INTEGRATION, the VLSI journal 55 (2016) 465–473

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Selective body biasing for post-silicon tuning of sub-threshold designs: A semi-infinite programming approach with Incremental Hypercubic Sampling

Hui Geng [a], Jianming Liu [a], Jinglan Liu [b], Pei-Wen Luo [c], Liang-Chia Cheng [c], Steven L. Grant [a], Yiyu Shi [b,*]

[a] Missouri University of Science and Technology, Rolla, MO 65409, USA
[b] University of Norte Dame, Notre Dame, IN 46556, USA
[c] Industrial Technology Research Institute, Hsin-Chu 31040, Taiwan, ROC

ARTICLE INFO

Available online 30 May 2016

Keywords:
Sub-threshold designs
Body biasing
Semi-infinite programming
Incremental Hypercubic Sampling

ABSTRACT

Sub-threshold designs have become a popular option in many energy constrained applications. However, a major bottleneck for these designs is the challenge in attaining timing closure. Most of the paths in sub-threshold designs can become critical paths due to the purely random process variation on threshold voltage, which exponentially impacts the gate delay. In order to address timing violations caused by process variation, post-silicon tuning is widely used through body biasing technology, which incurs heavy power and area overhead. Therefore, it is imperative to select only a small group of the gates with body biasing for post-silicon-tuning. In this paper, we first formulate this problem as a linear semi-infinite programming (LSIP). Then an efficient algorithm based on the novel concept of Incremental Hypercubic Sampling (IHCS), specially tailored to the problem structure, is proposed along with the convergence analysis. Compared with the state-of-the-art approach based on adaptive filtering, experimental results on industrial designs using 65 nm sub-threshold library demonstrate that our proposed IHCS approach can improve the pass rate by up to 7.3× with a speed up to 4.1×, using the same number of body biasing gates with about the same power consumption.

1. Introduction

Sub-threshold circuit has become a very compelling solution for energy efficiency design, and some successful sub-threshold designs have been proposed in the literature [1,2]. However, considering the exponential relationship between drive current and threshold voltage V_{th}, process variation induced V_{th} variation has become a major concern for sub-threshold designs [5–7], which can jeopardize timing or even lead to functional failures. As such, most sub-threshold designs have been limited to timing-insensitive applications.

Specifically, it has been established that V_{th} variation, mainly caused by random dopant fluctuations (RDF), is purely random [3]. Therefore, the critical paths in different chips of the same design are likely to be completely different after the fabrication. According to the analysis in [7], most paths in a sub-threshold design may become critical because of the V_{th} variation. For example, Fig. 1 shows the probability of each path being critical (out of 10 K Monte Carlo runs) in a sub-threshold design using 65 nm commercial library. From the figure we can see that the highest possibility is less than 1.4%, and more than 90% of the paths have nonzero probability of being critical. Due to this fact, it is almost impossible to perform traditional timing analysis/optimization or to use online measurement and adjustment by replicating critical paths such as [8] for sub-threshold designs, post-silicon tuning is necessary to correct the timing issues for each individual chip.

Body biasing is a very effective post-silicon tuning approach to conquer process variation. It uses body effect to modulate the V_{th} of transistors. Body biasing technology is initially proposed to decrease the leakage power in super-threshold designs, and is later applied to sub-threshold designs to fix timing violations caused by process variation [4–7].

Most body biasing studies assumed the presence of multiple biasing voltage domains [8–10], and these domains increase not only the routing but also the control complexity. It makes these

* Correspondence to: Department of Computer Science and Engineering, University of Norte Dame, Norte Dame, IN 46556, USA. Tel.: +1 574 631 6520.
E-mail addresses: hgrfd@mst.edu (H. Geng), jltx5@mst.edu (J. Liu), jliu16@nd.edu (J. Liu), peiwen@itri.org.tw (P.-W. Luo), aga@itri.org.tw (L.-C. Cheng), sgrant@mst.edu (S.L. Grant), yshi4@nd.edu (Y. Shi).

http://dx.doi.org/10.1016/j.vlsi.2016.05.007
0167-9260/

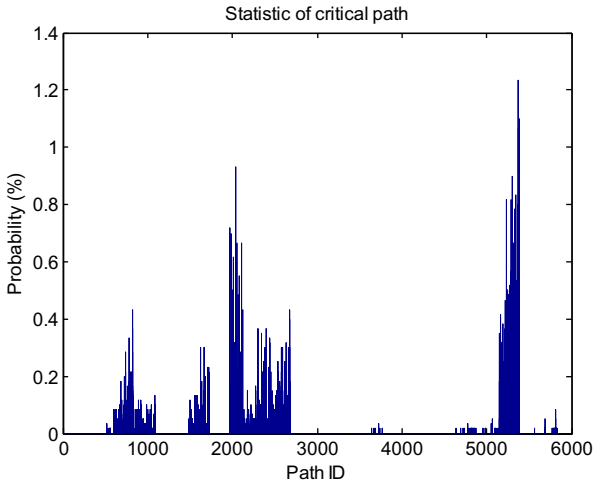

Fig. 1. Critical path distribution with process variation in a 65 nm sub-threshold design [7].

techniques hard to be used in large-volume commercial production. Furthermore, multiple body biasing schemes also suffer from less latch-up immunity, less threshold voltage controllability, more substrate noise vulnerability and less gate oxide reliability.

