A Timing Optimization Technique for Nanoscale CMOS Circuits Susceptible to Process Variations

Kumar Yelamarthi  
School of Engineering and Technology  
Central Michigan University  
Mt Pleasant, USA  
yelamarthii@ieee.org

Chien-In Henry Chen  
Department of Electrical Engineering  
Wright State University  
Dayton, USA  
henry.chen@wright.edu

Abstract—Performance variation is one of the primary concerns in nanoscale CMOS circuits. This performance variation is worse in circuits with multiple timing paths such as those used in microprocessors. In this paper, a Process Variation-aware Transistor (PVT) sizing algorithm is proposed, which is capable of reducing worst-case delay, delay uncertainty, and delay sensitivity to process variations in nanoscale CMOS circuits. The proposed algorithm is based on identifying the significance of timing paths in a design, and performing respective optimization for optimal design performance. Additional advantages in this algorithm include its simplicity, accuracy, independent of the transistor order, and initial sizing factors. Using 90 nm CMOS process, the proposed algorithm has demonstrated an average improvement in worst-case delay by 36.9%, delay uncertainty by 44.1%, delay sensitivity by 19.8%, and power-delay-product by 35.3% when compared to their initial performances.

Keywords: Transistor sizing, process variations, timing optimization

I. INTRODUCTION

Performance optimization of microprocessors and systems has been driven traditionally by nanometer scale CMOS logic circuits and can be further enhanced through circuit design and topology organization. Of the several methods available for performance optimization, transistor sizing is one of the most effective due to many reasons: 1) several sources of power consumption due to glitches and short-circuit currents can be minimized through transistor sizing, 2) it affects not only the resistance and timing constant, but also the propagation delay due to parasitic capacitances, and 3) it can help to maintain sufficient noise margins. However, technology scaling into the nanometer era has led to a lack of process uniformity in semiconductor manufacturing and has made process variations the primary cause of concern [1]. These variations cause significant unpredictability in the performance characteristics and introduce uncertainties at each step of the process in development, design, manufacturing, and test. This process and device variability challenge in the continued scaling of nanoscale CMOS circuit exacerbates the already-critical performance optimization problem, and is one of the most critical problems confronting designers. Accordingly, these process variations need to be taken into account during the design phase to ascertain the accurate estimation of performance characteristics [2].

One of the challenges in timing optimization of nanoscale CMOS circuits is delay uncertainty ($\Delta$) from process variations, $\Delta = T_{\text{max}} - T_{\text{min}}$, where $T_{\text{max}}$ and $T_{\text{min}}$ are the maximum and minimum delays of a timing path after process variations are taken into consideration. In the 180nm CMOS technology, these process variations have caused about 30% variation in chip frequency, along with 20x variation in current leakage [3]. Accordingly, a large number of chips with significantly high leakage have to be discarded, resulting in a considerable parametric yield loss [1]. Also, this magnitude of intra-die channel length variations has been estimated to increase from 35% of total variations in 130nm to 60% in 70 nm CMOS process, and variation in wire width, height, and thickness is also expected to increase from 25% to 35% [4]. In 65nm CMOS process, the parameters that affect timing the most are device length, threshold voltage, device width, mobility, and oxide thickness [5]. For process variation sensitive circuits such as SRAM arrays and dynamic logic circuits, these process variations may result in functional failure and yield loss [4].

There has been a wealth of research aiming to optimize timing performance through transistor sizing [6-10]. Timed LOGic Synthesizer (TILOS) [6] presented an algorithm of iteratively sizing transistors by a certain factor in the critical path. TILOS is not a deterministic approach, as it does not guarantee convergence in timing optimization. MINFLOTRANSIT [7] is another algorithm proposed for transistor sizing based on iterative relaxation method, but requires iterative generation of directed acyclic graphs in every step of timing optimization. Logical Effort computation is the one other method proposed for timing optimization [8]. However, it has two limitations. First, it requires estimation of input capacitance, of which circuits with complex branches or multiple paths have difficulty in accurate estimation. Second, it optimizes timing at the cost of increased area [9]. Simple Exact Algorithm [10] is another approach where series and parallel transistors are grouped for iterative parametric sweep analysis to identify optimal transistor sizes. Although this method works for dynamic CMOS logic, it is not a deterministic approach, as it does not guarantee timing convergence. Moreover, its simulation time increases at a quadratic rate with the number of transistors in the design. Most importantly, none of these algorithms account for process variations.

Addressing the challenges of timing optimization and delay uncertainty from process variations, this paper presents a process variation-aware transistor (PVT) sizing algorithm to reduce worst-case delay, delay uncertainty, and delay...
sensitivity due to process variations in nanoscale CMOS circuits that are primarily used in microprocessors and other logic elements, where the effect of random variation is more pronounced [11]. The proposed algorithm is validated through implementation on several ISCAS and other benchmark circuits using a 90nm CMOS process.

