

國立交通大學

電信工程學系

碩士論文

考慮溫度效應的三維積體電路
功率最佳化方法 – 雙電壓分配法

Power Optimization Method in 3D ICs
Considering Thermal Effects – Dual Supply
Voltage Assignment

研究生：魏書含

指導教授：李育民 教授

中華民國 九十九 年 八 月

考慮溫度效應的三維積體電路功率最佳化方法
- 雙電壓分配法
Power Optimization Method in 3D ICs
Considering Thermal Effects – Dual Supply Voltage Assignment

研究生：魏書含

Student : Shu-Han Whi

指導教授：李育民

Advisor : Yu-Min Lee

國立交通大學
電信工程學系
碩士論文



Submitted to Department of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Communication Engineering

Aug 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

考慮溫度效應的三維積體電路功率最佳化方法 - 雙電壓分配法

學生：魏書含

指導教授：李育民 博士

國立交通大學電信工程學系碩士班

摘 要



三維積體電路被視為一個有效解決二維積體電路上過長導線造成的問題之技術進而改善晶片效能。然而過高的溫度將成為三維積體電路嚴峻的挑戰並且減弱三維積體電路低功率特性的優點。因此，同時考慮溫度效應與功率最佳化是非常重要的。在這篇論文中，我們提出一個利用雙電壓源的方法來降低三維積體電路上的總功率消耗。此降低三維積體電路總功率損耗的方法：1) 所提出的雙電壓源法同時考慮三項因子作為電壓分配的標準，包含了功率延遲敏感度(power-delay sensitivity)、群聚效應和電壓轉換器的預算；2) 利用三維度電熱模擬分析器得到此三維積體電路的溫度；3) 採用具溫度相關性之邏輯閘延遲(gate delay)模型，並且完成一套考慮溫度的時序分析。實驗的結果驗證了我們方法的有效性，並且指出在電路分析中考慮熱效應(thermal effect)是極重要的。

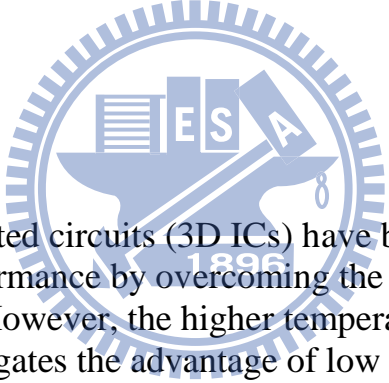
Power Optimization Method in 3D ICs Considering Thermal Effects – Dual Supply Voltage Assignment

Student: Shu-Han Whi

Advisor: Dr. Yu-Min Lee

Department of Communication Engineering
National Chiao Tung University

ABSTRACT



The three dimensional integrated circuits (3D ICs) have been viewed as an effective method to improve chip performance by overcoming the bottleneck of long interconnects in the 2D ICs. However, the higher temperature becomes a serious challenge for 3D ICs and mitigates the advantage of low power. Therefore, it is important to propose an effective method considering thermal effect and power optimization simultaneously. In this thesis, we present a methodology to minimize the total power consumption in the 3D ICs by employing a dual supply voltage technique. The proposed approach consider three main headings: 1) a voltage assignment process consider three main factor, which consists of sensitivity-based, proximity effect and level shifter budget factor, to be the voltage assignment criterion for power reduction; 2) a 3D electro-thermal simulation to get the temperature of chip; 3) a thermal aware static timing analysis to obtain the thermal-related delay of gate in the circuit. The experimental results demonstrate the effectiveness of our voltage assignment method and the thermal effect in circuit performance.

誌 謝

這篇論文能夠順利地完成，首先要感謝我的指導教授 李育民博士，當我遭遇困難時，老師總會適時指引我方向，讓我能夠繼續前進。兩年的學習過程中，老師的訓練和指導給了我莫大的幫助，讓我確實感受到自己的成長，我相信這對我未來不管是工作或進修都會有相當大的幫助。

研究的路上，特別感謝博士班學長培育。在研究方向給予指導，完成研究的過程更是時常與我一起討論實驗方法、程式等，給予我很多寶貴意見，這些對於我的研究有著莫大的幫助。還有感謝這兩年一同在實驗室打拼的夥伴們，學長柏毅、懷中、宗祐，同窗亭蓉、巧翎、麒文，學弟志升，你們給予的關心和幫助，豐富了我碩士班的生活。

我將最誠摯的感謝獻給我的家人、好友宜臻、冠群，因為有你們的支持與關懷，陪伴我度過研究過程的低潮期，給我更大的動力繼續向前進，謝謝你們。

最後，我將這份喜悅與快樂獻給所有關心我的人，並希望閱讀這份論文的讀者，能給予指教，謝謝。

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Our Contributions	2
1.3	Organization of the Thesis	2
2	Preliminaries	3
2.1	Three Dimensional Integrated Circuits (3D ICs)	3
2.1.1	Motivation of 3D ICs	3
2.1.2	Benefits of 3D ICs	3
2.1.3	Challenges of 3D ICs	4
2.2	Timing Analysis	6
2.2.1	Static Timing Analysis (STA)	6
2.3	Current Power Optimization in 2D ICs	8
2.3.1	Power Optimization	8
2.3.2	Multiple Supply Voltage (MSV)	9
3	Power Optimization in 3D ICs	12
3.1	Problem Formulation and Flowchart	12
3.2	3D Thermal Analysis	14
3.3	Thermal Aware Static Timing Analysis	14
3.4	Initial Voltage Assignment	15
3.5	3D ICs Voltage Assignment	16
3.5.1	Grid-Based Procedure	16
3.5.2	Voltage Re-Assignment	19
3.5.3	Incremental Update	20
3.6	Rescue Timing and LS Budget	21
3.6.1	Timing Rescue	21
3.6.2	LS Budget Rescue	22
4	Experimental Results	23
5	Conclusion	27

List of Figures

2.1	The schematic diagram of 3D ICs with three chip layers which using TSV to connect each layer	5
2.2	3D ICs chip schematic drawing [11]	5
2.3	An example illustrating block-based STA.	6
2.4	Schematic of Multiple-Supply Voltage Method [B. Sarker, cadence 2005]	9
2.5	Illustration of the static current flow in a VDDH gate when it is directly connected to a VDDL gate. [B. Sarker, cadence 2005]	10
2.6	Schematic of Conventional level Shifter [B. Sarker, cadence 2005]	10
3.1	Flowchart of the proposed power optimization for 3D ICs	12
3.2	Flowchart of initial voltage assignment of proposed power reduction method . .	15
3.3	Algorithm of 3D ICs Voltage Assignment.	16
3.4	A three-tier design example of the grid-based procedure for generating the voltage assignment. Each tier is first divided into many grids, and the grid with the high sensitivity has dark color. After compressing, the criteria are accumulated, the higher priority the grid has, and the darker the color is. When the grid with the highest priority is found, it is restored to the multi-layer structure to decide which tier should operate at the low supply voltage.	17
3.5	Schematic of Timing Rescue	21
3.6	Schematic of LS Budget Rescue	22
4.1	(a) Voltage assignment result on layer 1 of s35932. (b) Voltage assignment result on layer 2 of s35932.	26
4.2	(a) Voltage islands on layer 1 of s38584. (b) Voltage islands on layer 2 of s38584.	26

List of Tables

4.1	Our Proposed Voltage Assignment Method Result	24
4.2	Optimization Result	25



Chapter 1

Introduction

1.1 Introduction

Because of the drastically increasing both wire lengths and interconnect density in 2D ICs, the delay and the power consumption of circuit have become one of the most important concerns in VLSI design. In recent years many researchers have shown that 3D ICs is an effective method to solve these problems. The 3D ICs technology not only reduces the wire lengths but also improves the system integration. However, the heat removal is a great challenge in the 3D ICs design due to the higher power density and the low thermal conductivity inter-layer dielectrics. Moreover, the high temperature has serious impacts on the timing, power and reliability of designs [14]. Therefore, it is necessary to reduce the power consumption of circuit for solving the thermal problem of circuit.

Generally, the power consumption in CMOS digital circuits consists of dynamic power and static power caused by leakage current. The dynamic power consists of switching power caused by charging and discharging capacitive loads and short-circuits power. In practical, the short circuit power is insignificant compared to the other two. In this thesis, we focus on the power minimization for the dynamic power (switching power) and the leakage power by considering power and temperature simultaneously. The switching power consumption in CMOS circuits is proportional to the square of the supply voltage (V_{DD}). Among the existing dynamic power reduction techniques, MSV [19, 20, 3] is an effective method for scaling voltage to reduce dynamic power. Therefore, we apply dual supply voltage method in 3D ICs for power reduction. During scaling voltage process, if a low supply voltage gate drives a high supply voltage gate, a level shifter (LS) must be inserted to eliminate the undesirable static current. The additional

LS would increase the cost in area, delay and power; hence, LS overhead must be considered in our work.

