization based on the thermal-cycling fatigue [8]. Secondly, the constant steady-state temperature assumption, considered, e.g., in [10], [11], [13], [15], can rarely be justified since power profiles are not invariant in reality. In addition, one can frequently encounter the assumption that power and/or temperature follow *a priori* known probability distributions, for instance, Gaussian and log-normal distributions are popular choices as in [2], [11], [17]. However, this assumption often fails in practice (also noted in [11] regarding the normality of the leakage power) due to: (a) the strict nonlinearities between the process-related parameters, power, and temperature; (b) the nonlinear interdependency of temperature and the leakage power [3]. To illustrate this, we simulated the example given in Sec. I 10^4 times assuming the widespread Gaussian model for the variability of the effective channel length; the rest of the experimental setup was configured as it will be described in Sec. VI and Sec. VII. Then we applied the Jarque-Bera test of normality to the collected data (temperature) directly as well as after processing them with the log transformation. The null hypothesis that the data are from an unspecified Gaussian distribution was firmly rejected in both cases at the significance level of 5%. Therefore, the two distributions are neither Gaussian nor log-normal, which can also be seen in Fig. 6 described in the experimental results, Sec. VII.

To conclude, the prior stochastic PTA techniques for electronic system design are restricted in use due to one or several of the following traits: based on MC simulations (potentially slow) [12], limited to power analysis [12], [16], [17], [18], [19], [20], ignoring the leakage-temperature interplay [13], [16], [17], [18], [19], [20], limited to the assumption of the constant steady-state temperature [10], [11], [13], [15], exclusive focus on the maximal temperature [10], and *a priori* chosen distributions of power and temperature [2], [11], [17]. Consequently, there is a lack of flexible stochastic PTA techniques, which we aim to eliminate.

III. Our Contribution

Our work makes the following main contribution. We develop a probabilistic framework for the analysis of the transient power and temperature profiles of electronic systems subject to the uncertainty due to process variation. The proposed technique is flexible in modeling diverse probability distributions, specified by the user, of the uncertain parameters, such as the effective channel length and gate oxide thickness. Moreover, there are no assumptions on the distributions of the resulting power and temperature traces as these distributions are unlikely to be known *a priori*. The proposed technique is capable of capturing arbitrary joint effects of the uncertain parameters on the system since the impact of these parameters is introduced into the framework as a "black box," which is also defined by the user. In particular, it allows for the leakage-temperature interdependence to be taken into account with

[2] There are other sampling techniques that have better convergence rates than the one of the classical MC sampling, e.g., quasi-MC sampling; however, due to additional restrictions, their applicability is often limited [4].

no effort. Our approach is founded on the basis of polynomial chaos (PC) expansions, which constitute an attractive alternative to Monte Carlo (MC) sampling. This is due to the fact that PC expansions possess much faster convergence properties and provide succinct and intuitive representations of system responses to stochastic inputs. In addition, we illustrate the framework considering one of the most important parameters affected by process variation: the effective channel length. Note, however, that our approach is not bounded by any particular source of variability and, apart from the effective channel length, can be applied to other process-related parameters, e.g., the gate oxide thickness.

IV. Problem Formulation

The probability space that we shall reside in is defined as a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a set of outcomes, $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra on Ω, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure [4]. A random variable is a function $\zeta : \Omega \to \mathbb{R}$ which is $\mathcal{F}$-measurable. A random vector (matrix) is then a vector (matrix) whose elements are random variables. In what follows, the space $(\Omega, \mathcal{F}, \mathbb{P})$ will always be implied.

Consider a heterogeneous electronic system that consists of n_p processing elements and is equipped with a thermal package. The processing elements are the active components of the system identified at the system level (ALUs, FPUs, caches, etc.). Let $\mathcal{S}$ be a thermal specification of the system defined as a collection of temperature-related information: (a) the floorplans of the active layers of the chip; (b) the geometry of the thermal package; and (c) the thermal parameters of the materials that the chip and package are made of (e.g., the silicon thermal conductivity and specific heat).

A (transient) power profile $(\mathbf{P}, \tau[\mathbf{P}])$ is defined as a tuple composed of a data matrix $\mathbf{P} \in \mathbb{R}^{n_\mathrm{p} \times n_\mathrm{t}}$ that captures the power dissipation of the n_p processing elements at n_t moments of time and a (column) vector $\tau[\mathbf{P}] = (t_i) \in \mathbb{R}^{n_\mathrm{t}}$ with positive and strictly increasing components that specifies these moments of time. The definition of a (transient) temperature profile $(\mathbf{Q}, \tau[\mathbf{Q}])$ is the same as the one for power except that the data matrix $\mathbf{Q}$ contains temperature values.

The system depends on a set of process parameters that are uncertain at the design stage. These parameters are denoted by a random vector $\mathbf{u} : \Omega \to \mathbb{R}^{n_\mathrm{u}}$. Once the fabrication process yields a particular outcome, $\mathbf{u}$ takes (potentially) different values across each fabricated chip individually and stays unchanged thereafter. This variability leads to deviations of the actual power dissipation from the nominal values and, therefore, to deviations of temperature from the one corresponding to the nominal power consumption.

The goal of this work is to develop a system-level probabilistic framework for transient power-temperature analysis (PTA) of electronic systems where the actual power dissipation and temperature are stochastic due to their dependency on the uncertain parameters $\mathbf{u}$.[3] The user is required to: (a) provide a thermal specification of the platform $\mathcal{S}$; (b) have prior knowledge (or belief) about the probability distribution of the uncertain parameters; and (c) specify a power model, in which $\mathbf{u}$ is an input. The framework should provide the user with the tools to analyze the system under a given workload, without imposing any constraints on the nature/origins of this workload, and obtain the corresponding stochastic power $(\mathbf{P}, \tau[\mathbf{P}])$ and temperature $(\mathbf{Q}, \tau[\mathbf{Q}])$ profiles with a desired level of accuracy and at low costs.

V. Proposed Framework

The main idea of our framework is to construct a surrogate model for the joint power and thermal models of the considered system using PC expansions. Having constructed this surrogate, such quantities as cumulative distribution functions (CDFs) and probability density functions (PDFs) can be easily estimated. Moreover, the representations, which we compute, provide analytical formulae for probabilistic moments, i.e., the expected value and variance are readily available.

The major stages of our technique are depicted in Fig. 2.

Stage 1. *Parameter Preprocessing (Sec. V-A).* The PC approach operates on mutually independent random variables. The uncertain parameters $\mathbf{u}$ might not satisfy this requirement and, thus, should be preprocessed; we denote the corresponding independent random variables by $\boldsymbol{\xi}$.

Stage 2. *Power Modeling (Sec. V-B).* The user specifies the power model of the system via a "black-box" functional Π, which computes the total power $\mathbf{p}(t, \mathbf{u})$ for a particular temperature $\mathbf{q}(t, \mathbf{u})$ and an outcome of the parameters $\mathbf{u}$.

Stage 3. *Thermal Modeling (Sec. V-C).* With respect to the thermal specification $\mathcal{S}$ (defined in Sec. IV), a mathematical formulation of the thermal system is attained. The thermal model closely interacts with the power model from **Stage 2** and produces the corresponding temperature profile.

Stage 4. *Surrogate Modeling (Sec. V-D).* The surrogate model is obtained by traversing the desired time span and gradually constructing polynomial expansions (in terms of the processed uncertain parameters $\boldsymbol{\xi}$ from **Stage 1**) of the stochastic power and temperature profiles. The output is essentially a substitute for the model produced at **Stage 3** with respect to the power model determined at **Stage 2**.

Stage 5. *Post-processing (Sec. V-E).* The computed PC expansions are analyzed in order to obtain the needed characteristics of the system, e.g., CDFs, PDFs, and moments.

In the forthcoming subsections, Sec. V-A–Sec. V-E, the proposed framework is presented. We shall pursue generality such that the user can easily adjust the technique to a particular application, characterized by specific uncertain parameters.

A. Parameter Preprocessing

Independence of the parameters is a prerequisite for PC expansions. In general, however, $\mathbf{u}$ can be correlated and, therefore, should be preprocessed in order to fulfill the requirement. To this end, an adequate probability transformation should be undertaken [21]. Denote such a transformation by $\mathbf{u} = \mathbb{T}[\boldsymbol{\xi}]$, which relates the n_u dependent uncertain parameters $\mathbf{u}$ with n_ξ independent random variables $\boldsymbol{\xi}$.

Correlated random variables can be transformed into uncorrelated ones via a linear mapping based on a factorization

[3] Although the focal point of this paper is process variation, there can be other uncertainties such as those related to the system load and environment.

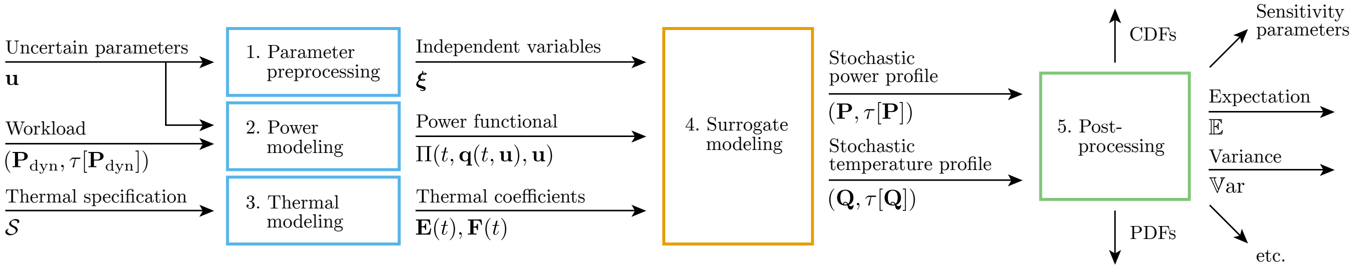

Figure 2. The structure of the proposed framework.

procedure of the covariance matrix or covariance function of $\mathbf{u}$; the procedure is known as the Karhunen-Loève (KL) decomposition [14]. If, in addition, the correlated variables form a Gaussian vector then the uncorrelated ones are also mutually independent. In the general case (non-Gaussian), the most prominent solutions to attain independence are the Rosenblatt [22] and Nataf transformations [23].[4] Rosenblatt's approach is suitable when the joint probability distribution function of the uncertain parameters $\mathbf{u}$ is known; however, such information is rarely available. The marginal probability distributions and correlation matrix of $\mathbf{u}$ are more likely to be given, which are already sufficient for perform the Nataf transformation.[5] The Nataf transformation produces correlated Gaussian variables, which are then turned into independent ones by virtue of the KL decomposition mentioned earlier.

