Total Leakage Reduction by Observance of Parameter Variations

Frank Sill, Dirk Timmermann

Department of Computer Science and Electrical Engineering, University of Rostock
{frank.sill, dirk.timmermann}@uni-rostock.de

Abstract:

Leakage power dissipation and parameter variations are main topics in current research. The problem of parameter variations leads to modified timing analysis. Traditionally, this is done with corner-case simulations, which are quite conservative and pessimistic approaches. This paper proposes a new statistical static timing analysis (SSTA) to improve performance predictions. Furthermore, the developed SSTA and a Dual-V_{th}*/Dual-*T_{ox} *CMOS (DTTCMOS) design technique are combined to reduce the total leakage within designs. Compared to traditional corner-case timing analyses the proposed approach reduces leakage by an average of 70% for raw designs and by an average of 25% in pre-optimized DTTCMOS designs.*

1. Introduction

Aggressive downscaling of CMOS devices in each technology generation resulted in higher integration density and performance. At the same time, parameter variations strongly increase [1]. This resulted in decreasing yield, which is the ratio of flawless versus all fabricated chips. Parameter variations are divided into intra-die and inter-die variations. Due to the latter, same circuit might have different characteristics on different dies. Intra-die variations mean the variations of transistor characteristics within a single chip. The parameter variations are based on different effects, such as fluctuations in process parameters, temperature, or supply voltage. These fluctuations lead to changes in transistor characteristics, which might result in longer delay and higher power dissipation.

Established static timing analysis is very conservative [2]. At design time, this leads to a high effort on circuits to achieve the desired performance. Statistical static timing analysis (SSTA) presents a solution, which does not guarantee 100% security that the desired timing is reached in all cases. But, SSTA drastically reduces timing effort at design time and the assured security is still high. Orshansky presents a general framework for SSTA [2]. Agarwal et. al. proposes approaches to handle SSTA on multiple input gates [3].

Leakage represents another serious problem. It increased exponential and has become a major contributor to total power dissipation on chips. Today, up to 50% of IC power dissipation is due to leakage currents [1]. The main reason for strong increase of leakage is the lowering of the transistor threshold voltage V_{th}. This is necessary to meet desired performance because supply voltage V_{DD} is scaled down. As V_{th} is not only proportional to delay but also to channel leakage, sub-threshold current increases exponentially. In addition, in current technologies the thickness of gate oxide corresponds to just some atom layers. Thus, tunneling currents through the gate increase, known as gate leakage current. Since the last decade, the reduction of leakage currents is a main part in low-power research, whereas a lot of technologies where developed. A very common solution for lowering sub-threshold leakage current is the application of two or more device types, which distinguish in their threshold voltage [4]. Thus, the devices differ in performance and leakage. This Dual-V_{th} CMOS (DTCMOS) design technique applies fast devices in critical paths while slower devices with lower leakage are used in noncritical paths. A related design technique is the application of two transistors types, which differ in thickness of gate oxide layer T_{ox} [5]. This Dual-T_{ox} CMOS (DTOCMOS) design technique allows the reduction of gate oxide leakage currents.

The combination of SSTA and optimization for leakage [6] or size [7] is a new research field. Two of the main problems are high characterization-effort for gates and that only the variation of some parameters is observed. The purpose of this paper is the improvement of leakage reduction under the consideration of parameter variations. Section 2 proposes a method to model gate delay under several parameter variations. Section 3 presents an approach for statistical static timing analysis. Section 4 introduces the idea of DTTCMOS design technique and describes an algorithm which reduces the leakage of a design for a desired yield. Simulation results of ISCAS circuits are presented in section 5. Section 6 draws the conclusion.

2. Modeling Variations of Gate Delay

The delay T_d of a CMOS device can be modeled with alpha-power law model as [9]:

$$T_d = \frac{k' \cdot V_{dd} C_L}{(W/L) \cdot (V_{dd} - V_{th})^{\alpha}}$$

where, *k'* is a technology constant, V_{dd} is the supply voltage, W_{eff} is the width, and L_{eff} the effective gate length, C_L is the load, and α models the short channel effects. The threshold voltage V_{th} can be modeled as:

$$V_{th} = \frac{1}{n\beta}\left(V_{th0} + \gamma' \cdot \sqrt{NDEP} \cdot T_{ox} V_{bs} + \eta' \frac{T_{ox}}{L_{eff}^2 \cdot \sqrt{NDEP}} V_{ds} \right)$$

$$\beta = \frac{kT}{q}, \eta' \approx \frac{\left(E_{TA0} + E_{TAB} \cdot V_{bs}\right) \varepsilon_{Si}^{3/2}}{D_{SUB}{}^{2} \cdot \varepsilon_{ox} \sqrt{q}}, \gamma' = \frac{\sqrt{2q\varepsilon_{Si}}}{\varepsilon_{ox}}$$

