A Comprehensive Stochastic Design Methodology for Hold-Timing Resiliency in Voltage-Scalable Design

Zhengyu Chen, *Student Member, IEEE*, Huanyu Wang, *Student Member, IEEE*, Geng Xie, *Student Member, IEEE*, and Jie Gu, *Member, IEEE*

***Abstract*— In order to fulfill different demands between ultralow energy consumption and high performance, integrated circuits are being designed to operate across a large range of supply voltages, in which resiliency to timing violation is the key requirement. Unfortunately, traditional timing analysis which focuses on setup timing tolerance for higher performance cannot model the hold violation efficiently across different voltages. In this paper, we proposed a complete flow of computationally efficient methodology for guaranteeing hold margin, which is particularly important for low-power devices, e.g., Internet-of-Things devices. Leveraging both the conventional static timing analysis and a most probable point (MPP) theory, we develop a new hold-timing closure methodology across voltages eliminating expensive Monte Carlo simulation. To improve the efficiency of locating MPP, a novel MPP search method is proposed that employs a set of approximation eliminating the time-consuming iterative search. Several novel modeling techniques, such as the incorporation of nonlinear behaviors of circuit operations and correlation coefficient modeling, are also proposed in this paper to significantly improve the accuracy of the design. With the proposed modeling techniques, a novel hold resilient design technique equipped with the proposed variation-aware timing resilient flip-flop is developed. The proposed design technique eliminates the excessive hold-fixing operation at low voltage and its associated performance degradation at high voltage, whereas still being compatible with the conventional design closure flow. Design example on a voltage-scalable digital signal processor was used to demonstrate the potential of the technique. The result in a 45-nm technology shows that the elimination of more than 20 000 hold buffers, 23% performance improvement at high voltages, and 7% area saving are achieved using the proposed technique compared with the conventional digital design technique.**

***Index Terms*— Hold and setup violations, lognormal distribution, most probable point (MPP), resilient design, ultralow-voltage operation.**

I. Introduction

ULTRALOW-POWER design has recently drawn tremendous interest from consumer electronic industry due to the rapid growth of mobile device market. Unfortunately, the applications utilized by such devices raise significant challenges due to the conflicting requirements between low power consumption and high performance. To achieve low power consumption, supply voltage V_{DD} is typically reduced to near-threshold voltages, e.g., around 0.5 V. At such a voltage, stochastic variation associated with transistor threshold voltages becomes a major factor in determining logic timing which makes timing analysis extremely time consuming [1]. Conventional solution which introduces extra timing margin to avoid timing violation at low-voltage region leads to performance degradation at high-voltage region. In addition, the inefficiency of conventional static timing analysis (STA) cannot meet the demand of shorter time-to-market cycle requirement.

Process variation, including local variation, global variation, and systemic variation, has been continuously exacerbated throughout technology. Among them, local variation holds the largest threat to the circuit timing closure due to difficulties of: 1) controlling the threshold voltage of extremely small channel length of the transistors and 2) being captured by conventional corner base STA. At nominal voltage ($V_{DD} > 0.9$ V), it is usually accurate to assume that the circuit performance is linear in transistor variation [2]–[4]. Under this circumstance, the stochastic circuit delay follows Gaussian, and the standard deviation can be easily calculated from the standard deviations of the transistor parameters. However, at low voltage ($V_{DD} < 0.4$ V), circuit delay is a nonlinear function of the transistor random variables. This greatly complicates the statistical analysis because the probability density function (pdf) of the circuit delay is no longer Gaussian [16].

As a result, either excessive pessimism is built into conventional corner-based design or a time-consuming statistical-based analysis has to be utilized. For example, fast Monte Carlo-based statistical static timing analysis (SSTA) has been proposed to estimate the delay variation with relative large computing expense [11], [12]. Principle component analysis-based SSTA has been proposed to capture the non-Gaussian parameters in the manufacturing process, but it is not clear if such an approach can be extended into ultralow-voltage region where delay is lognormal [11]. For low-voltage operation, an operating point-based analysis is demonstrated with high accuracy to predict the delay variation of the critical paths [12]. However, the iterative search used in the technique is rather expensive and requires large computing efforts to close a design with large number of paths [12], [13]. It is important to stress that the issue of mismatch is particularly elevated for hold timing because the hold critical

Manuscript received December 4, 2017; revised March 13, 2018 and April 27, 2018; accepted May 27, 2018. Date of publication July 2, 2018; date of current version September 25, 2018. This work was supported by NSF under Grant CCF-1533656. *(Corresponding author: Zhengyu Chen.)*

The authors are with the Electrical Engineering and Computer Science Department, Northwestern University, Evanston, IL 60208 USA (e-mail: zhengyuchen2015@u.northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2018.2847622

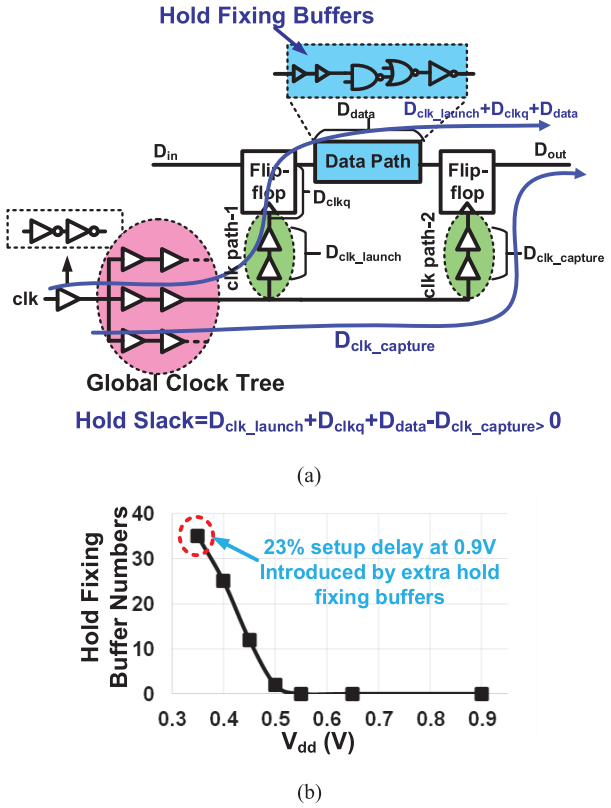


Fig. 1. (a) Schematic of a typical hold-timing paths. (b) Number of hold-fixing buffers required in one of the critical paths at various low voltages.

path is short and suffers from large amount of variation in comparison with setup timing path where the amount of variation is reduced due to a deeper logic depth. Fig. 1(a) shows the schematic drawing of hold-timing path extracted in our example fast Fourier transform (FFT) processor design in a 45-nm process. Fig. 1(b) shows the amount of hold-fixing buffer that is needed to account for the worst case delay variation of the critical path and hold-fixing buffer. To allow the design to work down to 0.35 V, 35 buffers need to be inserted which will in turn reduce the performance as well as power efficiency at high voltage. Similar observation has been shown in [8] and [14] where the min short-path delay has to increase to 60% of clock period due to local mismatch causing 2.2X of logic area increase due to hold-fixing buffers for voltage operating down to 0.35 V. Besides the performance degradation at nominal voltage introduced the extra fixing buffers, the determination of the number of hold-fixing buffers requires excessive amount of statistical-based timing analysis which becomes bottleneck of the chip design activity and also suffers from accuracy issues. Although a two-phase latch-based design was proposed with error detection built into pipeline stages to resolve the hold-fixing problem in [14], a latch-based design deviates substantially from conventional design methodology leading to complexity of the adaptation of the scheme.

A. Previous Work

To incorporate the challenges at low-voltage operation especially under process–voltage–temperature (PVT) variation, several error resilient designs have drawn significant efforts from industry to academia in the past decades [5]–[8], [22], [27]. The concept is that by incorporating error detection mechanism, the designers can remove some of safety margin to achieve further energy saving. For example, the "Razor" design includes additional latch to detect error within a detection timing window and flush the pipeline when an error is detected [5]. Several improved designs of the error resilient system have been proposed recently.

1) A bubble razor technique was introduced to stall a clock cycle and intelligently propagate the error message throughout the pipeline [6].
2) A latch-based error detection design along with the voltage boosting technique was also introduced to create delay variation resilience at low voltages [8].
3) A PVT-variation-aware error-detecting latch design is presented in [22], [25], and [26].

The abovementioned techniques can effectively create certain amount of tolerance to setup violation and improves the speed of the design. More recently, several latch-based designs with multiphase clock were proposed to provide a viable migration to the hold-timing issues at low voltages [9], [10].

However, it is important to mention that previous latch-based techniques sacrifice the hold design margin and require significant amount of hold verification and fixing effort. In addition, even if a hold violation can be detected in a similar fashion of Razor technique, the violation cannot be fixed by flushing the pipeline or slowing down the clock as proposed in previous techniques. As a result, previous solutions are more applicable for high-performance design where setup timing is more important. But it may not be a viable solution for aggressive voltage scaling down to near-threshold or subthreshold voltage region where hold violation is harder to be modeled as will be shown in this paper.

Thus, there still lacks a comprehensive methodology of hold-timing closure for low-voltage operation, which requires that it can maintain performance at high voltage while being able to function properly at ultralow-voltage without timing violation. Such a requirement favors a scheme that is not only compatible with conventional timing closure methodology but also does not sacrifice performance from enabling low-power operation.

B. Contributions of This Work

As extended from [21], this paper develops a computationally efficient methodology that can perform accurate timing analysis in the low-voltage regime where delay is a highly nonlinear function of the random variables and/or the PDFs of the random variables are non-Gaussian. Also, with the help of the proposed timing resilient flip-flop (TRFF) solution, degradation of performance at high voltage is eliminated. This paper also incorporates a most probable point (MPP) theory with conventional STA to achieve an efficient statistical timing analysis. More specifically, this paper presents the following.

1) A complete theoretical analysis on MPP-based hold-timing closure technique under nonlinear scenario.

2) A complete modeling of correlation coefficient (CC) of a buffer chain and the physical explanation of the modeling.
3) Variation-aware tracking sensor-based circuit solution.
4) A complete implementation of the proposed methodology which is integrated seamlessly into the traditional timing closure methodology for logic at low voltage.