1.2 Our Contributions

Our major contributions of this thesis are summarized in following terms:

1. A post-placement MSV method in 3D ICs is developed for the power reduction by considering level shifter and thermal effects simultaneously. Compared with the previous works which ignore level shifter issue and use the thermal-unrelated models, the proposed method is more flexible and practical.
2. In work [16], they assume the temperature and the circuit design are independent and each location temperature was equal; however, the placement can affect the temperature distribution. Therefore, in this thesis, the temperature is regarded as a variable related to the circuit and can be added into the STA become a thermal aware STA for timing analysis.
3. The investigation of MSV techniques has been studied particularly in 2D ICs; however, fewer researches deal with MSV technique problem in 3D ICs design. Therefore, we propose a new voltage assignment method for 3D ICs design.

1.3 Organization of the Thesis

The rest of the thesis is organized as follows. In chapter 2, some important background, 3D IC technology, statistical timing analysis, and the multiple supply voltage technique, are illustrated. Moreover, the proposed power optimization methodology, experiment flowchart, 3D thermal analysis, thermal aware timing analysis and 3D ICs voltage assignment method, is presented in chapter 3. Then, the experimental results are given in chapter 4. Finally, some conclusions are drawn in chapter 5.

Chapter 2

Preliminaries

In this chapter, we first introduce the background of 3D ICS in section 2.1. The method of timing analysis of circuit and some power optimization studies in 2D ICs are surveyed in section 2.2 and section 2.3, respectively. Finally, a great power optimization method, MSV, is illustrated in section 2.3.2.

2.1 Three Dimensional Integrated Circuits (3D ICs)

2.1.1 Motivation of 3D ICs

The VLSI technologies have been rapidly scaled down to meet the more and more portable electronic production demand at less cost and power consumption. With the VLSI technology scaling down, it is reducing gate delays but rapidly increasing interconnect delays. Firstly, the long interconnect problem in 2D ICs will cause significantly parasitic effects which decelerate the circuit speed. Then, it also leads to higher power consumption and make the power management become more difficult than ever. The higher power consumption causes serious thermal effect. Besides, it also results in some problem like signal integrity and routing congestion. Moreover, increasing demand for the integration of disparate signals and technologies is introducing various system-on-chip (SoC) design concepts. However, existing planar IC (2D IC) design may not be suitable for above needs.

2.1.2 Benefits of 3D ICs

3D ICs have been comprehensively studied in order to come up with new approaches that could effectively address both miniaturization and integration demand for advanced and portable elec-

tronic products. 3D ICs technique is a desirable solution to long interconnect of 2D ICS by utilizing the shorter vertical interconnect (Through Silicon Via; TSV) to decrease the interconnect of the worst case wire length [1] as shown in Fig. ???. Meanwhile, the interconnect power and the chip area also can be reduce. The most important benefit of 3D ICs is that it can highly integrate different systems by vertically stacking and connecting various materials, technologies and functional components together [11], as illustrated in Figure 2.2. In this way, the total power consumption is reduced, packaging size, weight and cost are reduced and the packing density can be increased.

2.1.3 Challenges of 3D ICs

Although 3D ICs have been comprehensively studied, it is reviewed as a desirable solution to long interconnect of 2D ICs. However, 3D IC still has some challenge such as fabrication, production yield, heat removal and process variations. On the one hand, the vertical interconnect, TSV, is the main challenge of fabrication and production yield because building TSV is complicated in design. On the other hand, 3D ICs have the much higher temperature than 2D ICs due to the higher power density and the low thermal conductivity. Therefore, it extend more problem such as thermal management, reliable co-design and simulation tools, low-cost TSV structures and via fill processes needed to solved. In our work, we focus on the problem of power saving by thermal effect.

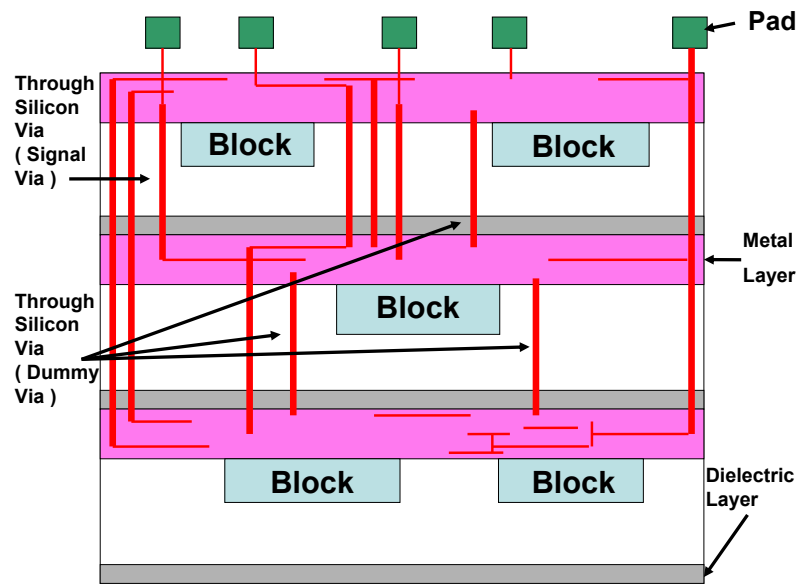


Fig. 2.1: The schematic diagram of 3D ICs with three chip layers which using TSV to connect each layer

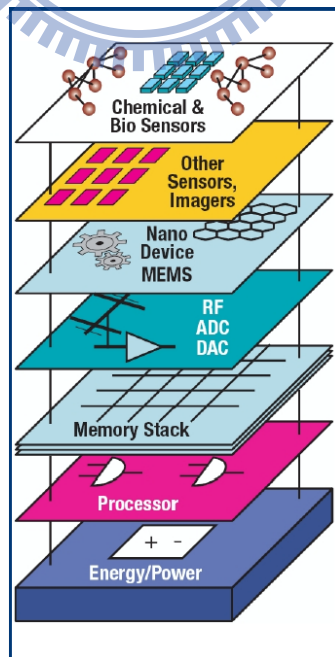
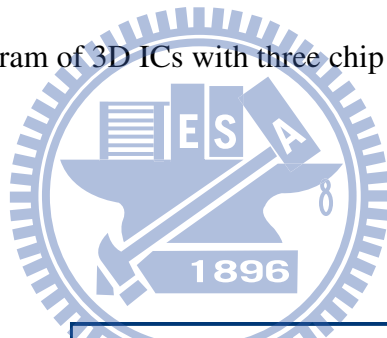


Fig. 2.2: 3D ICs chip schematic drawing [11]

2.2 Timing Analysis

2.2.1 Static Timing Analysis (STA)

STA provide a good solution to ensuring IC quality issues by computing the expected timing of a circuit without requiring simulation to judge whether the timing of circuit can work under user-defined timing constrain. The simple definition of STA is applying the specific timing model to a particular circuit by analyzing whether the timing of circuit violate the timing constrain of circuit. There are two analysis methods: path-based STA and block-based STA. In our work, we utilize the block-based STA to check the timing of circuit. The aim of block-based STA is to calculate the slack of each gate in the circuit. If the slack of any gate is negative then the circuit is violated.

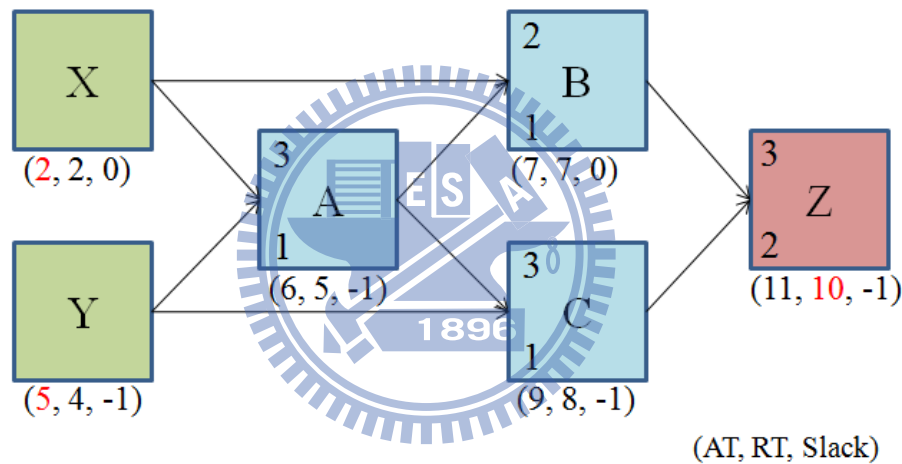
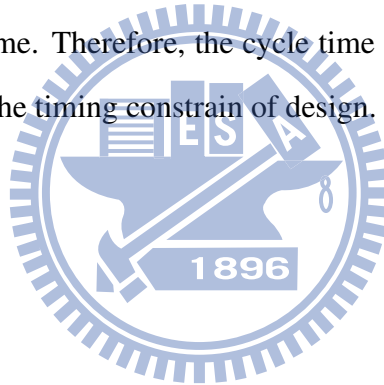


Fig. 2.3: An example illustrating block-based STA.