where, T is the operating temperature, n is the sub-threshold swing coefficient, V_{th0} is the zero-bias threshold voltage, V_{gs} is the gate-source voltage, V_{bs} is the bulk-source voltage, V_{ds} is the drain-source voltage. D_{SUB} and E_{TA0} are technology dependent drain induced barrier lowering (DIBL) coefficients, and E_{TAB} is a body-bias coefficient. The terms q, k, ε_{ox}, and ε_{Si} correspond to physical constants (respectively, charge on an electron, Boltzmann's constant, gate dielectric constants of gate oxide and silicon). *NDEP* labels the channel doping concentration, and T_{ox} the thickness of the oxide layer. As usual, the variation is described as Gaussian distribution, because variations are expected to be truly random in nature [7]:

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

where, μ is the mean value and σ the standard deviation. 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}. Then, 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 in common approaches.

The new idea is the modeling gate delay variation under consideration of more than one or two parameters. The demand is an easy to handle model. Thus, we choose a Gaussian distribution description, whereas values for μ and σ are extracted from *Monte-Carlo* spice simulations of the gates. We verified the accuracy of this approach for an inverter, based on a predictive 65nm technology library [8]. First, *Monte-Carlo* spice simulations were applied, where all parameters, which can vary, were described with Gaussian distribution. This is used, to approximate realistic parameter variations on a chip. We assumed 10% variation of L_{eff}, $W_{eff,}$ *NDEP*, T_{ox}, T, and V_{dd} for each transistor, whereas parameter variations at each transistor are different, except T and V_{dd}. As the distributions of resulted delays are similar to Gaussian distribution, we extract expected value μ and standard deviation σ of gate delay. Finally, we describe gate delay as function of gate length L_{eff}, where variance has the same distribution as in previous simulations.

The results are depicted in figure 1. As expected, the new model approach is much closer to approximate realistic variations than common approaches with one modified parameter. These results based on the behavior of Gaussian distributions, whereas convolution of Gaussian distributions results in new Gaussian distributions. As evaluation of gate delay is based on multiplication of varied parameters, the distribution of gate delay is a convolution of Gaussian distributions.

3. Timing Analysis

Corner-case Static Timing Analysis

The concern of worst-case static timing analysis (STA) is the evaluation of maximum circuit performance. This knowledge is necessary to integrate circuits into complex design environments.

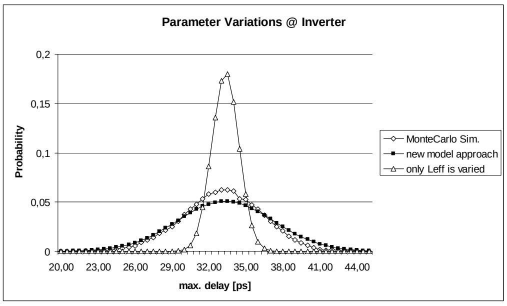

Figure 1: Distribution of Inverter Delay considering parameter variations

STA can be performed at different design levels, but analysis of gate level offers best trade-off between evaluation time for STA and accuracy. In established corner-case STA approaches, gate libraries with corner-case models are applied, to account for parameter variations. At corner-case timing analysis each gate is set to its worst-case delay value. This is followed by longest path computation and determination of critical path delay. 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. In addition, perfect correlation of delay response sensitivity of each gate to variation in each parameter with all other delay elements is assumed. But, intra-die variations cannot be ignored in technologies with gate length below 100nm [1]. In addition, the delay response of the different gates is not perfectly correlated [2]. Hence, traditional corner-case STA is quite pessimistic and underestimates the value for typical performance and overestimates the worst-case timing behavior [3].

Statistical Static Timing Analysis

In contrast, statistical static timing analysis (SSTA) considers intra-die variations. Gate delays are based on probabilistic functions (see section 2). 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 delay has the value x:

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

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

where, σ is the standard deviation, σ² is the variance, and μ is the expected value. As described in section 2, we model the gate delays and data arrival times as Gaussian distributions with expected value μ and variance σ². At single input gates the output signal arrival times are:

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

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

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

Timing Analysis on Multi Input Gates

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 combination [7]. 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 currently probability is considered for each input signal, that it has arrived. The result is a CDF for time of evaluation start. The estimation of this CDF can be divided into two main cases. In the first case, the signal arrives much later than other input signals. Then, its CDF is nearly equal to CDF of evaluation start time. Consequently, PDF of last arriving input signal and evaluation start time $eval_{begin}$ are nearly equal.

In the second case, overlap of inputs 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 of $eval_{begin}$, we simplify the complex problem by an approximation.