Apart from the extraction of the independent parameters $\boldsymbol{\xi}$, an essential operation at this stage is model order reduction since the number of stochastic dimensions of the problem directly impacts the complexity of the rest of the computations. The intuition is that, due to the correlations possessed by the random variables in $\mathbf{u}$, some of them can be harmlessly replaced by combinations of the rest, leading to a smaller number of the random variables in $\boldsymbol{\xi}$. This operation is often treaded as a part of the KL decomposition.

In Sec. VI-A, we shall demonstrate the Nataf transformation together with the discrete KL decomposition. A description of the latter can also be found in Appendix B.

B. Power Modeling

As stated in Sec. IV, the user of the framework is supposed to decide on the power model for the system under consideration. Such a model can be generally expressed as the following n_{p}-dimensional functional Π:

$$\mathbf{p}(t, \mathbf{u}) = \Pi\left(t, \mathbf{q}(t, \mathbf{u}), \mathbf{u}\right) \tag{1}$$

where n_{p} is the number of processing elements in the system, and $\mathbf{p}(t, \mathbf{u}) \in \mathbb{R}^{n_{\text{p}}}$ and $\mathbf{q}(t, \mathbf{u}) \in \mathbb{R}^{n_{\text{p}}}$ are random vectors of power and temperature, respectively, at time t.

The user can choose any Π. It can be, for instance, a closed-form formula, a piece of code, or an output of a system/power simulator that takes in, for some fixed $\mathbf{u} \equiv \mathbf{u}(\omega)$, $\omega \in \Omega$, the temperature vector $\mathbf{q}(t, \mathbf{u})$ and uncertain parameters $\mathbf{u}$ and computes the corresponding total power $\mathbf{p}(t, \mathbf{u})$. The only assumption we make about Π is that the function is smooth in $\boldsymbol{\xi}$ and has a finite variance, which is generally applicable to most physical systems [4]. Note also that the operation performed by this "black box" is purely deterministic. It can be seen that the definition of Π is flexible enough to account for such effects as the interdependency between leakage and temperature [2], [3], which is discussed in Sec. VI-B.

C. Thermal Modeling

Given the thermal specification $\mathcal{S}$ of the system at hand (see the second paragraph of Sec. IV), an equivalent thermal RC circuit with n_{n} thermal nodes is constructed [24]. The structure of the circuit depends on the intended level of granularity and, therefore, impacts the resulting accuracy. Without loss of generality, we assume that each processing element is mapped onto one corresponding node, and the thermal package is represented as a set of additional nodes.

The thermal behavior of the constructed circuit is modeled with the following system of differential-algebraic equations (see Appendix A for a derivation):

$$\begin{cases} \dfrac{d\,\mathbf{s}(t, \mathbf{u})}{dt} = \mathbf{A}\,\mathbf{s}(t, \mathbf{u}) + \mathbf{B}\,\mathbf{p}(t, \mathbf{u}) & \text{(2a)} \\ \mathbf{q}(t, \mathbf{u}) = \mathbf{B}^T \mathbf{s}(t, \mathbf{u}) + \mathbf{q}_{\text{amb}} & \text{(2b)} \end{cases}$$

where $\mathbf{p}(t, \mathbf{u})$ and $\mathbf{q}(t, \mathbf{u})$ are the input power and output temperature vectors of the processing elements, respectively, and $\mathbf{s}(t, \mathbf{u}) \in \mathbb{R}^{n_{\text{n}}}$ is the vector of the internal state of the system. Note that, as shown in (1), $\mathbf{p}(t, \mathbf{u})$ is an arbitrary function of $\mathbf{q}(t, \mathbf{u})$. Therefore, in general, the system in (2a) is nonlinear and does not have a closed-form solution.

Recall that the power and temperature profiles we work with are discrete-time representations of the power consumption and heat dissipation, respectively, which contain n_{t} samples, or steps, covering a certain time span (see Sec. IV). As detailed in Appendix A, we let the total power be constant between neighboring power steps and reduce the solution process of (2) to the following recurrence, for $k = 1, \dots, n_{\text{t}}$,

$$\mathbf{s}_k = \mathbf{E}_k\,\mathbf{s}_{k-1} + \mathbf{F}_k\,\mathbf{p}_k \tag{3}$$

where $\mathbf{s}_0 = \mathbf{0}$. In the deterministic case, (3) can be readily employed to perform deterministic transient PTA [7], [8]. In the stochastic case, however, the analysis of (3) is substantially different since $\mathbf{p}_k$ and, consequently, $\mathbf{s}_k$ and $\mathbf{q}_k$ are probabilistic quantities. The situation is complicated by the fact that, at each step of the iterative process in (3), (a) $\mathbf{p}_k$ is an arbitrary transformation of the uncertain parameters $\mathbf{u}$ and stochastic temperature $\mathbf{q}_k$ (see Sec. V-B), which results in a multivariate

[4] Only a few alternatives are listed here, and such techniques as independent component analysis (ICA) are left outside the scope of the paper.

[5] The transformation is an approximation, which operates under the assumption that the copula of the distribution is elliptical.

random variable with a generally unknown probability distribution, and (b) $\mathbf{u}$, $\mathbf{p}_k$, $\mathbf{s}_k$, and $\mathbf{q}_k$ are dependent random vectors as the last three are functions of the first. Hence, the operations involved in (3) are to be performed on dependent random vectors with arbitrary probability distributions, which, in general, have no closed-form solutions. To tackle this difficulty, we utilize PC expansions as follows.

D. Surrogate Modeling

The goal now is to transform the "problematic" term in (3), i.e., the power term defined by (1), in such a way that the recurrence in (3) becomes computationally tractable. Our solution is the construction of a surrogate model for the power model in (1), which we further propagate through (3) to obtain an approximation for temperature. To this end, we employ polynomial chaos (PC) [4], which decomposes stochastic quantities into infinite series of orthogonal polynomials of random variables. Such series are especially attractive from the post-processing perspective as they are nothing more than polynomials; hence, PC expansions are easy to interpret and easy to evaluate. An introduction to orthogonal polynomials, which we rely on in what follows, is given in Appendix C.

1) Polynomial basis: The first step towards a polynomial expansion is the choice of a suitable polynomial basis, which is typically made based on the Askey scheme of orthogonal polynomials [4]. The step is crucial as the rate of convergence of PC expansions closely depends on it. Although there are no strict rules that guarantee the optimal choice [25], there are best practices saying that one should be guided by the probability distributions of the random variables that drive the stochastic system at hand. For instance, when a random variable follows a beta distribution, the Jacobi basis is worth being tried first; on the other hand, the Hermite basis is preferable for Gaussian distributions. In multiple dimensions, which is the case with the n_ξ-dimensional random variable $\boldsymbol{\xi}$, several (possibly different) univariate bases are to be combined together to produce a single n_ξ-variate polynomial basis, which we denote by $\{\psi_i : \mathbb{R}^{n_\xi} \to \mathbb{R}\}_{i=1}^{\infty}$; see [4].

2) Recurrence of polynomial expansions: Having chosen an appropriate basis, we apply the PC expansion formalism to the power term in (3) and truncate the resulting infinite series in order to make it feasible for practical implementations. Such an expansion is defined as follows:

$$\mathcal{C}_{n_\text{po}}^{n_\xi}\left[\mathbf{p}_k\right] := \sum_{i=1}^{n_\text{pc}} \hat{\mathbf{p}}_{ki}\, \psi_i(\boldsymbol{\xi}) \tag{4}$$

where $\{\psi_i : \mathbb{R}^{n_\xi} \to \mathbb{R}\}_{i=1}^{n_\text{pc}}$ is the truncated basis with n_pc polynomials in n_ξ variables, and $\hat{\mathbf{p}}_{ki} \in \mathbb{R}^{n_\text{p}}$ are the coefficients of the expansion, which are deterministic. The latter can be computed using spectral projections as it is described in Sec. V-D3. n_po denotes the order of the expansion, which determines the maximal degree of the n_ξ-variate polynomials involved in the expansion; hence, n_po also determines the resulting accuracy. The total number of the PC coefficients n_pc is given by the following expression, which corresponds to the total-order polynomial space [21], [26]:

$$n_\text{pc} = \binom{n_\text{po} + n_\xi}{n_\xi} := \frac{(n_\text{po} + n_\xi)!}{n_\text{po}!\, n_\xi!}. \tag{5}$$

It can be seen in (3) that, due to the linearity of the operations involved in the recurrence, $\mathbf{s}_k$ retains the same polynomial structure as $\mathbf{p}_k$. Therefore, using (4), (3) is rewritten as follows, for $k = 1, \dots, n_\text{t}$:

$$\mathcal{C}_{n_\text{po}}^{n_\xi}\left[\mathbf{s}_k\right] = \mathbf{E}_k\, \mathcal{C}_{n_\text{po}}^{n_\xi}\left[\mathbf{s}_{k-1}\right] + \mathbf{F}_k\, \mathcal{C}_{n_\text{po}}^{n_\xi}\left[\mathbf{p}_k\right]. \tag{6}$$

Thus, there are two PC expansions for two concurrent stochastic processes with the same basis but different coefficients.

Using (4), (6) can be explicitly written as follows:

$$\sum_{i=1}^{n_\text{pc}} \hat{\mathbf{s}}_{ki}\, \psi_i(\boldsymbol{\xi}) = \sum_{i=1}^{n_\text{pc}} \left(\mathbf{E}_k\, \hat{\mathbf{s}}_{(k-1)i} + \mathbf{F}_k\, \hat{\mathbf{p}}_{ki}\right) \psi_i(\boldsymbol{\xi}).$$

Multiplying the above equation by each polynomial from the basis and making use of the orthogonality property (given in (18) in Appendix C), we obtain the following recurrence:

$$\hat{\mathbf{s}}_{ki} = \mathbf{E}_k\, \hat{\mathbf{s}}_{(k-1)i} + \mathbf{F}_k\, \hat{\mathbf{p}}_{ki} \tag{7}$$

where $k = 1, \dots, n_\text{t}$ and $i = 1, \dots, n_\text{pc}$. Finally, (2b) and (7) are combined together to compute the coefficients of the PC expansion of the temperature vector $\mathbf{q}_k$.