The simple concept of block-based STA is illustrated in Fig. 2.3. In block-based STA, a gate is called as a node/block which could be a logic gate or combinational block, and each block have three timing information for timing analysis. The timing information consists of arrival time (AT), required time (RT) and slack. AT is the real time that the signal arrives, RT is the time of user-defined timing constrain that when the signal should arrive and slack is the difference between AT and RT. In the following, we will describe the STA analysis process in detail using the example of Fig. 2.3.

The method is commonly referred to as PERT (Program Evaluation and Review Technique) which is widely used in STA [15]. In block-based STA, we only have AT of node X and Y

(primary input/flip-flop) and RT of node Z (primary output/flip-flop) timing information. Based on known AT information ($AT_X = 2, AT_Y = 5$), the AT of A is determined from X and Y . The AT from X is computed as $2+3=5$; while the AT from Y is computed as $5+1=6$. Because STA considers the worst case, the AT of A should choose the maximum of 5 and 6. Through a forward traversal, the arrival time of all blocks in the circuit are obtained by above step. Similarly, the RT is calculated by a backward traversal from the primary output Z . Based on known RT information ($RT_Z = 10$), the RT of B and C are computed as $10-3=7$ and $10-2=8$, respectively. Particularly, the RT of A is determined from B and C . The RT from B is computed as $7-1=6$; while the RT from C is computed as $8-3=5$. The RT of A should choose the minimum of 6 and 5 due to STA considers the worst case. After calculating AT and RT of each gate, by subtracting the AT from RT and we can get the slack of each gate. In this example, the critical path is the path $Y-A-C-Z$ which is defined as the path between an input and an output with the maximum arrival time. Therefore, the cycle time is obtained by the maximum arrival time which is referred to as the timing constrain of design.



2.3 Current Power Optimization in 2D ICs

Generally, the power consumption in CMOS digital circuits consists of three factors: dynamic power, short circuit power and leakage power. Usually, the short circuit power is insignificant compared to the other two factors. Therefore, in the following section, we will introduce some studies about the dynamic and the leakage power.

2.3.1 Power Optimization

Leakage Power

For the leakage power optimization, [21] works utilize gate-sizing and dual-threshold voltage to reduce the leakage power but most of them neglect the importance of dynamic power. Some studies consider both the dynamic and the leakage power reduction at the same time, by employing both multiple supply and multiple threshold voltage method [13]. An algorithm based on the linear programming is presented in [13]. The genetic algorithm is employed in [7]. In [13], they utilize the integer linear programming approach to formulate the problem.

Dynamic Power

For the dynamic power optimization, it can be computed by

$$P_{dynamic} = k \cdot f \cdot C_{load} \cdot V_{DD}^2 \quad (2.1)$$

where k is switching rate, f is clock frequency, C_{load} is the load capacitance, V_{DD} is supply voltage. Among the existing dynamic power reduction techniques, MSV is an effective method for scaling voltage to reduce dynamic power with the quadratic relation between the supply voltage and dynamic power. Many researches [19] have been studied this method comprehensively; nevertheless, the leakage power is ignored in their experiments.

2.3.2 Multiple Supply Voltage (MSV)

Multiple-Supply Voltage (MSV) is a popular and effective way to power reduction. The basic idea is that assign high voltage to cell on timing critical paths to maintain performance and assign low voltage to cell on timing non-critical paths to reduce power. As the Fig. ??, assign low voltage to the gate with "slow" label for power reduction due to it can work without operating at high voltage, another one gate with "fast" label must be operated at high voltage so it is assigned high voltage for maintain performance.

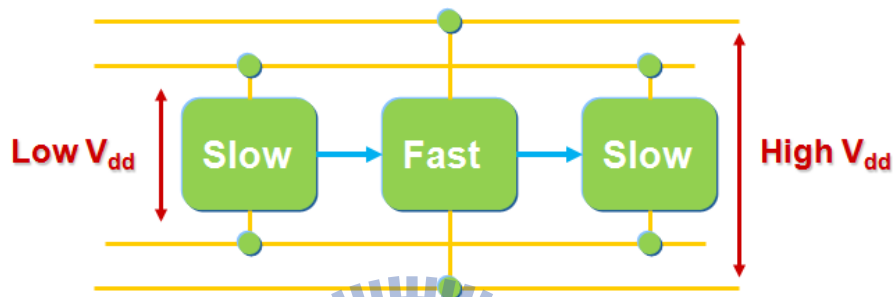


Fig. 2.4: Schematic of Multiple-Supply Voltage Method [B. Sarker, cadence 2005]

In MSV design, voltage assignment is a very important phase because an undesirable voltage assignment result makes the power-supply system complex and cause higher design cost such as more routing resource and heavy human intervention. Beside, another important issue must be concerned in voltage assignment is level shifter (LS). LS is an essential interface between low and high voltage for voltage conversion when a gate with lower supply voltage drives a gate with higher supply voltage. A example is shown in Fig. 2.5. In Fig. 2.5, the right inverter is operated at high voltage and this inverter is driven by a gate with low voltage. Because the voltage of the input signal of the inverter will not be higher than VDDL even when the input signal is at the HIGH level, the pMOS in this inverter may not be cut-off. In this way, a static current flow from drain to source through the pMOS to nMOS. Therefore, LS is needed between a low and a high voltage gate to avoid the phenomenon of a static current. However, the additional LS would increase the cost in area, delay and power; hence, the number of LS must be controlled. The Fig 2.6 shows an example of an LS.

Existing works [17] propose clustered voltage method to reduce LS overhead. Clustered voltage scaling (CVS) [17] limits the LS position only at sequential element output to reduce

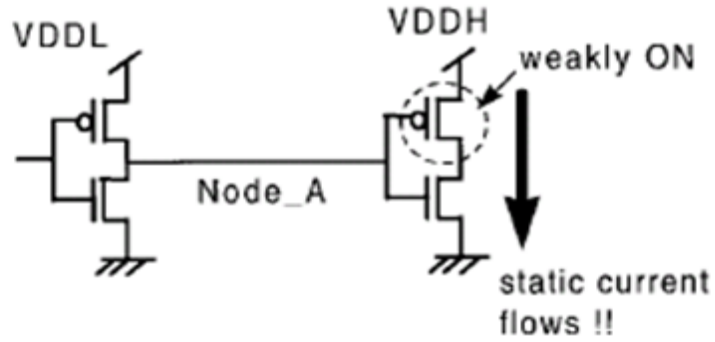


Fig. 2.5: Illustration of the static current flow in a VDDH gate when it is directly connected to a VDDL gate. [B. Sarker, cadence 2005]

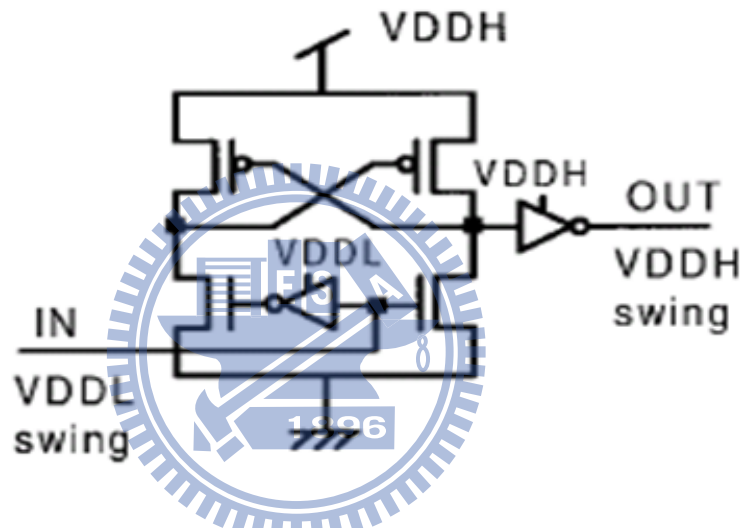


Fig. 2.6: Schematic of Conventional level Shifter [B. Sarker, cadence 2005]

the number of LS. Extended CVS (ECVS) [18] relaxes CVS topological constraint and LS is allowed anywhere. Thus, ECVS can provide more power saving than CVS but the delay penalty tends to be larger too. Another effective solution to LS overhead reduction: voltage island generation technique [19, 20], a group gates with same supply voltage is located on a contiguous space, can effectively reduce the number of LS and complexity of power-supply system.

