SIGNAL-PATH LEVEL ASSIGNMENT FOR DUAL-V_t TECHNIQUE

Yu Wang, Huazhong Yang, Hui Wang

Circuits and Systems Laboratory, Department of Electronic Engineering,
Tsinghua University, 100084, Beijing, People's Republic of China
E-mail: wangyuu99@mails.tsinghua.edu.cn

ABSTRACT

Along with the fast development of dual threshold voltage (dual-V_t) technology, it is possible to use it to reduce static power in low-voltage high-performance circuits. In this paper we present a new signal-path level circuit model and an algorithm based on the new circuit model which introduces the concept of extracting sub-circuits. Experimental results show that, for the ISCAS85 benchmark circuits, our algorithm produces a significant leakage-power reduction similar to the transistor level dual-V_t assignment, but the computational cost is comparative to gate level dual-V_t assignment.

1. INTRODUCTION

With the development of the fabrication technology, leakage power dissipation has become comparable to switching power dissipation [1]. At the 90nm technology node, leakage power may make up 42% of total power [2]. Inevitably, techniques are necessary for reducing the increasing leakage power. These leakage control methods can be broadly categorized into two main categories: process level and circuit level techniques. At the process level, leakage reduction can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profile in transistors. There are also four major circuit design techniques, namely, transistor stacking, supply voltage scaling, dynamic V_t and multiple V_t for leakage reduction in digital circuits. The two main methods in multiple V_t technique are sleep transistor insertion and dual V_t CMOS assignment. The extra area and delay due to the insertion of sleep transistors have considerable influence on the circuit performance. Furthermore, with the supply voltage scaling, it is becoming harder to turn on under a very low supply voltage. A higher V_t can be assigned to some of the transistors in the non-critical paths and a lower V_t is assigned to other transistors in the dual V_t CMOS assignment for a logic circuit in order to achieve both high performance and low power simultaneously. Recently, a dual-V_t MOSFET process was developed [3], which makes the implementation of dual-V_t logic circuits more feasible. Dual-V_t method results in a significant reduction in total power dissipation and energy. Therefore, determining which gate should be the high V_t has already become a major emphasis in the research field. Traditionally, gate level dual-V_t assignment [4][5][6][7] suppresses less leakage power than transistor level dual-V_t assignment [8] while saving much more computation time.

In this paper, we assume that all the gates are using the low threshold voltage in order to get the best performance (timing characteristic). The new signal-path level circuit model we used is different from the circuit models which consider a gate as a vertex or an edge in a graph. This is to make our algorithms useful for transistor level leakage control. We use look up table method in static timing analysis to get the critical paths and non-critical paths of the circuit much faster and with more accuracy. The gates in the critical paths will remain unchanged to maintain the performance; and the gates in the non-critical paths are extracted into several sub-circuits. Without reiterating the whole circuit, we can focus solely on dealing with the sub-circuits in which we use a new developed hierarchy based algorithm to get an optimal result faster. Our new signal-path level dual V_t assignment aims to have more leakage power saving and a similar or even less computation complexity than gate level dual-V_t assignment.

2. DUAL V_t ASSIGNMENT

2.1 Gate level circuit model

A combinational circuit is represented by a directed acyclic graph (DAG) $G = (V, E)$. Traditionally a vertex $v \in V$ represents a CMOS transistor network which realizes a single output logic function (a logic gate), while an edge $(i,j) \in E$, $i,j \in V$ represents a connection from vertex i to vertex j. In this way, the transistors within a vertex that are driven by the same logic signal will be assigned to the same threshold.

The assignment of threshold voltages to the transistors in the circuit can be represented as one of assigning a threshold voltage to a vertex $v \in V$ [3][4][5], or one of assigning a threshold voltage to an edge $(i,j) \in E$ [6]. Thus this allows us to treat the dual-V_t optimization problem as a kind of graph problem. It greatly simplifies delay analysis and standby power estimation during V_t assignment. The effects on delay when a V_t change is made can be easily modeled by static timing analysis (STA) methods. In Figure 1, a combinational circuit is presented at the left side; the traditional circuit model is at the right side.

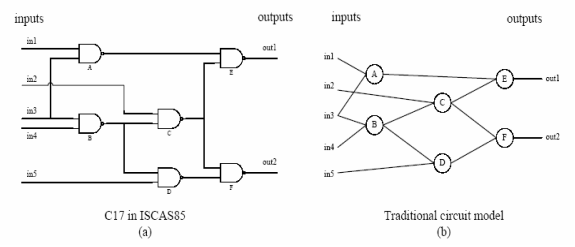

Figure 1. Original circuit C17 in ISCAS85 and traditional gate level circuit model.