II. Unified Statistical Hold-Timing Modeling

In this section, we discuss a modeling methodology for subtraction (SUB) using a MPP analysis. The hold slack is defined as the difference between data arrival time and data required time as shown in Fig. 1(a)

$$\text{Slack\_neg} = D_{\text{clk\_capture}} - (D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}}) \quad (1)$$

where $D_{\text{clk\_capture}}$ is the delay of capture clock path, D_{data} is the delay of data path, $D_{\text{clk\_launch}}$ is the delay of launch clock path, and D_{clkq} is the delay from clock to output of a flip-flop. Here, we calculate the negative slack because we are only interested in finding the maximum negative slack of the design. The goal of our analysis is to predict the stochastic hold slack value of (1) at a target percentile, e.g., 3-sigma of 99.73%. Although numerous efforts have been given in predicting a stochastic distribution of SUM and MAX operations of timing path, there is a lack of discussion on the SUB operation, which is critical for hold analysis. In this paper, we adapt an MPP analysis where the vectors representing the cell-level contribution to the joint probability of the entire paths are located and computed to find out the entire target numbers of sigma slack of the path [15], [16]. The use of MPP method compared with other arithmetic equation-based analysis allows us to take advantage of the existing STA result and lookup table (LUT)-based stochastic analysis to achieve dramatic reduction of characterization and modeling efforts. One of the previous works utilizes an iterative search method to converge on the required MPP of each delay component including data path and clock path [13]. However, it requires extensive iterative search to find the MPP leading to days of computation for design closure of a large design. Instead, this paper leverages STA results and proposes a simplified model that can accurately capture the stochastic distribution of the entire timing path without iterative search of MPP.

A. Most Probable Point (MPP) Approach Analysis

MPP-based methods are widely used for engineering reliability analysis and reliability-based design. Their major advantage is the good balance between accuracy and efficiency. In this section, we will discuss how to develop a computationally efficient algorithm based on MPP theory that can perform accurate path-based timing analysis in the regime where delay is a highly nonlinear function of the random variables, i.e., the PDFs of the random variables are non-Gaussian [13]–[15].

1) Linear-Gaussian Subtraction: For simplicity and the practical scenario that we are dealing with, the theory will be developed for the case of two variables. The results can easily be extended to an arbitrary number of variables. Let $D(x_1, x_2)$ be a function of two random variables x_1 and x_2. Here, D stands for stochastic delay. The random variables are statistically independent and follow the Gaussian, i.e., the individual PDFs $P_1(x)$ and $P_2(x)$ are Gaussian with different mean and standard deviation (μ_1, σ_1) and (μ_2, σ_2), respectively

$$D(x_1, x_2) = a_1 x_1 + a_2 x_2. \quad (2)$$

Invoking the theorem that the PDF of the sum to statistically independent random variables is the convolution of the PDFs of the respective variables, the PDF of $D(x_1, x_2)$ can be written as

$$P_D(D) = \int_{-\infty}^{\infty} P_{z1}(z) P_{z2}(D - z) dz \quad (3)$$

where $z_1 = a_1 x_1$ and $z_2 = a_2 x_2$

$$P_D(D) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} a_1 \sigma_1} \times e^{\left(\frac{-z^2}{2(a_1\sigma_1)^2}\right)} \times \frac{1}{\sqrt{2\pi} a_2 \sigma_2} \times e^{\left(\frac{-(D-z)^2}{2(a_2\sigma_2)^2}\right)} dz \quad (4)$$

$$= \frac{1}{2\pi a_1 \sigma_1 a_2 \sigma_2} \times e^{\left(\frac{-D^2}{2((a_1\sigma_1)^2+(a_2\sigma_2)^2)}\right)} \times \int_{-\infty}^{\infty} e^{\left(\frac{-((a_1\sigma_1)^2+(a_2\sigma_2)^2)}{2(a_1\sigma_1)^2(a_2\sigma_2)^2} \times (z - \frac{(a_1\sigma_1)^2}{(a_1\sigma_1)^2+(a_2\sigma_2)^2})\right)} dz \quad (5)$$

$$= \frac{1}{2\pi a_1 \sigma_1 a_2 \sigma_2} \times e^{\left(\frac{-D^2}{2((a_1\sigma_1)^2+(a_2\sigma_2)^2)}\right)} \times c \quad (6)$$

where c is a constant value.
From (5), the integrand peaks at point (z_1, z_2), where

$$z_1 = \frac{(a_1\sigma_1)^2}{(a_1\sigma_1)^2 + (a_2\sigma_2)^2} \quad z_2 = \frac{(a_2\sigma_2)^2}{(a_1\sigma_1)^2 + (a_2\sigma_2)^2} \quad (7)$$

$$x_1 = \frac{z_1}{a_1} = \frac{a_1\sigma_1^2}{(a_1\sigma_1)^2 + (a_2\sigma_2)^2}$$

$$x_2 = \frac{z_2}{a_2} = \frac{a_2\sigma_2^2}{(a_1\sigma_1)^2 + (a_2\sigma_2)^2} \quad (8)$$

and falls sharply as $\exp[((-z^2((a_1\sigma_1)^2+(a_2\sigma_2)^2))/(2(a_1\sigma_1)^2 (a_2 r\sigma_2)^2))]$. Point (x_1, x_2) in (8) is known as the MMP which contributes the most to the PDF of $D(x_1, x_2)$ [15], [16]. Fig. 2(a) shows the graphic illustration in x_i-space of the convolution integrand and the MPP under the linear-Gaussian condition. All points of (x_1, x_2) and that contribute to the ε-sigma value of the stochastic delay lie on the line $D(x_1, x_2) = a_1 x_1 + a_2 x_2 = D_{\varepsilon-\sigma}$ [16]. In geometry, this MMP which contributes the most to the integrand of PDF is also the shortest distance point from the origin to the line $D(x_1, x_2) = D_{\varepsilon-\sigma}$ [15], [16], [19]. Then, we normalize the variables as

$$\omega_1 = \frac{x_1 - \mu_1}{\sigma_1} \quad \omega_2 = \frac{x_2 - \mu_2}{\sigma_2}. \quad (9)$$

Equation (2) can be converted to standard normal variable space with ε-sigma value of the stochastic delay

$$a_1(\sigma_1\omega_1 + \mu_1) + a_2(\sigma_2\omega_2 + \mu_2) = D_{\varepsilon-\sigma} \quad (10)$$

$$a_1\sigma_1\omega_1 + a_2\sigma_2\omega_2 = D^*_{\varepsilon-\sigma} \quad (11)$$

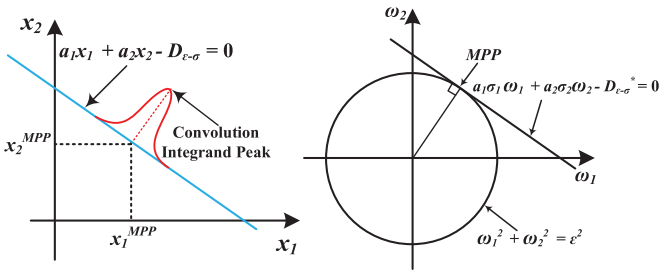


Fig. 2. (a) MPP concept and graphic illustration in x_i-space of the convolution integrand (b) Geometric illustration of MPP in normalized ω_i-space.

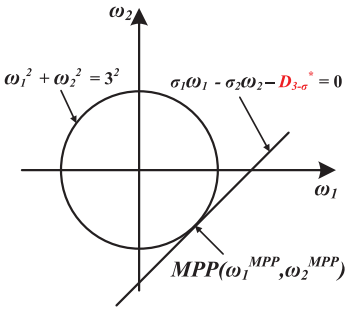


Fig. 3. MPP of linear-Gaussian SUB case.

where $D^*_{\varepsilon-\sigma}$ is the ε-sigma value of the stochastic delay in ω_i-space. From geometry, that line (11) is perpendicular to the radius of the circle shown in (12) at the MMP $(\omega_1^{\text{MPP}}, \omega_2^{\text{MPP}})$. Its geometric illustration is shown in Fig. 2(b)

$$\omega_1^2 + \omega_2^2 = \varepsilon^2 \tag{12}$$

in order to fulfill our goal which is to predict the stochastic hold slack value of equation at a target percentile, e.g., 3-sigma of 99.73% for the real case shown by Fig. 1. We can adapt (9)–(11) by making $a_1 = 1$, $a_2 = -1$, and $\varepsilon = 3$. Here, x_1 represents the stochastic delay of capture clock ($D_{\text{clk\_capture}}$) which follows Gaussian with mean and standard deviation (μ_1, σ_1); x_2 represents the stochastic delay summation (SUM) of launch clock data path ($D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}}$) which also follows Gaussian with mean and standard deviation (μ_2, σ_2)

$$D_{3-\sigma} = x_1 - x_2 \tag{13}$$

$$\omega_1^2 + \omega_2^2 = 3^2. \tag{14}$$

By solving (9), (13), and (14), the stochastic hold slack value at 3-sigma ($D_{3-\sigma}$) can be obtained easily. The graphic illustration for (13) and (14) is shown in Fig. 3.

2) Nonlinear-Gaussian Subtraction and Summation: In Section II-A1, we discussed how to estimate stochastic hold slack value at 3-sigma under linear-Gaussian assumption. This is true when the supply voltage is high, e.g., at nominal voltage [2]–[4]. Under this condition, the circuit delay is Gaussian and whose mean and standard deviation can be either calculated by transistor parameters or obtained from Monte Carlo-based simulation. However, at low voltage,

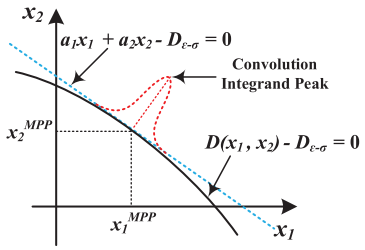


Fig. 4. Graphic illustration in x_i-space of nonlinear-Gaussian case.

e.g., at near-threshold or subthreshold voltage, the delay of cells in timing paths follows the lognormal model. Hence, the negative hold slack at low voltage could be formulated as a SUB of two lognormal items as shown in the following equation:

$$D(x_1, x_2) = e^{x_1} - e^{x_2}. \tag{15}$$

where x_1 represents the transformed random variable of the stochastic delay of capture clock, $\ln(D_{\text{clk\_capture}})$, which follows Gaussian with mean and standard deviation (μ_1, σ_1), and x_2 represents the transformed random variable of stochastic delay of launch clock data path, $\ln(D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}})$, which also follows Gaussian with mean and standard deviation (μ_2, σ_2).

