Gate delay variability estimation method for parametric yield improvement in nanometer CMOS technology

Digeorgia da Silva, André I. Reis, Renato P. Ribas

PGMICRO, Federal University of Rio Grande do Sul, Brazil
Instituto de Informática, Federal University of Rio Grande do Sul, Av. Bento Gonçalves 9500, Bloco IV, CEP 91501-970, Porto Alegre, RS, Brazil

Article info
Article history:
Received 1 July 2010
Accepted 16 July 2010
Available online 17 August 2010

Abstract
In digital CMOS circuits, parametric yield improvement may be achieved by reducing the variability of performance and power consumption of individual cell instances. Such improvement of variation robustness can be attained by evaluating parameter variation impact at gate level. Statistical characterization of logic gates are usually obtained by computationally expensive electrical simulations. An efficient gate delay variability estimation method is proposed for variability-aware design. The proposed method has been applied to different topologies (transistor network arrangements) and CMOS gates, and it has been compared to Monte Carlo simulations for data validation, resulting in computation time savings.

1. Introduction

Fluctuations of device characteristics are very pronounced in deep submicron technology. As a consequence, it is necessary to take transistor characteristics variability into account in order to properly calculate the delay of logic gates. Physical parameters are susceptible to random variations, representing an important challenge in performance of circuit built in advanced nanometer processes [1,2]. The statistical nature of process characteristics makes it possible to consider the process parameters and their variations as random variables represented by their probability distribution functions (PDF).

Detailed electrical simulations of circuits are able to provide precise information about their behavior, performance and power dissipation, but at expense of computation time. An efficient way to perform fast circuit timing analysis is by using simplified timing models. These models may provide enough data accuracy of gates under certain conditions of input slope and output load, as well as voltage supply and operating temperature. Timing models can be generated from simulated data or provided by analytical estimation methods [3–6].

On the other hand, logic gates are basically composed by switch networks to execute a specific functional behavior. Switch network building is an important synthesis issue for digital integrated circuits, even if standard cells design methodology is targeted. Different switch (transistor) arrangements can represent the same logic behavior, but presenting distinct electrical characteristics (signal delay propagation and power consumption) as well as physical construction (layout) [7,8]. Technology mapping tools will choose the most adequate available cell implementations considering design costs and constraints.

In this paper, it is proposed a fast and straightforward way to predict the CMOS gate delay variability by considering the transistor threshold voltage variation. This method is suitable for any topology of single stage gates as well as for multi-level circuits. The goal of the proposed method is to provide variability aware information at the cell level to choose among alternative implementations during technology mapping.

In Section 2, some related works are briefly discussed. In Section 3 is presented the proposed method, while results and method validation is discussed in Section 4. Section 5 presents a brief analysis of computation runtime. Conclusions are outlined in Section 6.

2. Related works

Static timing analysis (STA) is usually applied to find the critical paths in terms of signal delay propagation through the circuit. Technology mapping (and sizing) tools chose the most appropriated gates (and drive strengths) in order to optimize area and power consumption, while respecting the specified timing constraints [9]. However, due to random process parameter variation, a large number of fabricated circuits may not meet the required delays, resulting in parametric yield loss. Parametric yield is the probability of achieving timing and power dissipation characteristics, taking into account the variability of process parameters [10,11].

The traditional design effort, guided just by the best, the worst and the nominal case models for device parameters (corners), over- or underestimates the impact of variations. Approaches in
the field of statistical static timing analysis (SSTA) have appeared to overcome the issues of corner-based methodologies [12,13].

In [14,15], the delay distributions of some logic gates are estimated by considering threshold voltage variations due to random-dopant fluctuations. Regarding the threshold voltage, its standard deviation is modeled as depending on the transistor dimensions (channel length and width, gate oxide thickness) and doping concentration.

