國立交通大學

電信工程學系

碩士論文

考慮製程變異與溫度效應的三維積體電路功率最佳化方法

Power Optimization in 3D ICs Considering Process Variations and Thermal Effect

研究生：于斯安

指導教授：李育民 教授

中華民國九十七年七月

考慮製程變異與溫度效應的三維積體電路
功率最佳化方法

學生: 于斯安　　　　　　　　　　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的方法來解決二維積體電路上過長導線造成的進步瓶頸，但是過高的溫度也成為三維積體電路的挑戰。晶片上的溫度會對效能造成嚴重的影響，因此有必要降低電路的功率消耗。同時，另一個在奈米製程中，對電路設計有重大影響的議題則是製程變異。在這篇論文中，我們提出一個利用雙電壓源的統計型方法來降低三維積體電路上的總功率消耗。利用卡洛展開(Karhunen-Loeve expansion)將通道長度(channel length)和氧化層厚度(oxide thickness)這類具有空間相關隨機過程的物理參數轉換成一組無相關性的隨機變數。因為製程變異的關係，晶片上的靜態功率是一個隨機過程，我們利用一個統計型的溫度分析方法來得到溫度平均值跟變異量的分布。為了強調溫度的影響，我們採用具溫度相關性的漏電流(leakage current)與邏輯閘延遲(gate delay)模型，並且完成一套考慮溫度的統計型時序分析方法。所提出降低功率的方法利用功率延遲敏感度(power-delay sensitivity)作為最佳化的標準，並使用一個以切格子的(grid-based)方式來處理整個三維積體電路的結構。演算法中使用一個有效的觀念取代每次的統計型時序分析來增加運作效率。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 in 3D ICs Considering Process Variations and Thermal Effect

Student: Shih-An Yu 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 methodology to overcome the bottleneck caused by the long interconnects in the 2D IC. However, the higher temperature becomes a big challenge for 3D ICs. On-chip temperature can significantly affect the circuit performance so it is necessary to reduce the power dissipation in the circuit. Meanwhile, the process variations, which have a serious influence on the circuit design, are another important issue for the nanometer IC design. In this thesis, we present a approach to statistically minimize the total power consumption on the 3D ICs by using the dual supply voltage technology. By Karhunen-Loeve expansion, the random processes of physical parameters such as the channel length and the oxide thickness with spatial correlations are transformed to a set of uncorrelated random variables. Since the leakage power on the chip is a random process due to the process variations, we employ a statistical thermal simulation method to get the mean and variance of temperature distribution. To emphasize the impact of temperature, the leakage current and gate delay models are temperature related and we implement a thermal aware statistical timing analysis method. The proposed power reduction approach uses power-delay sensitivity as the optimization criterion, and a grid-based method for handling the whole structure of the 3D ICs. Instead of executing statistical timing analysis every time, a potent concept is used in the algorithm to achieve the runtime efficiency. The experimental results demonstrate the effectiveness of our method and indicate that considering the thermal effect in the circuit simulation is imperative.

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。

實驗的部份，感謝周景揚老師提供我們聯電 90 奈米製程參數，感謝鍾菁哲博士和陳志龍學長在實驗設定的協助，感謝 Renato 教授提供的軟體，感謝 Markov 教授對實驗流程的指導，還有同儕建螢給予的幫助，沒有你們，要完成這個實驗是相當困難的。我在此對你們至上最大的謝意。

在實驗室裡，感謝培育學長和柏毅學長在知識與技術方面的指導與傳承，以及國富、志康、炳熏、至鴻、佳鴻、懷中、庚達、宗祐、阿文給予的關心與幫助，同窗焯基的相互勉勵，豐富這二年的生活。

特別感謝好友宣聆，給予我許多支持與鼓勵，陪伴我度過最低潮的日子。

最後要深深地感謝我的父母，你們無微不至的照顧並且提供安穩的環境讓我可以毫無顧慮地在學業上衝刺，讓我能夠順利完成碩士學業。僅在此將本論文獻給你們，共享這份喜悅與榮耀。