2.2 Signal-path level circuit model

In our signal-path level circuit model, a vertex $v \in V$ represents a pin of logic gates or a primary input/output; an edge $(i,j) \in E$ represents a connection from vertex i to vertex j. In our model, an edge is the abstraction of a wire connecting two pins of two different gates or a *signal-path* in a logic gate from one of its input pin to an output pin. Furthermore, we have added a virtual input vertex and a virtual output vertex to our model. The virtual input vertex is connected to all the primary inputs and the virtual output vertex is connected to all the primary outputs. The fan-in of a logic gate's input pin refers to the number of pins which connect with this input pin. The fan-out of a logic gate's output pin refers to the number of pins which is connected with this output pin. Furthermore, vertexes which have a fan-in of zero constitute primary inputs; similarly vertexes which have a fan-out of zero constitute primary outputs. Figure 2 shows our signal-path circuit model of circuit C17 from ISCAS85 bench mark. If vertex $i \in V$ represents one of gate A's input pins and vertex $j \in V$ represents gate A's output pin, we define edge $(i,j) \in E$ as a "signal-path" and this signal-path belongs to gate A. And actually each signal-path consists of one of the parallel transistors and the transistors in series in a simple classical CMOS gate.

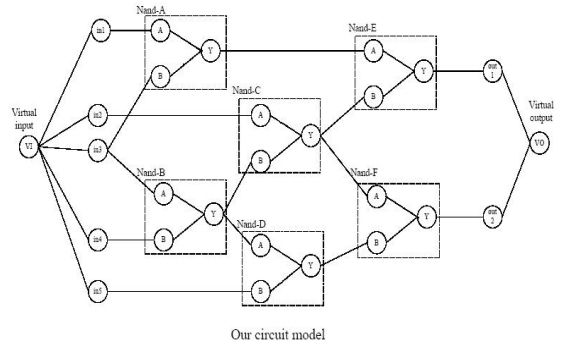

Figure 2. Signal-path level circuit model (C17).

There are several reasons for using this new circuit model. Firstly, the signal arrival time may be different for every input pin of a gate. More detailed delay information for every gate is presented since the delay information for every pin of the gate is computed by STA. Secondly, through the definition of signal-path, it is possible to have transistors with different V_t in a single gate at the same time, which means transistors in one gate may have different V_t. If we neglect the possibility of assigning different threshold voltage to signal-paths which belong to the same gate, it will get the same solution as previous methods [4][5]. Finally it leads to a much less computation complexity than transistor level assignment [8], yet achieving a similar accuracy.

The edge E in the graph represents two kinds of connections. One is "signal-path", the other is the connection of two pins belonging to different gates respectively which represents a wire between two pins in most cases. Hence it is possible to consider the interconnect delay during STA in order to get more accurate model of the circuit.

2.3 Delay model

We create our signal-path delay model and get arrival time, require time and slack time information for each vertex and signal-path in the signal-path level circuit model.

In order to get the delay attributes, we levelize the vertexes in the graph, making sure every two vertexes belongs to the same level have no edges between them. Each pin's fan-ins are not at the same level as itself, its fan-outs are not either; thus an edge $(i,j) \in E$'s two vertexes $i,j \in V$ are not at the same level. The delay of an edge $(i,j) \in E$, $i,j \in V$ is denoted by $d_{i,j}$.

We define three attributes for every vertex $v \in V$, they are namely, the arrival time $t_a(v)$, the required time $t_{req}(v)$, and the slack time $t_{slk}(v)$. The arrival time $t_a(v)$ is the worst case of delay from the primary inputs to pin v. $t_{req}(v)$ is the latest time the signal needs to arrive at pin v. We define them as:

$$t_a(v) \equiv \begin{cases} given\_time\_of\_arrival, & \text{if } v \text{ is the vitual input} \\ \max\limits_{i \in fan\_in(v)} \{t_a(i) + d_{i,v}\}, & \text{otherwise} \end{cases}$$

$$t_{req}(v) \equiv \begin{cases} t_a(v), & \text{if } v \text{ is the virtual output} \\ \min\limits_{i \in fan\_out(v)} \{t_{req}(i) - d_{v,i}\} \end{cases}$$

By comparison to traditional circuit model, the arrival time of a gate is the maximum of its input pins' arrival time, and the required time of a gate is its output pin's required time (if the gate is a CMOS transistor network which realizes a single output logic function). The slack time of a gate is also defined as the difference of its arrival time and required time. The *critical path* of the circuits is constituted by the set of gates that has the minimum slack time value.