The work described in [16], in turn, proposes a model for calculating statistical gate delay variation caused by intra- and inter-chip variability, based on a response surface methodology. It introduces sensitivity constants to facilitate the calculation of intra-gate variability without assigning variables to every individual transistor.

The work in [17] proposes a statistical design technique considering both inter- and intra-die variation of process parameters. The effects of process variations on gate delay are pre-characterized and accessed on-the-fly during statistical timing analysis. The goal is to resize the transistor widths with minimal increase in area and power consumption while improving the confidence that the circuit meets the delay constraint under process variations. A sizing tool based on Lagrangian Relaxation (LR) algorithm has been developed for global optimization of transistor widths.

The authors in [4] propose an analytical modeling of the speed performance of CMOS gates that is based on the average transfer of charges across the switching nodes under consideration and explicitly use the threshold voltage of the involved transistors in the calculation of delay. Although that model presents some interesting features, such as to consider the input slope, the input-to-output capacitance coupling (Miller effect) and the short circuit current effects, an even simpler but not less efficient model is aimed.

A simple way to analyze signal propagation delay is to replace the transistors by their ‘on-resistances’ and calculate (or extract) the intrinsic capacitances on the cell in order to find the RC time constant of the network. There are several approaches that use such strategy for delay analysis [18,19]. Thus, it seems reasonable to extend the use of this principle to include consideration of circuit delay variability analysis.

The general objective of this investigation is to provide means to parametric yield improvement of integrated circuits through a variability-aware design strategy. The generation of more robust CMOS gates, in relation to variability and improvement of the parametric yield of a given circuit, requires the knowledge of how the parameters variations impact the performance of the cell. Preliminary results about transistor arrangements behavior in respect to transistor threshold variation has been presented in [15].

The authors in [11] describe a new technology mapping algorithm that performs library binding to maximize parametric yield limited by timing and power constraints. It proposes a statistical technology mapping that find a circuit mapping such that the yield at a required power target is maximized while meeting the specified timing constraints at a pre-defined yield level. In the next section, a gate delay variability estimation method, based on electrical DC simulations and straightforward mathematical manipulations, is described.

3. Gate delay variability estimation method

The present work proposes a fast, straightforward and efficient procedure to predict the delay variability of logic gates. It is performed here by assigning a single variable – the threshold voltage variation \( \Delta V_{th} \) – to each transistor at the gate under analysis. The gate topology can be divided into pull-up and pull-down logic planes (or networks), presenting only the devices in pull-up and pull-down paths. The analysis is done separately for pull-up and pull-down. Fig. 1 shows the flow of the delay variability method.

Initially, the \( V_{th} \) of transistors are varied ±10% of their nominal values and transformed into coded variables with values \(-1\) and \(+1\), which represent the minimum and maximum values assumed by \( V_{th} \), respectively. Factorial designs [20] are used to identify the main effects of the factors \( \Delta V_{th} \) on the targeted variable, i.e., the transistors ‘on-resistance’. Since each variable of interest is given by two levels (the coded values \(-1\) and \(+1\)), each variant of such a design has \( 2^k \) experimental runs being called a \( 2^k \) factorial design, where \( k \) is the number of transistors in the gate under test (i.e., ‘k’ threshold voltage values).

As an example, Table 1 shows factorial values when a network with three transistors is considered. DC type electrical simulations extract the ‘on-resistances’ of transistors. The resistances are provided by \( 2^3 \) (three transistors) DC simulation with HSPICE, by dividing the voltage drop in each transistor by the current through it.

The effect estimated in a \( 2^k \) factorial design is then converted into a regression model (first-order function for the resistance of the device) that can be used to analyze the response at any point in the space spanned by the factors (coded \( \Delta V_{th} \) in the design). The procedure of fitting regression models to the responses of the simulation model evaluated at several points is known as response surface methodology [20].