Therefore, considering the relative importance of reliability and cost, Geng et al. proposed a more practical selective body biasing scheme with only one body biasing voltage domain to fix the timing violation [7]. However, though the formulated problem used uncertain constraints with probability distribution, the proposed adaptive algorithm does not explicitly prioritize the constraints with larger probability (i.e., constraints that are more likely to occur). In addition, the power optimization was achieved with power penalty heuristic, and the integer constraints in the formulation were conquered by the binary attractor heuristic. These heuristics lack theoretical justification and have no guaranteed convergence rate. All these issues will significantly degrade the performance of the algorithm.

In this paper, we notice the interesting fact that the problem of selective body biasing with single biasing voltage domain can be formulated alternatively as a linear semi-infinite programming (LSIP) problem. In addition, the structure of the problem, associated with the physical meaning of the design, can lead to a novel Incremental Hypercubic Sampling (IHCS) algorithm. The algorithm solves the LSIP problem through a number of finite mixed-integer linear programming. Finally, we are able to demonstrate many nice properties of the algorithm through rigid mathematical derivations. Experimental results on industrial designs using 65 nm sub-threshold library demonstrate that, compared with the state-of-art adaptive filtering approach in [7], our proposed approach can improve the pass rate by up to 7.3 × with a speed up to 4.1 ×, using the same number of body biasing gates with about the same power consumption.

The remainder of this paper is organized as follows: Section 2 reviews the previous works, and Section 3 gives the new LSIP problem formulation. The proposed IHCS algorithm is presented in Section 4. Section 5 demonstrates the experimental results and concluding remarks are given in Section 6.

2. Literature review

In this section, we briefly review the literature aiming at addressing timing issues in sub-threshold designs.

Zhai et al. [5] showed that relative delay variation reduces as size of the transistors or logic depth increases. Based on the statistical delay and energy model of sub-threshold designs with process variation, they tried to minimize the energy by choosing optimal pipelined depth. However, increasing the sizes of transistors will induce area overhead, and large logic depth will limit the speed of design. Furthermore, the timing analysis is still based on the existence of critical paths, and for sub-threshold design where up to 90% of the paths can become critical, this approach is less practical.

Liu et al. [6] studied the possibility of body biasing in sub-threshold designs, and proposed to use a fuzzy logic controller to decide the optimal biasing voltage at runtime based on the measured performance. However, it lacks the critical component of body biasing gate selection. Apparently, the number of gates used to perform body biasing affects both the power and routing overhead drastically, and should thus be minimized. This problem has been adequately addressed in super-threshold designs [8–10] where the dominant L_{eff} variations exhibit strong spatial correlations. But in sub-threshold designs, it is very hard as the dominant V_{th} variations are purely random.

Most works on body biasing using multiple body biasing voltages applied to all gates in the design [6,8–10], with an objective to minimize the power consumption. However, the existence of multiple body biasing voltage domains and the need to make the body of every gate biasable introduce the complexity of control logic as well as the large area overhead. To alleviate the overhead, Geng et al. [7] explored a scheme that selects only a small portion of gates with one body biasing voltage to fix the timing violations. The problem is formulated as a very interesting integer programming model with statistical linear inequality constraints, and an adaptive affine projection algorithm with binary attractors and power penalty (APA-BA-PP) was proposed to solve it. Considering that the problem formulation in this paper is improved based on [7], its formulation will be reviewed briefly below.

For a combinational logic with M paths and N gates, the problem of selected body biasing is to choose the optimal gate set, which can fix the timing violation at given probability with minimal power overhead. It can be formulated as the following optimization problem:

$$\begin{aligned} &\textbf{min} && (\mathbf{p}_b-\mathbf{p})\mathbf{x} \\ &\text{subject to} && \text{prob}\{\mathbf{A}_{tt'}\mathbf{x}\geq\mathbf{b}\}>1-\zeta. \end{aligned} \tag{1}$$

in which ζ is a small positive number. The $\mathbf{p}$ and $\mathbf{p}_b$ are two $1\times N$ vectors, in which each element represents n-th gate's average power p_n without body biasing, and average power $p_{b,n}$ with body biasing. The decision vector $\mathbf{x}$ is a binary $N\times 1$ vector, in which each element represents the decision of body biasing control on that gate, i.e., 1 means applying body biasing to that gate and 0 means no body biasing.

$\mathbf{A}_{tt'}$ in (1) is the $M\times N$ delay improvement matrix, which is the delay improvement under process variation when body biasing is applied to the design. The detailed definition of $\mathbf{A}_{tt'}$ will be given later in Section III. Considering $\mathbf{x}$ is the decision vector, $\mathbf{A}_{tt'}\mathbf{x}$ is the overall delay improvement for paths. Meanwhile, the vector $\mathbf{b}$ $(M\times 1)$ is the required delay improvement vector under process variation.

Problem (1) is very hard to solve directly if not impossible. Accordingly, the adaptive filtering approach (APA-BA-PP) was proposed to solve it heuristically [7]. The constraint is not explicitly considered, but rather translated into a penalty term added in the objective. The algorithm has two inherent drawbacks that degrade its efficacy. First, without explicitly considering the constraint, it is never possible to predict the solution quality. For example, the algorithm does not explicitly prioritize the scenarios with larger probability (i.e., $\mathbf{V}_{th}$ variations that are more likely to occur). Second, the power optimization is achieved with power