For smoothly varying functions, the convolution integrand still has a maximum at the MMP [16], [19]. This point is also the maximum in the joint probability density of and on the curve [16]. Fig. 4 shows the graphic illustration in x_i-space of the convolution integrand and the MPP in the nonlinear-Gaussian case. We then normalize the $D(x_1, x_2)$ as

$$D(\omega_1, \omega_2) = e^{\sigma_1\omega_1+\mu_1} - e^{\sigma_2\omega_2+\mu_2} \tag{16}$$

$$\omega_1 = \frac{x_1 - \mu_1}{\sigma_1} \quad \omega_2 = \frac{x_2 - \mu_2}{\sigma_2}. \tag{17}$$

With the first-order Taylor series around MMP $(\omega_1^{\text{MPP}}, \omega_2^{\text{MPP}})$, the $(D(x_1, x_2)$ is approximated by

$$D(\omega_1, \omega_2)^* = D\left(\omega_1^{\text{MPP}}, \omega_2^{\text{MPP}}\right) + \sum_{i=1}^{2}\left(\frac{\partial e^{\sigma_i\omega_i+\mu_i}}{\partial\omega_i}|_{\omega_i^{\text{MPP}}}\left(\omega_i - \omega_i^{\text{MPP}}\right)\right) \tag{18}$$

$$= \left(e^{\mu_2+\sigma_2\omega_2^{\text{MPP}}} - e^{\mu_2+\sigma_2\omega_2^{\text{MPP}}}\right) + \sigma_1 e^{\mu_1+\sigma_1\omega_1^{\text{MPP}}}\left(\omega_1 - \omega_1^{\text{MPP}}\right) - \sigma_2 e^{\mu_2+\sigma_2\omega_2^{\text{MPP}}}\left(\omega_1 - \omega_1^{\text{MPP}}\right) = D_{\varepsilon-\sigma}. \tag{19}$$

The linear approximation is shown in Fig. 5. The error between the linear approximation and original curve in ω_i-space is small in the vicinity of the MPP and increases away from the MPP. But only the points in the vicinity of the MPP contribute significantly toward convolution. In ω_i-space, all points that contribute to the ε-sigma delay lie on the curve $D(\omega_1, \omega_2)^* = D_{\epsilon-\sigma}$ are perpendicular to the radius of the circle shown in the following equation:

$$\omega_1^2 + \omega_2^2 = \varepsilon^2. \tag{20}$$

Thus, if the original curve is linear enough, the approximated MPP$(\omega_1^{\text{MPP}*}, \omega_2^{\text{MPP}*})$ can be calculated by

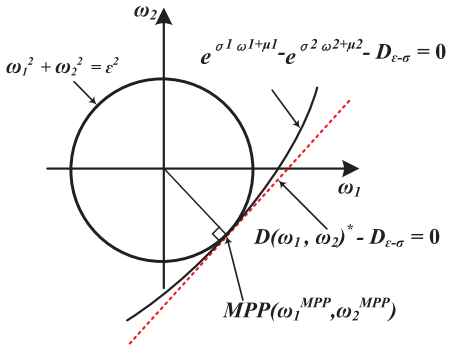


Fig. 5. MPP of nonlinear-Gaussian SUB case.

solving (19) and (20). Because the MPP values are located at the tangent point between the 3-σ cycle and delay contour, additional constraint equations can be obtained by taking differential operation to (16) to find out the tangentof the delay contour

$$\frac{d\omega_2}{d\omega_1} = \frac{\sigma_1 e^{\sigma_1\omega_1+\mu_1}}{\sigma_2 e^{\sigma_2\omega_2+\mu_2}}. \tag{21}$$

Considering the tangency condition to the cycle

$$\frac{d\omega_2}{d\omega_1} \times \frac{\omega_2}{\omega_1} = -1. \tag{22}$$

Finally, we obtain the additional equation for solving ω_1 and ω_2

$$\frac{\omega_1}{\omega_2} = -\frac{\sigma_1 e^{\sigma_1\omega_1+\mu_1}}{\sigma_2 e^{\sigma_2\omega_2+\mu_2}}. \tag{23}$$

Hence, combining (20) and (23), the exact values of ω_1 and ω_2 can be calculated. We calculated the MPP at ε-sigma of ω_1 and ω_2 from (20) and (23) with different (μ_1, σ_1) and (μ_2, σ_2) whose values are extracted from the Spice simulation on standard cells across voltages from 0.35 to 0.9 V. Then, the 3-sigma value of negative hold slack can be obtained by substituting the value of ω_1 and ω_2 into (16).

Similarly, for the SUM of lognormal nonlinear-Gaussian case, the MPP and 3-sigma stochastic delay value can be calculated by solving (17) and the following equations:

$$D(\omega_1, \omega_2) = e^{\sigma_1\omega_1+\mu_1} + e^{\sigma_2\omega_2+\mu_2} = D_{3-\sigma} \tag{24}$$

$$\frac{\omega_1}{\omega_2} = \frac{\sigma_1 e^{\sigma_1\omega_1+\mu_1}}{\sigma_2 e^{\sigma_2\omega_2+\mu_2}} \tag{25}$$

$$\omega_1^2 + \omega_2^2 = 3^2. \tag{26}$$

3) Accuracy and Limitation of Proposed MPP-Based Method: As discussed in Section II-A2, when the stochastic hold slack is a nonlinear Gaussian case: 1) we can use (16), (20), and (23) to estimate the ε-sigma SUB delay value and 2) we can use (24)–(26) to calculate the ε-sigma SUM delay value. This approach will be accurate under the condition that the function $D(\omega_1, \omega_2)$ is approximately linear in the vicinity of the operating point. We can rewrite (16) as

$$\omega_2 = \frac{\ln(e^{\sigma_1\omega_1+\mu_1} - D_{\varepsilon-\sigma}) - \mu_2}{\sigma_2}. \tag{27}$$

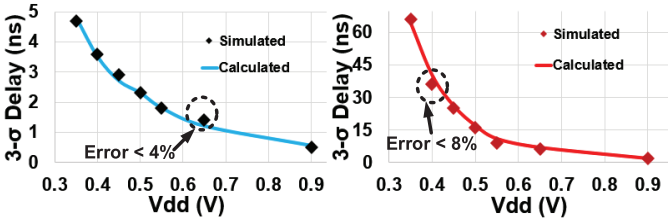


Fig. 6. (a) Accuracy of SUB case. (b) Accuracy of SUM case.

Taking the second derivative of (27), we get

$$\text{Linearity}_{\text{sub}} = \frac{\partial^2\omega_2}{\partial\omega_1^2} = \frac{\partial^2\left(\frac{\ln(e^{\sigma_1\omega_1+\mu_1} - D_{\varepsilon-\sigma}) - \mu_2}{\sigma_2}\right)}{\partial\omega_1^2} \tag{28}$$

$$= \frac{\sigma_1^2}{\sigma_2}\frac{e^{\sigma_1\omega_1+\mu_1} D_{\varepsilon-\sigma}}{(e^{\sigma_1\omega_1+\mu_1} - D_{\varepsilon-\sigma})^2}. \tag{29}$$

Then, the linearity of (16) at MPP can be estimated by substituting the real value of $\omega_1^{\text{MPP}}, \sigma_1, \sigma_2, \mu_1, D_{\varepsilon-\sigma}$ from Monte Carlo simulation into (29). Under real circuit condition, the calculated values of (29) are smaller than 0.1 at different supply voltages which makes our model for SUB quite accurate.

Similar approach can be taken to verify the model accuracy of lognormal SUM case.

Equation (24) can be rewrite as

$$\omega_2 = \frac{\ln(D_{\varepsilon-\sigma} - e^{\sigma_1\omega_1+\mu_1}) - \mu_2}{\sigma_2}. \tag{30}$$

Taking the second derivative of (30), we get

$$\text{Linearity}_{\text{sum}} = \frac{\partial^2\omega_2}{\partial\omega_1^2} = \frac{\partial^2\left(\frac{\ln(D_{\varepsilon-\sigma} - e^{\sigma_1\omega_1+\mu_1}) - \mu_2}{\sigma_2}\right)}{\partial\omega_1^2} \tag{31}$$

$$= \frac{\sigma_1^2}{\sigma_2}\frac{e^{\sigma_1\omega_1+\mu_1} D_{\varepsilon-\sigma}}{(D_{\varepsilon-\sigma} - e^{\sigma_1\omega_1+\mu_1})^2}. \tag{32}$$

Then, the linearity of (24) at MPP can be calculated by substituting the real value of $\omega_1^{\text{MPP}}, \sigma_1, \sigma_2, \mu_1, D_{\varepsilon-\sigma}$ from Monte Carlo simulation into (32). The values of (32) are smaller than 0.3 at different supply voltages which also makes our model for SUM quite accurate.

Fig. 6(a) shows the accuracy of calculating the 3-sigma delay with the proposed method in SUB case, in which the error is smaller than 4%; Fig. 6(b) shows the accuracy of calculating the 3-sigma delay in SUM case, in which the error is smaller than 8%. Fig. 7(a) shows real case that the location of MPP points for both SUM and SUB operations based on 100 000 Monte Carlo simulation with the μ and σ of lognormal delay extracted from real circuits using Spice simulation on standard cell buffers operating at 0.45 V. Each point represents a pair of ω_1 and ω_2 points that provides the same 3σ values for SUM and SUB operations. The group of points forms an equal delay contour for SUM and SUB in the hyper-space of ω_1 and ω_2.

Two major observations are highlighted here which testify the method we introduced in Section II-A3 including: 1) the MPP for both SUM and SUB happens near the tangent points of the equal delay contour and the sphere with a radius of

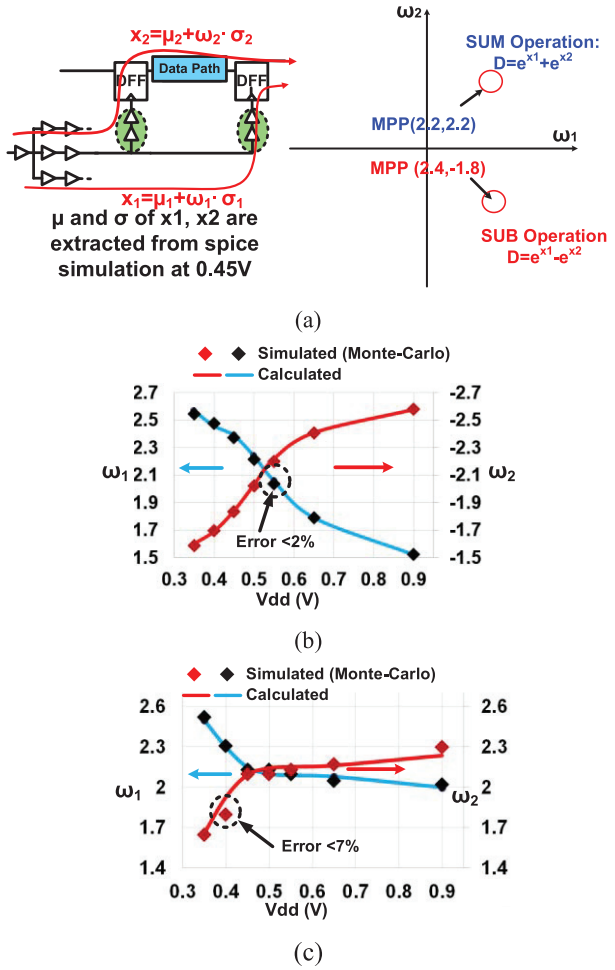


Fig. 7. (a) Monte Carlo simulated equal delay contour for SUM and SUB equations with parameters based on Spice simulation at 0.45 V. Simulated and calculated MPP of ω_1 and ω_2 for (b) SUB operation and (c) SUM operation.

the target sigma of 3 matching the theoretical expectation and 2) the MPP values of ω_1 and ω_2 represent a "balance" of the two random variables ω_1 and ω_2. For SUM, both values contribute equally and thus ω_1 and ω_2 have similar values. For SUB, the MPP settles toward unequal values, i.e., $\omega_1 = 2.4$ and $\omega_2 = -1.8$ because at the far-out tail of 3-sigma slack, the contribution from ω_1 dominates the contribution from ω_2 due to the lognormal behavior of the delay, i.e., positive tail outruns negative tail.

Fig. 7(b) and (c) shows the calculated values of ω_1 and ω_2 using (20), (23), and (25) in comparison with the Monte Carlo simulation. The calculated values match with the Monte Carlo simulation value within 2% error for SUB and 7% for SUM. This confirms that we could analytically calculate the 3-sigma value of negative hold slack by finding out the ω_1-sigma value of $D_{\text{clk\_capture}}$ minus the ω_2-sigma value of $D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}}$ as shown in (33). It is interesting to observe that the ω_1 and ω_2 values reverse the trend at high voltage, e.g., 0.9 V. This is because the delay distribution becomes Gaussian distribution at high voltage. The SUM of launch clock path and data path has a longer delay than does the capture clock path (launch and capture clock path are balanced in clock design) and thus starts to dominate the overall hold slack at high voltages

$$D_{3-\sigma} = (D_{\text{clk\_capture}})_{\omega 1} - (D_{\text{clk\_launch}} + D_{\text{clk\_to\_q}} + D_{\text{data}})_{\omega 2}. \tag{33}$$

Although it is possible to predict the MPP values of ω_1 and ω_2, in reality, the MPP values depend on the circuit configuration, i.e., values of μ and σ. As a result, a large number of circuit characterization still needs to be performed to obtain MPP values. To simplify the analysis and characterization, we leverage the following conditions to reduce the analysis space.

1) Corner-based STA can be utilized to provide the results for the negative portion of the analysis, i.e., ω_2. This eliminates the majority effort of characterization and modeling. This means that characterization of the entire standard cell library and the large numbers of data path delay is no longer needed with the help of STA.
2) Since only the capture clock delay needs to be stochastically predicted, the design space has been dramatically reduced by only characterizing the limited variety of clock buffers and depths of clock paths.

Section II-B explains our approach.

B. Subtraction Using Corner-Based STA

In this section, we discuss the methodology of utilizing corner-based STA of the data path to simplify our analysis. Note that, in our analysis, we focused more of the local random variation instead of the global variation since it holds the largest threat to the timing closure. In this paper, the local variation is applied on top of global corners which are already located at 0-, 1-, 2-, 3-sigma at global variation space. If global corner is too pessimistic at low voltage, the global corner should be set at close to 0-sigma and more variation should be given to local variation, which is handled by our method.

Based on the Spice simulation using global corner, the STA corner value of delay for a combination circuit, e.g., a chain of buffer, is always located at a negative sigma location when compared with Monte Carlo simulation. This observation deviates from the general expectation of corner location at 0-sigma but can be well explained from SUM operation of lognormal variables.

To prove the theoretical foundation of this observation, we adapt a widely used the Wilkinson model for SUM of lognormal operation in this analysis [20]. Below summarizes the main concept of the Wilkinson operation. In the Wilkinson method, the sum of lognormal items $\sum_{i=1}^{N} (1/N) e^{x_i}$ can be approximated as another lognormal e^y, where y is a new Gaussian variable with calculable mean and standard deviation. This approximation is completed by matching the first and second moment of both equations. Ignoring the detailed derivation, we list the formula as follows:

$$u_1 = E(s) = \sum_{i=1}^{N} \frac{1}{N} e^{\mu_{xi} + \sigma_{xi}^2/2} = e^{\mu_y + \sigma_y^2/2} \tag{34}$$

$$u_2 = E(s^2) = \frac{1}{N}\left(\sum_{i=1}^{N} e^{2\mu_{xi}+2\sigma_{xi}^2} + 2\sum_{i=1}^{N-1} \times \sum_{j=i+1}^{N} e^{\mu_{xi}+\mu_{xj}} e^{\frac{\sigma_{xi}^2+\sigma_{xj}^2+2r_{ij}\sigma_{xi}\sigma_{xj}}{2}}\right) \quad (35)$$

$$\mu_y = 2\ln u_1 - 1/2 \ln u_2 \quad (36)$$

$$\sigma_y^2 = \ln u_2 - 2\ln u_1 \quad (37)$$

where $(\mu_{x_i}, \sigma_{x_i})$ are the mean and standard deviation of the original Gaussian variables x_i and (μ_y, σ_y) are the mean and standard deviation of new Gaussian variable y. r_{ij} is the CC of each random variable, and N represents the number of stages in the data or clock path. The detailed modeling of CC r_{ij} is presented in Section II-C. The process can be modeled as

$$Ne^y = \sum_{i=1}^{N} e^{x_i}. \quad (38)$$

Each stage's delay is modeled as one lognormal item (e^{x_i}), and the sum of N stages is also a lognormal item (Ne^y). The lumped value Ne^{μ_y} represents the corner delay value reported from STA. The corner we mentioned here refers to the delay reported by STA from Spice simulation when local random mismatch parameter is set at 0. (Meanwhile, global variable is fixed at a certain value which corresponds to global slow, fast corners, etc.) Matching the corner location (e^{μ_x}) of the right-hand side of (39) with the left-hand side gives the difference between μ_y and μ_x

$$Ne^{\mu_y+\beta\sigma_y} = Ne^{\mu_x} \quad (39)$$

$$\beta = (\mu_x - \mu_y)/\sigma_y. \quad (40)$$

Due to the shift of the μ_y from μ_x in the SUM operation, the delay sum of a series of gates reported from STA (Ne^{μ_x}) is no longer located at the median location of the stochastic delay (Ne^{μ_y}) reported from Monte Carlo simulation. Instead, a negative shift at $\beta\sigma_y$ is observed due to the SUM operation of lognormal variables. Fig. 8(a) shows the histogram of Monte Carlo Spice simulation of a series of 10-stage buffers. The random variables at each buffer stage that contributes to the corner results are also annotated using a similar approach as MPP. The overall corner location has been shifted to -1.05σ despite the fact that the delay at each stage stays at near 0σ. This observation matches exactly with our mathematical explanation in (34)–(40).

Because we target to utilize the STA results for our MPP values of $D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}}$ in (33), we could recalculate the MPP value for ω_1 based on the fact that ω_2 obtained from STA is centered around $-\beta\sigma$. Fig. 9 shows that the MPP becomes (3, −1) at 0.45 V by using a STA result. Therefore, we only need to obtain a 3σ delay for the capture clock to complete the hold slack analysis. As clock tree has been well balanced and contains less variety of configurations than does the data path, the analysis has been significantly simplified. We obtained β value under different voltages for a particular configuration of the circuit to represent other circuit configurations. It is true that different circuit configurations may result in slight different values of β, e.g., 1 versus 1.1.

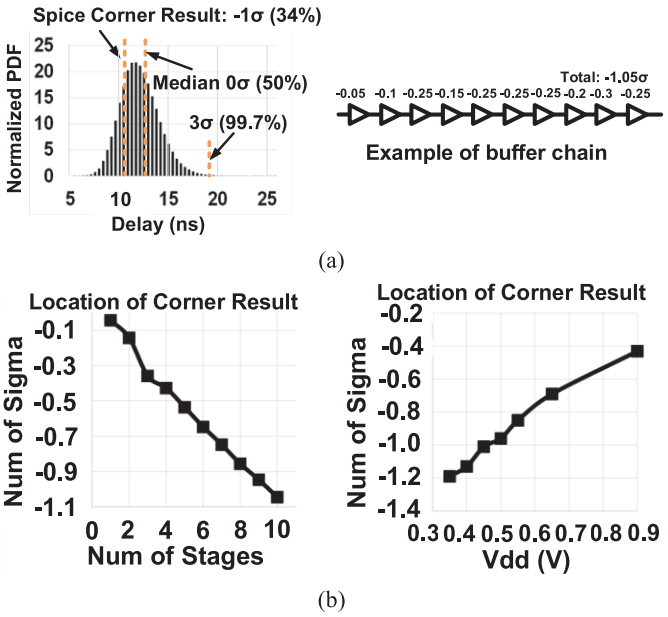


Fig. 8. (a) Monte Carlo simulation PDF compared with the corner delay value. (b) Corner location versus stages (left) and supply voltages (right).

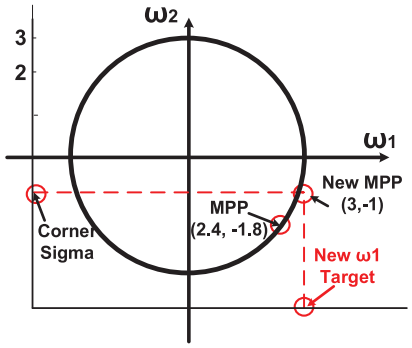


Fig. 9. MPP of SUM operation using STA results.

However, this slight difference of estimation for β would not affects the accuracy of the stochastic hold slack since 1-sigma delay of data path are similar to 1.1-sigma delay of the short data paths. Fig. 8(b) shows the simulated variation of corner location versus supply voltages and number of stages. At each supply voltage, the target sigma value for ω_1 is adjusted to account for the impact of supply voltages and depths of clock capture paths. A LUT-based approach is used to calculate the stochastic delay of the capture clock at various sigma targets as will be discussed in Section II-F. The following equation summarizes the hold slack calculation in this paper where $D_{\text{launch\_data\_STA}}$ is the corner-based STA result of launch data path

$$S_{3\sigma} = e^{\sigma_1 \times \omega 1 + \mu_1} - D_{\text{launch\_data\_STA}}. \quad (41)$$

C. Correlation Coefficient Modeling

In this section, we develop a CC model used to quantify the correlation of buffer delays between stages which was introduced in (35). Before analyzing the buffer, we first put efforts on developing the model of the inverter which is more intuitive and easy to extend.