We use the approximation, that the rising edge of a CDF can be described as straight line *s(x)* with:

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

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

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

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:

$$t_{\max} = \max(\mu_1 + 3\sigma_1, \mu_2 + 3\sigma_2, ..., \mu_n + 3\sigma_n)$$

$$\sigma_{new} = \left(t_{\max} - \mu_{new}\right) / 3$$

Hence, the approximated CDF and PDF of the evaluation start time results from the new expected value μ_{new} and the new standard deviation σ_{new}. Figures 2 depicts CDF of input signals *in1* and *in2* with overlapping arrival time probability, the resulting probability for evaluation start time $eval_{begin}$, the new generated probability for the evaluation start time $new\text{-}eval_{begin}$, 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 maximum μ_{input}. Furthermore, σ_{new} is based on σ of the latest arriving signal, which has not the greatest μ in any case. But, estimation of timing is much more accurate than approaches, where behavior of multi input gates is ignored and the effort is much lower than at library based approaches.

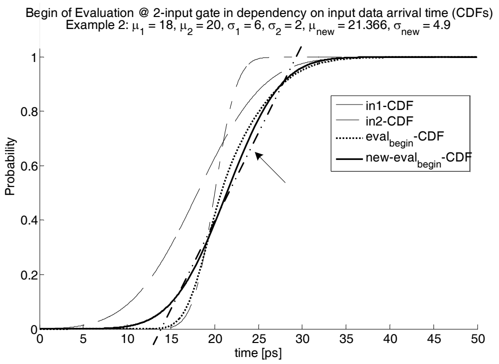

Figure 2: CDFs of overlapping input signals and resulting worst-case time for start of evaluation

4. SSTA and Leakage Reduction

DTTCMOS

The Dual Threshold CMOS (DTCMOS) design techniques are based on the dependency of delay and sub-threshold leakage of a gate from threshold voltage V_{th} of its transistors. High V_{th} results in a relative long delay and low leakage currents, while low V_{th} results in a relative short delay and high leakage currents. Thus, common DTCMOS approaches use LVT gates, which consist of low-V_{th} transistors, and HVT gates, which consist of high-V_{th} transistors. The purpose of DTCMOS optimization algorithms is to increase the number of HVT gates at constant performance. Hence, only gates in noncritical paths will be transformed into HVT gates, while gates in critical paths stay LVT type.

The Dual Oxide Thickness CMOS (DTOCMOS) design techniques are similar to DTCMOS. The difference is the varied parameter, which is gate oxide thickness T_{ox}. Gate delay and gate leakage current depend on this parameter [4]. Thus, gates with thick T_{ox}, which have low gate leakage and long delay, are applied to noncritical paths. On the other hand, gates with relatively thin T_{ox} and high gate leakage and short delay, are used in critical paths.

DTTCMOS design techniques are a combination of gates, which consist of transistors with two different T_{ox} and V_{th}. The first resulting gate type is LVTT (low-T_{ox}, low-V_{th}), which is fast and has relatively high leakage. The second gate type is HVTT (high-T_{ox}, high-V_{th}), which is slower and has much lower leakage. Based on stack effect and gate leakage, total leakage of a gate depends on the input vector [4]. Hence, total leakage of each gate has to be characterized for all possible input vectors.

In addition, leakage is sensitive to parameter variations. The proposed approach models the variation of total gate leakage in the same way as variation of the gate delay. Thus, Gaussian distribution for total leakage for each gate input vector with *Monte-Carlo* spice simulations are determined.

Algorithm

The purpose of the proposed algorithm is the reduction of leakage power dissipation at desired probability for circuit delay. Therefore, the new approach is a combination of the DTTCMOS approach and SSTA. At first, signal probabilities are estimated. Then, the expected average leakage $I_{exp,LVTT}(G)$ of LVTT type of each gate is taken from library. Next, the weighting delay factor $t_{wd,HVTT}(G)$, which is the sum of expected delay value μ_{HVTT} *(G)* and variance var_{HVTT} *(G)*, is evaluated for each gate. Subsequently, for each gate the weight factor $\Psi(G)$ is estimated. $\Psi(G)$ is the product of the inverse leakage $I_{exp,LVTT}(G)$ and weighted delay $t_{wd,HVTT}(G)$ of the gate. Next, all gates are set to HVTT type and SSTA as described in section 3. This step is followed by detection of the critical path. In this path, the gate with highest $\Psi(G)$ will be transferred to LVTT type and the new critical path will be detected. This step is repeated until all critical paths are completely optimized. Finally, at each gate a reordering of input signals is performed, so that the lowest possible total leakage is generated at the input vector with highest probability.

