Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

An efficient bi-objective optimization framework for statistical chip-level yield analysis under parameter variations

Xin LI[†1,2], Jin SUN[3], Fu XIAO[2], Jiang-shan TIAN[3]

([1]*Technology Innovation Center, Jiangsu Academy of Safety Science and Technology, Nanjing 210042, China*)

([2]*Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks,*
Nanjing University of Posts and Telecommunications, Nanjing 210013, China)

([3]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*)

[†]E-mail: lin65002@hotmail.com

Received May 25, 2015; Revision accepted Aug. 24, 2015; Crosschecked Nov. 24, 2015

Abstract: With shrinking technology, the increase in variability of process, voltage, and temperature (PVT) parameters significantly impacts the yield analysis and optimization for chip designs. Previous yield estimation algorithms have been limited to predicting either timing or power yield. However, neglecting the correlation between power and delay will result in significant yield loss. Most of these approaches also suffer from high computational complexity and long runtime. We suggest a novel bi-objective optimization framework based on Chebyshev affine arithmetic (CAA) and the adaptive weighted sum (AWS) method. Both power and timing yield are set as objective functions in this framework. The two objectives are optimized simultaneously to maintain the correlation between them. The proposed method first predicts the guaranteed probability bounds for leakage and delay distributions under the assumption of arbitrary correlations. Then a power-delay bi-objective optimization model is formulated by computation of cumulative distribution function (CDF) bounds. Finally, the AWS method is applied for power-delay optimization to generate a well-distributed set of Pareto-optimal solutions. Experimental results on ISCAS benchmark circuits show that the proposed bi-objective framework is capable of providing sufficient trade-off information between power and timing yield.

Key words: Parameter variations, Parametric yield, Multi-objective optimization, Chebyshev affine, Adaptive weighted sum
http://dx.doi.org/10.1631/FITEE.1500168 **CLC number:** TP312

1 Introduction

Continuous process scaling has led to a large increase in process, voltage, and temperature (PVT) variability and a wide spread fluctuation in integrated circuit (IC) performance. This increasing variability brings significant impact on the parametric yield of today's chip design (Mani *et al.*, 2005; Radfar and Singh, 2014; Banerjee and Chatterjee, 2015). To be specific, 30% variation in effective channel length could cause over 20× fluctuation in leakage power (Rao *et al.*, 2004a; Kanj *et al.*, 2010). In addition, Srivastava *et al.* (2008) pointed out the negative correlation between power dissipation and timing performance of a design. This relationship causes significant yield loss when considering both power and timing limits and leads to a two-sided constraint over the design region.

Most of the previous yield estimation works have been limited to predicting either timing or leakage yield (Orshansky and Bandyopadhyay, 2004; Rao *et al.*, 2004b; Xie and Davoodi, 2008). Dealing with only timing yield optimization will result in yield loss due to the power constraint (Srivastava *et al.*, 2008). On the other hand, all the power yield analyses neglect the correlation between power and timing metrics. As mentioned above, in a chip design, the leakage power and delay are negatively correlated. This situation will consequently bring on a conflict between these two objectives during the optimization

ORCID: Xin LI, http://orcid.org/0000-0002-4859-2477

procedure and cause designers to be in a dilemma. Specifically, this situation has been more serious at a 20-nm technology node. Thus, there is a critical requirement to develop an effective approach that performs parametric yield optimization considering both power and timing constraints.

There is recent research focusing on considering power and timing metrics simultaneously in yield analysis and optimization. Hwang *et al.* (2003) proposed a novel statistical leakage minimization method using the timing yield slack for a gate change metric. This method can help improve not only the performance of leakage optimization but also the efficiency by providing valuable information to guide statistical leakage optimization. Based on optimal delay budgeting and slack utilization, Mani *et al.* (2007) presented a two-phase approach to solve the statistical leakage power minimization problem under timing yield constraints. The first phase is delay budgeting, which is formulated as a robust version of the power-weighted linear program that assigns slacks based on power-delay sensitivities of gates. The second phase consists of a local search among gate configurations in the library, such that slacks assigned to gates in the previous phase are used for power reduction. However, these approaches mentioned above fail to take into account the close correlation between leakage power and delay. They do not perform parametric yield optimization incorporating leakage and delay considerations, but optimize the power yield under timing constraints in the presence of variability.

Several research efforts have been made on optimizing yield in a multi-objective design fashion. For example, Liu *et al.* (2013) proposed a new time-domain performance bound analysis method for analog circuits, considering process variations. The method can give transient lower and upper bounds of the performance variations affected in analog circuits accurately and reliably. However, their approach requires additional computational cost for estimating yield specification from the predicted performance bounds. Additionally, it cannot handle parameter variations that are partially specified. Also, Guerra-Gómez *et al.* (2015) proposed several evolutionary algorithms to solve the multi-objective yield optimization problem. In their work, a strategy based on the optimal computing budget allocation approach was presented to reduce the simulation cost in the yield optimization of analog integrated circuits. However, their method cannot provide more flexibility in design trade-offs. In contrast, our work is discussed under the assumption of partially specified PVT parameter variations. It provides more flexibility and a simple optimization procedure with lower computation cost.

This study aims at solving the power-delay optimization problem by using multi-objective optimization techniques. The proposed optimization method incorporates leakage and delay considerations. We introduce a new power and timing yield optimization framework using Chebyshev affine arithmetic (CAA) and the adaptive weighted sum (AWS) method for multi-objective optimization. This framework treats both timing and power yield as objective functions and optimizes these two goals simultaneously. Additionally, because AWS is used for optimization in multi-domain, our framework can include extra objectives, e.g., area and thermal metrics. Different from traditional multi-objective optimization methods, our optimization methodology distributes the optimal solutions uniformly upon the Pareto front. As a result, it can provide the designers with multiple solutions distributed over the optimal design spectrum, giving designers the flexibility to choose the most appropriate solution(s) according to power and timing requirements.

The contributions of the new approach include: (1) maintaining the correlation between leakage power and delay by explicitly expressing both metrics in terms of the same parameter variations; (2) allowing arbitrary correlations among PVT parameters, because the yield prediction scheme for leakage power and delay is under the assumption of uncertain parameter correlations; and (3) providing designers with trade-off information between power and timing yield to find the best solution(s). The final result is a set of Pareto-optimal solutions uniformly distributed over the design region. The flexibility obtained by the new multi-objective framework was demonstrated on various ISCAS benchmark circuits. For each circuit, well-distributed sets of Pareto-optimal solutions were obtained by the proposed methodology.

2 Statistical leakage and delay model

This section discusses in detail the statistical models for leakage power and delay under the influence of parameter variations, which will be incorporated into the bi-objective model for optimization. Here, the variability in leakage and delay will be expressed as a function of several key PVT parameters. In this way, the correlation between power and delay is preserved for yield estimation, because they both depend on the identical underlying parameter variations.

