Modeling of Total Parameter Variations

Frank Sill
College of CSEE
University of Rostock
Germany

Dirk Timmermann
College of CSEE
University of Rostock
Germany

Aggressive downscaling of CMOS devices in every technology generation resulted in higher integration density and performance. At the same time, yield, which is the ratio of flawless versus all fabricated chips, drastically decreased. Failed chips are divided in defect devices (defect yield) and devices, which failed the desired performance (parametric yield). Parameter variations, which strongly increase with reduced technology sizes, are responsible for decreasing parametric yield [1]. Parameter variations are divided into intra-die and inter-die variations. Due to the latter, the same circuits might have different characteristics on different dies. Intra-die variations are the variations of transistor characteristics within a single chip. Both kinds of variations are expected to be truly random in nature [2]. The parameter variations are based on different effects, such as variations in process parameters, temperature, or supply voltage. These variations lead to changes in transistor characteristics, which might result in longer delays.

In established static timing analysis (STA), which is used to determine circuit performance, the effect of parameter variation is modeled with corner-case models. Each gate is set to its worst-case delay value at this corner-case timing analysis. Signal arrival times at the output of a gate are estimated by adding the gate delay to the signal arrival time at the inputs. Corner-case STA is based on assumptions of inter-die variations only. But, intra-die variations cannot be ignored in technologies with gate length below 100nm [1]. Hence, traditional corner-case STA is quite pessimistic and underestimates the value for typical performance and overestimates the worst-case timing behavior [3].

In contrast, statistical static timing analysis (SSTA) considers intra-die variations. The gate delay is based on probability functions. Hence, signal arrival times are modeled as probabilistic functions. The delay variability can be described with cumulative probability distribution function (CDF) or probability density function (PDF). A CDF describes the probability that the delay is lower than a given value *x*. In contrast, a PDF describes the probability that the delay has the value *x:*

$$CDF(x) = \int_0^x \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

$$PDF(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

σ is the standard deviation, σ^2 is the variance, and μ is the expected value. We model the gate delay and data arrival time as Gaussian distributions with expected value μ and variance σ^2. At single input gates the output signal arrival time result from:

$$\mu_{out} = \mu_{gate} + \mu_{in}$$

$$\sigma_{out} = \sqrt{\sigma_{gate}^2 + \sigma_{in}^2}$$

μ_{out}, μ_{in}, and μ_{gate} are the expected values of the output and input arrival time, and the gate delay, respectively. σ_{out}, σ_{in}, and σ_{gate} are the variances of output and input signal, and gate delay, respectively.

A very common approach for evaluating output signal arrival time at multi-input gates is the creation of tables, which includes the results for different input signal arrival time combinations [2]. In [3], gates with multiple inputs are divided in single input gates. Both approaches considerably increase the complexity or require extensive library characterization.

We assume as worst-case that a gate needs all input signals to generate an output signal. Hence, the worst-case time for starting the gate evaluation cannot start before the latest input signal arrived. As shown, CDF can be used to describe the probability that a signal has arrived. Thus, the probability that all signals have arrived results from multiplication of all input signal arrival time CDFs. So, at each time the probability is considered for each arriving input signal. The result is a CDF for the time of the evaluation start. The estimation of this CDF can be divided into two main cases. In the first case, one signal arrives much later than the other input signals. Then, its CDF is nearly equal to the CDF of the evaluation start time. Consequently, PDF of last arriving input signal and evaluation start time $eval_{begin}$ are nearly equal.

In the second case, the overlap of input arrival time CDFs has to be considered. Then, the probability function of evaluation start time depends on different inputs. As the purpose is a manageable calculation of CDF and PDF for $eval_{begin}$, we simplify the complex problem by an approximation.

The rising edge of a CDF can be approximated as a straight line *s(x)* with:

$$s(x) = \frac{1}{(2 \cdot 1.5 \cdot \sigma)} \cdot (x - \mu + 1.5 \cdot \sigma)$$

This simplification of CDFs can be used to find μ_{new} and σ_{new} for an approximated description of $eval_{begin}$. A CDF(x) has the value 0.5, if $x = \mu$. Hence, the approach aims at the determination of the time instance where the product of the straight line approximations of all input arrival times is 0.5. That means the solution of:

$$0.5 = \prod_{i=0}^{input\_signals} \frac{1}{(3 \cdot \sigma_i)} \cdot (\mu_{new} - \mu_i + 1.5 \cdot \sigma_i)$$

As these are equations of higher order, more than one solution is possible. If μ_{new} is known, we calculate σ_{new} from the difference between μ_{new} and the point in time t_{max}, where the last signal arrives with a probability of 0.99. That means:

$$t_{\max} = \max(\mu_1 + 3\sigma_1, \mu_2 + 3\sigma_2, ..., \mu_n + 3\sigma_n)$$
$$\sigma_{new} = (t_{\max} - \mu_{new}) / 3$$

Hence, the approximated CDF and PDF of the evaluation start time result from the new expected value μ_{new} and the new standard deviation σ_{new}.