5. Simulation Results

A standard cell library consisting of high-V_{th} and thick-T_{ox} (HVTO) cells and low-V_{th} and thin-T_{ox} (LVTO) cells was created [10]. ISCAS circuits are used to benchmark the approach. Firstly, a corner-case LVTO-cc version of the circuits is implemented. There, the maximum operating frequency f_{max} of the circuits, which consist of LVTO cells only, results from the worst-case delays of cells and corner-case STA. The implemented of a corner-case DTTCMOS–cc version of all circuits followed, where the maximum frequency f_{max} is equal to the corresponding LVTO-wc version of the circuits. In this case, f_{max} results from worst-case of cells again. Then, two DTTCMOS-SSTA versions of the circuits were implemented, where f_{max} of LVTO-cc version is reached by 95% and 99%, respectively. The leakage for different design input vectors is measured and the average leakage of all implemented versions is evaluated. Lastly, the implementation of two DTTCMOS-SSTA versions of the circuits followed, where f_{max} is reached by 95% and 99%, respectively, without any restrictions to f_{max}. The results are depicted in figure 3.

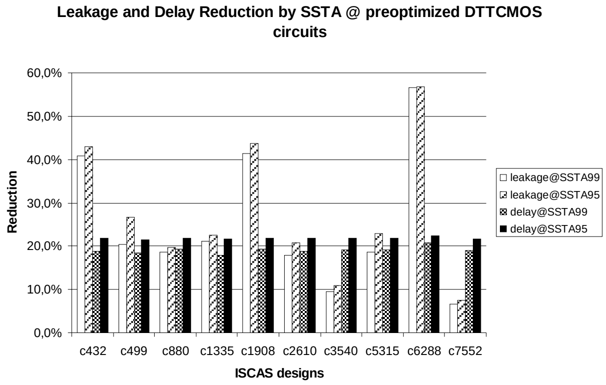

Figure 3: Results from optimizations of ISCAS circuits

As expected, the application of DTTCMOS reduces the total leakage by 57% on average at constant performance. In addition, depending on the improvement by DTTCMOS, application of statistical static timing analysis can improve the reduction of total leakage by more than 50%. The comparison of performance of corner-case STA and SSTA shows an average improvement of 19% and 22% for 99% and 95% reliability, respectively. The difference in total leakage between circuits, where the probability of fitting desired performance is 95% and 99% is average 2%. Thus, we recommend the application of 99% reliability.

6. Conclusion

This paper proposed an approach for reducing total leakage under consideration of intra-chip parameter variations. A pragmatic but effective method to model all parameter variation effects on gate delay was presented. Furthermore, the paper presented a worst-case statistical static timing analysis (SSTA) to handle the problem of performance underestimation. Therefore, a new approach for estimation of statistical distribution of evaluation start time on multi-input gates was proposed. The new SSTA was combined with a Dual-V_{th}/Dual-T_{ox} CMOS (DTTCMOS) optimization algorithm to reduce total design leakage. Benchmark simulations with ISCAS designs show an average reduction of total circuit leakage of 70% if SSTA and DTTCMOS are applied in contrast to traditional corner-case simulated raw designs. The total leakage of corner-case simulated and DTTCMOS optimized designs can be reduce by an average of 20% by application of SSTA. Furthermore, the paper demonstrated that the difference between 95% and 99% reliability at total leakage and performance optimization is nearly negligible.

7. References

[1] S. Borkar et. al., “Design and reliability challenges in nanometer technologies”, *DAC,* USA, 2004.

[2] M. Orshansky et. al., “A general probabilistic framework for worst-case timing analysis”, *DAC*, USA, 2002.

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

[4] F. Sill et. al., “Reducing Leakage with Mixed-Vth“, *18th Conf. on VLSI Design*, India, 2005.

[5] A.K. Sultania et. al., “Transistor and Pin Reordering for Gate Oxide Leakage Reduction in Dual-T_{ox} Circuits”, *ICCD*, USA, 2004.

[6] A. Srivastava et. al., “Statistical Optimization of Leakage Power Considering Process Variations using Dual-V_{th} and Sizing”, *DAC*, USA, 2004.

[7] S.H. Choi et. al., “Novel Sizing for Yield Improvement under Process Variation in Nanometer Technology”, *DAC*, USA, 2004.

[8] Device Group at UC Berkeley: “Berkeley Predictive Technology Model”, 2002.

[9] T. Sakurai and A. Newton, “Alpha-Power Law MOSFET Model and its Application to CMOS Inverter Delay and other Formulas”, *JSSC*, no. 2, 1990, pp. 584-594.

[10] F. Sill et. al., “Total leakage power optimization with improved Mixed Gates “, *SBCCI*, Brazil, 2005.