Statistical library characterization using belief propagation across multiple technology nodes

Citation: Li Yu, Sharad Saxena, Christopher Hess, Ibrahim (Abe) M. Elfadel, Dimitri Antoniadis, and Duane Boning. 2015. Statistical library characterization using belief propagation across multiple technology nodes. In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE '15). EDA Consortium, San Jose, CA, USA, 1383-1388.

As Published: http://dl.acm.org/citation.cfm?id=2757012.2757134

Publisher: Association for Computing Machinery (ACM)

Persistent URL: http://hdl.handle.net/1721.1/96913

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Statistical Library Characterization Using Belief Propagation across Multiple Technology Nodes

Li Yu, Sharad Saxena[1], Christopher Hess[1], Ibrahim (Abe) M. Elfadel[2], Dimitri Antoniadis, Duane Boning
Massachusetts Institute of Technology, [1]PDF Solutions, [2]Masdar Institute of Science and Technology
Email: yul09@mit.edu

***Abstract*—In this paper, we propose a novel flow to enable computationally efficient statistical characterization of delay and slew in standard cell libraries. The distinguishing feature of the proposed method is the usage of a limited combination of output capacitance, input slew rate and supply voltage for the extraction of statistical timing metrics of an individual logic gate. The efficiency of the proposed flow stems from the introduction of a novel, ultra-compact, nonlinear, analytical timing model, having only four universal regression parameters. This novel model facilitates the use of maximum-a-posteriori belief propagation to learn the prior parameter distribution for the parameters of the target technology from past characterizations of library cells belonging to various other technologies, including older ones. The framework then utilises Bayesian inference to extract the new timing model parameters using an ultra-small set of additional timing measurements from the target technology. The proposed method is validated and benchmarked on several production-level cell libraries including a state-of-the-art 14-nm technology node and a variation-aware, compact transistor model. For the same accuracy as the conventional lookup-table approach, this new method achieves at least 15x reduction in simulation runs.**

I. Introduction

A standard cell library capturing statistical information of delay and output slew variations is at the core of statistical static timing analysis (SSTA), and, cost efficient statistical characterization of such libraries has become essential. The most widely used statistical library cell characterization method is based on the look-up table (LUT) approach where gate propagation delay (t_d), output transition time (S_{out}) and their variations are stored in a look-up table with different combinations of inputs such as cell types, input slew (S_{in}), load capacitance (C_{load}), supply voltage (V_{dd}), and other parameters [1].

The runtime complexity required for such a statistical LUT-based approach is $O(N_{sample} \cdot N_{LUT})$, where N_{sample} is the number of SPICE runs needed to obtain each mean and variance value and N_{LUT} is the number of input vector combinations. This approach will quickly become infeasible as either N_{LUT} or N_{sample} in a technology increases. Historically, circuit level Monte Carlo (MC) simulation has been employed to generate a number of samples in the process parameter probability space [2]. Such approach allows variability-aware analysis to be implemented with minor changes on top of existing characterization tools but requires a large number of MC runs. To address this challenge, several approaches based on sensitivity analysis for library characterization have been proposed by EDA vendors. For instance, Composite Current Source (CSC) is adopted by the Synopsys PrimeTime SSTA tool and sensitivity-based effective-current-source-model (S-ECSM) is adopted by the Cadence statistical tool. All of these approaches aim at modelling the statistical impact of process parameter variations as a linear superposition of the impact of each parameter in the response model of the affected metric. Several Response Surface Methodologies (RSMs) have also been proposed to explore the sparsity of the process regression coefficients. An example of such a strategy is Least-Angle Regression (LAR) which uses L_1-norm regularization [3]. One major benefit of regularizing with the L_1-norm is that it results in sample complexity that is logarithmic in the number of features (e.g., principal components). For statistical characterization of standard cells, an error propagation technique using linear sensitivity analysis and Response Surface Methodology (RSM) using Brussel Design of Experiments (DoE) was proposed for library characterization in [4]. The Brussel DoE performs statistical feature selection keeping only those features that are most relevant to the response under consideration. Then it uses a model selection algorithm to build a suitable regression model for all the responses. More Recently, several statistical circuit simulator based on uncertainty quantification have been successfully applied to avoid the huge number of repeated simulations in conventional Monte Carlo flows [5]–[9].

On the other hand, the expensive simulation cost of the statistical LUT-based approach is not only due to high dimensionality of the process space, but also due to high dimensionality of the cell input space (e.g., cell type, input slew S_{in}, load capacitance, supply voltage V_{dd}, $etc.$). This problem is further exacerbated as more design options are provided in recent technologies (e.g., multi-V_t, multi-V_{dd}). While most of the existing work focuses on exploring the sparsity of the regression coefficients of the process space with a reduced process sample size for each input space vector, correlations between different cells and different input vectors within the same cell have not been considered in the open literature, to the best of our knowledge. This has been the main motivation of this work which proposes a novel acceleration method that operates in the library input space rather than its process space and that can be added to any acceleration used in the process space.

This is achieved through the systematic use of recent advances in statistics and semiconductor metrology that we apply to the development of computationally efficient statistical characterization algorithms for standard cell libraries. We propose two key techniques to explore correlations in library input space. The first is a novel ultra-compact, analytical model for gate timing characterisation, and the second is a Bayesian learning algorithm for the parameters of the aforementioned timing model using past library characterizations along with a very small set of additional simulations from the target technology. Bayesian approaches were initially introduced in the area of VLSI design for post-Silicon validation and parameter extraction [10]–[15]. The intrinsic simplicity of the proposed timing model combined with the Bayesian learning [16] framework is capable of building very accurate circuit response representations.