We define every edge $(i,j) \in E$, $i,j \in V$ in the graph G also has the attribute $s_{i,j}$ which represents the slack time of the edge:

$$s_{i,j} \equiv t_{req}(j) - t_a(i) - d_{i,j}$$

Finally, the slack time of a vertex v∈V is defined as the minimum slack time of its fan-in edges:

$$t_s(v) = \min_{i \in fan\_in(v)} s_{i,v}$$

In our delay model we define the *critical path* of the circuits as the set of the edges that has the minimum slack time value. If there is no negative slack in the circuit, then timing constraints are satisfied [9].

Notice that every signal-path in the same gate can have different delay difference when its V_t changes; and when several signal-paths can be simultaneously changed in one gate, the delay difference is even more complicated because of the infections between the changed signal-paths. Here we select the largest delay difference of all the signal-paths' change schemes as the reference delay difference of the signal-path in this kind of gate. The signal-path delay data are then derived from the look up table of standard cells and HSPICE simulation.

2.4 Leakage power model

We find out that the leakage power change due to only one signal path's change is always the same and furthermore, if there are k signal-paths which can change their threshold voltage in a gate with w signal paths ($k< w$), no matter how to choose the k signal paths, the power change due to k signal paths' threshold voltage change is always the same. The leakage power saving due to k signal paths' threshold voltage change is nearly the same as k times the leakage power saving due to only one signal path's change. However, if all the w signal paths in the gate is changed, the leakage power saving is larger than w times the leakage power saving due to only one signal path's change. Finally we use two values to represent each signal path's leakage power attributes: the larger one is for all the signal paths in that gate can change into high threshold voltage, and it equals to the leakage power saving due to the gate's threshold voltage change divided by the number of signal-paths in the gate; and the smaller one is for other conditions which equals the leakage power saving due to only one signal path's change. Using HSPICE and a typical library for each circuit scheme of the signal-path, we can create a table of leakage power for the signal-path's threshold voltage change.

We do not consider the signal probability at each pin of the gates, and we may use logic simulation or local probability propagation in our future work to make it possible to combine transistor stacking effects with the circuit analysis to get a more accurate leakage power estimation table.

2.5 Algorithm

The dual-V_t optimization is defined as a problem to assign one of two threshold voltages to each signal-path to get most leakage power reduction with no performance reduction.

In our algorithm, we assumed the DAG representation $G(V,E)$ of a signal-path level combinational circuit. The graph is firstly levelized to indicate the depth of the vertexes and signal-paths in the graph. We initialize the circuit by assigning a low threshold voltage (V_{THlow}) to all the signal-paths of the circuit, i.e. it essentially configures the circuit to have the minimum delay. In the initialization procedure, we decide the delay attributes of every vertex and edge in the graph: the arrival time $t_a(v)$, the required time $t_{req}(v)$and the slack time $t_{slk}(v)$, the edge slack time $s_{i,j}$, the edge propagation delay $d_{i,j}$. All these attributes can be calculated using static timing analysis and the formula we have denoted before. The fan-ins of a vertex are the former level vertexes which are connected with this vertex; the fan-outs of a vertex are the next level vertexes which are connected with this vertex. Since every edge has a slack time, we extract all the non-zero slack time edges to construct a set of sub-graphs G_{sub1}, G_{sub2}, …, G_{subn}. The critical paths' delay attributes are not affected when the V_t of some signal-paths on non-critical paths are changed.

Therefore the assignment of dual V_t in the whole circuit is decomposed into several small problems, which have much smaller solution space and thus are more easier to get the optimal assignment of V_t. Without reiterating the whole circuit, we can focus solely on dealing with the sub-circuits in which we use a new developed hierarchy priority based algorithm to get an optimal result faster.

If we do not consider the condition that signal-paths in the same gate can have different threshold voltages, we can get the solution for gate-level dual-V_t optimization. Therefore, during the sub-graph extraction, we will only consider the gates in which all the signal-paths' slack times are positive. It could be easily realized by mapping a whole gate to a single vertex in the graph. The arrival time of the gate is the maximum of the arrival times of the gate's input pins. The required time of the gate is the output pin's required time. The slack time of the gate is the difference between the arrival time and the required time of the gate. Through a little change in the algorithm, we can get the gate level optimization of the circuits.