3) Expansion coefficients: The general formula of a truncated PC expansion applied to the power term in (3) is given in (4). Let us now find the coefficients $\{\hat{\mathbf{p}}_{ki}\}$ of this expansion, which will be propagated to temperature (using (7) and (2b)). To this end, a spectral projection of the stochastic quantity being expanded—that is, of $\mathbf{p}_k$ as a function of $\boldsymbol{\xi}$ via $\mathbf{u} = \mathbb{T}\left[\boldsymbol{\xi}\right]$ discussed in Sec. V-A—is to be performed onto the space spanned by the n_ξ-variate polynomials $\{\psi_i\}_{i=1}^{n_\text{pc}}$, where n_pc is the number of polynomials in the truncated basis. This means that we need to compute the inner product of (1) with each polynomial from the basis as follows:

$$\langle \mathbf{p}_k, \psi_i \rangle = \left\langle \sum_{j=1}^{n_\text{pc}} \hat{\mathbf{p}}_{kj}\, \psi_j, \psi_i \right\rangle$$

where $i = 1, \dots, n_\text{pc}$, $k = 1, \dots, n_\text{t}$, and $\langle \cdot, \cdot \rangle$ stands for the inner product (see Appendix C for a definition), which should be understood elementwise. Making use of the orthogonality property of the basis, we obtain

$$\hat{\mathbf{p}}_{ki} = \frac{1}{\nu_i} \langle \mathbf{p}_k, \psi_i \rangle \tag{8}$$

where $\{\nu_i = \langle \psi_i, \psi_i \rangle\}_{i=1}^{n_\text{pc}}$ are normalization constants.

In general, the inner product in (8), given in (19) in Appendix C, should be evaluated numerically. This task is accomplished by virtue of a quadrature rule, which is a weighted summation over the integrand values computed at a set of prescribed points. These points along with the corresponding weights are generally precomputed and tabulated since they do not depend the quantity being integrated. Denote such a quadrature-based approximation of (8) by

$$\hat{\mathbf{p}}_{ki} = \frac{1}{\nu_i} \mathcal{Q}_{n_\text{ql}}^{n_\xi}\left[\mathbf{p}_k\, \psi_i\right] \tag{9}$$

where n_{ql} is the level of the quadrature utilized. The procedure is detailed in Appendix D; for the development in this section, we only need to note that n_ξ and n_{ql} dictate the number of quadrature points, which we shall denote by n_{qp}. Also, it is worth emphasizing that, since power depends on temperature as shown in (1), at each step of the recurrence in (7), the computation of $\hat{\mathbf{p}}_{ki}$ should be done with respect to the PC expansion of the temperature vector $\mathbf{q}_{k-1}$.

4) Computational challenges: The construction process of the stochastic power and temperature profiles, implemented inside our prototype of the proposed framework, has been estimated to have the following time complexity:

$$\mathcal{O}(n_{\text{t}}\, {n_{\text{n}}}^2\, n_{\text{pc}} + n_{\text{t}}\, n_{\text{p}}\, n_{\text{qp}}\, n_{\text{pc}} + n_{\text{t}}\, n_{\text{qp}}\, \Pi\,(n_{\text{p}}))$$

where $\mathcal{O}(\Pi\,(n_{\text{p}}))$ denotes the complexity of the computations associated with the power model in (1). The expression can be detailed further by expanding n_{pc} and n_{qp}. The exact formula for n_{pc} is given in (5), and the limiting behavior of n_{pc} with respect to n_ξ is $\mathcal{O}({n_\xi}^{n_{\text{po}}}/n_{\text{po}}!)$. For brute-force quadrature rules, $\log(n_{\text{qp}})$ is $\mathcal{O}(n_\xi)$, meaning that the dependency of n_{qp} on n_ξ is exponential. It can be seen that the theory of PC expansions suffers from the so-called curse of dimensionality [4], [21]. More precisely, when n_ξ increases, the number of polynomial terms as well as the complexity of the corresponding coefficients exhibit a growth, which is exponential without special treatments. The problem does not have a general solution and is one of the central topics of many ongoing studies. In this paper, we mitigate this issue by: (a) keeping the number of stochastic dimensions low using the KL decomposition as we shall see in Sec. VI-A and (b) utilizing efficient integration techniques as discussed in Appendix D. In particular, for sparse integration grids based on Gaussian quadratures, $\log(n_{\text{qp}})$ is $\mathcal{O}(\log(n_\xi))$, meaning that the dependency of n_{qp} on n_ξ is only polynomial [27].

To summarize, let us recall the stochastic recurrence in (3) where, in the presence of correlations, an arbitrary functional $\mathbf{p}_k$ of the uncertain parameters $\mathbf{u}$ and random temperature $\mathbf{q}_k$ (see Sec. V-B) needs to be evaluated and combined with another random vector $\mathbf{s}_k$. Now the recurrence in (3) has been replaced with a purely deterministic recurrence in (7). More importantly, the heavy thermal system in (2) has been substituted with a light polynomial surrogate defined by a set of basis functions $\{\psi_i\}_{i=1}^{n_{\text{pc}}}$ and the corresponding sets of coefficients, namely, $\{\hat{\mathbf{p}}_{ki}\}_{i=1}^{n_{\text{pc}}}$ for power and $\{\hat{\mathbf{q}}_{ki}\}_{i=1}^{n_{\text{pc}}}$ for temperature, where k traverses the n_{t} intervals of the considered time span. Consequently, the output of the proposed PTA framework constitutes two stochastic profiles: the power and temperature profiles denoted by $(\mathbf{P}, \tau[\mathbf{P}])$ and $(\mathbf{Q}, \tau[\mathbf{Q}])$, respectively, which are ready to be analyzed.

Finally, note the ease and generality of taking the uncertainty into consideration using the proposed approach: the above derivation is delivered from any explicit formula for any particular uncertain parameter. In contrast, a typical solution from the literature related to process variation is based on ad hoc expressions and should be tailored by the user for each new parameter individually; see, e.g., [13], [18], [20]. Our framework provides a great flexibility in this regard.

E. Post-processing

Due to the properties of PC expansions—in particular, due to the pairwise orthogonality of the basis polynomials as discussed in Appendix C—the obtained polynomial traces allow for various prospective analyses to be performed with no effort. For instance, consider the PC expansion of temperature at the kth moment of time given by

$$\mathcal{C}^{n_\xi}_{n_{\text{po}}}\,[\mathbf{q}_k] = \sum_{i=1}^{n_{\text{pc}}} \hat{\mathbf{q}}_{ki}\psi_i(\boldsymbol{\xi}) \tag{10}$$

where $\hat{\mathbf{q}}_{ki}$ are computed using (2b) and (7). Let us, for example, find the expectation and variance of the expansion. Due to the fact that, by definition [4], the first polynomial ψ_1 in a polynomial basis is unity, $\mathbb{E}\,(\psi_1(\boldsymbol{\xi})) = 1$. Therefore, using the orthogonality property in (18), we conclude that $\mathbb{E}\,(\psi_i(\boldsymbol{\xi})) = 0$ for $i = 2, \ldots, n_{\text{pc}}$. Consequently, the expected value and variance have the following simple expressions solely based on the coefficients:

$$\mathbb{E}\,(\mathbf{q}_k) = \hat{\mathbf{q}}_{k1} \quad \text{and} \quad \mathbb{V}\text{ar}\,(\mathbf{q}_k) = \sum_{i=2}^{n_{\text{pc}}} \nu_i\, \hat{\mathbf{q}}_{ki}^2 \tag{11}$$

where the squaring should be understood elementwise. Such quantities as CDFs, PDFs, probabilities of certain events, etc. can be estimated by sampling (10); each sample is a trivial evaluation of a polynomial. Furthermore, global and local sensitivity analyses of deterministic and non-deterministic quantities can be readily conducted on (10).

VI. Illustrative Example

So far we have not made any assumptions regarding the cause of the variability of the power term in the thermal system given by (2). In this section, we shall consider a particular application of the proposed framework. To this end, we begin with the problem formulation of this application.

The total dissipation of power is composed of two major parts: dynamic and static. The influence of process variation on the dynamic power is known to be negligibly small [2]; on the other hand, the variability of the static power is substantial, in which the subthreshold leakage current contributes the most [10], [11]. Hence, we shall focus on the subthreshold leakage and, more specifically, on the effective channel length, denoted by L, since it has the strongest influence on leakage and is severely deteriorated by process variation [1]. In particular, L also affects the threshold voltage [10].

It is well known that the dispersion due to process variation of the effective channel length around the nominal value resembles a bell shape, which is similar to the ones owned by Gaussian distributions. Therefore, such variations are often conveniently modeled using Gaussian variables [2], [10], [11], [12], [13], [15], [16], [17], [20]. In this work, due to both the underlying physics and demonstration purposes, we make a step further and bake right into the model the fact that the effective channel length—occupying the space between the drain and source of a nanoscopic transistor—cannot be arbitrarily large or take a negative value, as Gaussian distributions allow it to do. In other words, we require the

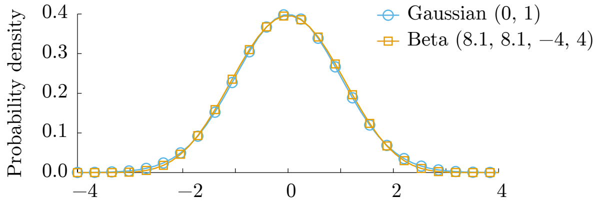

Figure 3. The standard Gaussian distribution and a fitted beta distribution.

model of L to have a bounded support. With this in mind, we propose to model physically-bounded parameters using the four-parametric family of beta distributions: Beta(a, b, c, d) where a and b are the shape parameters, and c and d are the left and right bounds of the support, respectively. a and b can be chosen in such a way that the typically found bell shape of the distribution is preserved. An illustration is given in Fig. 3 where we fitted a beta distribution to the standard Gaussian distribution.[6] It can be seen that the curves are nearly indistinguishable, but the beta one has a bounded support $[-4, 4]$, which can potentially lead to more realistic models.