The rest of this paper is organized as follows. Section 2 introduces basic notation and formulates the problem of statistical characterization in *library input space*. Section 3 describes prior work on gate delay modelling and presents our novel ultra-compact analytical model for gate delay and slew. Section 4 presents our Bayesian algorithm which learns timing model parameters from past library characterizations and a very small set of additional simulation runs in the target technology. The foundation of this algorithm is the use of maximum-a-posteriori (MAP) estimation. In Section 5, our new methods are validated on the library characterization in state-of-the-art 14-nm and

28-nm technology and compared with the LUT method. Our conclusions are given in Section 6.

II. Problem Formulation

In library characterization, an accurate model for cell delay (T_d) and output slew (S_{out}) is developed given the following input data: a cell type, input slew (S_{in}), output load capacitance (C_{load}), transition direction (RISE/FALL), and supply voltage (V_{dd}). To formalize the library characterization problem, we consider an individual logic gate with multiple inputs and one output, and for simplicity, we start from the standard assumption that only one timing arc is modelled at a time, which implies that we do not consider simultaneous input switching. For p input variables ($\xi = \{\xi_1, \xi_2, ..., \xi_p\}$), such as $S_{in}, V_{dd}, C_{load}, etc.$, the cell response is modeled as the following two functions:

$$T_d = f_T(\xi_1, \xi_2, ..., \xi_p) \quad (1)$$

$$S_{out} = f_S(\xi_1, \xi_2, ..., \xi_p) \quad (2)$$

The problem of nominal library characterisation is to estimate f_T and f_S given k input vectors $\{\xi\} = \{\xi^{(1)}, \xi^{(2)}, ..., \xi^{(k)}\}$ and k output observations $\{T_d^{(1)}, T_d^{(2)}, ..., T_d^{(k)}\}$ and $\{S_{out}^{(1)}, S_{out}^{(2)}, ..., S_{out}^{(k)}\}$, such that the timing prediction error with respect to a baseline case is minimized under the condition that k is very small. The nominal baseline case is defined by SPICE simulations under n different input vectors ($n >> k$) sampled randomly within the input space ξ.

Let us now denote by $\{T_d\}$ an ensemble of delay observations. This ensemble has been generated for a given input vector but under varying process parameters. Now we formulate the problem of statistical library characterisation in *input space* as that of estimating f_T and f_S given k input vectors $\{\xi\} = \{\xi^{(1)}, \xi^{(2)}, ..., \xi^{(k)}\}$ and k *ensembles* of output observations $\{\{T_d^{(1)}\}, \{T_d^{(2)}\}, ..., \{T_d^{(k)}\}\}$ and $\{\{S_{out}^{(1)}\}, \{S_{out}^{(2)}\}, ..., \{S_{out}^{(k)}\}\}$, such that the prediction error for the statistical metrics with respect to a statistical baseline case is minimized under the condition that k is very small. The statistical baseline case is defined by statistical SPICE simulations using the same n different input vectors ($n >> k$) as the nominal baseline case, where the SPICE simulations are now executed according to the Monte Carlo method in process space. The metrics of the statistical baseline case include the mean and standard deviation of delay and output slew at each input vector $i \in \{1, ..., n\}$. They are denoted as $\mu_{T_d}^{(i)}$, $\mu_{S_{out}}^{(i)}$ and $\sigma_{T_d}^{(i)}$, $\sigma_{S_{out}}^{(i)}$ $(i = 1, 2, ..., n)$, respectively.

III. Model for Delay and Output Slew

Accurate gate level modeling for delay and slew estimation has become a major challenge for nanometric technologies. Historically, the transistor delay has been simply approximated by $C_{load}V_{dd}/I_{dsat}$, where I_{dsat} is the drain current at $V_{gs} = V_{ds} = V_{dd}$. A more accurate model, named the alpha-power law, was later proposed in the early 1990s [17] where a closed-form expression was derived for the delay of an inverter. A simplified version of the alpha-power law was proposed in [18]. More recently, a simple analytical expression for the intrinsic MOSFET delay, using physics-based models for the effective current and the total gate switching charge, was proposed to better describe nanometric technologies [19], [20].

Although these advanced delay models provide accurate description of transition activity in the cell, they are still quite complex, and detailed process information is required to fit the entire model.

Our first goal therefore is to contribute an ultra compact timing model that is at once a generalisation of older models but whose parameters allow a sparse representation of input space vectors. Fig. 1 (a) shows the key factors that affect the delay and output slew of an inverter. In this work, we consider the impact of input slew (S_{in}), output load capacitance (C_{load}), supply voltage (V_{dd}), and driving strength (I_{eff}).

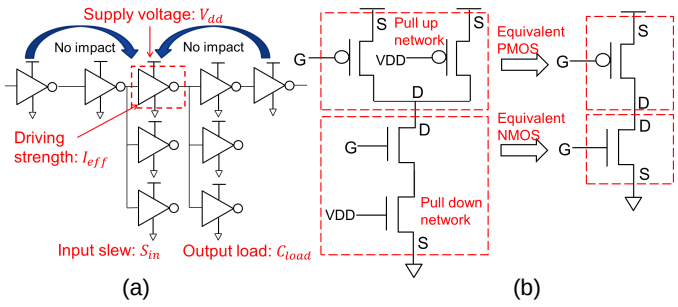

Fig. 1. (a) Key factors that affect the delay and output slew of an inverter; (b) NAND2 equivalent inverter: The pull-up network is replaced with an "equivalent" PMOS while the pull-down network is replaced with an "equivalent" NMOS device.