3. IMPLEMENTATION AND RESULTS

The assignment algorithm has been implemented in C++ under signal-path level static timing analysis environment. The value of various transistor parameters have been taken from the TSMC library, the effect channel length is 0.13μm and the gate oxide thickness is 2.4nm. The circuit temperature is assumed to be 110℃. The leakage power table and delay look up table is created by HSPICE simulation. In our analysis, the low threshold voltage and the supply voltage of the original circuits are assumed to be 0.2V and 1.2V, and high threshold voltage during the dual V_t optimization is assumed to be 0.3V.

In Figure 3, the signal-paths which are labeled out can be changed into high threshold voltage; meanwhile only NAND_A can be changed by gate level assignment. The leakage power saving of C17 is respectively 16.3% and 28.7% for gate level and signal-path level optimization.

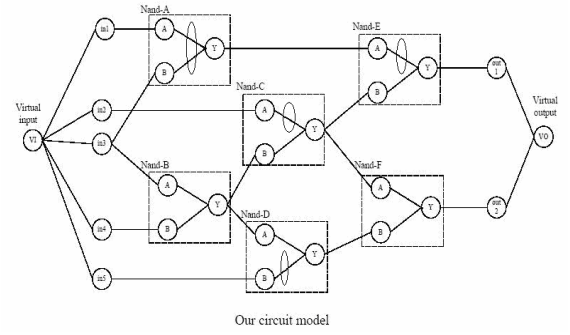

Figure 3. Changeable signal-paths in C17.

Table. 1. shows that the gate level and signal-path level optimization leads to different results for leakage power saving of ISCAS85 circuits, and obviously more leakage reduction can be achieved through signal-path level optimization since there are actually more transistors in the implementation of the circuit which can be assigned to high threshold voltage.

Table 1. Comparison of leakage power reduction.

ISCAS85 Benchmark Circuits	Gate level Reduction (%)	Signal-path level Reduction (%)
C432	34.5	43.3
C499	29.2	53.1
C880	71.8	77.4
C1355	45.2	68.1
C1908	55.4	66.8
C2670	77.3	85.5
C3540	81.9	89.7
C5315	69.6	78.5
C6288	40.3	56.5
C7522	69.2	75.4
Average	57.44	69.43

4. CONCLUSION

In this paper we have proposed a new circuit model for combinational circuit during dual-V_t assignment. The assignment algorithm is sped up by the proper extraction of sub graphs. By using a delay look-up table and a leakage power table generated by HSPICE simulation, we find that approximately 12% more leakage power savings can be achieved under the signal-path level optimization than the gate-level optimization.

5. REFERENCES

[1] G. Moore, "No exponential is forever: But forever can be delayed," *IEEE ISSCC Dig. Tech. Papers*, 2003, pp. 20 - 23.

[2] Kao J., Narendra S., Chandrakasan A., "Subthreshold Leakage modeling and reduction techniques", *ICCAD*, 2002, pp 141 - 149

[3] Z. Chen et. al., "0.18 um dual Vt MOSFET process and energy-delay measurement," in *IEDM Dig.*, 1996, p. 851.

[4] L. Wei, Z. Chen, and K. Roy, "Design and Optimization of Dual Threshold Circuits for Low Voltage, Low Power Applications", *IEEE Transaction on VLSI Systems*, Vol.2. 17, NO. 1, 1999, pp. 16-24.

[5] Vijay Sundararajan, and Keshab K. Parhi, "LOW Power Synthesis of Dual Threshold Voltage CMOS VLSI Circuits", *Proc. ISLPED*, 1999, pp. 363-368.

[6] Nikhil Tripathi, Amit Bhosle, Debasis Samanta, and Ajit Pal; "Optimal Assignment of High Threshold Voltage for Synthesizing Dual Threshold CMOS Circuits", *VLSI Design, 2001. Fourteenth International Conference on* ,3-7 Jan. 2001 pp. 227 - 232.

[7] Qi Wang; Vrudhula, S.B.K.; "Algorithms for minimizing standby power in deep submicrometer, dual-Vt CMOS circuits", *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* ,Volume: 21 ,Issue: 3 ,March 2002 pp. 306 - 318 .

[8] Liqiong Wei; Zhanping Chen; Roy, K.; Yibin Ye; De, V.; "Mixed-Vth (MVT) CMOS circuit design methodology for low power applications", *Design Automation Conference*, 1999. Proceedings. 36th ,21-25 June 1999 pp. 430 - 435 .

[9] S. Devadas, A. Ghosh, and K. Keutzer, Logic Synthesis. New York: McGraw-Hill, 1994.