Without loss of generality, the current study takes into account the variability in several key PVT parameters: effective transistor channel length L, threshold voltage V_{th}, oxide thickness T_{ox}, power-supply voltage V_{dd}, and on-chip temperature T. If a common notation P is used to represent all these PVT parameters, the variation deviated from the nominal value of process parameter ΔP may be expressed as

$$\Delta P = \Delta P_{\text{inter}} + \Delta P_{\text{intra}}, \tag{1}$$

where ΔP_{inter} denotes the inter-chip process variation, and ΔP_{intra} the intra-chip counterpart (Mande *et al.*, 2013). All process variations are assumed to follow Gaussian distributions, which is in agreement with empirical data (Visweswariah, 2003). The relative magnitudes of the intra- and inter-chip components can be controlled by adjusting their variances while satisfying the following equation (Mani *et al.*, 2005):

$$\sigma_P^2 = \sigma_{P_{\text{inter}}}^2 + \sigma_{P_{\text{intra}}}^2. \tag{2}$$

Based on above basic models, let us take a look at leakage power and timing, respectively. Leakage power can be expressed as the product of its nominal value and a multiplicative function representing the perturbation around the nominal leakage value (Rao *et al.*, 2004a):

$$\text{Leakage} = I_{\text{nom}} \cdot f(\Delta P), \tag{3}$$

where deviation ΔP represents the impact from parameter variation. To be more specific, the leakage power can be written as its nominal value I_{nom} multiplied with an exponential function in terms of effective transistor channel length variation (ΔL), threshold voltage variation (ΔV_{th}), oxide thickness variation (ΔT_{ox}), power-supply voltage variation (ΔV_{dd}), and on-chip temperature (ΔT). As ΔL imposes a significant influence on sub-threshold leakage, a quadratic exponential expression rather than a linear exponential model is adopted here. Besides, the super-linear dependency of leakage power on variability in threshold voltage, oxide thickness, power-supply voltage, and on-chip temperature can be well approximated using a linear exponential function according to SPICE simulations (Wang and Orshansky, 2006). In conclusion, the leakage power can be represented explicitly as follows:

$$\begin{aligned}\text{Leakage} &= I_{\text{sub}} + I_{\text{gate}} \\ &= I_{\text{sub,nom}} \cdot \exp(a\Delta L^2 + b\Delta L + c\Delta V_{\text{th}} + d\Delta V_{\text{dd}} + e\Delta T) \\ &\quad + I_{\text{gate,nom}} \cdot \exp(f\Delta T_{\text{ox}} + g\Delta V_{\text{dd}}),\end{aligned} \tag{4}$$

where $I_{\text{sub,nom}}$ and $I_{\text{gate,nom}}$ are the nominal values of the sub-threshold leakage power and the gate leakage current, respectively. The coefficients a, b, c, d, e, f, and g can be in fact regarded as the sensitivities of leakage power (in logarithm form) to corresponding PVT parameters under consideration. The values of the coefficients in Eq. (4) can be determined by nonlinear regression based on HSPICE simulation data.

On the other hand, for the timing issue, gate delay needs to be modeled as a function in terms of a set of PVT parameters. We assume that a first-order Taylor expansion is adequate to model the gate delay function (Sheng *et al.*, 2013). The delay function under parameter variations can be approximated linearly as

$$\text{Delay} = D_{\text{nom}} + \sum_i \left(\frac{\partial D}{\partial P_i}\right)\Delta P_i, \tag{5}$$

where D_{nom} is the nominal gate delay calculated at the nominal PVT parameter values and $\partial D/\partial P_i$ is the delay sensitivity of a specific parameter computed around its nominal value. The delay function is written more specifically as

$$\text{Delay} = D_{\text{nom}} + h\Delta L + k\Delta V_{\text{th}} + l\Delta T_{\text{ox}} + r\Delta V_{\text{dd}} + s\Delta T, \tag{6}$$

where h, k, l, r, and s are the corresponding parameter sensitivities.

Having established the statistical models of leakage power and delay in expressions of parameter variations, we are able to develop the power-delay bi-objective optimization framework, which will be described in the subsequent part. Note that leakage and delay are correlated due to their common dependence on identical PVT parameters.

3 Bi-objective optimization procedure

Guerra-Gómez *et al.* (2013) proposed a sensitivity analysis in the multi-objective optimization of analog circuits. The approach can achieve good accuracy for small design parameter perturbations or relatively linear behaviors in analog circuit performances. However, the leakage power in our optimization framework is highly nonlinear to design parameters. Thus, we must seek other yield prediction approaches that can handle nonlinear dependency upon parameter variations and limited descriptions of parameter variations.

This section applies the CAA method to address the above two issues and discusses how to formulate the proposed power-delay bi-objective optimization model to obtain a well-distributed set of Pareto-optimal solutions. First, the CAA methodology is applied to predict a guaranteed cumulative distribution function (CDF) bound for leakage power and delay based on the models described in Section 2. The distribution function directly provides the functional relationship between power/delay metrics and design parameters. Then leakage yield and timing yield functions can be established as two objective functions. Finally, the bi-objective optimization model for power and timing yield is proposed, which will be optimized in the subsequent part.

3.1 CAA-based probability bound prediction

The PVT parameter variations are assumed to be partially specified; i.e., only the mean and variance information may be available. As suggested in much literature, some PVT parameters tend to be uncertain or even have unknown distributions (Gong *et al.*, 2011; Ukhov *et al.*, 2014). Under this assumption, this study applies the CAA method to predict parametric yield robustly with fully or partially specified parameter variations.

According to the CAA theory (Sun *et al.*, 2008; Zhu and Wu, 2014), an uncertain random variable can be represented as

$$x' = x_{\mathrm{nom}} + \sum_{i=1}^{n} x_i \varepsilon_i, \tag{7}$$

where x_{nom} denotes the nominal value of the PVT parameter, ε_i an independent component representing the total fluctuation, and x_i the magnitude of corresponding ε_i. Note that in our bi-objective optimization framework, ε_i represents arbitrary parameter variations, including effective channel length variation (ΔL), threshold voltage variation (ΔV_{th}), oxide thickness variation (ΔT_{ox}), power-supply voltage variation (ΔV_{dd}), and on-chip temperature (ΔT).

Considering the bivariate affine arithmetic (AA) operation $z' \leftarrow f(x', y')$, an affine operation f can directly provide a first-order affine form without any computation error (de Figueiredo and Stolfi, 2004). For example,

$$\begin{cases} x' \pm y' = (x_{\mathrm{nom}} \pm y_{\mathrm{nom}}) + \sum_{i=1}^{n} (x_i \pm y_i)\varepsilon_i, \\ ax' = a(x_{\mathrm{nom}} + \sum_{i}^{n} x_i \varepsilon_i), \\ x' \pm b = x_{\mathrm{nom}} + \sum_{i}^{n} x_i \varepsilon_i \pm b. \end{cases} \tag{8}$$

However, when f is not affine, we need to choose an affine function to approximate z' over a given domain:

$$f^{\mathrm{a}}(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) = z_{\mathrm{nom}} + \sum_{i=1}^{n} z_i \varepsilon_i + z_k \varepsilon_k. \tag{9}$$

Here, $z_k \varepsilon_k$ indicates the approximation error.