The dynamic response of a MOS represents the output transition as a function of the time. It takes into account the charging and discharging of parasitic capacitances that are intrinsic to the device, and the capacitances introduced by the interconnect lines and the load. The intrinsic capacitances originate from three sources: the basic MOS structure, the channel charge, and the depletion regions of the reversed-biased pn-junction of transistor drain and source regions [18]. The intrinsic capacitances are calculated for specific technology nodes and are considered as constants in the proposed method. The delay is calculated through the RC constant of the network by using again factorial design and a first-order regression model represents the delay as a function of the coded \( \Delta V_{th} \).

Two different delay models have been used to calculate the delay variation for the transistors networks, the Elmore delay [21] and the asymptotic waveform evaluation (AWE) [22]. The resistance equations obtained in the first steps are used to perform the following analysis. The \( \Delta V_{th} \) of the transistors are considered as random variables represented by probability density functions (PDF) with mean values equal to zero and normalized standard deviations \( \text{N}(0,1) \). Since the \( V_{th} \) variations are treated as coded variables, the coefficients of the final delay equation is considered as the standard deviation (\( \sigma \)) resulting from each \( V_{th} \) variation. It is then performed a sum of normal distributions, and the result is a PDF that represents the delay of the network with the two aspects considered: mean and standard deviation (square root of variance). The normalized standard deviation (this one divided by the mean) of the delay is the focus of the analysis of different networks because it makes possible to compare the variability of arrangements with different mean delays. The goal of the proposed method is to provide variability aware information at the cell level to choose among alternative implementations during technology mapping.

4. Results and discussion

Experimental data, shown in the next sections, were obtained taking into account the 45 nm bulk CMOS PTM technology [24], using transistor channel length \( L \) of 45 nm and width \( W \) varying from 90 nm to 180 nm. Electrical simulations were carried out through HSPice considering the BSIM4 model.
4.1. Series networks

Tables 2 and 3 present the normalized delay deviation for NMOS and PMOS transistor stacks, with three and four devices (stack-03 and stack-04, respectively), according to the position of the switching transistor in relation to the output node, as illustrated in Fig. 2a.

The application of the AWE procedure presented better predicted delay deviations for the switching device far from the output node than for transitions in devices close to the gate output. One can also observe a small reduction in delay variability when more stacked transistors are present, for both simulated and

Table 1
Combinations of min and max values considered for the threshold voltages variations of devices in a 3-transistors network.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Stack-03</th>
<th>V_{th1}</th>
<th>V_{th2}</th>
<th>V_{th3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–1</td>
<td>–1</td>
<td>–1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>–1</td>
<td>–1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>–1</td>
<td>1</td>
<td>–1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>–1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>–1</td>
<td>–1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>–1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>–1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

4.1. Series networks

Tables 2 and 3 present the normalized delay deviation for NMOS and PMOS transistor stacks, with three and four devices (stack-03 and stack-04, respectively), according to the position of the switching transistor in relation to the output node, as illustrated in Fig. 2a.

The application of the AWE procedure presented better predicted delay deviations for the switching device far from the output node than for transitions in devices close to the gate output. One can also observe a small reduction in delay variability when more stacked transistors are present, for both simulated and
estimated results, but not for the case of ‘close switching devices’. According to the statistical simulations performed, ‘close switching’ causes higher delay deviation independently on the number of transistors in the stack.

4.2. Parallel networks

Tables 4 and 5 present the normalized delay deviation for NMOS and PMOS transistors placed in parallel arrangement in the networks, as illustrated in Fig. 2b. The metrics are presented considering 1, 2 or 3 transistors switching at the same time, with the other(s) device(s) turned-off. The equivalent resistance of the parallel arrangement is applied to get the RC time constant.

4.3. CMOS inverter topologies

The delay variability estimation method has also been applied to different CMOS inverter topologies, illustrated in Fig. 3, and the results are presented in Table 6. The total area of the pull-down NMOS and pull-up PMOS networks was kept constant for all configurations presented. In order to investigate the influence of the area on the standard deviation of $V_{th}$ and consequently on the delay of the logic gate, different variations were considered for different sizes by applying the relation between the parameter variation and the area provided by the Pelgrom’s model [23].