To find our ultra compact model, we first study gate delay in a simple inverter and generalize it to any combinatorial logic cell. Recent studies [21]–[23] show that the simple $C_{load}V_{dd}/I_{dsat}$ metric follows the experimental inverter delay much better if the on-current in the denominator is replaced with an effective current I_{eff} representing the average switching current. In line with the intrinsic transistor delay defined in [19], we model cell delay as

$$T_d = k_d \frac{\Delta Q}{I_{eff}} \quad (3)$$

where k_d is a scaling factor used to obtain a good fit to the actual cell delays. I_{eff} is defined as

$$I_{eff} = \frac{I_d(V_{gs} = V_{dd}, V_{ds} = \frac{V_{dd}}{2}) + I_d(V_{gs} = \frac{V_{dd}}{2}, V_{ds} = V_{dd})}{2} \quad (4)$$

and can be evaluated easily through performance modeling or through a circuit simulation that takes into account process variations [19], [24]. Since our focus is to model delay and output slew as functions of input variables, (1) and (2), we assume we know I_{eff} for each input vector. Note that the direct link between process parameters and delay is still preserved in the I_{eff} current. To generalize the above model to any combinatorial logic cell, we simply next each gate onto an "equivalent inverter" and use the inverter characterization to estimate delays and output slews [25]–[27]. Fig. 1(b) shows the equivalent inverter of a NAND2 where the pull-up network is replaced with a PMOS and the pull-down network is replaced with an NMOS device. The charge transferred to or from the load capacitance during switching is equal to

$$\Delta Q = (V_{dd} + V')(C_{load} + C_{par} + \alpha S_{in}) \quad (5)$$

where C_{par}, V' and α are all fitting parameters. Compared with the simple $C_{load}V_{dd}/I_{dsat}$ metric, several effects have been considered: (1) C_{par} is introduced to account for parasitic capacitance, such as those associated with junctions and interconnects, which are not included in C_{load}; (2) V' is introduced to compensate for the inaccuracy of the delay model at low V_{dd}; and (3) a linear coefficient α is introduced to account for S_{in}'s impact on delay. The estimates of f_T and f_S are then converted to parameter extraction problems for $\{k_d, C_{par}, V', \alpha\}$.

A special feature of this simple delay model is that the same format is used to describe not only delay but also output slew S_{out} albeit with a different set of values for the fitting

parameters $\{k_d, C_{par}, V', \alpha\}$. To validate the proposed model, $T_d \cdot I_{eff}/(V_{dd}+V')$ and $S_{out} \cdot I_{eff}/(V_{dd}+V')$ versus different V_{dd} values are shown in Fig. 2, where T_d and S_{out} are simulated through SPICE using a 14-nm industrial design kit and two separate V' values are extracted for T_d and S_{out}. For different groups of C_{load} and S_{in} combinations, a constant value of $T_d \cdot I_{eff}/(V_{dd}+V')$ and $S_{out} \cdot I_{eff}/(V_{dd}+V')$ is observed under different V_{dd}.

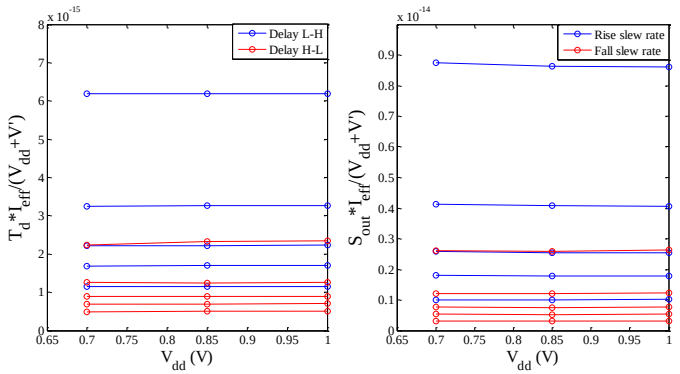

Fig. 2. For a NOR2 cell designed in a commercial state-of-the-art 14-nm technology, a constant value of $T_d \cdot I_{eff}/(V_{dd}+V')$ and $S_{out} \cdot I_{eff}/(V_{dd}+V')$ is observed versus different V_{dd} and RISE/FALL combinations.

Fig. 3 shows $T_d/(C_{load}+C_{par}+\alpha S_{in})$ and $S_{out}/(C_{load}+C_{par}+\alpha S_{in})$ versus different C_{load} and S_{in} combinations. A similar result is observed here that for different V_{dd} and transition (RISE/FALL) combinations, $T_d/(C_{load}+C_{par}+\alpha S_{in})$ and $S_{out}/(C_{load}+C_{par}+\alpha S_{in})$ are approximately constant.

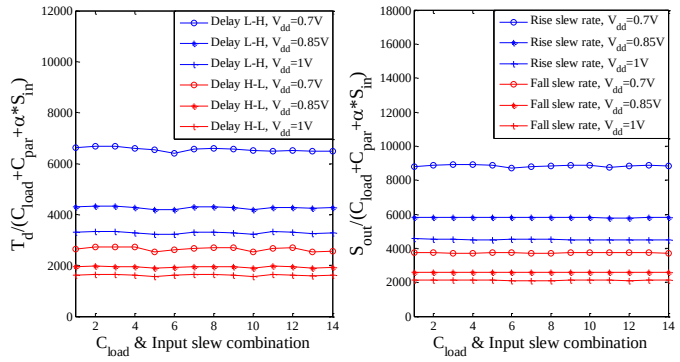

Fig. 3. For a NOR2 cell designed in a commercial state-of-the-art 14-nm technology, a constant value of $T_d/(C_{load}+C_{par}+\alpha S_{in})$ and $S_{out}/(C_{load}+C_{par}+\alpha S_{in})$ is observed versus different C_{load}, S_{in} and RISE/FALL combinations.

TABLE I
EXTRACTED PARAMETERS FOR DELAY MODEL FROM INV, NAND2 AND NOR2 IN THREE DIFFERENT TECHNOLOGIES WITH THEIR FITTING ERROR.