Many previous works addressed on voltage assignment problem [20, 2, 23, 9]. Considering timing-constrained and voltage island aware voltage assignment algorithm is proposed in [20] present, but it does not consider LS cost such as area, delay and power. In [23], they construct a general formulation of voltage assignment problem, and integrate it into floorplanner phase. [9] provide a new voltage assignment algorithm based on ECVS to further improve the power

optimization. In [2], a two-phase voltage assignment algorithm on gate-level is proposed. Although [23] works consider the effect of LS into their voltage assignment procedure, the number of LS is limited to voltage assignment result. Moreover, those assignment methods only focus on power variation, but they ignore the temperature-related power consumption.

MSV has been studied in different design phase. In the physical design, MSV is considered at various stages, include during floorplanning [6]; post-floorplanning [12]; and post-placement [19, 3]. [12] propose voltage assignment and voltage island partition method to optimize power consumption and traditional floorplanning objectives such as chip area and wire length, but [6] is limited to core-based SOC design. The idea of post-placement voltage island generator [19] is utilized in our work.

Although MSV has been studied comprehensively for power reduction in 2D ICs, it still has fewer studies of MSV in 3D ICs so far. Intuitively, we can extend any MSV technique of 2D ICs to 3D ICs and utilize MSV technique to do voltage scaling considering tier by tier. However, without considering every tier in voltage assignment procedure simultaneously would make the power consumption distribution of each layer unbalance; hence, it would cause the thermal problem.

Based on [22], we apply this grid-based total power optimization approach for 3D ICs to our work and extend this idea to explore more improvement and more flexibility in voltage assignment method. Therefore, we propose a grid-based *3D IC Voltage Assignment* method considering each tier at the same time to deal with the interaction of voltage assignment with above mentioned problem.

In [10], they propose a two phase voltage scaling algorithm to minimize the total power, which is composed of a greedy voltage assignment phase and an iterative voltage re-assignment refinement phase. LS is determined by voltage assignment results, it would not effectively reduce LS overhead. Besides, the voltage assignment criterion is only determined by power difference between V_{DDH} and V_{DDL} , without considering temperature-related timing and leakage power. Therefore, our voltage assignment method considers the LS overhead before assigning voltage and our main voltage assignment criterion is determined by modifying [9] sensitivity for considering the above mentioned concerns at the same time.

Chapter 3

Power Optimization in 3D ICs

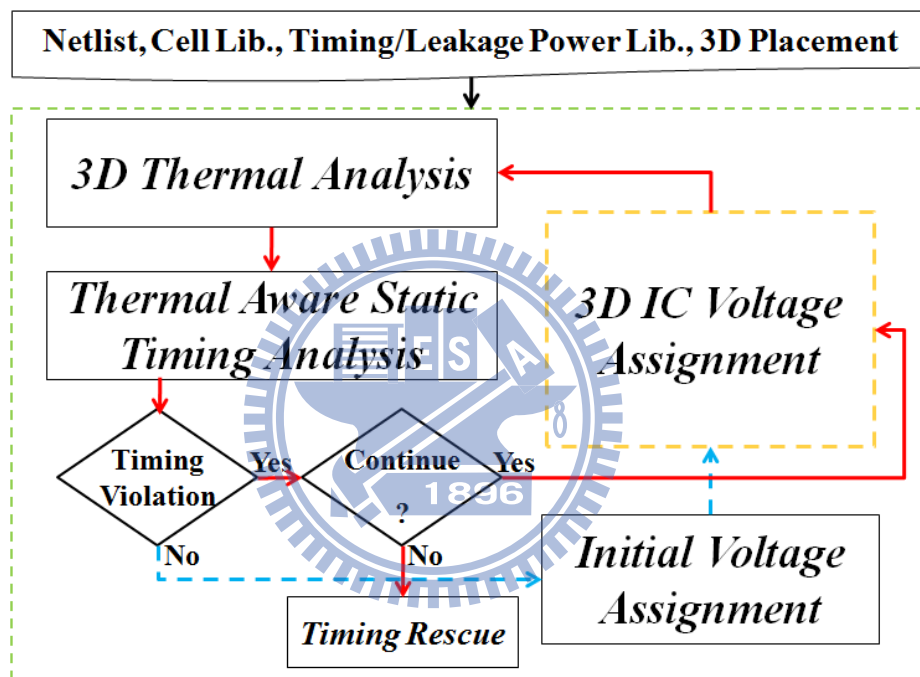
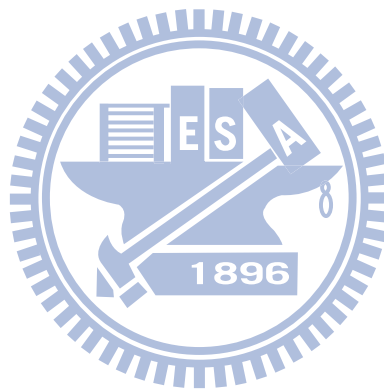


Fig. 3.1: Flowchart of the proposed power optimization for 3D ICs

3.1 Problem Formulation and Flowchart

The proposed a grid-based dual supply voltage assignment design flow for power reduction in 3D ICs is shown in Fig. 3.1. Given a known design placement of 3D ICs, design netlist, cell library and timing/leakage power cell library. For grid-based procedure, we partitioned each tier into n grids as the Fig. 3.4(a) and the number of grid is user definition. Firstly, the initial temperature of chip is obtained by *3D Thermal Analysis*. Then, a *Thermal Aware Static Timing Analysis* is performed with the temperature-related gate delay got from the initial temperature in *3D Thermal Analysis* step. After that, executing the *Initial Voltage Assignment* which makes the

circuit has timing violation due to aggressively assigning low V_{DD} to all gates, and then a grid-based procedure is developed for the *3D ICs Voltage Assignment* which assign an appropriate supply voltage for minimizing power saving penalty while enjoying delay gain. After executing voltage assignment procedure once, the power consumption and the delay of gates are changed; hence, the thermal and timing analysis should be done again to update the temperature-related gate delay and leakage power. After updating the temperature and timing, *3D ICs Voltage Assignment* is executed until no grid can be selected or timing violation is rescued. After voltage assignment, our voltage assignment can be applied to any voltage island generation in 2D ICs design. In the following, we will present the procedure of the proposed power optimization for 3D ICs in detail.



3.2 3D Thermal Analysis

Generally, the dynamic power is independent to temperature but the leakage power is significantly affected by temperature. Based on the empirical models [8, 4], the gate leakage current I_{gate} is related to the oxide thickness and the subthreshold current I_{sub} is related to the channel length and temperature. Fixing the oxide thickness and the channel length, and the subthreshold current model of gate can be built by utilizing the least square fitting method to the estimated results of HSPICE under the 90 nm technology as follow.

$$I_{sub} = s_0 \exp(s_1 T) \quad (3.1)$$

where s_i 's are fitting constants, L_{eff} is the effective channel length and T is the temperature of gate. Because the fitting constants and V_{DD} are dependent, we fit two pairs of coefficients for the high supply voltage (V_{DDH})/low supply voltage (V_{DDL}), respectively. A look-up table is set up to store them for V_{DDH} and V_{DDL} . The gate tunneling and the subthreshold leakage power are

$$P_{gate} = V_{DD} I_{gate}, \quad (3.2)$$

$$P_{sub} = V_{DD} I_{sub}. \quad (3.3)$$

Based on 3D ICs statistically thermal simulator [22], obtain 3D ICs deterministically thermal simulator, and combines the 3D ICs thermal simulator with the electro-thermal iterative updating loop.

3.3 Thermal Aware Static Timing Analysis

The thermal-aware STA is developed by the block-based STA considering thermal effect. The delay of gate is built as a canonical first-order form by applying the least square fitting method to the estimated results of HSPICE under 90nm technology. The expression of delay of gate is

$$Delay = (d_0 + d_1 T)(a_2 + a_3 C_{load}) \exp(a_4 C_{load}) \quad (3.4)$$

where a_i 's are V_{DD} dependent fitting constants and a look-up table is constructed to store those coefficients for V_{DDH} and V_{DDL} , T is temperature and C_{load} is load capacitance.