To obtain an optimal approximation to z', we usually consider that the approximation is only an affine combination of x' and y':

$$f^{\mathrm{a}}(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) = \alpha x' + \beta y' + \zeta + \delta \varepsilon_k, \tag{10}$$

where α and β denote the coefficients of x' and y', respectively, ζ is a constant, and $\delta \varepsilon_k$ represents the approximation error. In this study, Chebyshev approximation (de Figueiredo and Stolfi, 2004) is used to approximate z'. Chebyshev approximation can minimize the maximum absolute error δ in Eq. (10) better than other algorithms.

Besides the range information, an uncertain random variable can be represented by a set of CDFs or a p-box (Saad *et al.*, 2014). Thus, in this study, we represent the parameter variations as a set of CDFs. The CDF bounds for parameter variations can be constructed by the CAA method relying on mean and variance information and computed under affine and non-affine operations. To effectively predict the probability bound, we apply Chebyshev approximation on an uncertain random variable's CDFs to address its nonlinearity. Given an uncertain random variable already in the p-box representation, the whole range of an uncertain random variable is divided into several subintervals (Fig. 1). Chebyshev approximation is then performed on each interval, which clearly returns a linear function with the least perturbation according to Eq. (10). This provides the upper and lower bounds in piecewise linear form, enclosing the CDFs well.

The resulting CDF bounds obtained by Chebyshev approximation are named 'piecewise linear probability bounds' (PLPBs) (Sun *et al.*, 2008). Given random variables in PLPB representations, an efficient prediction scheme can be provided for correlating CDF bounds under operations upon random variables. This scheme transforms all the non-affine operations into affine forms by Chebyshev approximation, and then CDF bounds are predicted step by step under affine operations, handling arbitrary correlations among variations.

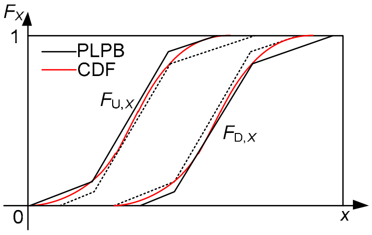


Fig. 1 Piecewise linear probability bounds (PLPB) representation of an uncertain random variable (the dotted lines are PLPB bounds, which are not focused on in this study)

3.2 Correlation CDF bound computation

Without loss of generality, we describe any two PVT parameters as random variables X and Y, which are used in PLPB representations. $F_{\mathrm{U},X}$ and $F_{\mathrm{D},X}$ are the upper bound and lower bound of X, respectively, while $F_{\mathrm{U},Y}$ and $F_{\mathrm{D},Y}$ of Y. Here, we denote Z as a binary function of X and Y with one of 'add', 'subtract', 'multiply', and 'divide' operations, and let F_{U}^{-1} and F_{D}^{-1} denote the upper and lower bounds of the inverse CDF, respectively. Based on Williamson and Downs (1990), for 'add' operation, these bounds can be derived as

$$\begin{aligned} F_{\mathrm{D},Z}^{-1}(p) &= F_{\mathrm{D},X+Y}^{-1}(p) \\ &= \min_{u\in[p,1]}[F_{\mathrm{D},X}^{-1}(u) + F_{\mathrm{D},Y}^{-1}(p-u+1)], \end{aligned} \tag{11}$$

$$\begin{aligned} F_{\mathrm{U},Z}^{-1}(p) &= F_{\mathrm{U},X+Y}^{-1}(p) \\ &= \max_{u\in[0,p]}[F_{\mathrm{U},X}^{-1}(u) + F_{\mathrm{U},Y}^{-1}(p-u)]. \end{aligned} \tag{12}$$

Similarly, we can obtain the bounds for 'subtract', 'multiply', and 'divide' operations according to Williamson and Downs (1990). As 'multiply' and 'divide' operations are not used in this study, we do not consider them here.

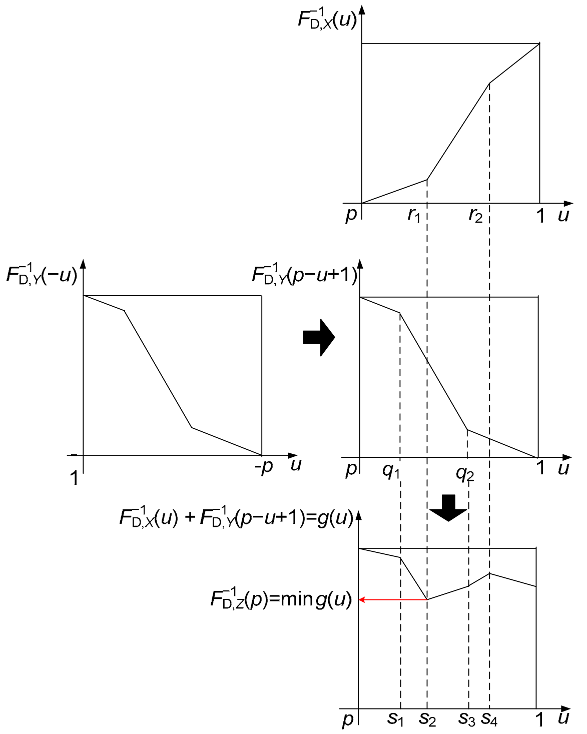
Once the abovementioned bounds are obtained, affine operations Z=$X\pm Y$ exhibit a functional relationship in the inverses of X' and Y's CDF bounds. Here, taking Eq. (11) as an example, for a fixed probability value p, if we assume $g(u) = F_{\mathrm{D},X}^{-1}(u)$ $+F_{\mathrm{D},Y}^{-1}(p-u+1)$, then $F_{\mathrm{D},X+Y}^{-1}(p)$ is obviously the minimum value of $g(u)$ in the interval $[p,1]$. To solve this optimization problem, we will represent random variables in PLPB formation. It can propose a simple optimization procedure with low computation cost.

Now let us take $F_{\mathrm{D},X+Y}$ as an example to show how to construct the lower bound. $F_{\mathrm{U},X+Y}$, $F_{\mathrm{D},X-Y}$, and $F_{\mathrm{U},X-Y}$ can be derived similarly. Here we have

$$\begin{aligned} F_{\mathrm{D},X+Y}^{-1}(p) &= \min_{u\in[p,1]}[F_{\mathrm{D},X}^{-1}(u) + F_{\mathrm{D},Y}^{-1}(p-u+1)] \\ &= \min_{u\in[p,1]} g(u). \end{aligned} \tag{13}$$

From the above discussion we can conclude that, for a fixed probability value p, $F_{\mathrm{D},Y}^{-1}(p-u+1)$ is a piecewise linear function of u, because it is simply an affine transformation of $F_{\mathrm{D},Y}^{-1}(u)$ (Fig. 2). Thus, we can find the minimum value of $g(u)$ in the interval

$[p, 1]$, where $F^{-1}_{\mathrm{D},X}(u)$ and $F^{-1}_{\mathrm{D},Y}(p-u+1)$ have transition points r_1, r_2 and q_1, q_2, respectively. As shown in Fig. 2, due to the continuity and monotonicity of $F^{-1}_{\mathrm{D},Y}(u)$ and $F^{-1}_{\mathrm{D},Y}(p-u+1)$ in the interval $[p,1]$, the piecewise linear function $g(u)$ has four transition points, i.e., $s_1=q_1$, $s_2=r_1$, $s_3=q_2$, and $s_4=r_2$. Intuitively, the minimum value of $g(u)$ is achieved at point s_2 corresponding to a fixed provability value p.