Tech	Cell	k_d	$C_{par}(fF)$	$V'(V)$	α	% error
A	INV	0.389	0.951	-0.266	0.092	1.56%
A	NAND2	0.372	1.328	-0.209	0.034	1.98%
A	NOR2	0.356	1.186	-0.241	0.102	0.91%
B	INV	0.416	1.046	-0.287	0.103	1.50%
B	NAND2	0.403	1.471	-0.228	0.034	2.05%
B	NOR2	0.374	1.276	-0.253	0.104	1.12%
C	INV	0.389	0.978	-0.272	0.107	1.84%
C	NAND2	0.383	1.12	-0.258	0.050	1.94%
C	NOR2	0.368	1.225	-0.264	0.117	1.47%

Table I shows extracted parameters for delay model from INV, NAND2 and NOR2 in three different technologies with their fitting errors. Strong similarities in extracted parameters are observed among different cells and technologies from different nodes, which serves as a basis for minimizing the required input combinations in statistical characterization in the next section. Although our proposed model captures major physical effects, for some technologies there might be an offset between the proposed model and circuit simulations. In those cases, extra fitting terms (e.g., $S_{in} \cdot C_{load}$) might be needed. The optimal model complexity will be given by a trade-off between model accuracy of degree of data compression.

IV. BAYESIAN INFERENCE WITH MAXIMUM A POSTERIORI (MAP) ESTIMATION

In this section, we present a Bayesian inference approach with maximum-a-posteriori (MAP) estimation where instead of computing $\{T_d, S_{out}\}$ at each input condition separately, we will estimate $\{k_d, C_{par}, V', \alpha\}$ globally by maximizing the joint probability of observing $(\xi^{(i)}, T_d^{(i)})$ or $(\xi^{(i)}, S_{out}^{(i)})$, $(i = 1, 2, ..., k)$.

The first step is to transfer observed training samples $(\xi^{(i)}, T_d^{(i)})$ or $(\xi^{(i)}, S_{out}^{(i)})$, $(i = 1, 2, ..., k)$ to parameter subspace $\{k_d, C_{par}, V', \alpha\}$ and use both to derive a probability distribution on the parameter space. The *pdf*'s on $\{k_d, C_{par}, V', \alpha\}$ for delay and output slew can then be calculated and the parameter extraction problem solved using maximum a posteriori (MAP) estimation.

Without loss of generality, we describe the MAP estimation for delay parameter subgroup $\mathbf{P_T} = \{k_d, C_{par}, V', \alpha\}$. Parameters for output slew are estimated in a similar manner.

First, we assume that $\mathbf{P_T}$ follows a Gaussian distribution $\mathbf{P_T} \sim \mathcal{N}(\mu_{\mathbf{P_T}}, \mathbf{\Sigma_{P_T}})$:

$$pdf(\mathbf{P_T}) = \frac{1}{4\pi^2\sqrt{|\Sigma_{P_T}|}} \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_T} - P_T)^{\mathbf{T}}\mathbf{\Sigma}_{P_T}^{-1}(\boldsymbol{\mu}_{P_T} - P_T)] \tag{6}$$

where $\mu_{\mathbf{P_T}}$ and $\mathbf{\Sigma_{P_T}}$ are the mean vector and covariance matrix of the parameter subgroup $\mathbf{P_T}$, respectively. Next, we assume that the $\boldsymbol{\mu}_{P_T}$ follows a conjugate Gaussian prior distribution $\boldsymbol{\mu}_{P_T} \sim \mathcal{N}(\boldsymbol{\mu}_{t0}, \mathbf{\Sigma}_{t0})$.

$$pdf(\boldsymbol{\mu}_{P_T}) = \frac{1}{4\pi^2\sqrt{|\mathbf{\Sigma_{t0}}|}} \cdot exp[-\frac{1}{2}(\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0})^{\mathbf{T}}\mathbf{\Sigma}_{t0}^{-1}(\boldsymbol{\mu}_{P_T} - \boldsymbol{\mu}_{t0})] \tag{7}$$

where $\mu_{\mathbf{t0}}$ and $\mathbf{\Sigma_{t0}}$ are the mean vector and covariance matrix of $\mu_{\mathbf{P_T}}$, respectively. We also define the delay model precision as $\beta_{f_{T_d}}$, which equals the inverse variance of modeling errors across different technologies. Given $\boldsymbol{\mu}_{P_T}$ and $\beta_{f_{T_d}}$, we calculate the likelihood of observing delay at ith input condition $T_d^{(i)}$ associated with subspace distribution $pdf(\boldsymbol{P_T})$ as

$$\begin{aligned} pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T}, \beta_{f_{T_d}}(\xi^{(i)})) &= \sqrt{\frac{\beta_{T_d}(\xi^{(i)})}{2\pi}} \\ &\cdot exp[-\frac{1}{2}(T_d^{(i)} - f_T(\xi^{(i)}, \boldsymbol{\mu}_{P_T}))^2\beta_{f_{T_d}}(\xi^{(i)})] \end{aligned} \tag{8}$$

The learning of precision $\beta_{f_{T_d}}$ is a key step in this method. In practice, $\beta_{f_{T_d}}$ represents our "uncertainty" on proposed delay model at different input conditions due to its inability of capturing certain physical effects. While they depend on the details of the technologies, these precisions show a strong systematic trend across different input conditions ξ. In this work, extracted parameters $\mu_{\mathbf{P_T}}$ from past technologies are used to learn the systematic precision $\beta_{f_{T_d}}$ at different input conditions. Characterizations from a variety of technology nodes enable us to propagate our historical belief to a new technology node. While generic or broad historical technologies can be used to learn approximate precisions, in order to achieve the highest applicable prior precision, the best historical technologies would

be those with the same design or process choices as the target technology. For example, if we intend to fit a library in a low power process, appropriate historical technologies would also be technologies in low power processes. Therefore a bias-variance tradeoff is needed in the selection of historical libraries.