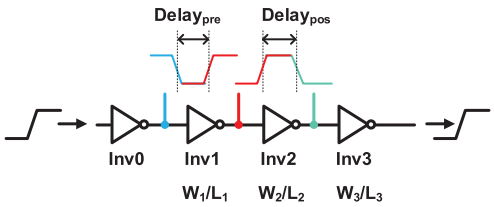


Fig. 10. Circuit schematic for inverter correlation model.

TABLE I
α AND β VALUES AT DIFFERENT SUPPLY VOLTAGES

V_{DD} (V)	α	β
0.4	0.88	24.1
0.5	0.80	6.6
0.6	0.72	3.5
0.7	0.60	2.5
0.8	0.53	2.2
0.9	0.49	2.0

1) Delay Model of Inverter: Fig. 10 shows the circuit schematic for inverter case. Both the first and last inverters (Inv0 and Inv3) are used to provide real circuit condition, e.g., slew rate and load. We first analyze the relation between the delay introduced by the second inverter (Inv1) and the third inverter (Inv2) by sweeping: 1) the slew rate of incoming input signal; 2) the size (W/L) of each inverter; and 3) the supply voltage from 0.4 to 0.9 V. Interestingly, from the simulation results, the delay of Inv2 ($\text{Delay}_{\text{pos}}$) is a linear combination of the delay of Inv1 (D_{pre}) and the size ratio between Inv2 and Inv3 $[(W_3/L_3)/(W_2/L_2)]$, which can be described in the following equation. Note that in practical, the lengths of the transistors are the same, i.e., $L_3 = L_2 = L_1$

$$D_{\text{pos}} = \alpha D_{\text{pre}} + \beta F \tag{42}$$

where D_{pos} and D_{pre} are the delays of Inv1 and Inv2 and F represents the ratio between Inv2 and Inv3. α is the coefficients which describes the delay contributes from previous stage's inverter. β is the coefficient which describes the delay contributes current stage's intrinsic effort. By using linear aggression algorithm, we obtained α and β values at different supply voltages which are shown in Table I. One observation from the data is that the $V_{\text{DD}} \times \alpha$ is approximately a constant value (0.4 V) which is about the V_{th} of the transistors.

The physical explanation behind (42) is provided as follows. The delay of current stage's inverter can be divided into two stages: 1) the rising the input voltage of current inverter to about V_{th} and 2) once the input voltage reaches V_{th}, the transistor (either nMOS or pMOS) of current inverter would be tuned ON and starts the charging/discharging process. The first stage can be described by the term αD_{pre}. Since D_{pre} also represents the slew rate of the output coming from previous inverter, the physical meaning of α is the percentage of V_{DD} that turns ON the current inverter. This fits the observation that $V_{\text{DD}} \times \alpha$ is approximately a constant value of V_{th}. This is also verified from the real simulation results shown in Fig. 11. The green waveform is the output from Inv1, and the blue one is the

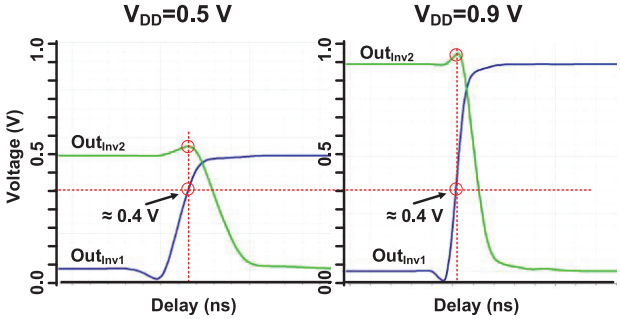


Fig. 11. Inverter output at different V_{DD}. $V_{\text{DD}} = 0.5$ V (left). $V_{\text{DD}} = 0.9$ V (right).

output from Inv2. As we can see, at both V_{DD} (0.5 and 0.9 V), the current inverter (Inv2) begins to discharge while the input signal (output from previous inverter) reaches about 0.4 V. The second stage can be represented by the term βF. In this stage, the speed of charging/discharging is related to the transistor size ratio between next stage and current stage. The smaller ratio the shorter delay will be generated.

2) Correlation Coefficient Model of Inverter: Based on (42), we can develop our CC model by introducing the random variables written as

$$D_{\text{pos}}(x_1, x_2) = \alpha x_1 + \beta F x_2/\mu_2 \tag{43}$$

where x_1 and x_2 are the two independent variables; here, x_1 represents the stochastic delay of previous inverter whose mean and standard deviation are μ_1 and σ_1. x_2 is another random variable that represents the intrinsic effort of the current inverter whose mean and standard deviation are μ_2 and σ_2. The CC between the delays of two connected investors can be derived as

$$\rho_{D_{\text{pos}}, D_{\text{pre}}} = \rho_{D_{\text{pos}}(x_1,x_2),x_1} = \frac{Cov(D_{\text{pos}}(x_1, x_2), x_1)}{\sigma_{D_{\text{pos}}(x_1,x_2)}\sigma_{x_1}} \tag{44}$$

$$= \frac{E((\alpha x_1 + \beta F x_2/\mu_2) \times x_1) - E(\alpha x_1 + \beta F x_2/\mu_2) \times E(x_1)}{\sigma_{D_{\text{pos}}(x_1,x_2)}\sigma_{x_1}}. \tag{45}$$

Since x_1 and x_2 are independent, $E(\beta F x_2/\mu_2 \times x_1) = E(\beta F x_2/\mu_2) \times E(x_1)$, (45) can be written as

$$\rho_{D_{\text{pos}}, D_{\text{pre}}} = \frac{E\left(\alpha x_1^2\right) - E(\alpha x_1) \times E(x_1)}{\sigma_{D_{\text{pos}}(x_1,x_2)}\sigma_{x_1}} \tag{46}$$

$$= \frac{\alpha\sigma_1^2}{\sqrt{(\alpha\sigma_1)^2 + (\beta F\sigma_2/\mu_2)^2} \times \sigma_1} \tag{47}$$

$$= \frac{\alpha\sigma_1}{\sqrt{(\alpha\sigma_1)^2 + (\beta F\sigma_2/\mu_2)^2}}. \tag{48}$$

Equation (48) provides the mathematical solution to calculate the CC by given the values of $\alpha, \beta, \mu_1, \sigma_1, \mu_2$, and σ_2. In order to verify the accuracy of (48), the following steps are taken.

1) We obtained the CC for different supply voltages based on Monte Carlo simulation.

TABLE II
CC COMPARISON FOR INVERTER CASE

Voltage (V)	Experimental CC	Calculated CC	Error (%)
0.4	0.41	0.38	7.3
0.5	0.36	0.35	2.8
0.6	0.35	0.32	8.6
0.7	0.33	0.30	9.1
0.8	0.30	0.28	6.7
0.9	0.28	0.26	7.1

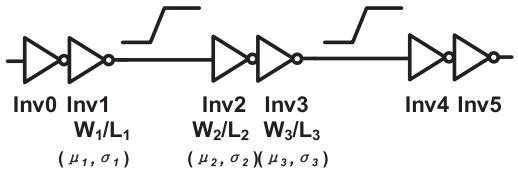


Fig. 12. Circuit schematic for buffer correlation model.

2) We extracted μ_1, σ_1, μ_2, and σ_2 at different supply voltages from the corresponding real circuits and use (48) and Table I to calculate the expected CC.
3) Compare the CC results between steps 2) and 3).

Table II summarizes the comparison results. From Table II, an error rate smaller than 10% is observed by using the proposed model to calculate the CC for inverter case. In this case, we can estimate the CC by using (48) which provides us both the accuracy and efficiency.

3) Correlation Coefficient Model of Buffer: Now we can develop the correlation coefficient model based on Sections II-C1 and II-C2. The circuit illustration is shown in Fig. 12. Since a buffer contains two inverters, the delay model of buffer is written based on (42)

$$D_{\text{Inv2}} = \alpha D_{\text{Inv1}} + \beta F \tag{49}$$
$$D_{\text{buf}} = \alpha D_{\text{Inv2}} + \beta F \tag{50}$$
$$= \alpha^2 D_{\text{Inv1}} + \alpha\beta F + \beta F. \tag{51}$$

Based on (51), we can develop our CC model by introducing the random variables written as

$$D_{\text{buf}}(x_1, x_2, x_3) = \alpha^2 x_1 + \alpha\beta F x_2/\mu_2 + \beta F x_3/\mu_3. \tag{52}$$

Similar to the inverter case, x_1, x_2, and x_3 are three independent variables, where x_1 represents the stochastic delay of input of the first inverter whose mean and standard deviation are μ_1 and σ_1. x_2 is a random variable represents the intrinsic effort of the first inverter whose mean and standard deviation are μ_2 and σ_2. x_3 is a random variable represents the intrinsic effort of the second inverter whose mean and standard deviation are μ_3 and σ_3. The CC between the input and output of the buffer derived as

$$\rho_{D_{\text{buf}}, D_{\text{Inv1}}} = \rho_{D_{\text{buf}}(x_1,x_2,x_3),x_1} = \frac{Cov(D_{\text{buf}}(x_1, x_2, x_3), x_1)}{\sigma_{D_{\text{buf}}(x_1,x_2,x_3)}\sigma_{x_1}} \tag{53}$$
$$= \frac{\alpha^2\sigma_1}{\sqrt{(\alpha^2\sigma_1)^2 + (\alpha\beta F\sigma_2/\mu_2)^2 + (\beta F\sigma_3/\mu_3)^2}}. \tag{54}$$

TABLE III
CC COMPARISON FOR BUFFER CASE

Voltage (V)	Experimental CC	Calculated CC	Error (%)
0.4	0.28	0.26	7.2
0.5	0.27	0.25	3.7
0.6	0.24	0.24	4.1
0.7	0.22	0.21	4.5
0.8	0.19	0.18	5.2
0.9	0.16	0.17	6.3

Similar approach introduced in Section II-C2 is used to evaluate the accuracy of the proposed CC model. Table III summarizes the comparison results.

4) Summary of the Proposed Correlation Coefficient Model: From Table III, an error rate smaller than 8% is observed by using the proposed model to calculate the CC for connected buffers. We can estimate the CC accurately and efficiently by the following steps: 1) find out the corresponding α and β values at the particular V_{DD}; 2) obtain the $\mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3$, and σ_3 from real circuit; and 3) calculate the CC based on (54).