$F^{-1}_{\mathrm{D},X}(u)$
p
r_1
r_2
1
u
$F^{-1}_{\mathrm{D},Y}(-u)$
$F^{-1}_{\mathrm{D},Y}(p-u+1)$
1
-p
u
p
q_1
q_2
1
u
$F^{-1}_{\mathrm{D},X}(u)+F^{-1}_{\mathrm{D},Y}(p-u+1)=g(u)$
$F^{-1}_{\mathrm{D},Z}(p)=\min g(u)$
p
s_1
s_2
s_3
s_4
1
u

Fig. 2 The computing process of the lower bound for the inverse CDF of variable *Z*=*X*+*Y*

To summarize generally, the probability ranges for $F^{-1}_{\mathrm{D},X}(u)$ and $F^{-1}_{\mathrm{D},Y}(p-u+1)$ are divided by sets $\{p, r_1, r_2, \ldots, r_n, 1\}$ and $\{p, q_1, q_2, \ldots, q_n, 1\}$, respectively. The probability range of $g(u)$ can be divided as

$$\begin{aligned} S &= \{s_1, s_2 \cdots, s_n\} \\ &= \{p, r_1, r_2 \cdots, r_n, q_1, q_2 \cdots, q_n, 1\}, \end{aligned} \tag{14}$$

where $\{p, r_1, r_2, \ldots, r_n, q_1, q_2, \ldots, q_n, 1\}$ are in ascending order. For each interval $[s_i, s_{i+1}]$, because $g(u)$ is monotonic over this interval, the minimum value must be determined by one of the end points. Therefore, the global minimum can be determined by choosing the most minimum value of the end points among these intervals.

3.3 Bi-objective optimization model

In chip-level parametric yield analysis, a reasonable assumption is that each device has a unique intra-chip variation ΔP_{intra} while sharing the same inter-chip variation ΔP_{inter} with all other devices. Therefore, global process variations may be regarded as fixed values for each device. All process variations are fully specified by corresponding CDFs, while all environmental variations are partially specified by the corresponding mean and variance values. The corresponding PLPB representations can be constructed conveniently by Chebyshev approximation.

According to Eqs. (4) and (6), the leakage power and gate delay for a chip design are represented as functions in terms of PVT parameter variations. Using the CAA methodology, we can finally obtain a guaranteed CDF bound for leakage power or delay distribution. Taking the delay model as an example, it is already in the affine form according to Eq. (8). Within several steps, CAA is able to predict the upper and lower probability bounds for delay distribution under parameter variations. Regardless of relationship among PVT parameters, any CDF generated under an arbitrary correlation situation will be enclosed by CAA predicted bounds. As our purpose is to optimize the guaranteed parametric yield, we consider only the lower probability bound, which is denoted by F_{D}. There will be a similar conclusion for power distribution. In the leakage model, two CAA approximations, quadratic and exponential operations, are required to reduce the leakage function to a series of affine operations on parameter variations. The guaranteed (lower) CDF bound for leakage distribution, generated in the same manner, is denoted by F_{L}.

To analyze parametric yield considering both power and timing limits, we now focus on the predicted distributions F_{L} and F_{D}. For example, leakage distribution F_{L} is actually a function that returns the cumulative probability at a given leakage value. In the opposite direction, given a specific yield probability, it is also able to provide the leakage value corresponding to the given particular yield level. Fig. 3 shows the relationship between F_{L} and the power yield.

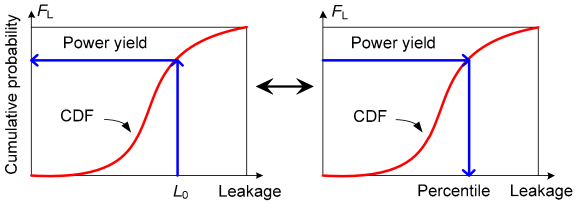


Fig. 3 Distribution function F_L provides yield information

If we define a specific leakage limit L_0 as a power yield criterion, F_L directly provides the yield information:

$$Y_L = F_L(L_0) = P(\text{Leakage} < L_0), \tag{15}$$

where Y_L denotes the power yield defined by leakage distribution. Timing yield Y_D can be defined by the same token:

$$Y_D = F_D(D_0) = P(\text{Delay} < D_0). \tag{16}$$

So far we have established the functional relationship between parametric yield and design parameters $x\in[L, V_{th}, T_{ox}, V_{dd}, T]$, because leakage and delay distributions both depend on the variability in parameters.

The algorithmic flow of CAA-based yield prediction is summarized in Algorithm 1:

Algorithm 1 CAA based yield prediction

```
Input: design parameter $x$, metric limit $M_0$
Output: Yield
fun←Model_Extraction()
(op_NonAff, op_Aff)←Non_Affine_Check(fun)
op_Appx←Chebyshev_Approx(op_NonAff)
op_Stack←Combine(op_Aff, op_Appx)
for $i$=1 to length(op_Stack) do
    dummy_CDF←CDF_Generation($x_i$)
    PUSH(dummy_CDF, CDF_Stack)
end for
$F_1$←POP(CDF_Stack)
while CDF_Stack≠∅ do
    $F_2$←POP(CDF_Stack)
    op←POP(op_Stack)
    $F_1$←CAA_Bound_Computation(op, $F_1$, $F_2$)
end while
Distribution←$F_1$
Yield←Prob(Distribution, $M_0$)
return Yield
```

In Algorithm 1, the 'Model_Extraction' subroutine returns the expressions demonstrated in Eqs. (4) and (6). Having the explicit expressions of leakage power and gate delay parameterized with design parameters, the 'Non_Affine_Check' subroutine identifies the affine operations and non-affine operations in analytical power and timing models, denoted by op_NonAff and op_Aff, respectively. The 'Chebyshev_Approx' and 'Combine' subroutines further translate the power or delay model into a sequence of affine operations and put them into a stack, op_Stack. The 'CDF_Generation' subroutine returns the CDF bounds represented by PLPB, denoted by dummy_CDF. The 'PUSH' subroutine is the push operation to push the dummy_CDF into the stack, CDF_Stack; the 'POP' subroutine is the pop operation. The 'CAA_Bound_Computation' subroutine is responsible for generating the correlation CDF bound under a specified affine operation. By repeatedly performing the 'CAA_Bound_Computation' subroutine, the algorithm predicts the distribution information for power and timing metrics which are represented by 'Distribution'. Then the 'Prob' subroutine returns the parametric yield by computing the CDF value at limit M_0. The resulting power yield Y_L and timing yield Y_D are determined as two objective functions in our bi-objective optimization framework.

After determining the objective functions in our proposed power-delay bi-objective optimization framework, the proposed bi-objective optimization model can be rigorously expressed as follows:

$$\begin{aligned}&\max\ Y_L = F_L(L_0) = P(\text{Leakage} \le L_0)\\ &\max\ Y_D = F_D(D_0) = P(\text{Delay} \le D_0)\\ &\text{subject to } x_L \le x \le x_U,\ x\in[L, V_{th}, T_{ox}, V_{dd}, T],\end{aligned} \tag{17}$$

where F_L and F_D are distribution functions with respect to design parameters, and x_L and x_U are the boundary values for PVT parameters over the design region. Metric limits L_0 and D_0 are predetermined values.