The detailed learning process proceeds as follows. First, a full set of standard cell libraries in N_{tech} fabrication processes and technology nodes ($N_{tech} = 6$ in this paper, including technologies from 14-nm to 45-nm, with both bulk-Silicon and SOI technologies and non-FINFET and FINFET technologies) are employed as "historical data" to improve our confidence in predicting $\beta_{f_{T_d}}$ on an unknown library. This assumes that although a new technology introduces different lithography, structures and materials, parameters from our proposed delay model do not change much, as is shown in Section 3. After selection of a group of historical libraries, each cell is fitted into the proposed delay model with different input conditions ξ. $\beta_{f_{T_d}}$ is then calculated by the inverse variance of relative difference between measurements and delay model predictions using extracted parameters.

$$\beta_{f_{T_d}} = \frac{1}{\frac{1}{N_{tech}}\sum_{j=1}^{N_{tech}}(\frac{T_d^{(j)}-f_T(P_T^{(j)})}{T_d^{(j)}})^2-(\frac{1}{N_{tech}}\sum_{j=1}^{N_{tech}}\left|\frac{T_d^{(j)}-f_T(P_T^{(j)})}{T_d^{(j)}}\right|)^2} \tag{9}$$

After the estimation of likelihood and precision, we are able to transfer delay characterization $\{T_d^{(1)},...,T_d^{(k)}\}$ to parameter subspace $\mathbf{P_T}$ and obtain the conditional probability of observing $T_d^{(i)}$ given $\boldsymbol{\mu}_{P_T}$ and $\beta_{f_{T_d}}(\xi^{(i)})$. We then combine this conditional probability with the prior distribution $pdf(\boldsymbol{\mu}_{P_T})$ in (7) to accurately estimate $\boldsymbol{\mu}_{P_T}$. Assuming each delay simulation is ideal, we can write the likelihood function $pdf(T_d|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}})$ as:

$$pdf(T_d|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}}) = \prod_{i=1}^{k} pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}}(\xi^{(i)})) \tag{10}$$

According to Bayes' theory, the conditional distribution $pdf(\boldsymbol{\mu}_{P_T}|T_d)$ is proportional to the product of the prior $pdf(\boldsymbol{\mu}_{P_T})$ and the likelihood function $pdf(T_d|\boldsymbol{\mu}_{P_T})$:

$$pdf(\boldsymbol{\mu}_{P_T}|T_d) \propto pdf(\boldsymbol{\mu}_{P_T}) \cdot pdf(T_d|\boldsymbol{\mu}_{P_T}) \tag{11}$$

The precision $\beta_{f_{T_d}}$ is learned from historical cell delay characterization and is therefore independent of the observation T_d. Consequently,

$$pdf(T_d|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}}) = pdf(T_d|\boldsymbol{\mu}_{P_T}) \tag{12}$$

Substituting (10) and (12) into (11) yields:

$$pdf(\boldsymbol{\mu}_{P_T}|T_d) \propto pdf(\boldsymbol{\mu}_{P_T}) \cdot \prod_{i=1}^{k} pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}}(\xi^{(i)})) \tag{13}$$

The last step is maximum-a-posteriori (MAP) estimation to find optimal estimates of $\boldsymbol{\mu}_{P_T}$ that maximize the log likelihood of the posterior distributions $\ln pdf(\boldsymbol{\mu}_{P_T}|T_d)$. It can be mathematically formulated as an optimization problem

$$\underset{\boldsymbol{\mu}_{P_T}}{maximize} \ln pdf(\boldsymbol{\mu}_{P_T}) + \sum_{i=1}^{k} \ln pdf(T_d^{(i)}|\boldsymbol{\mu}_{P_T},\beta_{f_{T_d}}(\xi^{(i)})) \tag{14}$$

Substituting (7) and (8) into (14) and removing the constant items yield:

$$\underset{\boldsymbol{\mu}_{P_T}}{minimize} \frac{1}{2}(\boldsymbol{\mu}_{P_T}-\boldsymbol{\mu}_{t0})^{\mathbf{T}}\boldsymbol{\Sigma}_{t0}^{-1}(\boldsymbol{\mu}_{P_T}-\boldsymbol{\mu}_{t0}) + \frac{1}{2}\sum_{i=1}^{k}(T_d^{(i)}-f_T(\xi^{(i)},\boldsymbol{\mu}_{P_T}))^2\beta_{f_{T_d}}(\xi^{(i)}) \tag{15}$$

where (15) is the summation of a concave quadratic function. Hence the optimization problem in (15) is also a convex programming problem and can be solved both efficiently and robustly.

So far we have achieved individual library cell characterization (no statistical characterization included). The detailed efficient statistical library cell characterization proceeds as follows. N_{sample} different seeds for each cell under process variation are generated through Monte Carlo (MC) simulation or Design of Experiments (DoE) [4]. For jth seed in each cell, $\{T_d\}$ and $\{S_{out}\}$ under k input conditions are simulated through a SPICE simulation using .ALTER statement. $P_T^{(j)}$ and $P_S^{(j)}$ are extracted through proposed Bayesian inference with maximum-a-posteriori (MAP) estimation for jth seed. For a targeted input condition ξ, the probability distribution of delay and output slew are calculated as $pdf(f_T(\xi,P_T))$ and $pdf(f_S(\xi,P_S))$.

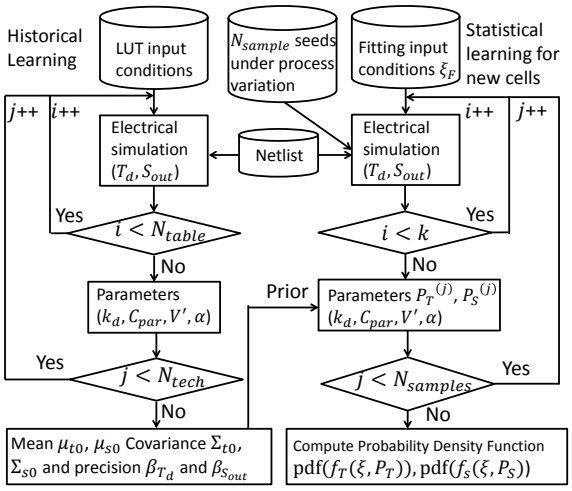

Fig. 4. Proposed flow for statistical characterization with both old and new libraries interacting and priors being passed from an old library to a new one.