3.4 Initial Voltage Assignment

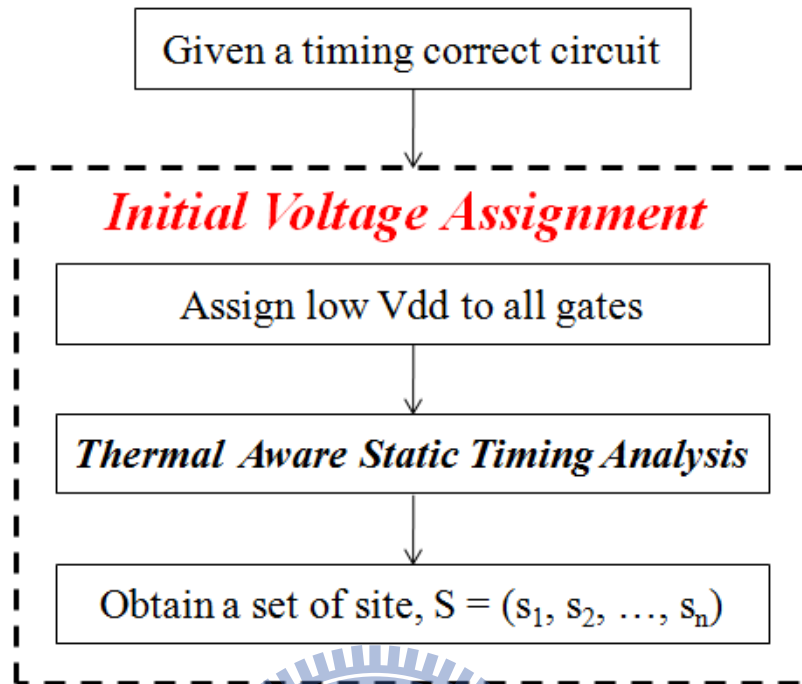


Fig. 3.2: Flowchart of initial voltage assignment of proposed power reduction method

As the flow of Fig 3.2, given a timing correct circuit and the initial supply voltage of all gates are V_{DDH} . Timing correct circuit means that it meets the timing constrain of the circuit. A *Initial Voltage Assignment* procedure is performed to produce best power saving but timing violation circuit. To begin with, all gates are assigned V_{DDL} . This step must cause some gates with negative slack. Then, a *Thermal Aware STA* step is performed to calculate the slack of each gate. The new temperature of gates are obtained by executing *3D Thermal Analysis* and using (3.4) to get the new delay of gate, and then new arrival and require time of each gate are obtained. After getting the new delay of gate, the slack data is calculated and the gate is referred to as site if the gate with negative slack is found. Finally, this step output is a set of site.

3.5 3D ICs Voltage Assignment

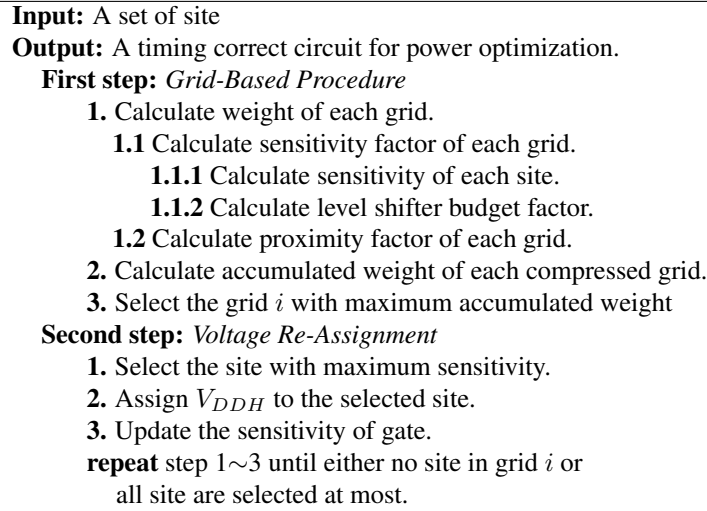


Fig. 3.3: Algorithm of 3D ICs Voltage Assignment.

The Algorithm of the proposed *3D ICs Voltage Assignment* method is shown in Fig 3.3. Given a set of site from *Initial Voltage Assignment*, firstly, the *Grid-Based Procedure* is executed to make a decision for deciding which grid is firstly picked, and then sites in this grid are assigned V_{DDH} to rescue timing under complex three dimensional structure. Then, given the selected, the *Voltage Re-Assignment* is performed by selecting some sites in the selected grid from *Grid-Based Procedure* to operate at high V_{DD} for timing rescue. Repeat these two steps until no site in the selected grid can be re-assign to operate at high V_{DD} or all sites in this grid have been rescued successfully. In the follow, we will introduce the *Grid-Based Procedure* and the *Voltage Re-Assignment* in section 3.5.1 and section 3.5.2 in detail, respectively.

3.5.1 Grid-Based Procedure

We improve the idea [22] to perform a grid-based procedure to make a decision for deciding which grid is firstly picked and then sites in this grid are assigned V_{DDH} to meet timing constrain under complex three dimensional structure. The aim is that the decision can obtain best benefit to timing rescue and power saving.

A three-tier design example of the grid-based procedure is illustrated in Fig 3.4. In the beginning, we consider the weight of each grid under three criteria: 1) the sensitivity factor, 2) proximity factor and 3) LS budget factor. Next, the three dimensional structure is vertically

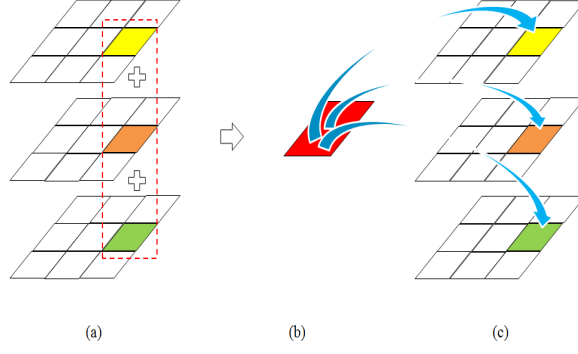


Fig. 3.4: A three-tier design example of the grid-based procedure for generating the voltage assignment. Each tier is first divided into many grids, and the grid with the high sensitivity has dark color. After compressing, the criteria are accumulated, the higher priority the grid has, and the darker the color is. When the grid with the highest priority is found, it is restored to the multi-layer structure to decide which tier should operate at the low supply voltage.

compressed into a two dimensional planar in the Fig 3.4(b), and the weight of each compressed grid is obtained by accumulating the weight of each grid of z-axis. After that, maximum accumulated weight of compressed grid is selected and this grid-based decision step is finished. Finally, as Fig. 3.4(c), the selected compressed grid is restored back to the original three layer structure. This procedure helps us to decide which grid is firstly assigned for next stage *Voltage Re-Assignment*.

Based on all sites are forced to operate at V_{DDH} assumption to compute the following voltage assignment decision weight W_i . The weight of grid i is defined as

$$W_i = c_1\alpha_i + c_2\beta_i \quad (3.5)$$

where α_i is the sensitivity factor and β_i is proximity factor. There are two main concerns for the α_i definition. The first is how to make a assignment decision can obtain the most rescue of timing and the least penalty of power saving. The second concern for the α_i definition is how to make a assignment decision can lead to fewer LS overhead like the number of LS. Therefore, we previously calculate the attainable white space in each grid to avoid available white space is not enough after really assigning supply voltage. The β_i factor consider clustering which we hope more gates in clusters in a single grid and this cluster operates at same V_{DD} . The sensitivity factor α_i and proximity factor β_i are defined as follows

$$\alpha_i = \frac{S_{grid_i}}{MAX\{S_{grid_i}, i = 1 \sim n\}} \lambda_i \quad (3.6)$$

$$\beta_i = \frac{N_i - N_i^H}{N_i} = \frac{N_i^L}{N_i} \quad (3.7)$$

where S_{grid_i} is the sum of site sensitivity S_{site} in grid i . N_i is the number of all gates in grid i , N_i^H is the number of gates with high V_{DDH} and N_i^L is the number of gates with low V_{DDL} . Once operating voltage of all sites in the selected grid are determined, the needed number of LS is determined by the relation between V_{DDL} gate and V_{DDH} gate. However, the LS budget may be not enough after assigning V_{DDH} to all sites in grid i . Where λ_i is a LS budget factor, it indicated that whether the estimated S_{grid_i} is too optimistic. The definition of λ_i is

$$\lambda_i = \min\left\{1, \frac{N_{LS}^a}{N_{LS}^n}\right\} \quad (3.8)$$

where N_{LS}^n is the needed number of LS if all sites are assigned V_{DDH} , N_{LS}^a is the available number of LS which is estimated by the attainable white space in each grid. Finally, the grid and the site sensitivity are defined as

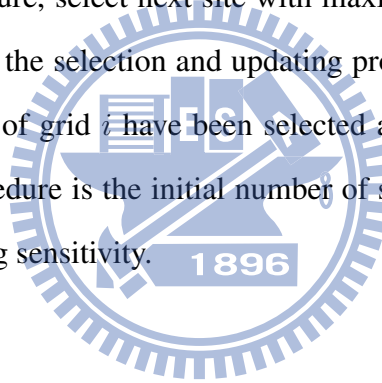
$$S_{grid_i} = \sum_{site_j \in grid_i} S_{site_j} \quad (3.9)$$

$$S_{site_j} = \frac{\Delta D}{\Delta P} |Slack_{site_j}|. \quad (3.10)$$

where $\Delta D = D_{site}^{V_{DDL}} - D_{site}^{V_{DDH}}$ and $\Delta P = P_{site}^{V_{DDH}} - P_{site}^{V_{DDL}} + P_{LS}$ are the delay and the power dissipation difference between V_{DDL} and V_{DDH} , $Slack_{site_j}$ is the timing slack of site j .