The variability of L is split into global $\delta L^{(g)}$ and local $\delta L^{(l)}$ parts [12], [16].[7] $\delta L^{(g)}$ is assumed to be shared among all processing elements whereas each processing element has its own local parameter $\delta L_i^{(l)}$. Therefore, the effective channel length the ith processing element is modeled as follows:

$$L_i = L_{\text{nom}} + \delta L^{(g)} + \delta L_i^{(l)} \tag{12}$$

where L_{nom} is the nominal value of the effective channel length. Hence, the uncertain parameters of the problem are

$$\mathbf{u} = \left(\delta L_1^{(l)}, \dots, \delta L_{n_\text{p}}^{(l)}, \delta L^{(g)}\right)^T. \tag{13}$$

Global variations are typically assumed to be uncorrelated with respect to the local ones. The latter, however, are known to have high spatial correlations, which we shall model using the following correlation function:

$$k(\mathbf{r}_i, \mathbf{r}_j) = \eta\, k_{\text{SE}}(\mathbf{r}_i, \mathbf{r}_j) + (1 - \eta) k_{\text{OU}}(\mathbf{r}_i, \mathbf{r}_j) \tag{14}$$

where $\mathbf{r}_i \in \mathbb{R}^2$ is the spatial location of the center of the ith processing element relative to the center of the die. The correlation function is a composition of two kernels:

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

which are known as the squared-exponential and Ornstein-Uhlenbeck kernels, respectively. $\eta \in [0, 1]$ is a weight coefficient balancing the kernels; ℓ_{SE} and $\ell_{\text{OU}} > 0$ are so-called length-scale parameters; and $\|\cdot\|$ stands for the Euclidean distance. The choice of the correlation function in (14) is guided by the observations of the correlations induced by the fabrication process [1], [28], [29]: k_{SE} imposes similarities between the spatial locations that are close to each other, and k_{OU} imposes similarities between the locations that are at the same distance from the center of the die (see also [13], [14], [15], [18], [20]). The length-scale parameters 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 (14) captures certain features inherent to the fabrication process, it is still an idealization. In practice, it can be difficult to make a justifiable choice and tune such a formula, which is a prerequisite for the techniques in Sec. II based on the (continuous) KL decomposition. A correlation matrix, on the other hand, can readily be estimated from measurements and, thus, is a more probable input to PTA. Thus, we use (14) with the only purpose of constructing a correlation matrix of $\{\delta L_i^{(l)}\}$. For convenience, the resulting matrix is extended by one dimension to pack $\delta L^{(g)}$ and $\{\delta L_i^{(l)}\}$ together. In this case, the correlation matrix obtains one additional non-zero element on the diagonal. Taking into account the variances of the variable, the final covariance matrix of the whole random vector $\mathbf{u}$ (see (13)) is formed, which we denote by $\mathbf{\Sigma_u}$.

To conclude, an input to our analysis is the marginal distributions of the parameters $\mathbf{u}$, which are beta distributions, and the corresponding covariance matrix $\mathbf{\Sigma_u}$.

A. Parameter Preprocessing

At **Stage 1**, $\mathbf{u}$ should be preprocessed in order to extract a vector of mutually independent random variables denoted by $\boldsymbol{\xi}$. Following the guidance given in Sec. V-A, the most suitable transformation for the ongoing scenario is the Nataf transformation. Here we describe the algorithm in brief and refer the interested reader to [23] for additional details. The transformation is typically presented in two steps. First, $\mathbf{u} \in \mathbb{R}^{n_\text{u}}$, $n_\text{u} = n_\text{p} + 1$, is morphed into correlated Gaussian variables, denoted by $\boldsymbol{\xi}' \in \mathbb{R}^{n_\text{u}}$, using the knowledge of the marginal distributions and covariance matrix of $\mathbf{u}$. Second, the obtained correlated Gaussian variables are mapped into independent standard Gaussian variables, denoted by $\boldsymbol{\xi}'' \in \mathbb{R}^{n_\text{u}}$, using one of several available techniques; see [23].

The number of stochastic dimensions, which so far is $n_\text{p}+1$, directly impacts the computational cost of PC expansions as it is discussed in Sec. V-D4. Therefore, one should consider a possibility for model order reduction before constructing PC expansions. To this end, we perform the second step of the Nataf transformation by virtue of the discrete Karhunen-Loève (KL) decomposition [14] as the reduction comes naturally in this way. A description of this procedure can be found in Appendix B. Let us denote the trimmed independent variables by $\boldsymbol{\xi}'''$ and their number by n_ξ. We also denote the whole operation, i.e., the reduction-aware Nataf transformation, by

$$\mathbf{u} = \mathbb{N}\text{ataf}^{-1}\left[\boldsymbol{\xi}'''\right]$$

where the superscript "−1" signifies the fact that we are interested in expressing $\mathbf{u}$ via $\boldsymbol{\xi}'''$ and, hence, need to perform all the operations in the reversed order.

At this point, we have n_ξ independent Gaussian random variables stored in $\boldsymbol{\xi}'''$, which already suffice the independence prerequisite for PC expansions (see Sec. V-A). However, we prefer to construct PC expansions in terms of bounded

[6] Alternatively, one can match the moments of the distributions.

[7] Without loss of generality, $\delta L^{(g)}$ can be treated as a composition of independent inter-lot, inter-wafer, and inter-die variations; likewise, $\delta L^{(l)}$ can be treated as a composition of independent and dependent local variations.

variables since such expansions will also be bounded. To this end, we undertake one additional transformation that yields a vector of (independent) random variables $\boldsymbol{\xi} \in \mathbb{R}^{n_\xi}$ whose distributions have bounded supports. This transformation is a standard technique based on the composition of the inverse CDF of $\boldsymbol{\xi}'''$ and the CDF of $\boldsymbol{\xi}$ denoted by $F^{-1}_{\boldsymbol{\xi}'''}$ and $F_{\boldsymbol{\xi}}$, respectively. The overall probability transformation $\mathbb{T}$ (see Sec. V-A) from $\mathbf{u}$ to $\boldsymbol{\xi}$ is then given as follows:

$$\mathbf{u} = \mathbb{T}\left[\boldsymbol{\xi}\right] = \mathbb{N}\text{ataf}^{-1}\left[F^{-1}_{\boldsymbol{\xi}'''}(F_{\boldsymbol{\xi}}(\boldsymbol{\xi}))\right].$$

The distributions of $\boldsymbol{\xi}$ can be chosen arbitrary as long as one can construct a suitable polynomial basis as described in Sec. V-D1. We let $\boldsymbol{\xi}$ have beta distributions, staying in the same family of distributions with the parameters $\mathbf{u}$.

B. Power Modeling

At **Stage 2** in Fig. 2, we need to decide on the power model with the identified uncertain parameters as an input. To this end, (1) is decomposed into the sum of the dynamic and static components denoted by $\Pi_{\text{dyn}}(t, \mathbf{u})$ and $\Pi_{\text{stat}}(\mathbf{q}(t, \mathbf{u}), \mathbf{u})$, respectively. As motivated earlier, we let $\Pi_{\text{dyn}}(t, \mathbf{u}) = \mathbf{p}_{\text{dyn}}(t)$ (does not depend on $\mathbf{u}$). We assume that the desired workload of the system is given as a dynamic power profile denoted by $(\mathbf{P}_{\text{dyn}}, \tau[\mathbf{P}_{\text{dyn}}])$. Without loss of generality, the development of the static part is based on SPICE simulations of a reference electrical circuit composed of BSIM4 devices (v4.7.0) [30] configured according to the 45-nm PTM (high-performance) [31]. Specifically, we use a series of CMOS invertors for this purpose. The simulations are performed for a fine-grained two-dimensional grid, the effective channel length vs. temperature, and the results are tabulated. The interpolation facilities of MATLAB (vR2013a) [32] are then utilized whenever we need to evaluate the leakage power for a particular point within the range of the grid, which is chosen to be sufficiently wide.

C. Thermal Modeling

We move on to **Stage 3** where the thermal model of the multiprocessor system is to be established. Given the thermal specification $\mathcal{S}$ of the considered platform (the floorplan of the die, the configuration of the thermal package, etc.), we employ HotSpot (v5.02) [24] in order to construct an equivalent thermal RC circuits of the system. Specifically, we are interested in the coefficient matrices $\mathbf{E}(t)$ and $\mathbf{F}(t)$ in (3) (see also Fig. 2), which HotSpot helps us to compute by providing the corresponding capacitance and conductance matrices of the system as described in Appendix A. In this case, thermal packages are modeled with three layers, and the relation between the number of processing elements and the number of thermal nodes is given by $n_{\text{n}} = 4\,n_{\text{p}} + 12$.

To conclude, the power and thermal models of the platform are now acquired, and we are ready to construct the corresponding surrogate model via PC expansions, which is the topic for the discussion in the following subsection.

D. Surrogate Modeling

At **Stage 4**, the uncertain parameters, power model, and thermal model developed in the previous sections are to be fused together under the desired workload $(\mathbf{P}_{\text{dyn}}, \tau[\mathbf{P}_{\text{dyn}}])$ to produce the corresponding stochastic power $(\mathbf{P}, \tau[\mathbf{P}])$ and temperature $(\mathbf{Q}, \tau[\mathbf{Q}])$ profiles. The construction of PC expansions, in the current scenario, is based on the Jacobi polynomial basis as it is preferable in situations involving beta-distributed parameters [4]. To give an example, for a dual-core platform (i.e., $n_{\text{p}} = 2$) with two stochastic dimensions (i.e., $n_\xi = 2$), the second-order PC expansion (i.e., $n_{\text{po}} = 2$) of temperature at the kth time moment is as follows:[8]

$$\begin{aligned} \mathcal{C}_2^2\left[\mathbf{q}_k\right] = {} & \hat{\mathbf{q}}_{k1}\,\psi_1(\boldsymbol{\xi}) + \hat{\mathbf{q}}_{k2}\,\psi_2(\boldsymbol{\xi}) + \hat{\mathbf{q}}_{k3}\,\psi_3(\boldsymbol{\xi}) \\ & + \hat{\mathbf{q}}_{k4}\,\psi_4(\boldsymbol{\xi}) + \hat{\mathbf{q}}_{k5}\,\psi_5(\boldsymbol{\xi}) + \hat{\mathbf{q}}_{k6}\,\psi_6(\boldsymbol{\xi}) \end{aligned} \tag{15}$$

where the coefficients $\hat{\mathbf{q}}_{ki}$ are vectors with two elements corresponding to the two processing elements,

$$\psi_1(\mathbf{x}) = 1, \quad \psi_2(\mathbf{x}) = 2x_1, \quad \psi_3(\mathbf{x}) = 2x_2, \quad \psi_4(\mathbf{x}) = 4x_1x_2$$
$$\psi_5(\mathbf{x}) = \frac{15}{4}x_1^2 - \frac{3}{4}, \quad \text{and} \quad \psi_6(\mathbf{x}) = \frac{15}{4}x_2^2 - \frac{3}{4}.$$

The expansion for power has the same structure but different coefficients. Such a series might be shorter or longer depending on the accuracy requirements defined by n_{po}.

Once the basis has been chosen, we need to compute the corresponding coefficients, specifically, $\{\hat{\mathbf{p}}_{ki}\}_{i=1}^{n_{\text{pc}}}$ in (4), which yield $\{\hat{\mathbf{q}}_{ki}\}_{i=1}^{n_{\text{pc}}}$. As shown in Appendix C, these computations involve multidimensional integration with respect to the PDF of $\boldsymbol{\xi}$, which should be done numerically using a quadrature rule; recall Sec. V-D3. When beta distributions are concerned, a natural choice of such a rule is the Gauss-Jacobi quadrature. Additional details are given in Appendix D.

To summarize, we have completed four out of five stages of the proposed framework depicted in Fig. 2. The result is a light surrogate for the model in (2). At each moment of time, the surrogate is composed of two n_{p}-valued polynomials, one for power and one for temperature, which are defined in terms of n_ξ mutually independent random variables; an example of such a polynomial is given in (15).

E. Post-processing

We turn to **Stage 5** in Fig. 2. It can be seen in, for example, (15) that the surrogate model has a negligibly small computational cost at this stage: for any outcome of the parameters $\boldsymbol{\xi} \equiv \boldsymbol{\xi}(\omega)$, we can easily compute the corresponding temperature by plugging in $\boldsymbol{\xi}$ into (15); the same applies to power. Hence, the constructed representation can be trivially analyzed to retrieve various statistics about the system in (2). Let us illustrate a few of them still retaining the example in (15). Assume that the dynamic power profile $(\mathbf{P}_{\text{dyn}}, \tau[\mathbf{P}_{\text{dyn}}])$ corresponding to the considered workload is the one shown in Fig. 4. Having constructed the surrogate with respect to this profile, we can then rigorously estimate, say, the PDF of temperature at some kth moment of time by sampling the surrogate and obtain curves similar to those shown Fig. 6 (discussed in Sec. VII). Furthermore, the expectation and variance of temperature are trivially calculated using the formulae in (11) where $n_{\text{pc}} = 6$. For the whole time span of the power

[8] The Jacobi polynomials have two parameters [4], and the shown $\{\psi_i\}_{i=1}^{6}$ correspond to the case where both parameters are equal to two.

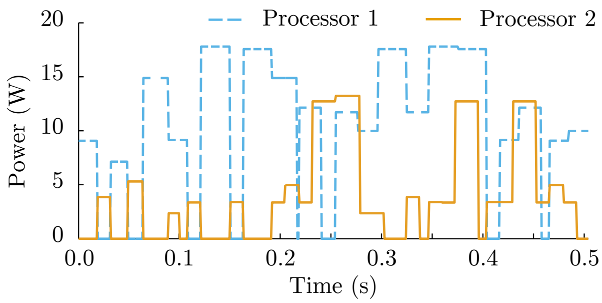

Figure 4. A dynamic power profile.

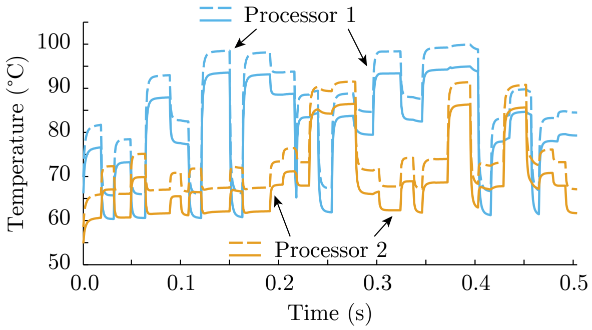

Figure 5. The expected temperature (the solid lines) and one standard deviation above it (the dashed lines).

profile $(\mathbf{P}_{\text{dyn}}, \tau[\mathbf{P}_{\text{dyn}}])$ depicted in Fig. 4, these quantities are plotted in Fig. 5. The displayed curves closely match those obtained via MC simulations with 10^4 samples; however, our method takes less than a second whilst MC sampling takes more than a day as we shall see next.

VII. Experimental Results

In this section, we report the results of the proposed framework for different configurations of the illustrative example in Sec. VI. All the experiments are conducted on a GNU/Linux machine with Intel Core i7 2.66 GHz and 8 GB of RAM.

Now we shall elaborate on the default configuration of our experimental setup, which, in the following subsections, will be adjusted according to the purpose of each particular experiment. We consider a 45-nanometer technological process. The effective channel length is assumed to have a nominal value of 17.5 nm [31] and a standard deviation of 2.25 nm where the global and local variations are equally weighted. Correlation matrices are computed according to (14) where the length-scale parameters ℓ_{SE} and ℓ_{OU} are set to half the size of the square die. In the model order reduction technique (see Sec. VI-A), the threshold parameter is set to 0.99 preserving 99% of the variance of the data. Dynamic power profiles involved in the experiments are based on simulations of randomly generated applications defined as directed acyclic task graphs.[9] The floorplans of the platforms are constructed in such a way that the processing elements form regular grids.[10] The time step of power and temperature traces is set to 1 ms (see Sec. IV), which is also the time step of the recurrence in (7). As a comparison to our polynomial chaos (PC) expansions, we employ Monte Carlo (MC) sampling. The MC approach is set up to preserve the whole variance of the problem, i.e., no model order reduction, and to solve (2) directly using the Runge-Kutta formulae (the Dormand-Prince method) available in MATLAB [32].

Since the temperature part of PTA is the main contribution of this work, we shall focus on the assessment of temperature profiles. Note, however, that the results for temperature allow one to implicitly draw reasonable conclusions regarding power since power is an intermediate step towards temperature, and any accuracy problems with respect to power are expected to propagate to temperature. Also, since the temperature-driven studies [10], [11], [13], [15] work under the steady-state assumption ([10] is also limited to the maximal temperature, and [13] does not model the leakage-temperature interplay), a one-to-one comparison with our framework is not possible.

A. Approximation Accuracy

The first set of experiments is aimed to identify the accuracy of our framework with respect to MC simulations. At this point, it is important to note that the true distributions of temperature are unknown, and both the PC and MC approaches introduce errors. These errors decrease as the order of PC expansions n_{po} and the number of MC samples n_{mc}, respectively, increase. Therefore, instead of postulating that the MC technique with a certain number of samples is the “universal truth” that we should achieve, we shall vary both n_{po} and n_{mc} and monitor the corresponding difference between the results produced by the two alternatives.

In order to make the comparison even more comprehensive, let us also inspect the effect of the correlation patterns between the local random variables $\{\delta L_i^{(l)}\}$ (recall Sec. VI). Specifically, apart from n_{po} and n_{mc}, we shall change the balance between the two correlation kernels shown in (14), i.e., the squared-exponential k_{SE} and Ornstein-Uhlenbeck k_{OU} kernels, which is controlled by the weight parameter $\eta \in [0, 1]$.

The PC and MC methods are compared by means of three error metrics. The first two are the normalized root mean square errors (NRMSEs) of the expectation and variance of the computed temperature profiles.[11] The third metric is the mean of the NRMSEs of the empirical PDFs of temperature constructed at each time step for each processing element. The error metrics are denoted by $\epsilon_{\mathbb{E}}$, $\epsilon_{\mathbb{V}\text{ar}}$, and ϵ_f, respectively. $\epsilon_{\mathbb{E}}$ and $\epsilon_{\mathbb{V}\text{ar}}$ are easy to interpret, and they are based on the analytical formulae in (11). ϵ_f is a strong indicator of the quality of the distributions estimated by our framework, and it is computed by sampling the constructed PC expansions. In contrast to the MC approach, this sampling has a negligible overhead as we discussed in Sec. V-E.

The considered values for n_{po}, n_{mc}, and η are the sets $\{n\}_{n=1}^{7}$, $\{10^n\}_{n=2}^{5}$, and $\{0, 0.5, 1\}$, respectively. The three cases of η correspond to the total dominance of k_{OU} ($\eta = 0$), perfect balance between k_{SE} and k_{OU} ($\eta = 0.5$), and total dominance of k_{SE} ($\eta = 1$). A comparison for a quad-core

[9] In practice, dynamic power profiles are typically obtained via an adequate simulator of the architecture of interest.

[10] The task graphs of the applications, floorplans of the platforms, configuration of HotSpot, which was used to construct thermal RC circuits for our experiments, are available online at [33].

[11] In the context of NRMSEs, we treat the MC results as the observed data and the PC results as the corresponding model predictions.

Table II

Error measurements for $\eta = 0$ and various numbers of MC samples n_{mc} and PC orders n_{po}

	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	ϵ_f, %	ϵ_f, %	ϵ_f, %	ϵ_f, %
n_{po}	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$
1	1.70	0.92	0.51	0.48	88.19	55.73	55.57	53.08	10.88	11.48	8.85	8.83
2	1.36	0.58	0.20	0.18	67.66	23.30	23.05	19.64	10.15	10.11	6.26	6.04
3	1.26	0.49	0.15	0.14	61.16	13.06	12.78	9.08	5.49	5.04	2.95	2.73
4	1.23	0.45	0.14	0.14	58.49	8.85	8.57	4.78	3.84	2.02	1.50	1.51
5	1.21	0.44	0.14	0.14	57.31	7.00	6.71	2.92	3.83	2.27	1.03	0.84
6	1.21	0.44	0.14	0.14	56.75	6.12	5.83	2.08	3.08	1.94	0.93	0.66
7	1.20	0.43	0.14	0.14	56.41	5.60	5.31	1.62	2.78	1.39	0.72	0.62

Table III

Error measurements for $\eta = 0.5$ and various numbers of MC samples n_{mc} and PC orders n_{po}

	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	ϵ_f, %	ϵ_f, %	ϵ_f, %	ϵ_f, %
n_{po}	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$
1	1.66	0.98	0.60	0.57	65.82	64.11	66.13	66.70	10.97	10.69	9.27	8.77
2	1.31	0.63	0.27	0.23	49.55	29.21	30.49	28.24	6.43	5.42	3.87	3.59
3	1.13	0.44	0.16	0.14	43.44	15.94	16.88	13.48	5.60	3.80	1.83	1.53
4	1.17	0.48	0.17	0.14	40.24	9.11	9.80	5.71	5.48	3.80	1.77	1.47
5	1.07	0.38	0.16	0.16	39.68	7.96	8.56	4.35	3.80	1.72	1.59	1.62
6	1.19	0.49	0.18	0.15	38.23	5.19	5.51	1.24	4.62	2.86	1.16	0.86
7	0.99	0.30	0.21	0.21	38.27	5.27	5.59	1.29	3.45	2.01	1.82	1.68

Table IV

Error measurements for $\eta = 1$ and various numbers of MC samples n_{mc} and PC orders n_{po}

	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{E}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	$\epsilon_{\mathbb{V}\text{ar}}$, %	ϵ_f, %	ϵ_f, %	ϵ_f, %	ϵ_f, %
n_{po}	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$	$n_{\text{mc}}=10^2$	$n_{\text{mc}}=10^3$	$n_{\text{mc}}=10^4$	$n_{\text{mc}}=10^5$
1	1.49	0.27	0.15	0.15	44.86	42.41	43.10	46.51	12.45	11.19	9.57	9.21
2	1.42	0.22	0.17	0.14	26.47	8.89	1.56	5.03	11.84	6.22	5.52	4.79
3	1.40	0.21	0.19	0.15	24.90	7.54	4.08	1.39	10.62	2.93	1.66	1.42
4	1.45	0.24	0.16	0.14	24.78	7.57	4.50	1.27	9.92	1.98	0.72	0.40
5	1.38	0.20	0.20	0.16	25.14	7.63	3.41	1.77	10.19	1.90	0.78	0.70
6	1.48	0.25	0.15	0.14	24.64	7.58	4.93	1.34	9.90	2.27	1.14	0.74
7	1.34	0.19	0.23	0.19	24.86	7.48	4.13	1.40	8.47	1.57	1.13	1.24

architecture with a dynamic power profile of $n_{\text{t}} = 10^2$ steps is given in Table II, Table III, and Table IV, which correspond to $\eta = 0$, $\eta = 0.5$, and $\eta = 1$, respectively. Each table contains three subtables: one for $\epsilon_{\mathbb{E}}$ (the left most), one for $\epsilon_{\mathbb{V}\text{ar}}$ (in the middle), and one for ϵ_f (the right most), which gives nine subtables in total. The columns of the tables that correspond to high values of n_{mc} can be used to assess the accuracy of the constructed PC expansions; likewise, the rows that correspond to high values of n_{po} can be used to judge about the sufficiency of the number of MC samples. One can immediately note that, in all the subtables, all the error metrics tend to decrease from the top left corners (low values of n_{po} and n_{mc}) to the bottom right corners (high values of n_{po} and n_{mc}), which suggests that the PC and MC methods converge. There are a few outliers, associated with low PC orders and/or the random nature of sampling, e.g., $\epsilon_{\mathbb{V}\text{ar}}$ increases from 66.13 to 66.70 and ϵ_f from 1.59 to 1.62 when n_{mc} increases from 10^4 and 10^5 in Table III; however, the aforementioned main trend is still clear.

For clarity of the discussions below, we shall primarily focus on one of the tables, namely, on the middle table, Table III, as the case with $\eta = 0.5$ turned out to be the most challenging (explained in Sec. VII-B). The drawn conclusions will be generalized to the other two tables later on.

First, we concentrate on the accuracy of our technique and, thus, pay particular attention the columns of Table III corresponding to high values of n_{mc}. It can be seen that the error $\epsilon_{\mathbb{E}}$ of the expected value is small even for $n_{\text{po}} = 1$: it is bounded by 0.6% (see $\epsilon_{\mathbb{E}}$ for $n_{\text{po}} \geq 1$ and $n_{\text{mc}} \geq 10^4$).

The error $\epsilon_{\mathbb{V}\text{ar}}$ of the second central moment starts from 66.7% for the first-order PC expansions and drops significantly to 5.71% and below for the fourth order and higher (see $\epsilon_{\mathbb{V}\text{ar}}$ for $n_{\text{po}} \geq 4$ and $n_{\text{mc}} = 10^5$). It should be noted, however, that, for a fixed $n_{\text{po}} \geq 4$, $\epsilon_{\mathbb{V}\text{ar}}$ exhibits a considerable decrease even when n_{mc} transitions from 10^4 to 10^5. The rate of this decrease suggests that $n_{\text{mc}} = 10^4$ is not sufficient to reach the same accuracy as the one delivered by the proposed framework, and $n_{\text{mc}} = 10^5$ might not be either.

The results of the third metric ϵ_f allow us to conclude that the PDFs computed by the third-order (and higher) PC expansions closely follow those estimated by the MC technique with large numbers of samples, namely, the observed difference in Table III is bounded by 1.83% (see ϵ_f for $n_{\text{po}} \geq 3$ and $n_{\text{mc}} \geq 10^4$). To give a better appreciation of the proximity of the two methods, Fig. 6 displays the PDFs computed using our

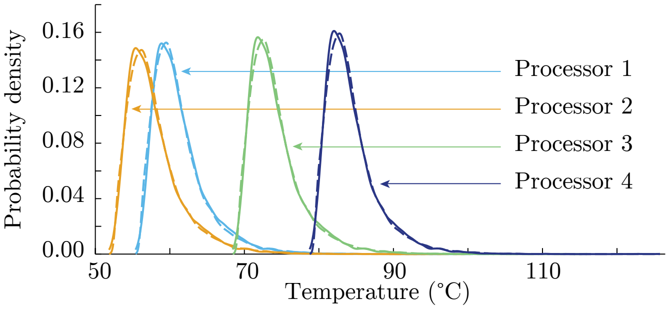

Figure 6. Probability density functions computed at time 50 ms using the proposed framework (the dashed lines) and MC sampling (the solid lines).

Table V
Scaling with respect to the number of processing elements n_{p}

	n_{p}	n_{ξ}	PC, seconds	MC, hours	Speedup, times
$\eta = 0$	2	2	0.16	38.77	8.76×10^5
	4	2	0.16	39.03	8.70×10^5
	8	3	0.27	39.22	5.29×10^5
	16	4	0.83	40.79	1.77×10^5
	32	7	11.02	43.25	1.41×10^4
$\eta = 0.5$	2	3	0.21	38.72	6.54×10^5
	4	5	0.62	38.92	2.28×10^5
	8	6	1.46	40.20	9.94×10^4
	16	10	23.53	41.43	6.34×10^3
	32	12	100.10	43.05	1.55×10^3
$\eta = 1$	2	3	0.20	38.23	6.88×10^5
	4	5	0.56	38.48	2.49×10^5
	8	7	2.47	39.12	5.71×10^4
	16	11	40.55	41.02	3.64×10^3
	32	8	21.41	43.82	7.37×10^3

Table VI
Scaling with respect to the number of steps n_{t}

	n_{t}	PC, seconds	MC, hours	Speedup, times
$\eta = 0$	10	0.01	0.51	1.77×10^5
	10^2	0.02	3.87	7.64×10^5
	10^3	0.16	38.81	8.72×10^5
	10^4	1.58	387.90	8.84×10^5
	10^5	15.85	3877.27	8.81×10^5
$\eta = 0.5$	10	0.02	0.39	6.10×10^4
	10^2	0.07	3.84	2.08×10^5
	10^3	0.52	38.41	2.66×10^5
	10^4	5.31	383.75	2.60×10^5
	10^5	54.27	3903.28	2.59×10^5
$\eta = 1$	10	0.02	0.39	6.15×10^4
	10^2	0.07	3.88	2.05×10^5
	10^3	0.54	38.86	2.60×10^5
	10^4	5.31	390.95	2.65×10^5
	10^5	53.19	3907.48	2.64×10^5

framework for time moment 50 ms with $n_{\mathrm{po}} = 4$ (the dashed lines) along with those calculated by the MC approach with $n_{\mathrm{mc}} = 10^4$ (the solid lines). It can be seen that the PDFs tightly match each other. Note that this example captures one particular time moment, and such curves are readily available for all the other steps of the considered time span.

Now we take a closer look at the convergence of the MC-based technique. With this in mind, we focus on the rows of Table III that correspond to PC expansions of high orders. Similar to the previous observations, even for low values of n_{mc}, the error of the expected values estimated by MC sampling is relatively small, namely, bounded by 1.19% (see $\epsilon_{\mathbb{E}}$ for $n_{\mathrm{po}} \geq 4$ and $n_{\mathrm{mc}} = 10^2$). Meanwhile, the case with $n_{\mathrm{mc}} = 10^2$ has a high error rate in terms of $\epsilon_{\mathbb{V}\mathrm{ar}}$ and ϵ_f: it is above 38% for variance and almost 3.5% for PDFs (see $\epsilon_{\mathbb{V}\mathrm{ar}}$ and ϵ_f for $n_{\mathrm{po}} = 7$ and $n_{\mathrm{mc}} = 10^2$). The results with $n_{\mathrm{mc}} = 10^3$ are reasonably more accurate; however, this trend is compromised by Table IV: 10^3 samples leave an error of more than 7% for variance (see $\epsilon_{\mathbb{V}\mathrm{ar}}$ for $n_{\mathrm{po}} \geq 4$ and $n_{\mathrm{mc}} = 10^3$).

The aforementioned conclusions, based on Table III ($\eta = 0.5$), are directly applicable to Table II ($\eta = 0$) and Table IV ($\eta = 1$). The only difference is that the average error rates are lower when either of the two correlation kernels dominates. In particular, according to $\epsilon_{\mathbb{V}\mathrm{ar}}$, the case with $\eta = 1$, which corresponds to k_{SE}, stands out to be the least error prone.

Guided by the observations in this subsection, we conclude that our framework delivers accurate results starting from $n_{\mathrm{po}} = 4$. The MC estimates, on the other hand, can be considered as sufficiently reliable starting from $n_{\mathrm{mc}} = 10^4$. The last conclusion, however, is biased in favor of the MC technique since, as we noted earlier, there is evidence that 10^4 samples might still not be enough.

B. Computational Speed

In this section, we focus on the speed of our framework. In order to increase the clarity of the comparisons given below, we use the same order of PC expansions and the same number of MC samples in each case. Namely, based on the conclusions from the previous subsection, n_{po} is set to four, and n_{mc} is set to 10^4; the latter also conforms to the experience from the literature [13], [15], [16], [18], [20] and to the theoretical results on the accuracy of MC sampling given in [9].

First, we vary the number of processing elements n_{p}, which directly affects the dimensionality of the uncertain parameters $\mathbf{u} \in \mathbb{R}^{n_{\mathrm{u}}}$ (recall Sec. VI). As before, we shall report the results obtained for various correlation weights η, which impacts the number of the independent variables $\boldsymbol{\xi} \in \mathbb{R}^{n_{\xi}}$, preserved after the model order reduction procedure described in Sec. VI-A and Appendix B. The results, including the dimensionality n_{ξ} of $\boldsymbol{\xi}$, are given in Table V where the considered values for n_{p} are $\{2^n\}_{n=1}^{5}$, and the number of time steps n_{t} is set to 10^3. It can be seen that the correlation patters inherent to the fabrication process [29] open a great possibility for model order reduction: n_{ξ} is observed to be at most 12 while the maximal number without reduction is 33 (one global variable and 32 local ones corresponding to the case with 32 processing elements). This reduction also depends on the floorplans, which is illustrated by the decrease of n_{ξ} when n_{p} increases from 16 to 32 for $\eta = 1$. To elaborate, one floorplan is a four-by-four grid, a perfect square, while the other an eight-by-four grid, a rectangle. Since both are fitted into square dies, the former is spread across the whole die whereas the latter is concentrated along the middle line; the rest is ascribed to the particularities of k_{SE}. On average, the k_{OU} kernel ($\eta = 0$) requires the fewest number of variables while the mixture of k_{SE} and k_{OU} ($\eta = 0.5$) requires the most.[12] It means that, in the latter case, more variables should be preserved in order to retain 99% of the variance. Hence, the case with $\eta = 0.5$ is the most demanding in terms of complexity; see Sec. V-D4.

It is important to note the following. First, since the curse of dimensionality constitutes arguably the major concern of the theory of PC expansions, the applicability of our framework primarily depends on how this curse manifests itself in the problem at hand, i.e., on the dimensionality n_{ξ} of $\boldsymbol{\xi}$. Second, since $\boldsymbol{\xi}$ is a result of the preprocessing stage depending on many factors, the relation between $\mathbf{u}$ and $\boldsymbol{\xi}$ is not straightforward, which is illustrated in the previous paragraph. Consequently, the dimensionality of $\mathbf{u}$ can be misleading when reasoning about the applicability of our technique, and n_{ξ}

[12] The results in Sec. VII-A correspond to the case with $n_{\mathrm{p}} = 4$; therefore, n_{ξ} is two, five, and five for Table II, Table III, and Table IV, respectively.

shown Table V is well suited for this purpose.

Another observation from Table V is the low slope of the execution time of the MC technique, which illustrates the well-known fact that the workload per MC sample is independent of the number of stochastic dimensions. On the other hand, the rows with $n_\xi > 10$ hint at the curse of dimensionality characteristic to PC expansions (see Sec. V-D4). However, even with high dimensions, our framework significantly outperforms MC sampling. For instance, in order to analyze a power profile with 10^3 steps of a system with 32 cores, the MC approach requires more than 40 hours whereas the proposed framework takes less than two minutes (the case with $\eta = 0.5$).

Finally, we investigate the scaling properties of the proposed framework with respect to the duration of the considered time spans, which is directly proportional to the number of steps n_t in the power and temperature profiles. The results for a quad-core architecture are given in Table VI. Due to the long execution times demonstrated by the MC approach, its statistics for high values of n_t are extrapolated based on a smaller number of samples, i.e., $n_\mathrm{mc} \ll 10^4$. As it was noted before regarding the results in Table V, we observe the dependency of the PC expansions on the dimensionality n_ξ of $\boldsymbol{\xi}$, which is two for $\eta = 0$ and five for the other two values of η (see Table V for $n_\mathrm{p} = 4$). It can be seen in Table VI that the computational times of both methods grow linearly with n_t, which is expected. However, the proposed framework shows a vastly superior performance being up to five orders of magnitude faster than MC sampling.

It is worth noting that the observed speedups are due to two major reasons. First of all, PC expansions are generally superior to MC sampling when the curse of dimensionality is suppressed [4], [21], which we accomplish by model order reduction and efficient integration schemes; see Sec. V-D4. The second reason is the particular solution process used in this work to solve the thermal model and construct PC expansions in a stepwise manner; see Sec. V-D2.

VIII. Conclusion

We presented a framework for transient power-temperature analysis (PTA) of electronic systems under process variation. Our general technique was then applied in a context of particular importance wherein the variability of the effective channel length was addressed. Note, however, that the framework can be readily utilized to analyze any other quantities affected by process variation and to study their combinations. Finally, we drew a comparison with MC sampling, which confirmed the efficiency of our approach in terms of both accuracy and speed. The reduced execution times, by up to five orders of magnitude, implied by the proposed framework allow for PTA to be efficiently performed inside design space exploration loops aimed at, e.g., energy and reliability optimization with temperature-related constraints under process variation.

Acknowledgments

We would like to thank Prof. Diana Marculescu and Da-Cheng Juan from Carnegie Mellon University for the valuable suggestions regarding the leakage modeling and the five anonymous reviewers for their precious feedback.

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] Y. Liu, R. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *DATE*, 2007, pp. 1526–1531.

[4] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

[5] R. Rao and S. Vrudhula, "Fast and accurate prediction of the steady-state throughput of multicore processors under thermal constraints," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 28, pp. 1559–1572, October 2009.

[6] D. Rai, H. Yang, I. Bacivarov, J.-J. Chen, and L. Thiele, "Worst-case temperature analysis for real-time systems," in *DATE*, 2011, pp. 631–636.

[7] L. Thiele, L. Schor, H. Yang, and I. Bacivarov, "Thermal-aware system analysis and software synthesis for embedded multi-processors," in *DAC*, 2011, pp. 268–273.

[8] 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.

[9] I. Díaz-Emparanza, "Is a small Monte Carlo analysis a good analysis?" *Statistical Papers*, vol. 43, pp. 567–577, October 2002.

[10] D.-C. Juan, S. Garg, and D. Marculescu, "Statistical thermal evaluation and mitigation techniques for 3D chip-multiprocessors in the presence of process variations," in *DATE*, 2011, pp. 383–388.

[11] 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.

[12] S. Chandra, K. Lahiri, A. Raghunathan, and S. Dey, "Variation-aware system-level power analysis," *IEEE Trans. VLSI Syst.*, vol. 18, pp. 1173–1184, August 2010.

[13] P.-Y. Huang, J.-H. Wu, and Y.-M. Lee, "Stochastic thermal simulation considering spatial correlated within-die proc. variations," in *ASP-DAC*, 2009, pp. 31–36.

[14] R. Ghanem and P. Spanos, *Stochastic Finite Element Method: A Spectral Approach*. Springer Verlag, 1991.

[15] Y.-M. Lee and P.-Y. Huang, "An efficient method for analyzing on-chip thermal reliability considering process variations," *ACM Trans. Design Automation of Elec. Sys.*, vol. 18, pp. 41:1–41:32, July 2013.

[16] R. Shen, N. Mi, S. Tan, Y. Cai, and X. Hong, "Statistical modeling and analysis of chip-level leakage power by spectral stochastic method," in *ASP-DAC*, 2009, pp. 161–166.

[17] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao, "Modeling of intra-die process variations for accurate analysis and optim. of nano-scale circuits," in *DAC*, 2006, pp. 791–796.

[18] S. Bhardwaj, S. Vrudhula, and A. Goel, "A unified approach for full chip statistical timing and leakage analysis of nanoscale circuits considering intradie process variations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, pp. 1812–1825, October 2008.

[19] S. Vrudhula, J. Wang, and P. Ghanta, "Hermite polynomial based interconnect analysis in the presence of process variations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, pp. 2001–2011, October 2006.

[20] P. Ghanta, S. Vrudhula, S. Bhardwaj, and R. Panda, "Stochastic variational analysis of large power grids considering intra-die correlations," in *DAC*, 2006, pp. 211–216.

[21] M. Eldred, C. Webster, and P. Constantine, "Evaluation of non-intrusive approaches for Wiener-Askey generalized polynomial chaos," in *AIAA Non-deterministic Approaches Conference*, 2008.

[22] M. Rosenblatt, "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, vol. 23, pp. 470–472, September 1952.

[23] H. Li, Z. Lü, and X. Yuan, "Nataf transformation based point estimate method," *Chinese Science Bulletin*, vol. 53, pp. 2586–2592, September 2008.

[24] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture: modeling and implementation," *ACM Trans. Architecture and Code Optimization*, vol. 1, pp. 94–125, March 2004.

[25] O. Knio and O. L. Maître, "Uncertainty propagation in CFD using polynomial chaos decomposition," *Fluid Dynamics Research*, vol. 38, pp. 616–640, September 2006.

[26] J. Beck, F. Nobile, L. Tamellini, and R. Tempone, “Implementation of optimal Galerkin and collocation approximations of PDEs with random coefficients,” *ESAIM*, vol. 33, pp. 10–21, October 2011.

[27] F. Heiss and V. Winschel, “Likelihood approximation by numerical integration on sparse grids,” *J. of Econometrics*, vol. 144, pp. 62–80, May 2008.

[28] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, “Modeling within-die spatial correlation effects for process-design co-optimization,” in *Symp. on Quality of Elec. Design*, 2005, pp. 516–521.

[29] 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,” *IEEE Trans. Comput.-Aided Design ICs Syst.*, vol. 30, pp. 388–401, March 2011.

[30] (2013, June) BSIM4. Berkeley Short-channel IGFET Model Group at the University of California, Berkeley. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/

[31] (2013, June) PTM. Nanoscale Integration and Modeling Group at Arizona State University. [Online]. Available: http://ptm.asu.edu/

[32] (2013, June) MATLAB. MathWorks. [Online]. Available: http://www.mathworks.com/products/matlab/

[33] (2013, June) Supplementary materials related to the experimental results. Embedded Systems Laboratory at Linköping University. [Online]. Available: http://www.ida.liu.se/~ivauk83/research/PAPT

[34] M. Hochbruck and A. Ostermann, “Exponential integrators,” *Acta Numerica*, vol. 19, pp. 209–286, May 2010.

[35] J. Burkardt. (2013, June) MATLAB software. The Florida State University. [Online]. Available: http://people.sc.fsu.edu/~jburkardt/

Appendix

A. Thermal Model

In this section, we provide additional details on the thermal model utilized by the proposed framework at **Stage 3** described in Sec. V-C. We use the widespread model based on Fourier’s heat equation [24], which, after a proper spacial discretization, leads to the following system:

$$\begin{cases} \mathbf{C}\,\dfrac{d\tilde{\mathbf{s}}(t)}{dt} + \mathbf{G}\,\tilde{\mathbf{s}}(t) = \tilde{\mathbf{B}}\,\mathbf{p}(t) & \text{(16a)} \\ \mathbf{q}(t) = \tilde{\mathbf{B}}^T\tilde{\mathbf{s}}(t) + \mathbf{q}_{\text{amb}} & \text{(16b)} \end{cases}$$

where the number of differential equations is equal to the number of thermal nodes denoted by n_n; $\mathbf{C} \in \mathbb{R}^{n_\text{n} \times n_\text{n}}$ and $\mathbf{G} \in \mathbb{R}^{n_\text{n} \times n_\text{n}}$ are a diagonal matrix of the thermal capacitance and a symmetric, positive-definite matrix of the thermal conductance, respectively; $\tilde{\mathbf{s}} \in \mathbb{R}^{n_\text{n}}$ is a vector of the difference between the temperature of the thermal nodes and the ambient temperature; $\mathbf{p} \in \mathbb{R}^{n_\text{p}}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{n_\text{n} \times n_\text{p}}$ are a vector of the power dissipation of the processing elements and its mapping to the thermal nodes, respectively; $\mathbf{q} \in \mathbb{R}^{n_\text{p}}$ is a vector of the temperature of the processing elements; and $\mathbf{q}_\text{amb} \in \mathbb{R}^{n_\text{p}}$ is a vector of the ambient temperature. $\tilde{\mathbf{B}}$ distributes power across the thermal nodes. Assuming that one processing element is mapped onto one thermal node, $\tilde{\mathbf{B}}$ is filled in with zeros except for n_p elements equal to unity that are located on the main diagonal. For convenience, we perform an auxiliary transformation of the system in (16) using [8]

$$\mathbf{s} = \mathbf{C}^{\frac{1}{2}}\tilde{\mathbf{s}}, \quad \mathbf{A} = -\mathbf{C}^{-\frac{1}{2}}\mathbf{G}\mathbf{C}^{-\frac{1}{2}}, \quad \text{and} \quad \mathbf{B} = \mathbf{C}^{-\frac{1}{2}}\tilde{\mathbf{B}}$$

and obtain the system in (2) where the coefficient matrix $\mathbf{A}$ preserves the symmetry and positive-definiteness of $\mathbf{G}$. In general, the differential part in (16) (and in (2)) is nonlinear due to the source term $\mathbf{p}(t)$ since we do not make any assumptions about its structure (see the discussion in Sec. V-B). Therefore, there is no closed-form solution to the system.