Fig. 4 summarizes the major steps of the proposed statistical library cell characterization method with both old and new libraries interacting and priors being passed from an old library to a new one. If we assume that library cell characterizations have been done in previous technologies, the total computation cost is O($k \cdot N_{sample}$), which is at least one order of magnitude smaller compared with O($N_{LUT} \cdot N_{sample}$) in prior work and several order of magnitudes smaller compared with O($N_{LUT} \cdot N_{MC}$) in standard method. The total computation cost is O($k \cdot N_{sample} + N_{Tech} \cdot N_{LUT}$) if we need to re-run characterization for old technologies, which is still a moderate speed up compared to the most cutting-edge techniques.

V. Validation

In this section, two library cell characterization examples in several cutting-edge CMOS technologies are used to demonstrate the efficiency of our proposed method. All test cases as well as the historical library cell characteristics are generated using different BSIM based industrial design kits reflecting real measurements. To test and compare with the prior part, we have also implemented both deterministic extraction and statistical extraction using a look-up table (LUT) approach.

The baseline characterization is defined in this work by a 1000 points Monte Carlo simulation sampled randomly within the whole input space $\xi = \{S_{in}, C_{load}, V_{dd}\}$. Note that these points only represent different operating conditions for a target cell while the effects of process variation are not included. Fig. 5 shows a scatter plot for 1000 points among the whole input space

where we will compare our characterization result with standard methods.

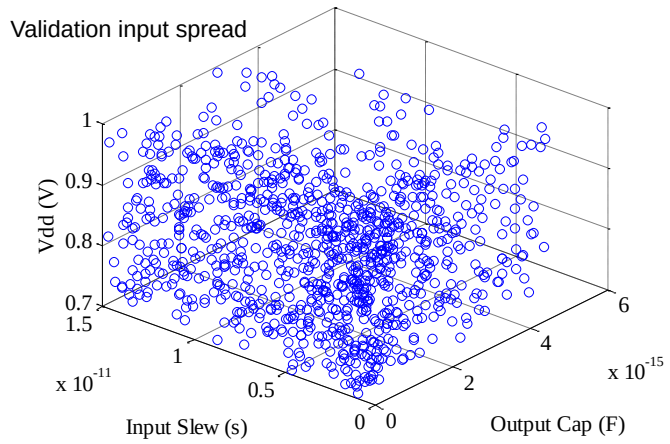

Fig. 5. A scatter plot of 1000 points among whole input space $\xi = \{S_{in}, C_{load}, V_{dd}\}$ used for comparing our characterization result with standard methods.

The first example is to conduct a nominal delay and output slew characterization for a library designed in a commercial state-of-the-art 14-nm FINFET technology. Both fitting and testing samples are generated through SPICE simulation using a well calibrated compact transistor model. Fig. 6 shows average prediction error compared with the baseline characterization using proposed model with Bayesian inference, proposed model with our least-square error function optimization, and look-up table approach. To achieve the same characterization accuracy on delay T_d, our proposed method achieves up to 15X runtime speedup compared to a traditional lookup table approach, where 6X speedup is contributed by our proposed timing model and an extra speedup of 2.5X is contributed by the Bayesian inference. Given the prior and two additional fitting input combinations, a 4.3% average error compared with the baseline characterization is achieved for all combinations of C_{load}, S_{in} and V_{dd}. This demonstrates the sparsity of effects across input vectors and the validity of the proposed delay model.

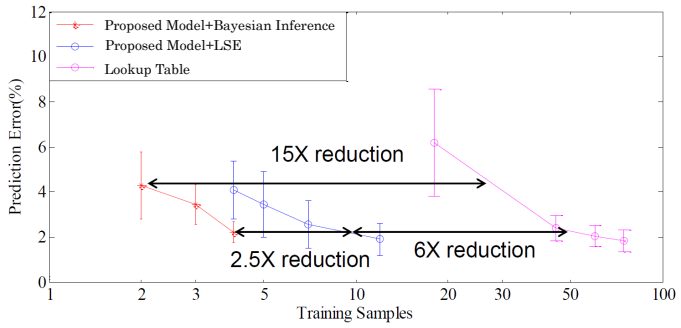

Fig. 6. Average testing error for delay T_d characterizing a library designed in a commerical state-of-the-art 14-nm technology. Error bars show one standard deviation of testing error for different cells and RISE/FALL.

The second example is to conduct statistical delay and output slew characterization for a library designed in a commercial state-of-the-art 28-nm bulk-Silicon technology, which is different from the model used in the first example. The baseline characterization is defined similar to previous example where 1000 input combinations are sampled randomly within the whole space $\xi = \{S_{in}, C_{load}, V_{dd}\}$. In this case 1000 seeds under process variation are generated for each cell to obtain statistical distributions for delay and output slew with different input combinations.