4 AWS-based bi-objective optimization

The adaptive weighted sum method (Kim and de Weck, 2005) is a methodology that effectively determines the Pareto front for a multi-objective optimization problem. It can produce well-distributed Pareto-optimal solutions by changing the weights adaptively. In this work, the AWS method is used to address the power-delay bi-objective optimization issue considering both leakage and delay limits.

4.1 Pareto-optimality

In a multi-optimization framework, the objective function $\boldsymbol{f}(\boldsymbol{x})=[f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})]$ often conflicts with each other (Kashfi *et al.*, 2011), such as the leakage power and delay in a circuit design. For conflicting objectives, it is not feasible to optimize the performance for all of them; improving one will result in deteriorating another. In such a case, we strive for Pareto-optimality that ensures the best overall performance. Here, a Pareto-optimal solution can be defined as follows (Srinivas and Deb, 1994; Li and Lian, 2008; Lourenco and Horta, 2012):

Definition 1 (Pareto-optimal solution) (Li and Lian, 2008) Given $\boldsymbol{u}^* \in U$, if $\neg\exists \boldsymbol{u} \in U$ s.t. $\boldsymbol{u} \prec \boldsymbol{u}^*$, $\boldsymbol{u}^*$ is said to be a Pareto-optimal solution.

The surface consisting of the complete set of Pareto-optimal solutions in the objective space is then called the Pareto-optimal front.

In this work, the optimization problem can be attributed as a bi-objective issue (it can, however, be extended to the multi-optimization case) whose two objectives are power and timing yield. AWS is an adaptive approach for multi-objective optimization. Different from the traditional weighted sum method, the weighting factor in AWS is not predetermined but evolves according to the nature of the Pareto front. By updating the weighting factor adaptively, AWS focuses on unexplored regions where no solution can be obtained by the traditional method; therefore, it is able to extract new Pareto-optimal solutions in these regions and generate a well-distributed Pareto front (Kim and de Weck, 2005).

4.2 Bi-objective optimization procedure

For the bi-objective problem (17), AWS starts with a traditional weighted sum optimization procedure performed on the objective functions normalized in the objective space. To be specific, given two objective functions, maximizing power yield Y_{L} and maximizing timing yield Y_{D}, and design parameters $x \in [L, V_{\mathrm{th}}, T_{\mathrm{ox}}, V_{\mathrm{dd}}, T]$, the optimization model (17) can be stated as

$$\begin{aligned} &\max\ \alpha Y_{\mathrm{L}}' + (1-\alpha) Y_{\mathrm{D}}' \\ &\text{subject to } x_{\mathrm{L}} \le x \le x_{\mathrm{U}},\ x \in [L, V_{\mathrm{th}}, T_{\mathrm{ox}}, V_{\mathrm{dd}}, T], \\ &\qquad \alpha \in [0,1], \end{aligned} \tag{18}$$

where Y_{L}' and Y_{D}' are the normalized objective functions of Y_{L} and Y_{D}, respectively, which are defined as

$$Y_{\mathrm{L}}' = \frac{Y_{\mathrm{L}} - Y_{\mathrm{L}}^{\mathrm{U}}}{Y_{\mathrm{L}}^{\mathrm{N}} - Y_{\mathrm{L}}^{\mathrm{U}}}, \quad Y_{\mathrm{D}}' = \frac{Y_{\mathrm{D}} - Y_{\mathrm{D}}^{\mathrm{U}}}{Y_{\mathrm{D}}^{\mathrm{N}} - Y_{\mathrm{D}}^{\mathrm{U}}}. \tag{19}$$

Take Y_{L}' as an example. Assume $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ are the optimal solutions for the single objective optimization of Y_{L} and Y_{D}, respectively. Then, $Y_{\mathrm{L}}^{\mathrm{U}}$ can be obtained by $Y_{\mathrm{L}}^{\mathrm{U}} = Y_{\mathrm{L}}(\boldsymbol{x}_1^*)$, and $Y_{\mathrm{L}}^{\mathrm{N}}$ is determined by $Y_{\mathrm{L}}^{\mathrm{N}} = \max[Y_{\mathrm{L}}(\boldsymbol{x}_1^*), Y_{\mathrm{D}}(\boldsymbol{x}_2^*)]$. Normalized Y_{D}' can be obtained in the same manner. The uniform step size of the weighing factor α is set as $\Delta\alpha=1/n_0$, where n_0 is the number of divisions (typically, n_0=5–10). By changing the weighting factor α according to the step size $\Delta\alpha$, a small set of optimal solutions for problem (18) will be obtained.

Generally, the optimal solutions obtained from problem (18) are not evenly distributed. Solutions may quite often appear only in some parts of the Pareto front, while no solutions are obtained in other parts. The distances between adjacent solutions differ much. To make the solutions well distributed on the Pareto front, the regions between adjacent solutions with long distances should be further explored. Fig. 4 shows an example of the Pareto front in the power-delay objective space for a specific design. Clearly, new optimal solutions need to be extracted from regions 1 and 2 to distribute all Pareto-optimal points uniformly on the Pareto front.

The regions in the power-delay objective space that need further refinement can be identified by computing the distances between adjacent solutions. If the distance is smaller than a preset value, no further refinement will be conducted in this region. Otherwise, the region with the long distance between adjacent solutions becomes a feasible region in which new solutions should be extracted. New solution extraction is implemented by imposing additional inequality constraints and solving a sub-optimization problem (Kim and de Weck, 2005).

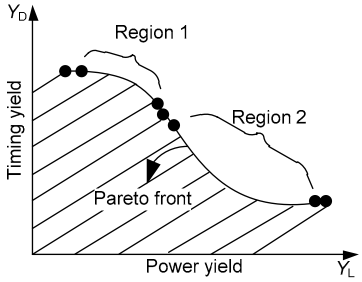


Fig. 4 An example of Pareto front for power and timing yield

The procedure is shown in Fig. 5. P_1 and P_2 are two end points of this region. δ is an offset distance defined by the user to control the final density of the Pareto solution distribution, with horizontal component δ_1 and vertical component δ_2 parallel to the Y_L and Y_D axes, respectively. The sub-optimization procedure to obtain new solutions in this region can be formulated as

$$\begin{aligned}\max\quad & \alpha_i Y_{\rm L}' + (1-\alpha_i) Y_{\rm D}' \\ \text{subject to}\quad & Y_{\rm L}(x) \le P_1^{Y_{\rm L}} - \delta_1, \\ & Y_{\rm D}(x) \le P_2^{Y_{\rm D}} - \delta_2, \\ & x_{\rm L} \le x \le x_{\rm U},\ x \in [L, V_{\rm th}, T_{\rm ox}, V_{\rm dd}, T], \\ & \alpha_i \in [0,1],\end{aligned} \tag{20}$$

where $P_i^{Y_{\rm L}}$ and $P_i^{Y_{\rm D}}$ (i=1,2) are the Y_L and Y_D positions of the end points P_1 and P_2, respectively. The weighting factor α_i for each feasible region is updated adaptively according to the relative length of this region. By solving the sub-optimization problem (20), new solutions can be identified in this region (Fig. 5c). The procedure described above is repeated in all feasible regions until a complete set of new solutions has been obtained.