The time intervals of the power and temperature profiles are assumed to be short enough such that the total power of a processing element can be approximated by a constant within one interval. In this case, (16a) (and (2a)) is a system of linear differential equations that can be solved analytically. The solution is as follows [8]:

$$\mathbf{s}(t) = \mathbf{E}(t)\,\mathbf{s}(0) + \mathbf{F}(t)\,\mathbf{p}(0) \tag{17}$$

where t is restricted to one time interval, $\mathbf{p}(0)$ is the power dissipation at the beginning of the time interval with respect to the corresponding temperature,

$$\begin{aligned} \mathbf{E}(t) &= e^{\mathbf{A}t} \in \mathbb{R}^{n_\text{n} \times n_\text{n}}, \quad \text{and} \\ \mathbf{F}(t) &= \mathbf{A}^{-1}(e^{\mathbf{A}t} - \mathbf{I})\,\mathbf{B} \in \mathbb{R}^{n_\text{n} \times n_\text{p}}. \end{aligned}$$

The procedure is to be repeated for all n_t time intervals starting from the initial temperature, which, without loss of generality, is assumed to be equal to the ambient temperature. Note that, when the power profile is evenly sampled, the coefficient matrices $\mathbf{E}(t)$ and $\mathbf{F}(t)$ are constant and can be efficiently computed using the technique in [8]. It is also worth noting that the described solution method belongs to the family of so-called exponential integrators, which have good stability properties; refer to [34] for an overview. Finally, taking into account $\mathbf{u}$, we obtain (3), operating on stochastic quantities.

B. Discrete Karhunen-Loève (KL) Decomposition

This section contains a description of the discrete Karhunen-Loève decomposition [14], which is utilized at **Stage 1**. We shall use the notation introduced in Sec. VI-A. Let $\boldsymbol{\Sigma}_{\boldsymbol{\xi}'}$ be the covariance matrix of the centered random vector $\boldsymbol{\xi}'$ (which is the result of the first step of the Nataf transformation discussed in Sec. VI-A). Since any covariance matrix is real and symmetric, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}'}$ admits the eigenvalue decomposition as $\boldsymbol{\Sigma}_{\boldsymbol{\xi}'} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ where $\mathbf{V}$ and $\boldsymbol{\Lambda}$ are an orthogonal matrix of the eigenvectors and a diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\xi}'}$, respectively. $\boldsymbol{\xi}'$ can then be represented as $\boldsymbol{\xi}' = \mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\xi}''$ where the vector $\boldsymbol{\xi}''$ is centered, normalized, and uncorrelated, which is also independent as $\boldsymbol{\xi}'$ is Gaussian.

The aforementioned decomposition provides means for model order reduction. The intuition is that, due to the correlations possessed by $\boldsymbol{\xi}' \in \mathbb{R}^{n_\text{u}}$, this vector can be recovered from a small subset $\boldsymbol{\xi}''' \in \mathbb{R}^{n_\xi}$ of $\boldsymbol{\xi}'' \in \mathbb{R}^{n_\text{u}}$, $n_\xi \ll n_\text{u}$. Such redundancies can be revealed by analyzing the eigenvalues λ_i stored in $\boldsymbol{\Lambda}$. Assume λ_i, $\forall i$, are arranged in a non-increasing order and let $\tilde{\lambda}_i = \lambda_i / \sum_j \lambda_j$. Gradually summing up the arranged and normalized eigenvalues $\tilde{\lambda}_i$, we can identify a subset of them that has the cumulative sum greater than a certain threshold. When this threshold is sufficiently high (close to one), the rest of the eigenvalues and the corresponding eigenvectors can be dropped as being insignificant, reducing the stochastic dimensionality of the problem.

C. Polynomial Chaos (PC) Expansions

Here we elaborate on the orthogonality property [4] of PC expansions, which is extensively utilized at **Stage 4** in

Sec. V-D. Due to the inherent complexity, uncertainty quantification problems are typically viewed as approximation problems. More precisely, one usually constructs computationally efficient surrogates of the initial models and then studies these light representations instead. PC expansions [4] are one way to perform such approximations, in which the approximating functions are orthogonal polynomials. A set of multivariate polynomials $\{\psi_i : \mathbb{R}^{n_\xi} \to \mathbb{R}\}$ is orthogonal if

$$\langle \psi_i, \psi_j \rangle = \nu_i \delta_{ij}, \qquad \forall i, j, \tag{18}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the Hilbert space spanned by the polynomials, δ_{ij} is the Kronecker delta function, and $\nu_i = \langle \psi_i, \psi_i \rangle$ is a normalization constant. The inner product with a weight function $f : \mathbb{R}^{n_\xi} \to \mathbb{R}$ is defined as the following multidimensional integral:

$$\langle h, g \rangle := \int h(\mathbf{x})\, g(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x}. \tag{19}$$

In our context, the weight function corresponds to the PDF of $\boldsymbol{\xi}$ (see Sec. V-A). For $h = \psi_i$ and $g = \psi_j$, the inner product yields the covariance of ψ_i and ψ_j, and the presence of orthogonality is equivalent to the absence of correlations.

Many of the most popular probability distributions directly correspond to certain families of the orthogonal polynomials given in the Askey scheme [4]. A probability distribution that does not have such a correspondence can be transformed into one of those that do have using the technique shown in Sec. VI-A. Another solutions is to construct a custom polynomial basis using the Gram-Schmidt process. In addition, apart from continuous, PC expansions can be applied to discrete distributions. Refer to [4] for further discussions.

D. Numerical Integration

As mentioned in Sec. V-D, Sec. VI-D, and Appendix C, the coefficients of PC expansions are integrals, which should be calculated numerically at **Stage 4**. In numerical integration, an integral of a function is approximated by a summation over the function values computed at a set of prescribed points, or nodes, which are multiplied by the corresponding prescribed weights. Such pairs of nodes and weights are called quadrature rules. A one-dimensional quadrature rule is characterized by its precision, which is defined as the maximal order of polynomials that the rule integrates exactly [27]. In multiple dimensions, an n_ξ-variate quadrature rule is formed by tensoring one-dimensional counterparts. Such a multidimensional rule is characterized by its accuracy level n_ql, which is defined as the index of the rule in the corresponding family of multidimensional rules with increasing precision.

It can be seen in (19) that the integrand can be decomposed into two parts: the weight function f and everything else. The former always stays the same; therefore, a rule is typically chosen in such a way that this "constant" part is automatically taken into consideration by the corresponding weights since there is no point of recomputing f each time when the other part, i.e., the functions that the inner product operates on, changes. In this regard, there exist different families of quadrature rules tailored for different weight functions. Define such a quadrature-based approximation of (19) by

$$\langle h, g \rangle \approx \mathcal{Q}^{n_\xi}_{n_\mathrm{ql}}[h\, g] := \sum_{i=1}^{n_\mathrm{qp}} h(\hat{\mathbf{x}}_i)\, g(\hat{\mathbf{x}}_i)\, w_i \tag{20}$$

where $\hat{\mathbf{x}}_i \in \mathbb{R}^{n_\xi}$ and $w_i \in \mathbb{R}$ are the prescribed points and weights, respectively; n_qp is their number; and n_ql is the accuracy level of the quadrature rule, which is said to be n_ξ-variate. It is important to note that $\hat{\mathbf{x}}_i$ and w_i do not change when the quantity being integrated changes. Thus, once the rule to use has been identified, it can be utilized to compute the inner product of arbitrary h and f with no additional computational effort. In our experiments in Sec. VII, we use the library of quadrature rules available at [35].

Since in the example in Sec. VI we need to compute the inner product with respect to beta measures, the Gauss-Jacobi quadrature rule is of particular interest. The rule belongs to a broad class of rules known as Gaussian quadratures. The precision of a one-dimensional Gaussian quadrature with $\tilde{n}_\mathrm{qp}$ points is $2\tilde{n}_\mathrm{qp} - 1$; this feature makes such quadratures especially efficient [27]. Using (20), we rewrite (8) as shown in (9) where $\{\nu_i\}_{i=1}^{n_\mathrm{pc}}$ are computed exactly, either by applying the same quadrature rule or by taking products of the one-dimensional counterparts with known analytical expressions [4]; the result is further tabulated. It is important to note that n_ql should be chosen in such a way that the rule is exact for polynomials of the total order at least $2n_\mathrm{po}$, i.e., twice the order of PC expansions, which can be seen in (8) [21]. Therefore, $n_\mathrm{ql} \geq n_\mathrm{po} + 1$ as the quadrature is Gaussian.

There is one more and arguably the most crucial aspect of numerical integration that we ought to discuss: the algorithm used to construct multidimensional quadratures. In low dimensions, the construction can be easily based on the direct tensor product of one-dimensional rules. However, in high dimensions, the situation changes dramatically as the number of points produced by this approach can easily explode. For instance [27], if a one-dimensional rule has only four nodes, i.e., $\tilde{n}_\mathrm{qp} = 4$, then in 10 stochastic dimensions, i.e., $n_\xi = 10$, the number of multivariate nodes becomes $n_\mathrm{qp} = \tilde{n}_\mathrm{qp}^{n_\xi} = 4^{10} = 1\ 048\ 576$, which is not affordable. Moreover, it can be shown that most of the points obtained in such a way do not contribute to the asymptotic accuracy and, therefore, are a waste of time. In order to effectively alleviate this problem, we construct so-called sparse grids using the Smolyak algorithm [21], [27], [35]. The algorithm preserves the accuracy of the underlying one-dimensional rules for complete polynomials while significantly reducing the number of integration nodes. For instance, in the example given earlier, the number of points computed by the algorithm would be only 1 581, which implies drastic savings of the computational time.