Figure 1 depicts CDF of input signals *in1* and *in2* with overlapping arrival time probability, the resulting probability for the evaluation start time $eval_{begin}$, the new generated probability for the evaluation start time *new-*$eval_{begin}$, and the approximated straight line for CDF of the latter.

The new approach is slightly pessimistic as due to applying straight-line multiplications, μ_{new} is greater than the highest expected value $\mu_{in,max}$ of the inputs. Furthermore, σ_{new} is based on standard deviation of the latest arriving signal which has not always the greatest $\mu_{in,max}$.

To model parameter variations, common approaches vary technology or transistor parameters, which strongly impact gate delay, like gate length L_{eff} or gate width W_{eff} [2][3]. Then, the gate delay is described as a function of varied parameters. This allows accurate mathematical formulation of the problem. But, evaluation effort increases drastically with each additional parameter. Thus, only one or two parameters are varied.

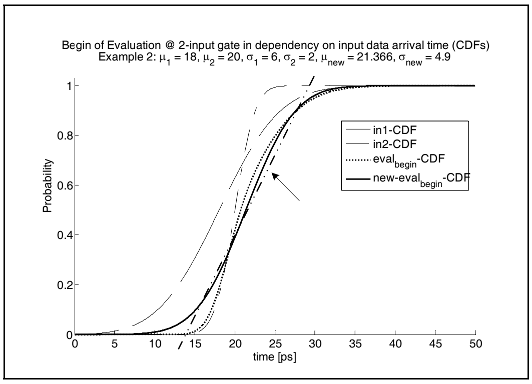

Figure 1: CDFs of overlapping input signals and resulting worst-case time for start of evaluation

The new idea is the modeling of the gate delay variation under consideration of more than one or two parameters. The demand is an easy to handle model. Thus, we chose a Gaussian distribution description, whereas its values are extracted from *Monte-Carlo* spice simulations. We verified the accuracy of this approach for an inverter, based on the BPTM predictive 65nm technology library. First, *Monte-Carlo* spice simulations were applied, where all parameters, which can vary, were described with Gaussian distribution. We assumed 10% variation of L_{eff}, W_{eff}, thickness of gate oxide layer T_{ox}, *temperature*, and supply voltage V_{dd} for each transistor. As the distributions of resulted delays are similar to Gaussian distribution, we extracted the expected value μ and standard deviation σ of gate delay. Finally, we described gate delay as function of gate length L_{eff}, where variance has the same distribution as in previous simulations. The result indicates, that the new approach for modeling of gate delay is closer to realistic distribution of gate delay than common approaches (see figure 2).

These results are based on the behavior of Gaussian distributions, whereas convolution of Gaussian distributions results in new Gaussian distributions. The evaluation of gate delay is based on multiplication of varied parameters. Hence, gate delay can be described as Gaussian distribution.

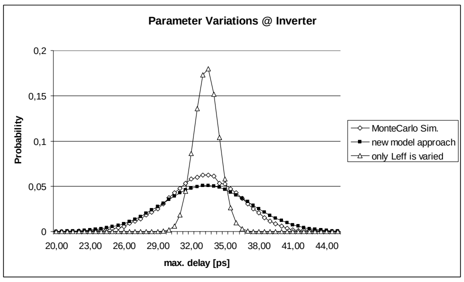

Figure 2: Distribution of Inverter Delay considering parameter variations

[1] S. Borkar, et. al., "Design and reliability challenges in nanometer technologies", Proc. DAC, USA, 2004.

[2] S.H. Choi, et. al., "Novel Sizing for Yield Improvement under Process Variation in Nanometer Technology", Proc. DAC, USA, 2004.

[3] A. Agarwal, et. al., "Statistical Gate Delay Model Considering Multiple Input Switching", Proc. DAC, USA, 2004.