D. Unified Stochastic SUM Operation Across Voltage

To compute the stochastic SUM operation of clock capture paths, we proposed a unified model based on (34)–(37) with μ and σ characterized from Spice simulation. Different from conventional timing analysis which assumes Gaussian operation for high-voltage calculation and lognormal operation for low voltages leading to two separate characterization and analyses across supply voltages, in this paper, we propose to only use lognormal distribution to model delay at entire voltage range. The reason lies in the fact that the lognormal distribution converges into Gaussian distribution when the ratio σ/μ becomes very small at a high voltage. This can be explained by utilizing the Taylor series expansion. Let D present the stochastic delay of a circuit path and $x = \ln(D)$ follows Gaussian whose mean and standard deviation are μ and σ

$$D(\omega) = e^{\sigma\omega+\mu}. \tag{55}$$

The Taylor series of $D(\omega)$ at 0 can be written as

$$D(\omega) = D(0) + \frac{D'(0)}{1!}(\omega - 0) + \frac{D'(0)}{2!}(\omega - 0)^2 + \cdots \tag{56}$$
$$= e^{\mu} + \sigma e^{\mu}\omega + \frac{\sigma^2}{2!}e^{\mu}\omega^2 + \cdots + \frac{\sigma^n}{n!}e^{\mu}\omega^n. \tag{57}$$

Under real circuit configuration, where $\sigma/\mu < 0.03$, which makes the nth-order coefficient, $(\sigma^n/n!)e^{\mu}$ negligible compares to the first-order coefficient σe^{μ}. Then, (55) can be approximated as

$$D(\omega) \approx e^{\mu} + \sigma e^{\mu}\omega. \tag{58}$$

Equation (58) explains the fact that lognormal distribution converges into Gaussian distribution when the ratio σ/μ becomes very small.

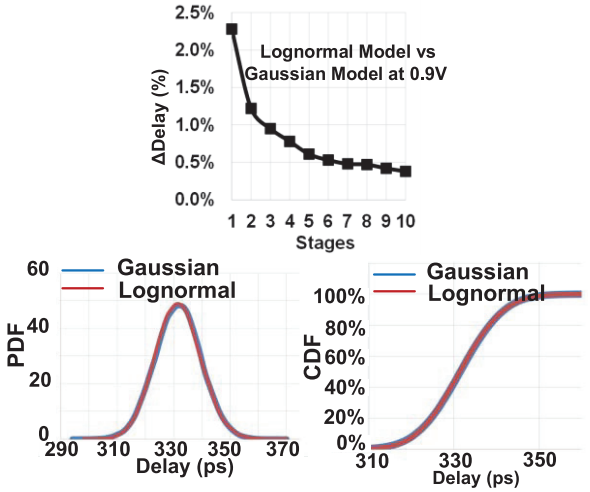


Fig. 13. Delay differences between lognormal model and Gaussian model at 0.9 V.

Fig. 13 shows the simulated delay differences between Gaussian model and lognormal model across various numbers of stages and the PDF and cumulative distribution function (CDF) distribution of the two models at 0.9 V. It can be observed that the CDF and PDF of Gaussian and lognormal model are almost overlapped. As a result, a unified model using lognormal model can be used for SUM operation across the voltage ranges. In this paper, we use the Wilkinson equation as presented in (34)–(37) to model the SUM with standard cells' delay μ and σ characterized into a LUT as will be shown in Section II-F. This unified model significantly simplifies the modeling and timing analysis of standard cells.

E. Hyper-Lognormal Region for Transistor-Level Cell Modeling

In this section, we discuss the practical issue and its solution of the stochastic SUM operation analysis across voltage. Most previous work has simplified the circuit delay as pure Gaussian or a lognormal delay based on the current relationship with threshold voltage variation [13]–[15]. However, using a simplified lognormal model to characterize a standard cell delay at near-threshold region can cause significant optimism issue. The optimism stems from the fact that at near-threshold region, the transistor traverses across subthreshold region and linear/saturation region when the threshold voltage varies. As a result, characterizing the cell delay based on mean and standard deviation of the Monte Carlo delay of a standard cell is likely to be optimistic because many data points are obtained when transistors operate at weak inversion rather than cutoff region. Fig. 14 shows an nMOS transistor current versus threshold voltage drawn in log scale. Instead of an ideal linear curve, the current flattens as the device moves into linear/saturation region.

Delay impact when a standard cell buffer is characterized at 0.45 V is presented in Fig. 15. An error of 26% is observed between the ideal lognormal model and real circuit simulation. We refer this effect as "hyper-lognormal" effect because it introduces additional nonlinear behavior beyond a conventional lognormal model. To model such effect, we propose

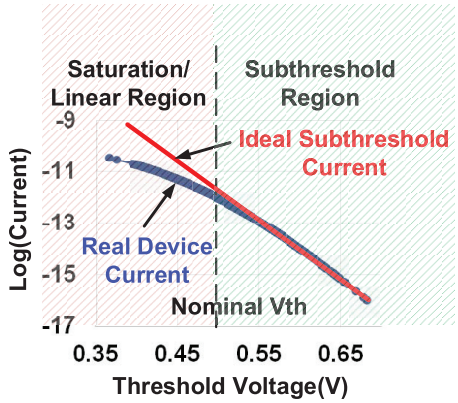


Fig. 14. Current versus threshold voltage in an nMOS transistor current at V_{DD} of 0.45 V. (V_{DS} is set at $V_{DD}/2$.)

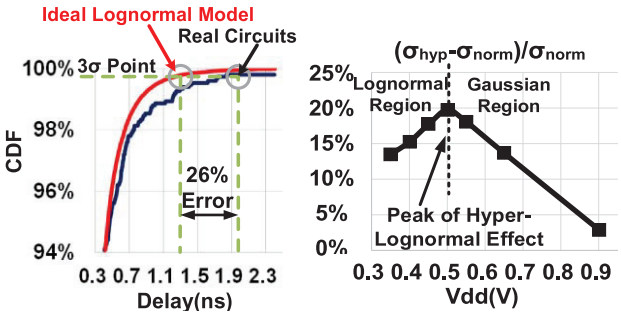


Fig. 15. CDF of buffer delay at 0.45 V versus ideal lognormal delay model (left). The deviation of σ_{hyp} from σ_{norm} across V_{DD} (right).

to use an additional σ_{hyp} to characterize the standard cell besides a normal σ_{norm}. While σ_{norm} quantify the overall delay distribution of the standard cell, σ_{hyp} captures the super-nonlinear tail of the delay distribution. Fig. 15 also shows the difference between σ_{hyp} and σ_{norm}, which characterized from a standard cell buffer across different supply voltages. As expected, at both linear/saturation region and deep subthreshold region, σ_{hyp} and σ_{norm} converges to be the same while at near-threshold region (~0.5 V), the hyper-lognormal effects reach the peak due to the crossing of operation region of transistors. In our model, we use an α value to present the impact of σ_{hyp} as shown in (59). An α value of 0.4 is used representing a balance of σ_{hyp} and σ_{norm}. Based on experiment and our analysis, the values of σ_{hyp} and α only matter for voltages at near-threshold region around 0.5 V and do not introduce significant difference at deep subthreshold and high voltages

$$\sigma_{new} = \alpha\sigma_1 + (1-\alpha)\sigma_2. \tag{59}$$

F. Summary of Overall Hold-Timing Modeling Analysis

Fig. 16 summarizes our stochastic hold-timing analysis flow. A 6×6 LUT of μ and σ with various loads and slew conditions is generated from Spice-level Monte Carlo simulation on standard cells related to clock paths at various supply voltages from 0.35 to 0.9 V.

Conventional STA is performed to find out the slew and load condition of the clock path as well as the corner delay for data

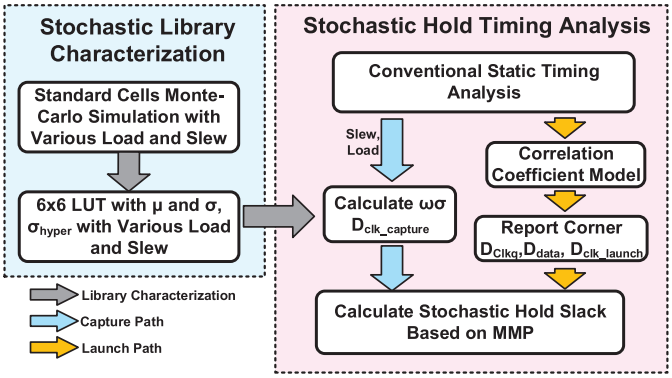


Fig. 16. Flowchart of the proposed stochastic hold analysis method.

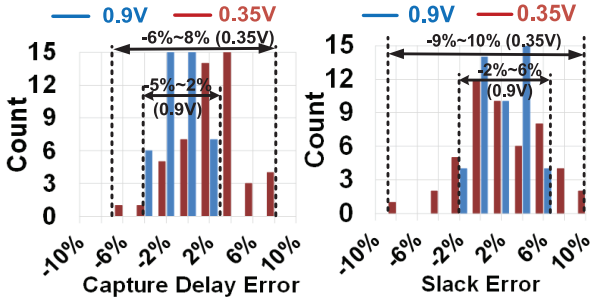


Fig. 17. Histogram of top 50 paths of the errors of the capture clock delay and overall hold slack at 0.35 and 0.9 V.

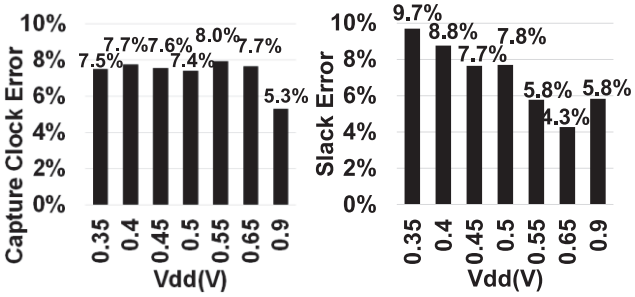


Fig. 18. Errors of the proposed methodology across large voltages for capture clock delay (left) and overall hold slack (right).

path and launch clock path. ω values for stochastic capture clock path delay are precharacterized from MPP analysis described in Sections II-A and II-B depending on supply voltages and circuit configurations. The stochastic SUM for capture clock is performed as described in Section II-D. Calculation following (41) is used to obtain the final hold slack of a particular path. Because the σ value has been characterized in an LUT, any target stochastic location $\omega\sigma$ of the delay can be easily calculated following the proposed methodology. Due to the simplicity of our scheme and compatibility with the existing timing analysis flow, the entire stochastic hold analysis can be performed with similar time as conventional STA, rendering orders of magnitude faster speed than the path-based iterative search approach reported in [16] and [17].

III. Evaluation on a DSP Design

In this section, we verify the proposed timing closure methodology. A 64-point 8-bit highly pipelined FFT processor was implemented using commercial synthesis and backend (P&R) tools in a 45-nm technology. Static timing libraries are generated across supply voltages from 0.35 to 0.9 V for STA. The backend design with routing parasitics from layout was sent to commercial STA engine for both STA and Spice netlist extraction. Although the clock tree has been well balanced in the design, due to exponential increase of delay variation at low voltages, significant hold-timing issues are observed from 0.55 V and below. We selected the worst 50 paths with minimum hold slack for evaluating the circuit-level Monte Carlo simulation. Due to the short data path and regularity of clock trees, the selected paths cover representative variety of clock paths and data paths. Transistor-level Spice netlist including both clock path and data path with extracted parasitics were simulated using Spice Monte Carlo simulation. Scripts with the generated LUT were used to perform the proposed timing analysis for comparison with the Spice Monte Carlo simulation results.

Fig. 17 shows histograms of errors on the stochastic capture clock delay and overall hold slack at 0.35 and 0.9 V. For the stochastic capture clock delay, the majority paths match within 5% with a maximum error of 8%. For overall hold stack, the maximum error is less than 10% while majority paths still match within 5%. The hold slack error is defined as the difference between the calculated stochastic hold slack and the Spice Monte Carlo-based simulated hold slack over the delay of capture clock path. Fig. 18 shows the overall accuracy of the capture clock delay and hold slack across the voltages from 0.35 to 0.9 V with worst case at 0.35 V. This result highlights the accuracy of the proposed unified model where the high voltage is also properly modeled with the lognormal equation. Fig. 18 also shows the accuracy improves with higher voltages due to much tighter stochastic distribution and the improved accuracy of the STA which also introduces errors compared with the Spice simulation.

In addition, Fig. 1(b) shows the numbers of buffers required for hold fixing from one of the worst case (min delay) paths in our design under various supply voltages. A worst case of 23% performance degradation was observed. Although it is possible to perform more sophisticated backend design improvement to avoid impacting the setup path, it still requires significant design modification and iterations of design verification. In Section IV, we present a novel hold resilient design scheme to remove the high-voltage impact leveraging the hold-timing analysis approach in this paper as shown.

IV. Hold-Free Scheme With Clock Duty-Cycle Modulation

A. Hold-Free Scheme

To accomplish the target of avoiding excessive hold buffers insertion for high-voltage operation, e.g., the FFT processor case shown in Section III, we propose to replace conventional flip-flop with a dual-mode TRFF as presented in Fig. 19(a). An additional hold-fixing latch is added in addition to the conventional flip-flop. At high voltage, the additional latch is bypassed and the whole design flow as well as timing closure is identical as the conventional design. The timing resilient mode of the proposed flip-flip is activated at low

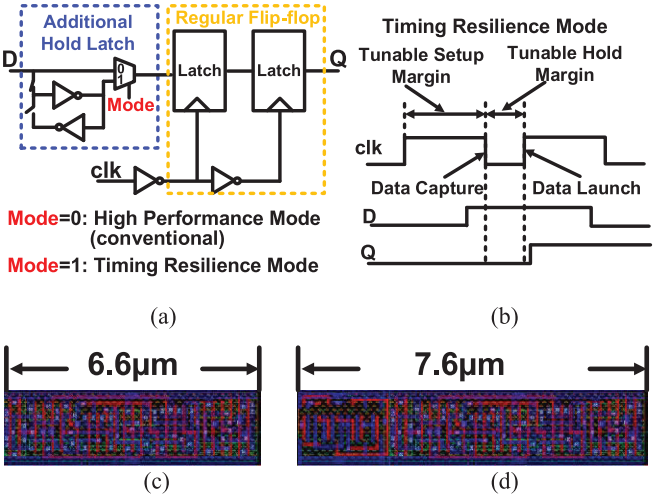


Fig. 19. Proposed TRFF design (a) schematic. (b) Waveform of operation. (c) Layout of conventional flip-flop. (d) Layout of the proposed flip-flop.

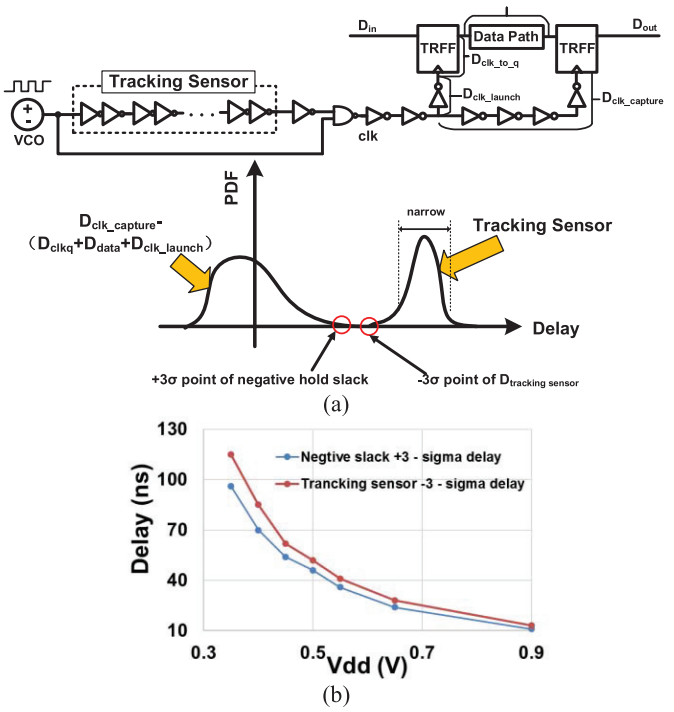


Fig. 20. (a) Illustration of tracking sensor-based solution. (b) Demonstration of tracking sensor performance.

voltage as shown in Fig. 19(b). The layouts of conventional and proposed flip-flop are shown in Fig. 19(c) and (d). The additional hold latch in timing resilient mode only passes the data when clock is high and gates the input from the main flip-flop when clock is low. As a result, the flip-flop can be considered as only latching the data at falling edge of the clock leaving the entire time of clock-low period as hold-timing margin. By modulating the clock duty cycle (defined as clock low/clock period), a programmable setup/hold-timing margin can be achieved. The downside of this scheme is that the setup time is sacrificed by requiring the data to arrive before the falling edge of the clock although the duty cycle can be kept as minimum to reduce the performance impact. As performance is less of an issue for the low-voltage operation mode, the proposed scheme provides an optimum tradeoff for the conflicting requirements between high voltage and low voltage. For clock duty-cycle control, a digital phase-locked loop or delay-locked loop with digital controlled oscillator can be used to generate multiple phases for variable duty cycle. Besides, a tracking sensor is proposed in Section IV-B to improve the performance under on-chip variation impact. In our design, the selection of clock duty cycle and the selected insertion of the TRFF are determined from the proposed timing analysis in Section II. As a result, we can accurately program the hold timing required across supply voltages without inserting excessive hold-fixing buffers. The TRFF has been simulated across voltages with Monte Carlo simulation for verifying functionality and timing at low voltages. As shown in Fig. 19(c) and (d), the area overhead of the proposed TRFF is around 15% of the conventional flip-flop. The energy consumption overhead is around 20%, and there is no delay overhead compared with conventional flip-flop. Although the proposed TRFF introduces extra area and energy consumption, it eliminates tons of hold-fixing buffers which not only compensate the overhead but also improve performance when circuit is operating at nominal voltage. Besides, the proposed flip-flop can be integrated seamlessly into the traditional timing synthesis and P&R flow. More detailed implementation will be discussed in Section IV-C.

B. Tracking Sensor

In this section, a tracking sensor is proposed to provide enough hold margin under the on-chip variation impact. In this paper, there are two types of variation we need to deal with: the global PVT variation and the local random variation. Since the tracking sensor is on the same die as digital circuits, both the sensor and the digital circuits will experience the same PVT corner. In this way, the PVT variation can be well tracked. The local variation which will have different effects on the sensor and digital circuits can be compensated by the low-variation tracking sensor. Under the local variation effect, we build the sensor which guarantees the lower bound (-3σ point) of stochastic delay of the tracking sensor is larger than upper bound ($+3\sigma$ point) of stochastic delay of the negative hold slack $D_{\text{clk\_capture}} - (D_{\text{clk\_launch}} + D_{\text{clkq}} + D_{\text{data}})$ which is illustrated in Fig. 20(a). Also, we need to design the tracking sensor with less variation compare with the clock chain, in another word, with narrower PDF. This can be achieved by enlarging the size of transistors of the tracking sensor. The simulated result is shown in Fig. 20(b). As we can see, the tracking sensor provides enough hold margin from low to high supply voltages under on-chip variation impact.

C. Case Study on DSP

We evaluated the proposed scheme in the digital signal processor (DSP). Fig. 21(b) shows the required minimum duty cycle for guaranteeing the hold timing without inserting hold-fixing buffers. As shown in Fig. 21(b), no hold violation is observed above 0.55 V and thus the design can be set

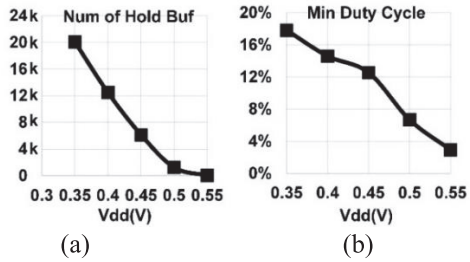


Fig. 21. (a) Numbers of hold-fixing buffers in conventional design. (b) Minimum duty cycle in the proposed scheme under different voltages.

TABLE IV
DESIGN SPEC AND STATISTICS OF THE DSP

Spec	Values	Spec	Value
Technology	45nm	Total Area (w/o hold fixing)	0.245 mm^2
Supply Voltages	0.35V~0.9V	Num of Flip-flops	16,004
Clock Freq	600MHz (0.9V) 2MHz (0.35V)	Total Num of Cells	72,300
Max Negative Slack	-96n (0.35V)	No. of Hold Buffers for 0.35V	20,058
Area of Conv. Design with hold buffers	0.271mm^2	Area of Proposed Design Scheme	0.253 mm^2

back into the conventional mode. The minimum duty cycle increases at lower voltage as the negative hold slack becomes larger and reaches 18% of the clock period at 0.35 V. Note that the minimum duty cycle is only the lower bound of the duty cycle in timing resilient mode. Other clock pulsewidth constraints required for reliable standard cell operations will likely limit the minimum duty cycle. Increasing duty cycle will increase performance degradation of the scheme at low voltage while gaining more hold margin to the design. We further evaluate all the timing paths (~44 737) in the design using our proposed timing analysis approach in the DSP example. The total runtime of our SSTA approach for this example is just less than 1.5 h. Fig. 21(a) shows the total number of hold buffers needed across supply voltages in conventional design.