3.5.2 Voltage Re-Assignment

Based on grid-level voltage assignment decision, a *Voltage Re-Assignment* procedure is performed to obtain the best timing rescue and the least power saving penalty within the selected grid i . These step has two constrains for selecting site to operate at V_{DDH} , include maximum sensitivity and LS budget. To start with, the site with maximum sensitivity is selected. Then, check the number of use of LS if the site is assigned to operate at V_{DDH} . If the number is over the LS budget of the selected grid, the selected site is not assigned to operate at V_{DDH} , and select next site with next maximum sensitivity until the site meets these tow constrain at the same time. After that, the selected site in grid i is assigned V_{DDH} for timing rescue. When the site is assigned with V_{DDH} the timing and sensitivity information of many gates are affected and should be updated. The number of site in $grid_i$ would be reduce during assignment procedure. After updating procedure, select next site with maximum sensitivity and do assignment step repeatedly. In this step, the selection and updating procedure is executed repeatedly until no sites in grid i or the site of grid i have been selected and assigned V_{DDH} . Therefore, the times of re-assignment procedure is the initial number of site in grid i at most that reduce the computation load of updating sensitivity.



3.5.3 Incremental Update

When a site i is selected to assign V_{DDH} , the timing information of many gates are affected and must be updated. In order to obtain the exact timing information of circuit, performing the STA after every voltage assignment is an instinctive but costly idea. We use an incremental approach to approximately update the timing information. Firstly, update the AT of the site i and the RT of site i is original value. Then, a backward traversal from the site i is performed to update the RT of gate which is the fan-in of the site i and a forward traversal from the site i is executed to update the AT of gate which is the fan-out of the site i . In the incremental update, the backward and the forward update execution stop until the the gates of second level from the site i is updated.



3.6 Rescue Timing and LS Budget

3.6.1 Timing Rescue

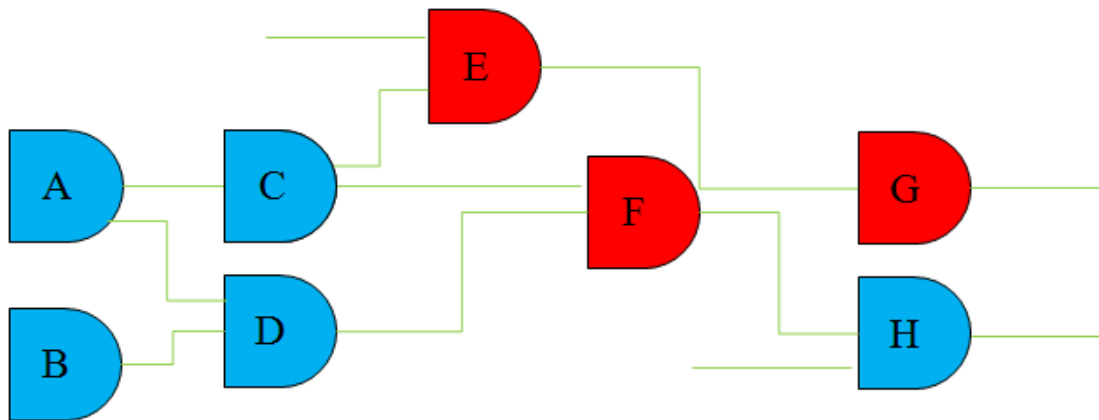


Fig. 3.5: Schematic of Timing Rescue

The 3D IC Voltage Assignment result would make the timing of circuit violated because the LS budget is limited, we can not arbitrarily assign V_{DDH} to the site. Therefore, a timing rescue is executed to rescue the timing of circuit. For example, a circuit schematic in Fig. 3.5, the gate with red color is a site and the gate with blue color is V_{DDL} gate. Firstly, we compute the gain value of each site, and sort them. The gain value is the delay difference between V_{DDH} and V_{DDL} . The site with maximum gain value is selected firstly and assigned to V_{DDH} . Then, select the fan-in gate of the site. We firstly select the fan-in gate C with maximum AT to assign V_{DDH} to it, and update the timing information of it. if the maximum AT is changed from the gate C to the gate D when the gate C is assigned V_{DDH} , we must assign V_{DDH} to gate D. In this way, all gates of fan-in of the site are checked at most until the dominate gate is found. Although the timing is rescued, the number of LS would be excess the LS budget due to executing the timing rescue procedure without considering LS budget. Therefore, we have to sacrifice the power saving by assigning V_{DDH} to many gates with V_{DDL} for reducing the number of LS. In the following section, we will performed our method for rescuing LS budget.

3.6.2 LS Budget Rescue

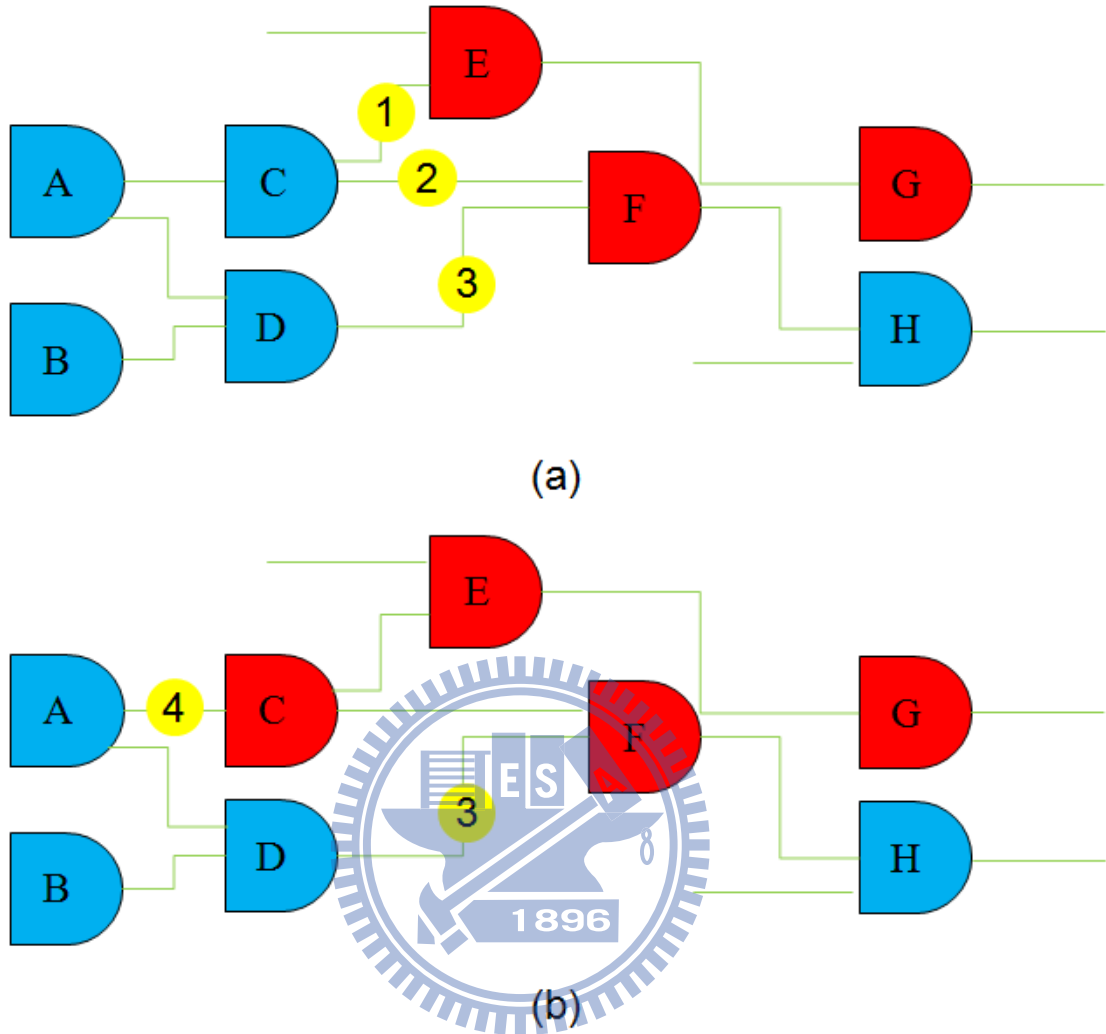


Fig. 3.6: Schematic of LS Budget Rescue

Intuitively, we can start with the fan-in gate of the LS, and assign V_{DDH} to it. This way not only reduce the number of LS but also maintain the correctness of circuit. In Fig 3.6, the gate with yellow color is LS, red is V_{DDH} gate and blue is V_{DDL} gate. Find out all LS, and check the fan-in of LS whether the gate with V_{DDL} can be changed to V_{DDH} by computing the gain of the gate. For example, the Fig 3.6(a) shown the gate C is the fan-in of LS of number 1 and 2, and the gate D is the fan-in of LS of number 3. The gain of the gate C is $-2+1=-1$ and the gain of the gate D is $-1+2=1$, that is, the gate C can reduce one LS but the gate will increase one LS. Therefore, we should select the gate C to operate at V_{DDH} for reducing the usage of LS as the Fig 3.6(b). At same way, the gate A and D in the Fig 3.6(b) are re-checked the gain, the rescue procedure stops when no gate can be assigned to V_{DDH} for reducing the usage of LS.