The error functions for statistical characterization of $\mathcal{E}(\mu_{T_d})$, $\mathcal{E}(\mu_{S_{out}})$, $\mathcal{E}(\sigma_{T_d})$ and $\mathcal{E}(\sigma_{S_{out}})$ are defined as

$$\mathcal{E}(\mu_{T_d}) = \frac{1}{n}\sum_{i=1}^{n}\left|\mu(f_T(\xi^{(i)}, P_T))) - \mu_{T_d}^{(i)}\right| \quad (16)$$

$$\mathcal{E}(\mu_{S_{out}}) = \frac{1}{n}\sum_{i=1}^{n}\left|\mu(f_S(\xi^{(i)}, P_S)) - \mu_{S_{out}}^{(i)}\right| \quad (17)$$

$$\mathcal{E}(\sigma_{T_d}) = \frac{1}{n}\sum_{i=1}^{n}\left|\sigma(f_T(\xi^{(i)}, P_T))) - \sigma_{T_d}^{(i)}\right| \quad (18)$$

$$\mathcal{E}(\sigma_{S_{out}}) = \frac{1}{n}\sum_{i=1}^{n}\left|\sigma(f_S(\xi^{(i)}, P_S)) - \sigma_{S_{out}}^{(i)}\right| \quad (19)$$

Fig. 7 and Fig. 8 show average prediction error for mean and standard deviation of delay and output slew characterizing a library designed in a commercial state-of-the-art 28-nm technology compared with the baseline characterization using proposed method and look-up table approach. Up to 20X runtime speedup is observed to achieve the same characterization accuracy in mean value and standard deviation of T_d and S_{out}.

Fig. 9 shows delay probability density simulated using baseline simulation, the proposed method with seven training input combinations, and an interpolation of look-up tables with 60 training input combinations together with baseline distribution using SPICE Monte Carlo simulation. The input combination is $V_{dd} = 0.734V$, $S_{in} = 5.09ps$, $C_{load} = 1.67fF$. The proposed method shows a much better prediction for delay distribution which correctly predicts the non-Gaussian distribution for low V_{dd}.

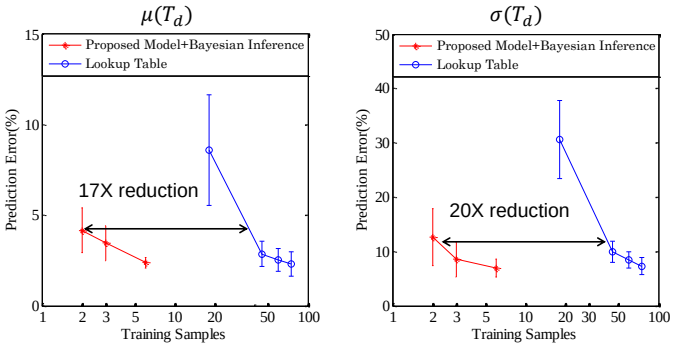

Fig. 7. Average testing error for mean and standard deviation of delay T_d characterizing a library designed in a commerical state-of-the-art 28-nm technology. Error bars show one standard deviation of testing error for different cell types and RISE/FALL combinations.

VI. Conclusion

In this paper we have presented an entirely different perspective on the acceleration of library characterizations. While previous authors have emphasized the use of statistical techniques to address the efficient design of variation-aware standard cell libraries by working in process space, in our work we use similar techniques for the efficient design of these libraries by working in the traditional library input space of input slew, output load, and voltage supply. The main insight that has enabled this shift in perspective is the contribution of a new ultra compact timing model for standard cells that is a powerful and accurate generalization of the simple $C_{load}V_{dd}/I_{dsat}$ metric. This new analytical timing model transfers the library characterization problem from one of input parameter sweep to one of machine learning and

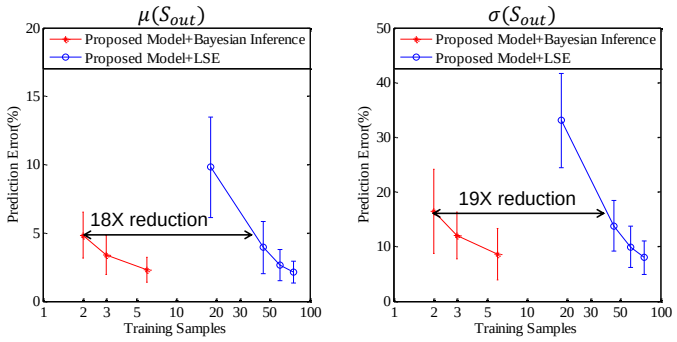

Fig. 8. Average testing error for mean and standard deviation of output slew S_{out} characterizing a library designed in a commerical state-of-the-art 28-nm technology. Error bars show one standard deviation of testing error for different cell types and RISE/FALL combinations.

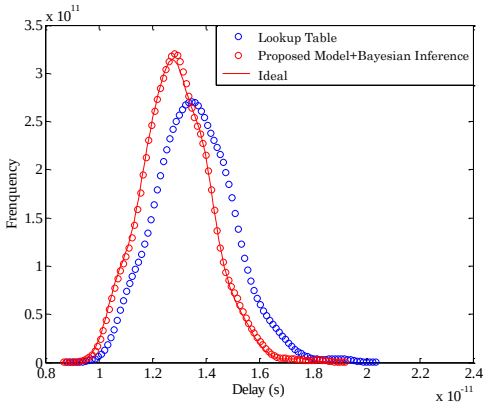

Fig. 9. Delay probability density simulated using baseline simulation, proposed method and an interpolation of look-up tables together with baseline distribution using SPICE Monte Carlo simulation with an input combination of $V_{dd} = 0.734V$, $S_{in} = 5.09ps$, $C_{load} = 1.67fF$.

sparse sampling. Machine learning is used to develop priors of timing model coefficients using old libraries while sparse sampling is used to provide the extra data points needed to build the new library in the target technology. Our methods have resulted in 15X reduction in simulation runs with respect to baseline techniques that use random sampling methods.

ACKNOWLEDGMENT

This work was funded by the Cooperative Agreement between the Masdar Institute of Science and Technology (Masdar Institute), Abu Dhabi, UAE and the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, Reference 196F/002/707/102f/70/9374.

REFERENCES

[1] L. Lavagno, L. Scheffer, and G. Martin. *EDA for IC Implementation, Circuit Design, and Process Technology*. Addision-Wesley, 2006.

[2] L. Yu, L. Wei, D. Antoniadis, I. Elfadel, and D. Boning. Statistical modeling with the virtual source MOSFET model. In *Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1454–1457, 2013.