The method presented delay variability values in agreement to the simulated results when it concerns to the topology less affected by $\Delta V_{th}$ for the rising- and falling-edge delays. In general, the method underestimated the delay variability for both NMOS and PMOS networks evaluated.

### Table 3

Delay deviation for PMOS transistors in series configuration, according to the position of the switching transistor in relation to the output node (1-close...4-far, as seen in Fig. 2a).

<table>
<thead>
<tr>
<th>Switching transistor</th>
<th>Stack-03 Method w/Elmore</th>
<th>Method w/AWE</th>
<th>Monte Carlo</th>
<th>Stack-04 Method w/Elmore</th>
<th>Method w/AWE</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0218</td>
<td>0.0216</td>
<td>0.0379</td>
<td>0.0188</td>
<td>0.0210</td>
<td>0.0403</td>
</tr>
<tr>
<td>2</td>
<td>0.0197</td>
<td>0.0198</td>
<td>0.0301</td>
<td>0.0172</td>
<td>0.0155</td>
<td>0.0297</td>
</tr>
<tr>
<td>3</td>
<td>0.0188</td>
<td>0.0176</td>
<td>0.0281</td>
<td>0.0163</td>
<td>0.0159</td>
<td>0.0261</td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.0160</td>
<td>0.0173</td>
<td>0.0249</td>
</tr>
</tbody>
</table>

### Table 4

Delay deviation for NMOS transistors in parallel association, according to the number of switching transistors.

<table>
<thead>
<tr>
<th>Number of switching transistor</th>
<th>Parallel-03 Method</th>
<th>Monte Carlo</th>
<th>Parallel-04 Method</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0560</td>
<td>0.0519</td>
<td>0.0560</td>
<td>0.0505</td>
</tr>
<tr>
<td>2</td>
<td>0.0351</td>
<td>0.0619</td>
<td>0.0351</td>
<td>0.0578</td>
</tr>
<tr>
<td>3</td>
<td>0.0351</td>
<td>0.0619</td>
<td>0.0351</td>
<td>0.0578</td>
</tr>
</tbody>
</table>

\* Only Elmore model were applied.

### Table 5

Delay deviation for PMOS transistors in parallel association, according to the number of switching transistors.

<table>
<thead>
<tr>
<th>Number of switching transistor</th>
<th>Parallel-03 Method</th>
<th>Monte Carlo</th>
<th>Parallel-04 Method</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0361</td>
<td>0.0429</td>
<td>0.0361</td>
<td>0.0430</td>
</tr>
<tr>
<td>2</td>
<td>0.0257</td>
<td>0.0278</td>
<td>0.0257</td>
<td>0.0276</td>
</tr>
<tr>
<td>3</td>
<td>0.0209</td>
<td>0.0218</td>
<td>0.0209</td>
<td>0.0213</td>
</tr>
</tbody>
</table>

\* Only Elmore model were applied.

...
4.4. Inverter chains

The delay deviations of chains with different number of inverters were measured and the results can be seen in Table 7. It is shown in Fig. 4 the test structures used in this exercise. Different chains, varying from one to five ‘X’ inverters placed on that, were evaluated and timing deviations were measured.

In both cases, estimated by proposed method and simulated by using Monte Carlo engine, the chains with the same numbers of inverters presented similar rise and fall delay deviations. The results provided by the method and by Monte Carlo agreed in an important factor: as the number of inverters in the chain increases, the normalized deviations of rise and fall delays also decrease.

### Table 6

<table>
<thead>
<tr>
<th>Inverter topology</th>
<th>Rise delay deviation</th>
<th>Fall delay deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Monte Carlo</td>
<td>Method</td>
</tr>
<tr>
<td>(a)</td>
<td>0.0359</td>
<td>0.0445</td>
</tr>
<tr>
<td>(b)</td>
<td>0.0201</td>
<td>0.0223</td>
</tr>
<tr>
<td>(c)</td>
<td>0.0363</td>
<td>0.0431</td>
</tr>
</tbody>
</table>

### Table 7

<table>
<thead>
<tr>
<th>Number of inverters</th>
<th>Rise delay deviation</th>
<th>Fall delay deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>1</td>
<td>0.0360</td>
<td>0.0407</td>
</tr>
<tr>
<td>2</td>
<td>0.0252</td>
<td>0.0293</td>
</tr>
<tr>
<td>3</td>
<td>0.0206</td>
<td>0.0266</td>
</tr>
<tr>
<td>4</td>
<td>0.0178</td>
<td>0.0224</td>
</tr>
<tr>
<td>5</td>
<td>0.0160</td>
<td>0.0209</td>
</tr>
</tbody>
</table>

4.5. Exclusive-OR with 4-inputs (XOR4)

It was also evaluated the delay deviation of 4-input exclusive-OR gates (XOR4), built in different topologies and logic styles: (1) differential cascode voltage switch logic (DCVSL) and (2) differential pass-transistor logic (DPTL), illustrated in Fig. 5.

The propagation signal path was simplified in order to apply the modeled resistances to the Elmore delay model in the calculation of the delay variability. This was performed by replacing parallel resistances with their equivalent ones. The AWE technique was also applied in some cases whose results obtained by Elmore presented the most significant errors. The input signals are correlated in the sense that a denied signal (!A) is collected at the output of an inverter that has normal signal (A) at its input. The $\Delta V_{th}$ of this additional inverter was not taken into account in the delay PDF calculation.

Table 8 compares the delay deviations provided by the proposed method with those achieved through statistical Monte Carlo simulations.

The best results (the most similar ones to the simulated values) presented by the method, considering both rising- and falling-edge delay variations, were observed for the DCVSL style. The method was able to predict the least robust configuration among those used to implement the XOR4 gate. The largest error presented by the method was for the falling-edge delay deviation of the DPTL style by considering the modeled resistances applied to the Elmore delay model. However, the AWE technique was then used for this case and a better result was achieved: 0.1314, with an error of 10.3%.

In Fig. 6 is observed the high rising-edge delay deviation presented by the tested XOR4 versions. The method presented good reliability. The method is able to estimate which configuration presents the highest variability between the alternative implementations.

---

**Fig. 3.** CMOS inverters: (a) conventional; (b) low leakage approach; (c) folded topology.

**Fig. 4.** Inverters chains varying from 1 to 5 ‘X’ gates.
This way, the method fulfills the goal of providing reliable variability aware information at the cell level to choose among alternative implementations during technology mapping.

5. Runtime analysis

The characterization of the logic gates by using Monte Carlo approach is computationally very expensive. It takes a long time to be completed; for 10 k runs it may take more than hours, what makes it prohibitive as an estimator in a technology mapping context. The characterization based on transient simulations requires an execution time that is at least one order of magnitude higher than the proposed method. The approach presented in this work to characterize the cells is based on fast simulations, which take only few seconds, since only one of them is a transient simulation and all the others are DC type analysis. A significant speed-up was observed for the evaluated gates, which is important on a mapping context as many evaluations will be performed.

6. Conclusions

This work proposed a new method for predicting the delay variability of CMOS logic gates. It has been applied to different logic families, including not only the conventional static CMOS, but also differential structures like DCVSL and DPTL structures. The method provides reliable variability aware information at the cell level to choose among alternative implementations during technology mapping. The method is also efficient from an execution time standpoint, which is important on a mapping context as many evaluations are performed during mapping.

Acknowledgements

Research partially funded by Nangate Inc. under a Nangate/UFRGS research agreement, by CNPq and Capes Brazilian funding agencies, and by the European Community’s Seventh Framework Programme under Grant 248538-Synaptic.

References