Table IV summarizes the design spec and timing analysis statistics. To allow operation down to 0.35 V, total 5857 flip-flops (37% of all flip-flops) are converted into the TRFFs. Note that in conventional design scheme, a total of 20 058 extra hold buffers need to be inserted for fixing hold-timing issues. Note that the area number reported in this paper is slightly different from our previous published one [21]. This is because the previous report in our published work used "report_area" command in encounter which is not accurate since it does not include all the layout overhead such as wiring and filters. In this paper, we corrected the reports by directly measuring the area from the encounter tool.

In our scheme, the hold-fixing buffers have been avoided rendering 23% performance improvement at 0.9 V. The area overhead of the new flip-flops is compensated by the saving of the hold-fixing buffers leading to a total area saving of 7.1% which is shown in Fig. 22. More importantly, the proposed scheme enables a "hold-free" design strategy that allows the supply voltage to freely operate into subthreshold regime

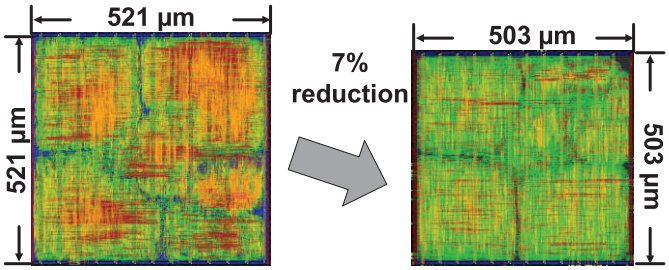


Fig. 22. Layout of the DSP (a) design with extra hold buffers. (b) Proposed design with TRFF.

without compromising the high-voltage performance. This result convinces us that the proposed methodology is not only compatible with conventional timing closure methodology but also does not sacrifice performance from enabling low-power operation.

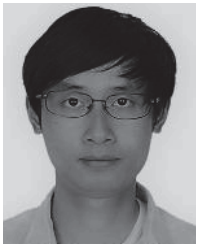
V. CONCLUSION

A comprehensive modeling and design methodology for voltage-scalable operation, based on MPP theory, has been proposed in this paper. The proposed scheme is computationally efficient for modeling statistical circuit tolerance across a wide range of supply voltages where delay is a nonlinear function of transistor random variables. The proposed MPP search method gets rid of the tedious iterative search. In addition, a theoretical modeling of CC with its physical explanation is presented to further improve the accuracy and efficiency of the design. The developed timing analysis approach features a unified voltage-scalable timing model and incorporates highly nonlinear effect of transistor behavior at near-threshold region to achieve high accuracy with computing effort similar with conventional STA. Leveraging the developed timing analysis approach, a novel "hold-free" variation-aware circuit solution is proposed. It has been implemented using commercial CAD tools and integrated into commercially used design flow in a 45-nm DSP design. The result shows that compared with conventional hold-fixing strategy, the proposed techniques not only accurately model the stochastic timing margin within 10% of Monte Carlo simulation but also eliminate the expensive hold-fixing efforts rendering a "hold-free" operation across large supply range and significant performance saving at high voltages. The proposed design methodology is demonstrated in a 45-nm DSP design enabling a voltage-scalable operation from 0.35 to 0.9 V for an area saving of 7% and eliminating more than 20 000 hold buffers as well as 23% performance degradation at high voltages.

REFERENCES

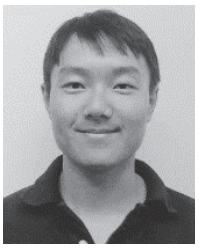
[1] A. P. Chandrakasan *et al.*, "Technologies for ultradynamic voltage scaling," *Proc. IEEE*, vol. 98, no. 2, pp. 191–214, Feb. 2010.
[2] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: From basic principles to state of the art," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 4, pp. 589–607, Apr. 2008.
[3] H. Mahmoodi, S. Mukhopadhyay, and K. Roy, "Estimation of delay variations due to random-dopant fluctuations in nanoscale CMOS circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1787–1796, Sep. 2005.

[4] S. Sundareswaran, J. A. Abraham, A. Ardelea, and R. Panda, "Characterization of standard cells for intra-cell mismatch variations," in *Proc. Int. Symp. Quality Electron. Design*, Mar. 2008, pp. 213–219.
[5] S. Das *et al.*, "RazorII: *In situ* error detection and correction for PVT and SER tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.
[6] K. A. Bowman *et al.*, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 49–63, Jan. 2009.
[7] M. Fojtik *et al.*, "Bubble Razor: An architecture-independent approach to timing-error detection and correction," in *IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, Feb. 2012, pp. 488–490.
[8] S. Kim and M. Seok, "R-Processor: 0.4 V resilient processor with a voltage-scalable and low-overhead *in-situ* error detection and correction technique in 65 nm CMOS," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2014, pp. 1–2.
[9] Y. Zhang and B. H. Calhoun, "Hold time closure for subthreshold circuits using a two-phase, latch based timing method," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (IEEE S3S)*, Oct. 2013, pp. 1–2.
[10] S. Kim and M. Seok, "Analysis and optimization of *in-situ* error detection techniques in ultra-low-voltage pipeline," in *Proc. Int. Conf. Low Power Electron. Design*, 2014, pp. 291–294.
[11] A. Singhee, S. Singhal, and R. A. Rutenbar, "Practical, fast Monte Carlo statistical static timing analysis: Why and how," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2008, pp. 190–195.
[12] V. Veetil, D. Sylvester, and D. Blaauw, "Efficient Monte Carlo based incremental statistical timing analysis," in *Proc. Design Autom. Conf.*, 2008, pp. 676–681.
[13] J. Singh and S. Sapatnekar, "Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis," in *Proc. Design Autom. Conf.*, 2006, pp. 155–160.
[14] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, and M. Sharma, "Correlation-aware statistical timing analysis with non-Gaussian delay distributions," in *Proc. Design Autom. Conf.*, 2005, pp. 77–82.
[15] C. Visweswariah *et al.*, "First-order incremental block-based statistical timing analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 10, pp. 2170–2180, Oct. 2006.
[16] R. Rithe *et al.*, "The effect of random dopant fluctuations on logic timing at low voltage," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 5, pp. 911–924, May 2012.
[17] N. Ickes *et al.*, "A 28 nm 0.6 V low power DSP for mobile applications," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 35–46, Jan. 2012.
[18] X. Du and W. Chen, "A most probable point-based method for efficient uncertainty analysis," *J. Design Manuf. Autom.*, vol. 4, no. 1, pp. 47–66, Oct. 2001.
[19] Y.-T. Wu, H. R. Millwater, and T. A. Cruse, "Advanced probabilistic structural analysis method for implicit performance functions," *J. Amer. Inst. Aeronaut. Astronaut.*, vol. 28, no. 9, pp. 1663–1669, 1990.
[20] A. A. Abu-Dayya and N. C. Beaulieu, "Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications," in *Proc. IEEE Veh. Technol. Conf.*, vol. 1, Jun. 1994, pp. 175–179.
[21] H. Wang, G. Xie, and J. Gu, "Comprehensive analysis, modeling and design for hold-timing resiliency in voltage scalable design," in *Proc. ISLPED*, 2016, pp. 22–27.
[22] J.-S. Wang and S.-N. Wei, "Process/voltage/temperature-variation-aware design and comparative study of transition-detector-based error-detecting latches for timing-error-resilient pipelined systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 10, pp. 2893–2906, Oct. 2017.
[23] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1440, Oct. 1989.
[24] B. H. Calhoun and A. P. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, Jul. 2006.
[25] Y. Zhang *et al.*, "iRazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 160–161.
[26] S. Kim and M. Seok, "Variation-tolerant, ultra-low-voltage microprocessor with a low-overhead, within-a-cycle *in-situ* timing-error detection and correction technique," *IEEE Solid-State Circuits*, vol. 50, no. 6, pp. 1478–1490, Jun. 2015.
[27] S. Das *et al.*, "A self-tuning DVS processor using delay-error detection and correction," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 792–804, Apr. 2006.



Zhengyu Chen (S'16) was born in Nanjing, China, in 1991. He received the B.S. degree from Southeast University, Nanjing, in 2013, and the M.S. degree in computer engineering from Cornell University, Ithaca, NY, USA, in 2015. He is currently working toward the Ph.D. degree in computer engineering at Northwestern University, Chicago, IL, USA.

He is currently an aspiring engineer at Northwestern University doing research in the area of ultralow-power design/algorithm for very large scale integration and mixed-signal ICs and emerging device. He is focusing on the interesting low-power algorithm design such as time-domain signal processing and reconfigurable architecture/circuit based on memristors.



Huanyu Wang (S'16) received the B.S. degree in electronic science and technology from the Huazhong University of Science and Technology, Wuhang, China, in 2014, and the M.S. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 2016. He is currently working toward the Ph.D. degree at the Florida Institute for CyberSecurity, University of Florida, Gainesville, FL, USA.

His current research interests include hardware security and trust.

Geng Xie (S'16) received the B.S. degree in electrical engineering from Jilin University, Changchun, China, in 2014, and the M.S. degree in electrical engineering from Northwestern University, Evanston, IL, USA.

He is currently a Digital Design Engineer with Omnivision Technology, Santa Clara, CA, USA, where he focuses on CMOS image sensor design and testing. His current research interests include the analysis and design of low-power and high-performance circuit design.

Jie Gu (M'10) received the B.S. degree from Tsinghua University, Beijing, China, the M.S. degree from Texas A&M University, College Station, TX, USA, and the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA.

He was an IC Design Engineer with Texas Instruments, Dallas, TX, USA, from 2008 to 2010, focusing on ultralow-voltage mobile processor design and integrated power management techniques. He was a Senior Staff Engineer with Maxlinear, Inc., Carlsbad, CA, USA, from 2011 to 2014, focusing on low-power mixed-signal broadband system-on-chip design. He is currently an Assistant Professor with Northwestern University, Evanston, IL, USA. His current research interests include ultralow-power mixed-signal very large scale integration circuit design, integrated power and clock management with hardware and software codesign, and emerging device/technology integration.

Dr. Gu has served as a Program Committee and Conference Co-Chair for numerous low-power design conference and journals, such as ISPLED, DAC, ICCAD, and ICCD.