[3] X. Li. Finding deterministic solution from underdetermined equation: Large-scale performance modeling by least angle regression. In *Design Automation Conference*, pages 364–369, 2009.

[4] L. Brusamarello, P. Wirth, G.and Roussel, and M. Miranda. Fast and accurate statistical characterization of standard cell libraries. *Microelectronics Reliability*, 51(12):2341 – 2350, 2011.

[5] Z. Zhang, T.A. El-Moselhy, I.M. Elfadel, and L. Daniel. Stochastic testing method for transistor-level uncertainty quantification based on generalized polynomial chaos. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(10):1533–1545, Oct. 2013.

[6] Z. Zhang, I.A.M. Elfadel, and L. Daniel. Uncertainty quantification for integrated circuits: Stochastic spectral methods. In *International Conference on Computer-Aided Design (ICCAD)*, pages 803–810, Nov. 2013.

[7] Z. Zhang, T.A. El-Moselhy, I.M. Elfadel, and L. Daniel. Calculation of generalized polynomial-chaos basis functions and gauss quadrature rules in hierarchical uncertainty quantification. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(5):728–740, May 2014.

[8] Z. Zhang, X. Yang, I. V. Oseledets, G. E. Karniadakis, and L. Daniel. Enabling high-dimensional hierarchical uncertainty quantification by ANOVA and tensor-train decomposition. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, PP(99):1–1, Nov. 2014.

[9] Z. Zhang, X Yang, G. Marucci, P. Maffezzoni, I.M. Elfadel, G. Karniadakis, and L. Daniel. Stochastic testing simulator for integrated circuits and mems: Hierarchical and sparse techniques. In *Custom Integrated Circuits Conference (CICC)*, pages 1–8, Sep. 2014.

[10] F. Wang, W. Zhang, S. Sun, X. Li, and C. Gu. Bayesian model fusion: large-scale performance modeling of analog and mixed-signal circuits by reusing early-stage data. In *Design Automation Conference (DAC)*, pages 64:1–64:6, 2013.

[11] S. Reda and S.R. Nassif. Analyzing the impact of process variations on parametric measurements: Novel models and applications. In *Design, Automation Test in Europe (DATE)*, pages 375–380, Apr. 2009.

[12] L. Yu, S. Saxena, C. Hess, I. Elfadel, D. Antoniadis, and D. Boning. Remembrance of transistors past: Compact model parameter extraction using bayesian inference and incomplete new measurements. In *Design Automation and Conference (DAC)*, 2014.

[13] L. Yu, S. Saxena, C. Hess, I. Elfadel, D. Antoniadis, and D. Boning. Efficient performance estimation with very small sample size via physical subspace projection and maximum a posteriori estimation. In *Design, Automation and Test in Europe (DATE)*, 2014.

[14] S. Sun, F. Wang, S. Yaldiz, X. Li, L. Pileggi, A. Natarajan, M. Ferriss, J. Plouchart, B. Sadhu, B. Parker, A. Valdes-Garcia, M. Sanduleanu, J. Tierno, and D. Friedman. Indirect performance sensing for on-chip analog self-healing via bayesian model fusion. In *Custom Integrated Circuits Conference (CICC)*, pages 1–4, Sep. 2013.

[15] S. Sun, X. Li, H. Liu, K. Luo, and B. Gu. Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space. In *International Conference on Computer-Aided Design (ICCAD)*, pages 478–485, Nov. 2013.

[16] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[17] T. Sakurai and A.R. Newton. Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, Apr. 1990.

[18] V. Stojanovic, D. Markovic, B. Nikolic, M.A. Horowitz, and R.W. Brodersen. Energy-delay tradeoffs in combinational logic using gate sizing and supply voltage optimization. In *European Solid-State Circuits Conference (ESSCIRC)*, pages 211–214, Sep. 2002.

[19] A. Khakifirooz and D.A. Antoniadis. MOSFET performance scaling - part I: Historical trends. *IEEE Transactions on Electron Devices*, 55(6):1391–1400, June 2008.

[20] L. Yu, O. Mysore, L. Wei, L. Daniel, D. Antoniadis, I. Elfadel, and D. Boning. An ultra-compact virtual source fet model for deeply-scaled devices: Parameter extraction and validation for standard cell libraries and digital circuits. In *Asia and South Pacific Design Automation Conference(ASPDAC)*, pages 521–526, 2013.

[21] M.-H. Na, E.J. Nowak, W. Haensch, and J. Cai. The effective drive current in cmos inverters. In *International Electron Devices Meeting (IEDM)*, pages 121–124, Dec. 2002.

[22] J. Deng and H. Wong. Metrics for performance benchmarking of nanoscale Si and carbon nanotube FETs including device nonidealities. *IEEE Transactions on Electron Devices*, 53(6):1317–1322, June 2006.

[23] E. Yoshida, Y. Momiyama, M. Miyamoto, T. Saiki, M. Kojima, S. Satoh, and T. Sugii. Performance boost using a new device design methodology based on characteristic current for low-power CMOS. In *International Electron Devices Meeting (IEDM)*, Dec. 2006.

[24] S. Rakheja and D. Antoniadis. MVS 1.0.1 nanotransistor model (Silicon), Nov. 2013.

[25] N. Weste and K. Eshraghian. *Principles of CMOS VLSI Design: A Systems Perspective*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1985.

[26] X. Lin, Y. Wang, and M. Pedram. Joint sizing and adaptive independent gate control for finfet circuits operating in multiple voltage regimes using the logical effort method. In *International Conference on Computer-Aided Design (ICCAD)*, pages 444–449, Nov. 2013.

[27] X. Lin, Y. Wang, S. Nazarian, and M. Pedram. An improved logical effort model and framework applied to optimal sizing of circuits operating in multiple supply voltage regimes. In *International Symposium on Quality Electronic Design (ISQED)*, pages 249–256, Mar. 2014.