Chapter 4

Experimental Results

We implement our proposed method in C++ and apply the algorithm to a set of ISCAS89 benchmark circuits and private designs. Firstly, utilizing Design Compiler to synthesize the benchmark circuits with the UMC 90 nm standard cell library. Next, generating an initial 2D placement by the SOC Encounter. Then, a 3D placement is obtained by transforming the 2D placement with Z-Place provided by professor Renato [5]. The timing/leakage power cell library with temperature effect is generated by evaluating the average leakage current and gate delay based on H-SPICE simulation for various types of logic gates. After getting the H-SPICE simulation results, based on the least square method, the fitting constants of the leakage current and gate delay models is obtained. We utilize one type of LS for reducing the complexity of problem.

Table 4.1: Our Proposed Voltage Assignment Method Result

Circuit	# Gates	# LS Budget	Initial	3D ICs		Timing		LS Budget		Final	
			Voltage Assignment	Voltage Assignment	Rescue	Rescue	Rescue	Rescue	Site	LS_excess	
			# Site	# Site	# LS	# Site	# LS	# VDDH	# LS	# Site	# LS_excess
s1488	288	11	86	39	10	0	34	42	20	0	9
s1494	294	11	14	0	7	0	7	8	7	0	0
s1423	343	16	83	76	15	0	60	58	42	0	26
s15850	438	32	79	46	28	0	83	112	36	0	4
s9234_1	596	32	60	30	27	0	45	26	43	0	11
s5378	710	46	26	3	16	0	33	24	33	0	0
s13207	919	63	20	3	16	0	15	10	15	0	0
s38417	5208	287	23	0	24	0	24	11	24	0	0
s35932	5496	290	15	14	15	0	1	19	1	0	0
s38584	5581	222	59	0	46	0	46	23	46	0	0

• Comparison of Voltage Assignment Result

Table 4.1 lists the result of voltage assignment in different phase. The number of site after the initial voltage assignment is listed in column 4. Column 5-6 shows the number of site and the usage of LS after *3D ICs Voltage Assignment*. The result of *Timing Rescue* and *LS Budget Rescue* are listed in column 7-8 and column 9-10, respectively. Finally, the summarization is listed in column 11-12.

Firstly, the results of *3D ICs Voltage Assignment* show that timing rescue is limited significantly by LS budget. Most circuit can not meet timing constrain after *3D ICs Voltage Assignment* procedure under limited LS budget constrain. Next, to deal with this problem, we try to rescue the site of circuit by *Timing Rescue*. After executing *Timing Rescue* procedure, all circuits meet timing constrain finally. However, the usage of LS is more than LS budget. Because the timing rescue procedure does not take LS budget into consideration, just considering the influence of voltage assignment on timing. Finally, we try to rescue the circuit again by *LS Budget Rescue*. The result of column 12 shows that it still has four circuits design which are rescued unsuccessfully.

For reducing the problem size, the *3D ICs Voltage Assignment* method limits the voltage assignment decision to the site, obviously, the number of site is minority of the circuit. Moreover, we think that the site is the most important gate for timing rescue. However, it is not enough just by the site for timing rescue due to considering LS budget. Obviously, the *3D IC Voltage Assignment* has some challenges can be solved. For example, extend the range of selected gate instead of a set of site in *3D ICs Voltage Assignment* procedure or change the voltage assignment method that assign V_{DDL} to a gate if it meet timing constrain and LS budget.

Table 4.2: Optimization Result

Circuit	Initial Power(μW) and Temperature			Optimized Power(μW) and Temperature				Improvement(%)			
	Dynamic	Leakage	Temp.	Dynamic	Leakage	LS	Temp.	Dynamic	Leakage	Total	Temp.
s1488	309.296	9.17531	318.472	176.989	2.90109	9.49963	48.2101	42.7769	68.3815	43.5146	35.774
s1494	309.297	9.4769	318.774	135.486	2.46178	5.89517	46.5284	56.1954	74.0233	56.7254	38.1006
s1423	395.74	17.1244	412.864	221.372	6.18041	24.5398	40.3195	44.0613	63.9087	44.8845	33.4557
s15850	470.18	16.9509	487.131	295.775	7.75729	17.9814	37.7831	37.0931	54.2366	37.6897	29.0719
s9234_1	795.462	31.2545	826.717	349.575	9.49689	25.5951	41.6037	56.0538	69.6143	56.5665	34.6316
s5378	1028.97	41.9583	1070.93	451.782	12.7489	19.5059	42.69	56.0938	69.6154	56.6236	34.9129
s13207	1375.25	57.609	1432.86	594.356	17.3737	15.2843	40.8162	56.7819	69.8421	57.307	33.6806
s38417	15410.3	619.59	16029.9	6262.64	112.99	33.4185	53.7564	59.3607	81.7637	60.2266	45.3949
s35932	17984.4	822.007	18806.4	7423.47	138.532	4.95071	56.2681	58.7228	83.1471	59.7903	46.9839
s38584	16002.8	1614.49	17617.3	6519.46	117.395	55.9655	60.0174	59.2605	92.7286	62.3276	52.9168
Avg.								57.74	78.52	58.83	42.00

• **Power Reduction**

Table 4.2 summarizes the optimization results of each circuit. Base on Table 4.1, the results shows that four circuits are fail and six circuits are successful after executing our power reduction method. Therefore, the average of improvement is the average of the six circuit results. The initial power and average temperature of the circuit are listed in columns 2-4, the power and average temperature of the circuit and the power of LS after optimization are listed in columns 5-8 and the improvement percentage of the dynamic power, the leakage power, the total power and the temperature decrement of circuit are listed in column 9-12, respectively. The improvement columns indicate that the proposed 3D ICs voltage assignment method can provide almost 58.83% total power saving and 42 degree decrement in temperature in average. Observe that the leakage power reduction is a great improvement due to the temperature decrement.

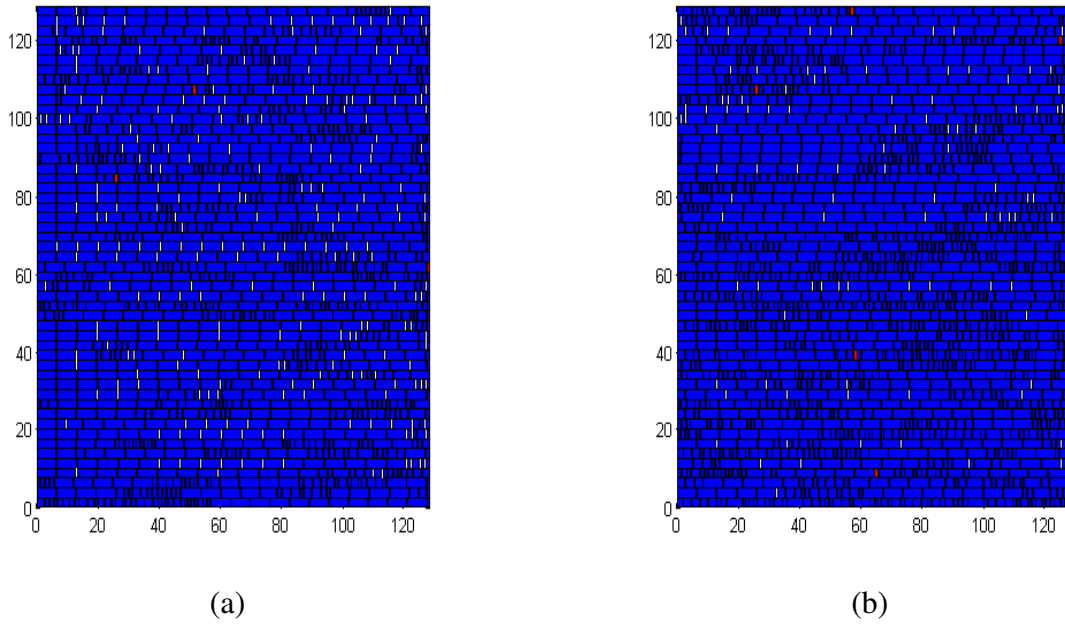


Fig. 4.1: (a) Voltage assignment result on layer 1 of s35932. (b) Voltage assignment result on layer 2 of s35932.