II. PERFORMANCE UNCERTAINTY DUE TO PROCESS VARIATIONS

Process variations are broadly classified into two types: inter-die variations representing variations from chip to chip for the same circuit, and intra-die variations representing the process variations at different locations on the same chip. Research has shown that intra-die variations primarily impact the mean delay, and inter-die variations impact the variance of delay [13]. To account for inter-die and intra-die variations in timing optimization process, both mean and variance should be considered. In Fig. 1, curve A represents the probability density function (PDF) of delay distribution in an un-optimized circuit. It can be seen that minimizing the mean delay of the circuit without any regard to its standard deviation results in an increased leakage current (curve B). Whereas, minimizing just the standard deviation without considering mean delay will not reduce the overall mean delay (curve C). Hence, the goal of maximizing performance improvement can be achieved by minimizing a linear combination of mean and the standard deviation of the delay distribution due to process variations (curve D). Thus, the new objective becomes reducing \( \mu + \sigma \) where \( \mu \) is the mean delay, and \( \sigma \) is the standard deviation of delay distribution. In addition to optimizing path delay, other parameters affected by process variations and considered during optimization are delay uncertainty \( \Delta = T_{\text{max}} - T_{\text{min}} \) and delay sensitivity \( \delta = \sigma / \mu \), where \( T_{\text{max}} \) and \( T_{\text{min}} \) are the maximum and minimum delays due to process variations.

Extensive research was performed to understand and reduce the significance of process variations. These algorithms deal with statistical variations and are not optimal for designs with a large number of parameter variations, or require the parameters to be correlated, which is the not case in nanoscale CMOS circuits where process variations occur in a highly uncorrelated and multi-dimensional space. The true (physics based) measurement and analysis involves tens of thousands, if not hundreds of thousands of variables [12]. On the other hand, performance uncertainty measurement and analysis using multiple corner method (slow, typical, fast) is impractical as some problems are almost uncaught in this method [12]. With the increasing number of process variations parameters, it’s impractical to use enough corners to estimate the performance uncertainty in nanoscale CMOS circuits. A better solution for this is to pick some sets of technology parameters, find the worst case paths in each case, and take the union of these paths to obtain the performance uncertainty characteristics such as \( \mu, \sigma, \Delta, \delta \) of the design under test as in Fig 2. Some of the parameters variations considered in this method are gate oxide thickness, threshold voltage, mobility variation due to dopant mismatch, gate length and width, variation due to gate orientation, drain overlap capacitance, and junction capacitance as these are parameters that have a significant effect on performance characteristics [5].

III. PROCESS VARIATION-AWARE TRANSISTOR SIZING

The delay of dynamic circuit is highly dependent on the number and size (width) of transistors in the critical path. Increasing width of transistors in a path will increase the discharging current and reduce the output pull-down path delay. However, increasing width of transistors to reduce one path delay may increase the capacitive load of channel-connected transistors on other paths, and substantially increase their delays. This complexity increases along with the number of paths present in the circuit. A 2-b Weighted Binary-to-Thermometric Converter (WBTC) used in high-performance binary adders is shown in Fig. 3. This design is used as an example to explain the path delay optimization complexity and the Process Variation-aware Transistor (PVT) sizing algorithm presented in Fig. 4.

Consider the timing path with a series of channel-connected NMOS transistors, \( T_{1b} T_{1} T_{2} T_{3} \). While \( T_{j} \) conducts the discharge current of load capacitance at just the output \( c_{oj} \), \( T_{b} \) has to conduct the discharge currents from several timing paths and outputs \( (c_{0b}, c_{1}, c_{2}, c_{3}, c_{4}, c_{5}) \), which are substantially large. Accordingly, discharge time of the transistor closer to the voltage source \( (T_{b}) \) is greater than that of transistor closer to the dynamic output node \( (T_{i}) \). Accounting for these discharge times, sizes of transistors on the discharging path needs to be made progressively larger. In the proposed PVT sizing algorithm, a transistor weight \( w \) in the range of 0.05-0.5 is

![Figure 1. Mean-standard deviation tradeoff for delay reduction](image)