Fig. 6 shows an example to explain the detailed procedures of this optimization framework.

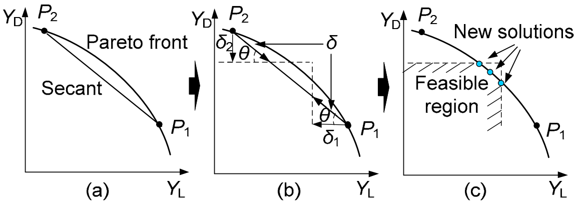


Fig. 5 Suboptimization procedure to extract new solutions: (a) initial solutions, P_1 and P_2; (b) feasible region determination; (c) new solutions

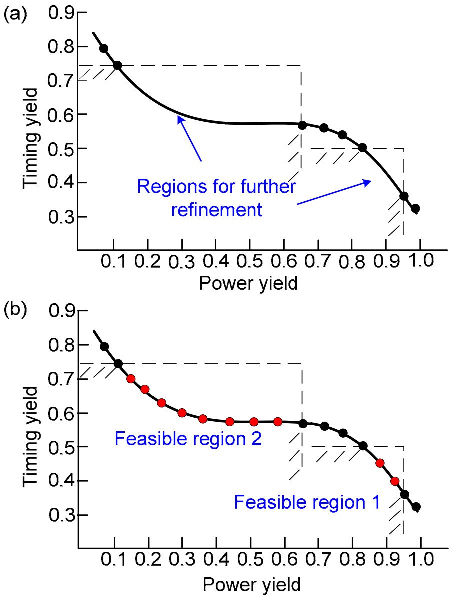


Fig. 6 First round solutions of optimization procedures (a) and second round refinement of optimization procedures (b)

The two objective functions in problem (18), power yield and timing yield, have been established by the yield prediction procedure in Section 3. The first step is to generate the first round solutions using the traditional weighted sum method. By setting $\Delta\alpha$, a small set of solutions is specified. These are not close enough to form a well-distributed Pareto front. By calculating the distances between adjacent solutions, we identify two feasible regions where extraction of a new solution is necessary (Fig. 6a).

The next step is further refinement in these two feasible regions by solving the sub-optimization problem (20). We need to determine the weighting factors α_i in (20) for each region. First, the number of further refinements required in each feasible region can be evaluated based on the relative length of the region (Kim and de Weck, 2005). We denote this number as n_i. Then, α_i can be updated adaptively with a uniform step size:

$$\Delta\alpha_i = 1/n_i. \tag{21}$$

In each region, with $\Delta\alpha_i$ substituted into Eq. (21), a set of new solutions is generated by solving this sub-optimization problem (Fig. 6b). Now the Pareto solutions are uniformly distributed on the Pareto front.

5 Experimental results

This section presents the results of the proposed bi-objective optimization framework. The computer used to perform all experiments has a quad-core 2.5 GHz CPU and a 4 GB RAM. The coefficients in the leakage model and delay model are determined by HSPICE simulations. Here, according to the empirical data in Visweswariah (2003), we model the process variations as truncated Gaussian distributions. The 3σ values of effective channel length, threshold voltage, and oxide thickness are 20%, 10%, and 8% of the nominal values, respectively. The inter- and intra-chip variations of the process parameters account for 50%, respectively. With regard to environmental parameters, power-supply voltage and on-chip temperature are assumed as being distributed uniformly. The nominal values of voltage and temperature are 1.1 V and 25 °C, respectively. The maximum voltage drop is 0.11 V (10% of the nominal value). The maximum deviation on on-chip temperature is 10 °C. The effectiveness of the algorithm is evaluated by using ISCAS benchmark circuits.

As mentioned above, the proposed framework is capable of handling arbitrary correlations among parameter variations when predicting the probability bounds for leakage power and gate delay. To verify this point, we choose circuit C432 and run Monte Carlo simulations under correlation assumptions. Positive, negative, and no correlations among PVT parameters are taken into account for comparison. Fig. 7 demonstrates that the CDFs obtained by correlation simulations are well enclosed by the guaranteed bound generated by the CAA method, both for leakage and delay distributions. The leakage and delay metrics are normalized to respective nominal values. The results also indicate the importance of taking parameters' correlation into account; without consideration of correlation, it tends to give an over-optimistic prediction of parametric yield.

Having verified the reliability of CAA predicted probability bounds, we can perform the proposed power-delay bi-objective optimization procedure based on the predicted leakage distribution F_L and delay distribution F_D. It needs to be indicated that leakage power exhibits a greater sensitivity than gate delay. Larger spread in leakage variability can be observed in Fig. 7. This difference is due to the exponential term in the leakage model, which propagates significant fluctuation in leakage power.

To describe optimization results, in this step, we take circuit C432 as an example. We set the specific delay limit, i.e., D_0 in Eq. (17) as 1.02× of the nominal delay, and L_0 is set as 1.13× of the nominal leakage power. The power and timing yield are defined as $Y_L=P\{\text{Leakage}\leq 1.13 I_{\text{nom}}\}$ and $Y_D=P\{\text{Delay}\leq 1.02 D_{\text{nom}}\}$, respectively. The proposed method generates sufficient solutions evenly distributed on the Pareto front, as shown in Fig. 8. Also, design values are randomly selected to generate the sample points in the

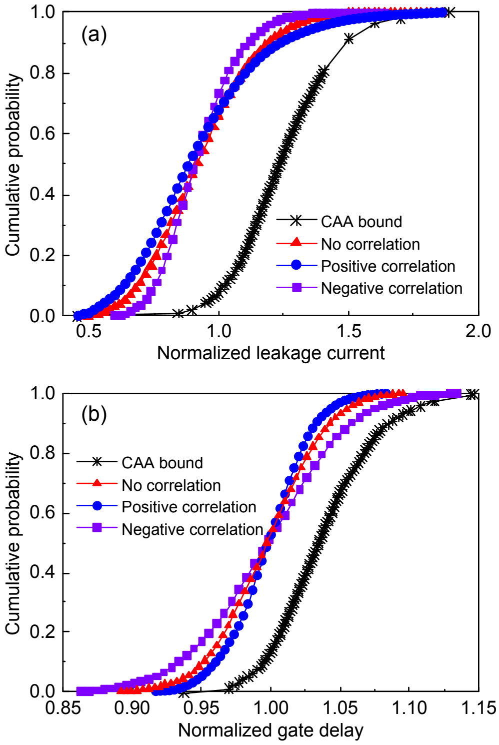


Fig. 7 Leakage power distributions (a) and gate delay distributions (b) for circuit C432

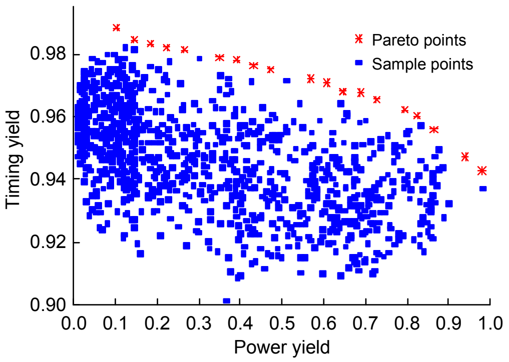


Fig. 8 Monte Carlo verification for circuit C432

power-delay objective space. Fig. 8 shows that the Pareto front predicted by the proposed method matches Monte Carlo results perfectly.

We now present the optimization results on various benchmark circuits. The experimental results demonstrate that about 30 solutions are obtained for each circuit (second column in Table 1). A few of the solutions, generated according to certain weighting factors, are listed in Table 1. On the other hand, considering a given yield level, we optimize the leakage and delay metrics that produce the given yield value. Optimization results on various ISCAS benchmark circuits are listed in Table 2, at 95% power and level of timing yield. Likewise, only a few of the solutions are provided. Both leakage power and delay values in Tables 1 and 2 have been normalized to their nominal values.

To further demonstrate the effectiveness of our bi-objective framework, we choose circuit C432 in particular to provide a set of Pareto fronts under different metric limits. When the timing constraint is fixed at 1.02× nominal delay, Fig. 9 shows the optimization results for various power yield criteria. Pareto fronts are generated by AWS according to different values of power limit L_0. The Pareto fronts for

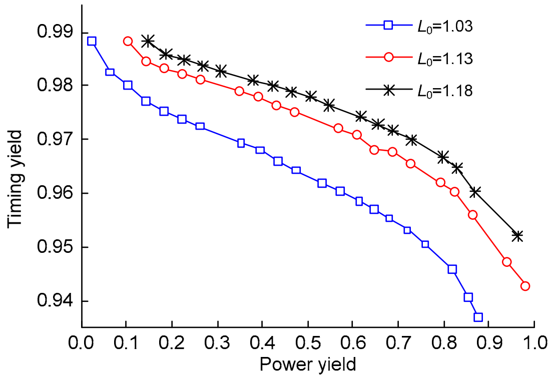


Fig. 9 The Pareto fronts for circuit C432 under different power limits

Table 1 A few Pareto-optimal solutions obtained by AWS for parametric yield

Circuit name	Number of solutions	Power yield			Timing yield			Runtime (s)
		α=1.0	α=0.5	α=0	α=1.0	α=0.5	α=0	
C432	21	0.9947	0.8176	0.1037	0.9370	0.9608	0.9883	41
C499	21	0.9947	0.8200	0.1037	0.9209	0.9479	0.9797	37
C880	18	0.9947	0.6987	0.0180	0.9568	0.9788	0.9844	81
C1335	20	0.9940	0.8171	0.1034	0.9474	0.9688	0.9893	85
C1908	23	0.9943	0.8507	0.0182	0.9054	0.9161	0.9446	106
C2670	20	0.9942	0.8200	0.0179	0.9326	0.9554	0.9672	148
C3540	22	0.9947	0.8199	0.1037	0.9369	0.9595	0.9506	201
C5315	23	0.9947	0.1037	0.0181	0.9163	0.9709	0.8741	253
C6288	23	0.9946	0.1037	0.0179	0.9283	0.9652	0.9611	361
C7552	21	0.9947	0.1036	0.0181	0.9359	0.9850	0.9707	373

Table 2 A few Pareto-optimal solutions obtained by AWS for yield percentile

Circuit name	Number of solutions	Leakage (95%)			Delay (95%)			Runtime (s)
		α=1.0	α=0.5	α=0	α=1.0	α=0.5	α=0	
C432	15	1.071	1.262	1.740	1.029	1.072	0.989	41
C499	15	1.070	1.197	1.738	1.039	1.025	0.998	37
C880	14	1.070	1.355	1.739	1.015	0.997	0.979	81
C1335	14	1.071	1.479	1.740	1.022	0.996	0.985	85
C1908	15	1.070	1.245	1.739	1.047	1.033	1.008	106
C2670	15	1.069	1.240	1.739	1.031	1.017	0.994	148
C3540	15	1.070	1.257	1.740	1.029	1.014	0.991	201
C5315	15	1.070	1.264	1.740	1.043	1.028	1.005	253
C6288	15	1.070	1.428	1.739	1.040	1.019	1.005	361
C7552	14	1.071	1.254	1.740	1.030	1.015	0.991	373

selecting different D_0 values under a fixed power limit L_0=1.13I_{nom} are described in Fig. 10. Each curve in the power-delay objective space represents a Pareto front, while each point in these curves denotes a particular Pareto-optimal solution. All these Pareto-optimal points are obtained by the AWS method, and they compose the well-distributed Pareto fronts, providing the designers with useful and flexible trade-off information between power and timing yield.

Finally, Fig. 11 provides the optimal power-delay curves for circuit C432 at different yield levels. Both power and timing yields are selected identically at 99%, 95%, and 85%, respectively. The respective Pareto-optimal curves of power and delay percentiles can be extracted by AWS accordingly.

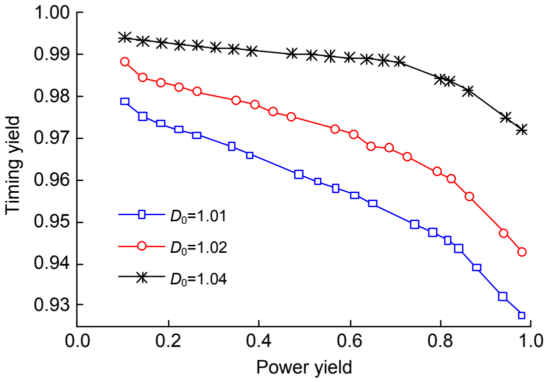


Fig. 10 The Pareto fronts for circuit C432 under different timing limits

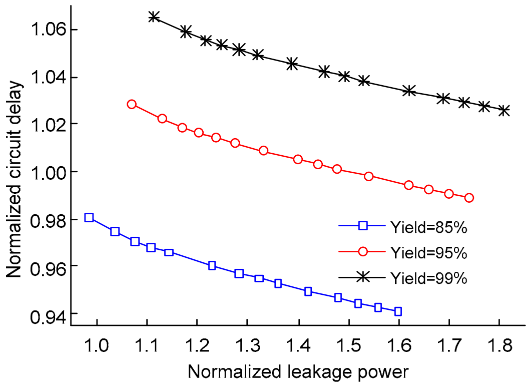


Fig. 11 Power-delay curves for circuit C432 at different yield levels

6 Conclusions

This paper proposes a novel power-delay bi-objective optimization methodology for statistical yield optimization. Regarding both power and timing yield as objective functions, an efficient bi-objective optimization framework is suggested to optimize these two goals simultaneously under PVT parameter variations. The proposed algorithm was verified using ISCAS benchmark circuits, demonstrating its efficiency.

References

Banerjee, A., Chatterjee, A., 2015. Signature driven hierarchical post-manufacture tuning of RF systems for performance and power. *IEEE Trans. VLSI Syst.*, **23**(2): 342-355. http://dx.doi.org/10.1109/TVLSI.2014.2309114

de Figueiredo, L.H., Stolfi, J., 2004. Affine arithmetic: concepts and applications. *Numer. Algor.*, **37**(1):147-158. http://dx.doi.org/10.1023/B:NUMA.0000049462.70970.b6

Gong, F., Yu, H., He, L., 2011. Stochastic analog circuit behavior modeling by point estimation method. Proc. Int. Symp. on Physical Design, p.175-182. http://dx.doi.org/10.1145/1960397.1960437

Guerra-Gómez, I., Tlelo-Cuautle, E., de la Fraga, L., 2013. Richardson extrapolation-based sensitivity analysis in the multi-objective optimization of analog circuits. *Appl. Math. Comput.*, **222**:167-176. http://dx.doi.org/10.1016/j.amc.2013.07.059

Guerra-Gómez, I., Tlelo-Cuautle, E., de la Fraga, L., 2015. OCBA in the yield optimization of analog integrated circuits by evolutionary algorithms. IEEE Int. Symp. on Circuits & Systems, p.1933-1936. http://dx.doi.org/10.1109/ISCAS.2015.7169051

Hwang, E.J., Kim, W., Kim, Y.H., 2013. Timing yield slack for timing yield-constrained optimization and its application to statistical leakage minimization. *IEEE Trans. VLSI Syst.*, **21**(10):1783-1796. http://dx.doi.org/10.1109/TVLSI.2012.2220792

Kanj, R., Joshi, R., Nassif, S., 2010. Statistical leakage modeling for accurate yield analysis the CDF matching method and its alternatives. ACM/IEEE Int. Symp. on Low-Power Electronics and Design, p.337-342.

Kashfi, F., Hatami, S., Pedram, M., 2011. Multi-objective optimization techniques for VLSI circuits. 12th Int. Symp. on Quality Electronic Design, p.156-163. http://dx.doi.org/10.1109/ISQED.2011.5770720

Kim, I.Y., de Weck, O.L., 2005. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Struct. Multidiscipl. Optim.*, **29**(2):149-158. http://dx.doi.org/10.1007/s00158-004-0465-1

Li, H., Lian, J., 2008. Multi-objective optimization of water-sedimentation-power in reservoir based on Pareto-optimal solution. *Trans. Tianjin Univ.*, **14**(4):282-288. http://dx.doi.org/10.1007/s12209-008-0048-0

Liu, X.X., Tan, S.X.D., Palma-Rodriguez, A.A., *et al.*, 2013. Performance bound analysis of analog circuits in frequency- and time-domain considering process variations. *ACM Trans. Des. Autom. Electron. Syst.*, **19**(1): 1-22. http://dx.doi.org/10.1145/2534395

Lourenço, N., Horta, N., 2012. GENOM-POF: multi-objective evolutionary synthesis of analog ICs with corners validation. Proc. 14th Int. Conf. on Genetic and Evolutionary Computation, p.1119-1126. http://dx.doi.org/10.1145/2330163.2330318

Mande, S.S., Chandorkar, A.N., Iwai, H., 2013. Computationally efficient methodology for statistical characterization and yield estimation due to inter- and intra-die process variations. 5th Asia Symp. on Quality Electronic Design, p.287-294. http://dx.doi.org/10.1109/ASQED.2013.6643602

Mani, M., Devgan, A., Orshansky, M., 2005. An efficient algorithm for statistical minimization of total power under timing yield constraints. Proc. Design Automation Conf., p.309-314. http://dx.doi.org/10.1109/DAC.2005.193823

Mani, M., Devgan, A., Orshansky, M., *et al.*, 2007. A statistical algorithm for power- and timing-limited parametric yield optimization of large integrated circuits. *IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst.*, **26**(10):1790-1802. http://dx.doi.org/10.1109/TCAD.2007.895797

Orshansky, M., Bandyopadhyay, A., 2004. Fast statistical timing analysis handling arbitrary delay correlations. Proc. 41st Annual Design Automation Conf., p.337-342. http://dx.doi.org/10.1145/996566.996664

Radfar, M., Singh, J., 2014. A yield improvement technique in severe process, voltage, and temperature variations and extreme voltage scaling. *Microelectron. Reliab.*, **54**(12): 2813-2823. http://dx.doi.org/10.1016/j.microrel.2014.07.138

Rao, R., Devgan, A., Blaauw, D., *et al.*, 2004a. Parametric yield estimation considering leakage variability. Proc. 41st Annual Design Automation Conf., p.442-447. http://dx.doi.org/10.1145/996566.996693

Rao, R., Srivastava, A., Blaauw, D., *et al.*, 2004b. Statistical analysis of subthreshold leakage current for VLSI circuits. *IEEE Trans. VLSI Syst.*, **12**(2):131-139. http://dx.doi.org/10.1109/TVLSI.2003.821549

Saad, A., Frühwirth, T., Gervet, C., 2014. The p-box CDF-intervals: a reliable constraint reasoning with quantifiable information. *Theory Pract. Log. Programm.*, **14**(4-5):461-475. http://dx.doi.org/10.1017/S1471068414000143

Sheng, Y., Xu, K., Wang, D., *et al.*, 2013. Performance analysis of FET microwave devices by use of extended spectral-element time-domain method. *Int. J. Electron.*, **100**(5): 699-717. http://dx.doi.org/10.1080/00207217.2012.720947

Srinivas, N., Deb, K., 1994. Multi-objective optimization using non-dominated sorting in genetic algorithms. *Evol. Comput.*, **2**(3):221-248. http://dx.doi.org/10.1162/evco.1994.2.3.221

Srivastava, A., Chopra, K., Shah, S., *et al.*, 2008. A novel approach to perform gate-level yield analysis and optimization considering correlated variations in power and performance. *IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst.*, **27**(2):272-285. http://dx.doi.org/10.1109/TCAD.2007.907227

Sun, J., Huang, Y., Li, J., *et al.*, 2008. Chebyshev affine arithmetic based parametric yield prediction under limited descriptions of uncertainty. Proc. Asia and South Pacific Design Automation Conf., p.531-536. http://dx.doi.org/10.1109/ASPDAC.2008.4484008

Ukhov, I., Eles, P., Peng, Z., 2014. Probabilistic analysis of power and temperature under process variation for electronic system design. *IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst.*, **33**(6):931-944. http://dx.doi.org/10.1109/TCAD.2014.2301672

Visweswariah, C., 2003. Death, taxes and falling chips. Proc. Design Automation Conf., p.343-347. http://dx.doi.org/10.1109/DAC.2003.1219021

Wang, W.S., Orshansky, M., 2006. Robust estimation of parametric yield under limited descriptions of uncertainty. Proc. IEEE/ACM Int. Conf. on Computer-Aided Design, p.884-890. http://dx.doi.org/10.1109/ICCAD.2006.320093

Williamson, R.C., Downs, T., 1990. Probabilistic arithmetic. I. numerical methods for calculating convolutions and dependency bounds. *Int. J. Approx. Reason.*, **4**(2):89-158. http://dx.doi.org/10.1016/0888-613X(90)90022-T

Xie, L., Davoodi, A., 2008. Robust estimation of timing yield with partial statistical information on process variations. 9th Int. Symp. on Quality Electronic Design, p.156-161. http://dx.doi.org/10.1109/ISQED.2008.4479718

Zhu, W., Wu, Z., 2014. The stochastic ordering of mean-preserving transformations and its applications. *Eur. J. Oper. Res.*, **239**(3):802-809. http://dx.doi.org/10.1016/j.ejor.2014.06.017