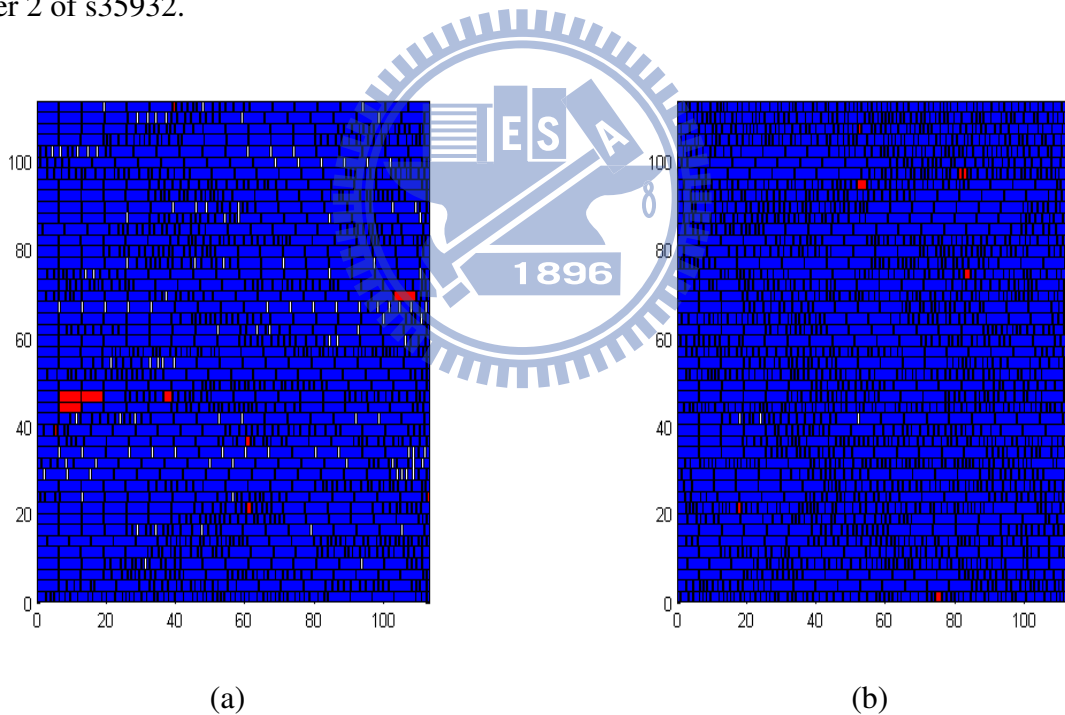


Fig. 4.2: (a) Voltage islands on layer 1 of s38584. (b) Voltage islands on layer 2 of s38584.

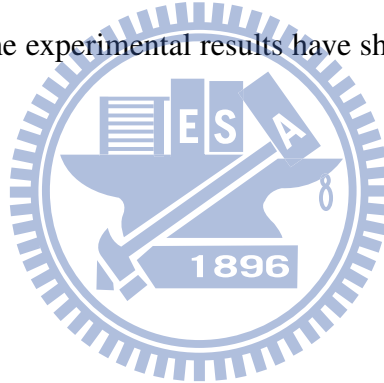
- **Voltage Assignment Result**

In this experiment, we use 1.1 and 0.9 as the high/low supply voltage, respectively. The grid with dark color means that the low supply voltage is used. Fig. 4.1 and Fig. 4.2 illustrate the voltage assignment result of Circuit s35932 and s38584.

Chapter 5

Conclusion

In this thesis, a *3D ICs Voltage Assignment* method with the combination of selecting grid by *Grid-Based Procedure* and *Voltage Re-Assignment* is proposed to minimize the total power consumption in 3D ICs design. By employing the temperature-related delay of gate and leakage power models obtain more accurate estimation of the circuit performance. Although it has four unsuccessful circuits The experimental results have shown a great power reduction by the proposed method.



Bibliography

- [1] K. Banerjee, S.J. Souri, P. Kapur, and K.C. Saraswat. 3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. *Proceedings of the IEEE*, 89(5):602–633, 2001.
- [2] J.C. Chi, H.H. Lee, S.H. Tsai, and M.C. Chi. Gate level multiple supply voltage assignment algorithm for power optimization under timing constraint. *IEEE transactions on very large scale integration (VLSI) systems*, 15(6):637–648, 2007.
- [3] RLS Ching, EFY Young, KCK Leung, and C. Chu. Post-placement voltage island generation. In *IEEE/ACM International Conference on Computer-Aided Design, 2006. IC-CAD'06*, pages 641–646, 2006.
- [4] HF Dadgour, S.C. Lin, and K. Banerjee. A statistical framework for estimation of full-chip leakage-power distribution under parameter variations. *IEEE Transactions on Electron Devices*, 54(11):2930–2945, 2007.
- [5] R. Hentschke, G. Flach, F. Pinto, and R. Reis. 3D-vias aware quadratic placement for 3D VLSI circuits. In *IEEE Computer Society Annual Symposium on VLSI, 2007. ISVLSI'07*, pages 67–72, 2007.
- [6] J. Hu, Y. Shin, N. Dhanwada, and R. Marculescu. Architecting voltage islands in core-based system-on-a-chip designs. In *Proceedings of the 2004 international symposium on Low power electronics and design*, pages 180–185. ACM, 2004.
- [7] W. Hung, Y. Xie, N. Vijaykrishnan, M. Kandemir, MJ Irwin, and Y. Tsai. Total power optimization through simultaneously multiple-Vdd multiple-Vth assignment and device sizing

- with stack forcing. In *Low Power Electronics and Design, 2004. ISLPED'04. Proceedings of the 2004 International Symposium on*, pages 144–149, 2004.
- [8] K.K. Kim. Accurate Macro-modeling for Leakage Current for IDDQ Test. 2007.
- [9] S.H. Kulkarni, A.N. Srivastava, and D. Sylvester. A new algorithm for improved VDD assignment in low power dual VDD systems. In *Proceedings of the 2004 international symposium on Low power electronics and design*, pages 200–205. ACM, 2004.
- [10] H.H. Lee, S.H. Tsai, J.C. Chi, and M.C. Chi. A Partition-Based Voltage Scaling Algorithm Using Dual Supply Voltages for Low Power Designs. In *VLSI Design, Automation and Test, 2006 International Symposium on*, pages 1–4, 2006.
- [11] J.Q. Lu, K. Rose, and S. Vitkavage. 3D Integration: Why, What, Who, When? *Future Fab International* (<http://www.future-fab.com/>), pages 25–27, 2007.
- [12] W.K. Mak and W. Chen Jr. Voltage island generation under performance requirement for SoC designs. In *Proceedings of the 2007 Asia and South Pacific Design Automation Conference*, pages 798–803. IEEE Computer Society, 2007.
- [13] D. Nguyen, A. Davare, M. Orshansky, D. Chinnery, B. Thompson, and K. Keutzer. Minimization of dynamic and static power through joint assignment of threshold voltages and sizing optimization. In *Proceedings of the 2003 international symposium on Low power electronics and design*, pages 158–163. ACM, 2003.
- [14] V. Reddy, A.T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan. Impact of negative bias temperature instability on digital circuit reliability. *Microelectronics Reliability*, 45(1):31–38, 2005.
- [15] S.S. Sapatnekar. *Timing*. Springer Netherlands, 2004.
- [16] D. Sinha, NV Shenoy, and H. Zhou. Statistical gate sizing for timing yield optimization. In *IEEE/ACM International Conference on Computer-Aided Design, 2005. ICCAD-2005*, pages 1037–1041, 2005.

- [17] K. Usami and M. Horowitz. Clustered voltage scaling technique for low-power design. In *Proceedings of the 1995 international symposium on Low power design*, pages 3–8. ACM, 1995.
- [18] K. Usami, M. Igarashi, F. Minami, T. Ishikawa, M. Kanzawa, M. Ichida, and K. Nogami. Automated low-power technique exploiting multiple supply voltages applied to a media processor. *IEEE Journal of Solid-State Circuits*, 33(3):463–472, 1998.
- [19] H. Wu, I.M. Liu, MDF Wong, and Y. Wang. Post-placement voltage island generation under performance requirement. In *IEEE/ACM International Conference on Computer-Aided Design, 2005. ICCAD-2005*, pages 309–316, 2005.
- [20] H. Wu, M.D.F. Wong, I. Liu, et al. Timing-constrained and voltage-island-aware voltage assignment. In *Proceedings of the 43rd annual Design Automation Conference*, page 432. ACM, 2006.
- [21] X. Ye, Y. Zhan, and P. Li. Statistical leakage power minimization using fast equi-slack shell based optimization. In *DAC'07*, pages 853–858, 2007.
- [22] S.A. Yu, P.Y. Huang, and Y.M. Lee. A multiple supply voltage based power reduction method in 3-D ICs considering process variations and thermal effects. In *ASPDAC*, pages 55–60. IEEE Press, 2009.
- [23] Y.F. Yu. Multi-Voltage Floorplan Design with Optimal Voltage Assignment. In *Proceedings of 2004 International Symposium on Physical Design (ISPD'04)*, volume 6, 2009.