![Figure 2. Induce parameter variations to estimate performance variations](image)
assigned to each transistor relative to its distance from the dynamic output node. For instance, the 2-b WBTC in Fig. 2 is comprised of seven transistor stacks relative to their distance from the dynamic output node. Stack-1, closest to the dynamic output node includes transistors $T_8$, $T_{10}$, $T_{16}$, $T_{31}$, $T_{25}$ and $T_{27}$. Stack-2 includes transistors $T_6$, $T_{13}$, $T_{18}$, $T_{23}$ and $T_{26}$. Stack-3 includes transistors $T_2$, $T_9$, $T_{15}$, $T_{20}$, and $T_{24}$. Stack-4 includes transistors $T_5$, $T_{12}$, $T_{17}$, and $T_{22}$. Stack-5 includes transistors $T_1$, $T_b$, $T_{14}$ and $T_{19}$. Stack-6 includes transistors $T_4$ and $T_{11}$. Stack-7 farthest from the dynamic output node includes transistors $T_0$ and $T_7$. Accordingly, transistor sizes are updated using the respective weights of 0.05, 0.1, 0.15, 0.2, 0.3, 0.4 and 0.5 respectively. For designs with different number of stacks, weights are evenly distributed in the range 0.05-0.5 relative to its distance from the dynamic output node with a weight of 0.05 assigned to stack closest to the dynamic output node, and a weight of 0.5 assigned to stack farthest from the dynamic output node.

As increasing width of transistor that appears in the most number of paths reduces overall delay, the number of paths a transistor is present in is computed and denoted as repeats. The initial step in PVT sizing algorithm is to compute the repeat and weight profiles of all transistors \{T_{0}, T_{1}, ..., T_{n}\}. Upon computing these profiles, all transistor sizes are initialized with $S_i = S_{min} \times f^\alpha$ for faster optimization convergence, where $f$ is a technology dependent initial sizing factor between 1.1 and 1.5, and $\alpha$ is the relative position of transistor from dynamic output node. With this initial transistor sizing, process variations are induced to obtain delay distribution of all paths. The transistors on the top 20% critical paths based on delay $(\mu + \sigma)$ are grouped to set-x, and their weights are increased by an optimization formula, $S_{new} = S_{old}(1 + \frac{r_{jw_j}}{(1 + r_{j})})$. As the delay of critical path is dependent on the capacitive load of channel-connected transistors, reducing this capacitive load from neighboring paths reduces the overall delay. The 1st order connection transistors in set-x are identified and grouped to set-$y$. Then transistors in set-$x$ are excluded from set-$y$ if any, and sizes of set-$y$ transistors are decreased by an optimization formula, $S_{new} = S_{old}(1 - \frac{r_{jw_j}}{(1 + r_{j})})$. Next, the circuit is updated with these new sizes and process variations are induced as in Fig. 2 to obtain new delay profiles for further timing optimization. This algorithm is repeated until the delay converges to an acceptable value.

### IV. PERFORMANCE EVALUATION OF PVT SIZING ALGORITHM

A 2-b WBTC used in high-performance binary adders is shown in Fig 3. With less than 50 transistors, the 2-b WBTC has 34 timing paths, and of which path delays change significantly with different transistor sizes. Prior to optimization, the worst-case delay of 2-b WBTC was 328 ps from path-1 ($T_{28}$, $T_{0}$, $T_{6}$, $T_{11}$, $T_{22}$, $T_{26}$). The top 20% critical paths are path-1, 2, 5, 8, 26, and 29. Widths of all transistors in these critical paths are initially increased by a ratio $S_i$ to their initial values. For example, the sizes of transistors ($T_{22}$, $T_{11}$, $T_4$ and $T_0$) in path-1 are increased to 120x1.1$^1$=132 nm, 120x1.1$^2$=145 nm, 120x1.1$^3$=160 nm, 120x1.1$^4$=176 nm, respectively. After initial transistor sizing, process variations are induced as in Fig.2 and delay distribution of each path is obtained. Accordingly, transistor sizes are updated using the respective formulas, and simulations are performed to obtain a new critical path order. After a few iterations of the PVT sizing algorithm, the worst-case delay of 2-b WBTC converged to an optimal delay of 190 ps, which accounts for a 42.1% improvement. Figure 5 shows the significance of PVT sizing algorithm in performance optimization through comparison of delay distribution in pre- and post-optimized 2-b WBTC design, where the mean delay reduced from 328 ps to 190 ps, and delay uncertainty reduced from 133 ps to 33, an improvement of 42.1% and 75% respectively.

1: Compute repeat ($r_j$) and weight ($w_j$) for $i$={$1, 2, 3, ..., m$}
2: Initialize all transistor sizes with $S_i = S_{min} \times f^\alpha$
3: do
4: induce process variations & obtain path-delay distribution ($d_1$,$d_2$,$d_3$,...,$d_n$) of all paths
5: sort and rank all paths based on their delay ($\mu + \sigma$)
6: list set-x ={$T_1,T_2,T_3,...,T_x$} transistors from top 20% paths
7: for $j \leftarrow 1$ to $x$
8: $S_i = S_{old}\left(1 + \frac{r_{jw_j}}{(1 + r_{j})}\right)$
9: end for
10: identify 1st order connection transistors \{set-$y$\} to set-x transistors
11: exclude set-x transistors from set-$y$ if any
12: for $k \leftarrow 1$ to $y$
13: $S_k = S_{new}(1 - \frac{r_{jw_j}}{(1 + r_{j})})$
14: end for
15: for all unaltered transistors, $S_{new} = S_{old}$
16: while // delay converges to an acceptable value

![Figure 3. 2-b weighted binary to thermometric converter](image)

![Figure 4. Process Variation-Aware Transistor (PVT) Sizing Technique](image)
Research presented in [2] shows that a drop in supply voltage degrades cell timing at a quadratic rate; a 5% drop in total rail-to-rail voltage may result in a 15% timing degradation. After the implementation of PVT sizing algorithm on this design, it was observed that a 20% drop in supply voltage results in only 4% variation in timing (much less than 15%), further demonstrating the algorithm’s robustness to variation in supply voltage. Furthermore, the PVT sizing algorithm was also implemented on several ISCAS benchmark circuits listed in Table I, and results obtained have shown an average worst-case delay improvement by 36.9%, delay uncertainty improvement by 44.1%, delay sensitivity improvement by 19.8% with an area increase by 50.8%. As power increases at a quadratic rate with reduction of delay in nanoscale circuits, when power-delay-product (PDP) is used as an evaluation metric, the PVT sizing algorithm has shown an improvement by 35.3%.

V. CONCLUSION

In this paper, it is shown that the significance and complexity in timing optimization for nanoscale CMOS circuits increases along with the number of timing paths and the number and magnitude of process variations. A solution addressing these issues is presented through a process variation-aware transistor (PVT) sizing algorithm. A 2-b WBTC and several ISCAS benchmark circuits have been used to validate the performance of the proposed algorithm with an average improvement in worst-case delay by 41%, delay uncertainty by 41%, delay sensitivity by 19.8%, and power-delay-product by 35.3%.

### Table I. Performance Optimization Results from PVT Sizing Algorithm

<table>
<thead>
<tr>
<th>Design</th>
<th>Pre-Optimized Delay (ps)</th>
<th>Post-Optimized Delay (ps)</th>
<th>Delay Improvement (%)</th>
<th>Area Increase (%)</th>
<th>Uncertainty Improvement (%)</th>
<th>Sensitivity Improvement (%)</th>
<th>PDP Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-b WBTC</td>
<td>328</td>
<td>190</td>
<td>42.1</td>
<td>61.9</td>
<td>75.0</td>
<td>67.7</td>
<td>42.1</td>
</tr>
<tr>
<td>74181-CLA</td>
<td>211</td>
<td>106</td>
<td>49.7</td>
<td>55.0</td>
<td>47.7</td>
<td>8.7</td>
<td>49.7</td>
</tr>
<tr>
<td>c2670-CLA</td>
<td>193</td>
<td>91</td>
<td>52.8</td>
<td>58.4</td>
<td>36.1</td>
<td>9.8</td>
<td>52.8</td>
</tr>
<tr>
<td>c3540-CC5</td>
<td>60.2</td>
<td>36.8</td>
<td>38.8</td>
<td>37.1</td>
<td>22.1</td>
<td>17.1</td>
<td>38.8</td>
</tr>
<tr>
<td>c3540-CC8</td>
<td>362</td>
<td>174</td>
<td>51.9</td>
<td>54.8</td>
<td>43.8</td>
<td>8.9</td>
<td>-</td>
</tr>
<tr>
<td>c3540-CC9</td>
<td>158</td>
<td>121</td>
<td>23.4</td>
<td>46.4</td>
<td>35.6</td>
<td>3.1</td>
<td>-</td>
</tr>
<tr>
<td>c5315-CalP2</td>
<td>105</td>
<td>80</td>
<td>23.8</td>
<td>37.0</td>
<td>41.1</td>
<td>31.1</td>
<td>51.9</td>
</tr>
<tr>
<td>c5315-GLC4_2</td>
<td>87.5</td>
<td>80</td>
<td>8.5</td>
<td>30.8</td>
<td>24.7</td>
<td>12.4</td>
<td>23.4</td>
</tr>
<tr>
<td>c5315-CB4</td>
<td>111.8</td>
<td>81.3</td>
<td>27.2</td>
<td>61.3</td>
<td>44.6</td>
<td>22.4</td>
<td>23.8</td>
</tr>
<tr>
<td>c7552-CGC34_4</td>
<td>218.4</td>
<td>105</td>
<td>51.9</td>
<td>66.1</td>
<td>64.4</td>
<td>19.0</td>
<td>8.5</td>
</tr>
<tr>
<td>c7552-CGC17</td>
<td>58.0</td>
<td>37.4</td>
<td>35.5</td>
<td>49.9</td>
<td>50.6</td>
<td>17.5</td>
<td>27.2</td>
</tr>
<tr>
<td>Average (%)</td>
<td>36.9</td>
<td>50.88</td>
<td>44.1</td>
<td>19.8</td>
<td>35.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES