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

As consumer demands on integrated circuits increase, the interconnect structure has become the bottleneck of the chip performance. Increased interconnection lengths produce serious parasitic effects which increase the circuit delay and the power consumption. Three-dimensional integrated circuits (3D ICs) have been proposed to solve this problem. The 3D ICs allow the designer to stack dies or wafers vertically in the same package and connect components on the different tier by through-silicon vias (TSVs) [1]. However, the heat removal is an important issue in the 3D IC design due to the higher power density and the low thermal conductivity inter-layer dielectrics between the device layers. The high temperature can reduce the life time and reliability of device and impact the circuit performance in the timing and leakage power [2, 3, 4, 5]. Numerous researches have been done in the physical design level to handle the temperature problem. Floorplanning [6, 7], placement [8], routing [9, 10, 11], and thermal via insertion [12, 13, 14] are used to arrange the power distribution and improve the thermal conductivity on the chip. However, if the designer wants to solve this issue fundamentally, making power reduction is a necessary step since the power consumption in the circuit is the source of thermal problem. In addition, due to the shrinking of device geometries, it is more difficult to control the device parameters. Growing process variability such as the effective channel length, the gate oxide thickness, the ILD thickness, the random doping concentration and the threshold voltage has made the deterministic analysis and optimization at a prescribed process corner be no longer effective. The traditional approaches can significantly overestimate or underestimate the impact of process variations. Overestimation leads to increased design time/effort and results in the performance loss while underestimation causes the yield loss. As shown in [15], 30% process variations can cause up to 20X leakage power variations. According to ITRS, the ratio of the leakage power to the total power has increased 10 times from the 180 nm process to the 90 nm process [16]. Furthermore, at the 90 nm process nodes, leakage power accounts for 25% to 40% of the total power, and it is expected that 50% to 70% of the total power will be lost through the leakage currents at the 65 nm process [17]. Hence, it is imperative to consider process variations in the power optimization work.

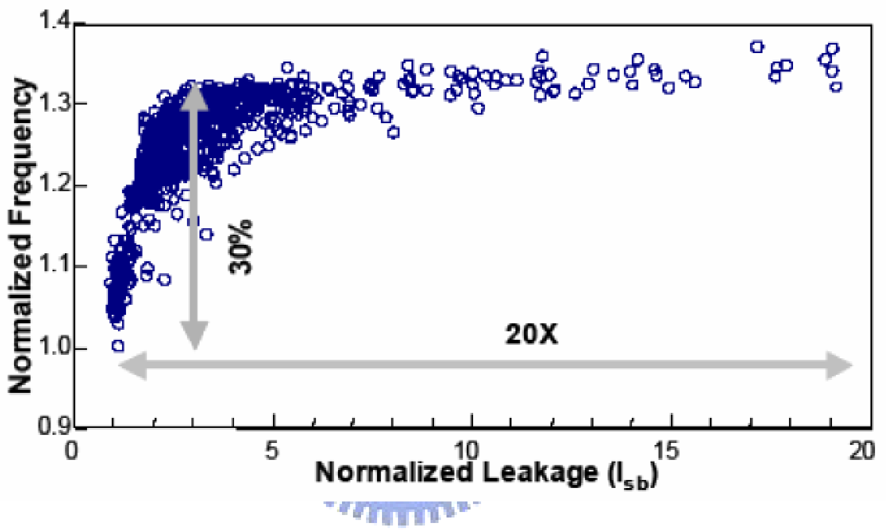


Fig. 1.1: Leakage variations [18]

1.2 Our Contributions

In this thesis, a power minimization methodology for 3D ICs is presented; the main contributions of the thesis are summarized in the following terms:

- We propose a power optimization approach for the 3D IC design which considers process variations and thermal effect to the timing and leakage power by using the dual supply voltages. Compared with the previous works which solve the deterministic problems and use the thermal unrelated models, our method is more flexible and practical. Furthermore, this is the first work that discusses the voltage island generation in 3D ICs.

- Besides the physical parameters such as the channel length and the oxide thickness, the temperature is treated as a variation parameter in the statistical static timing analysis (SSTA). A similar idea can be found in [19]. In that work, the temperature is assumed to be independent of the circuit design and the temperature on each location is equal; however, the placement of circuit can affect the power distribution, and the leakage power is impacted by the physical parameters. Therefore, the temperature should be viewed as a random variable which is a function of physical parameters to get a reasonable answer.

- No matter what method is used, the timing information of the circuit should be updated after each step or adjustment. The proposed algorithm uses a heuristic and efficient method to avoid the expensive cost of SSTA.

1.3 Organization of the Thesis

The organization of the rest of the thesis is as follows. Chapter 2 gives an overview of the necessary background for this work, including 3D IC technology, parameter modeling, Karhunen-Loeve expansion, leakage current modeling, statistical timing analysis, and the multiple supply voltage technique. Our experiment flowchart, statistical 3D thermal analysis, thermal aware statistical timing analysis, and power optimization method are addressed in chapter 3. Finally, the experimental results and conclusion are presented in chapter 4 and 5, respectively.

Chapter 2

Preliminaries

In this chapter, we first study the background knowledge of 3D IC. Second, the parameter modeling is presented in section 2.2. After that, we introduce the statistical leakage current modeling in section 2.3. The next section surveys the methods of static timing analysis and statistical timing analysis. Finally, some power optimization works are reviewed and the idea of post-placement voltage island is presented.

2.1 3D IC Technology

2.1.1 Motivation for 3D ICs

The unprecedented growth of the computer and the information technology industry are demanding very large scale integrated (VLSI) circuits with increasing functionality and performance at minimum cost and power dissipation. While the VLSI technology scales down, the circuit improvement is limited by the long interconnect. The long interconnect on 2D chip causes serious parasitic effects which slow down the circuit speed and require an increasing number of inserting buffers. The increasing interconnect loading also affects the power consumption in high-performance chips. On-chip global wires contribute about 34% to the total chip power dissipation in an Intel microprocessor [20]. Additionally, it results in other problems such as signal integrity and routing congestion. Furthermore, increasing drive for the integration of analog/digital signals and disparate technologies introduces various system-on-chip design concepts, for which existing planar IC design may not be suitable.

2.1.2 Benefits and Challenges of 3D Integration

3D ICs replace the long interconnect on 2D chip by the shorter vertical vias which can lead to over 25% decrease in the worst case wire length [21, 22]; at the same time, the interconnect power [23] and the chip area are also reduced [1]. This is especially important for processors as they access memory continuously. With 3D integration, the access time is reduced, and the system performance is improved. This improvement has been studied in many recent works [24, 25, 26]. In addition, 3D ICs allow the integration of different technologies such as memory, logic, RF and analog components on one chip. IC technologies in different active device tiers wouldn't face the technology and manufacturing incompatibilities and cross-contamination issues. Other advantages include reduced power consumption, increased packing density, decreased packaging size, weight, and cost. However, 3D integration has its own challenges in terms of fabrication, production yield, heat removal and process variations [27, 28, 29, 30]. Both the fabrication process and production yield are highly related to the through-silicon vias. Making these vertical vias is a complicated and difficult procedure in the design flow. On the other hand, due to the higher power density and the low thermal conductivity inter-layer dielectrics between the device layers, 3D ICs have a much higher temperature than 2D ICs. As shown in [29], the gate delay has a linear relation with the temperature and the leakage power has an exponential relation with the temperature. In brief, the temperature effect is an undeniable factor in the circuit simulation.

2.1.3 Current 3D Technology

3D IC fabrication technologies include multi-chip module (MCM) packaging, wafer bonding, solid-phase recrystallization, etc. Different fabrication technologies can greatly affect the circuit performance, manufacturing cost, on-chip temperature, etc. Wafer bonding is the most popular method currently. In wafer-level 3D integration, functional materials and components are prefabricated on separate wafers, followed by wafer aligning, bonding, and vertical inter-wafer interconnection to integrate these functional materials and components in a 3D stack. There are three main methods to achieve wafer-scale 3D integration [31]: *wafer-to-wafer, die-to-wafer, and die-to-die integration*. Wafer-to-wafer integration directly bonds entire wafers together.

Die-to-wafer uses a substrate wafer to integrate an already diced die on top of it. Die-to-die integration allows the same high yield as die-to-wafer but suffers from low-production throughput. Recent developments toward reliable and high yield adhesive bonding processes have made adhesive wafer bonding a good candidate for 3D integration platforms. In adhesive wafer bonding, an intermediate adhesive polymer layer is used to create a bond between two wafers. The main advantages of adhesive bonding are the compatibility with integrated circuit wafers, the relatively low bonding temperatures, the ability to join practically any kind of wafer material and an insensitivity to particles and structures at the wafer surfaces. In most commonly used adhesive wafer bonding processes, the polymer adhesive is applied to one or both of the wafer surfaces to be bonded. Often, the polymer coatings are heated after spin-coating to remove solvents and/or to partially cross-link the polymer coating. After aligning two wafers and joining the polymer-coated surfaces, pressure is applied to fore the wafer surfaces into intimate contact [32]. The topics about the application and design flow can be found in [33, 34, 35, 36, 37]. The assembly process and a 3D chip consisting of three tiers are illustrated in Fig. 2.1.

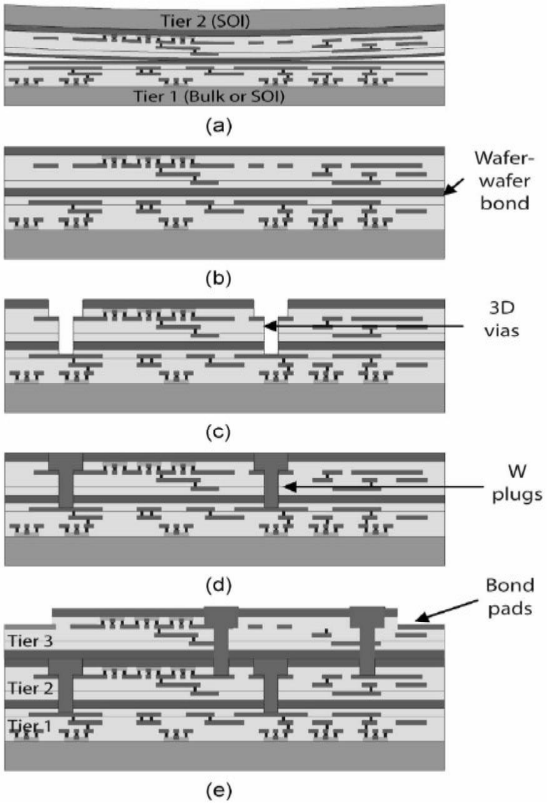


Fig. 2.1: Assembly process for a 3D chip [38]

2.2 Parameter Modeling

Process variations can be classified into inter-die variations and intra-die variations. Inter-die variations are the variations from die-to-die. All the cells in one die have the same variation type. On the other hand, intra-die variations correspond to the variability within a single chip. With continuous process scaling, intra-die variations have to consider the spatial correlation which the devices with close proximity are more likely to have similar process parameter values.

2.2.1 Karhunen-Loeve Expansion

Considering the spatial correlation of intra-die process variations is a significantly challenging task because of the severely increased variables and computational complexity. The physical parameter on one location is a random variable while all the parameters on the chip become a random process. Thus, we need infinite random variables to accurately model the spatial correlation on the chip. It is complicated and impossible. An alternative method is the use of Karhunen-Loeve (KL) Expansion. The concept of KL expansion is to find a set of mutually independent random variables as the bases for the on-chip variations, and express the physical parameter on any location as the combination of these bases. Hence, the original large number of correlated random variables can be modeled by the small number of uncorrelated random variables, which can reduce the complexity and improve the efficiency. The set size is decided by the user to control the approximation error. Moreover, the chip can be divided into many grids and the center point of each grid can be the reference point for computing the covariance value and distance. Therefore, all the gates in the same grid can share the same coefficients and the number of grids is relatively small compared with the number of gates on the chip.

The KL expansion of a second-order random process $\alpha(x,y,z,\vartheta)$ with a continuous spatial covariance function is expressed as follows [39]

$$\alpha(x,y,z,\vartheta) = \overline{\alpha}(x,y,z) + \sum_{k=1}^{\infty} \sqrt{\gamma_k}\,\phi_k(x,y,z)\eta_k(\vartheta), \tag{2.1}$$

where $\overline{\alpha}(x,y,z)$ is the mean value of $\alpha(x,y,z,\vartheta)$, and each γ_k and each $\phi_k(x,y,z)$ are the eigenvalue and the eigenfunction derived from the following Fredholm integral equation

$$\int_{D_0} C(\mathbf{x}_1,\mathbf{x}_2)\phi_k(\mathbf{x}_2)d\mathbf{x}_2 = \gamma_k\phi_k(\mathbf{x}_1). \tag{2.2}$$

Here, $C(\mathbf{x}_1, \mathbf{x}_2)$ is the covariance function of the random process $\alpha(x, y, z, \vartheta)$, $\mathbf{x_1} = (x_1, y_1, z_1)$ and $\mathbf{x_2} = (x_2, y_2, z_2)$ are locations, ϑ is the sampling event of the sample space Ω_α, and $\{\eta_k(\vartheta)\}$ is a set of uncorrelated random variables with each $\eta_k(\vartheta)$ being zero mean and unit variance. In contrast to the conventional 2D IC, we add the z- dimension to represent the tiers of 3D IC.

With the spatial covariance function, physical parameters such as the effective channel length and the gate oxide thickness can be modeled as random processes. It has been indicated that, for the covariance function of a physical parameter, the statistical covariance at two different points decreases as the distance between these two points increases [40]. Functions like the exponential form, the Gaussian form, the linear form, or a fitting form with experimental data are suggested to model this property [41, 42, 43]. By the use of the second-order property and the above covariance functions, the KL expansions of the physical parameters are valid.

2.2.2 Spatial Correlation Modeling

Researchers in [44] presented that the decreasing rate of the spatial covariance of physical parameters is different in the x- and y- directions on the chip. The following spatial covariance function is adopted to model this characteristic

$$C(\mathbf{x_1}, \mathbf{x_2}) = \sigma^2 \exp\left(-\frac{|x_1 - x_2|}{\eta_x}\right) \exp\left(-\frac{|y_1 - y_2|}{\eta_y}\right), \tag{2.3}$$

where η_x and η_y are correlation lengthes of the target random process in the x- and y- directions, respectively, σ is the standard deviation of the target random process, and this covariance kernel is defined in the rectangular domain D_0. Because the process variations on different tiers are mutually independent [30], the function does not consider the z- dimension and can still be utilized.

With the above function, the KL expansions of gate oxide thickness ($t_{ox}(x, y, z, \varpi)$) and the effective channel length ($L_{eff}(x, y, z, \theta)$) can be obtained as

$$\begin{aligned} t_{ox}(x, y, z, \varpi) &\approx \bar{t}_{ox}(x, y, z) + \triangle t_{ox}(x, y, z, \varpi), & (2.4) \\ L_{eff}(x, y, z, \theta) &\approx \overline{L}_{eff}(x, y, z) + \triangle L_{eff}(x, y, z, \theta), & (2.5) \end{aligned}$$

and

$$\triangle t_{ox}(x,y,z,\varpi) \equiv \sum_{n=1}^{N_{t_{ox}}} \sqrt{\beta_n} f_n(x,y,z)\varsigma_n(\varpi), \tag{2.6}$$

$$\triangle L_{eff}(x,y,z,\theta) \equiv \sum_{m=1}^{N_{L_{eff}}} \sqrt{\chi_m} q_m(x,y,z)\zeta_m(\theta). \tag{2.7}$$

where $\bar{t}_{ox}(x,y,z)$ and $\overline{L}_{eff}(x,y,z)$ are the expected values of $t_{ox}(x,y,z,\varpi)$ and $L_{eff}(x,y,z,\theta)$, $f_n(x,y,z)$ and $q_m(x,y,z)$ are eigenfunctions of $t_{ox}(x,y,z,\varpi)$ and $L_{eff}(x,y,z,\theta)$, β_n and χ_m are eigenvalues of $t_{ox}(x,y,z,\varpi)$ and $L_{eff}(x,y,z,\theta)$, respectively. The $N_{t_{ox}}$ and $N_{L_{eff}}$ are the truncated numbers of t_{ox} and L_{eff}. $\{\varsigma_n(\varpi)\}$ and $\{\zeta_m(\theta)\}$ are mutually independent standard normal random variables since $t_{ox}(x,y,z,\varpi)$ and $L_{eff}(x,y,z,\theta)$ can be assumed to be Gaussian processes [45], and they are assumed to be independent. Here, we use $\{\xi_n(\varpi,\theta)\}_{n=1}^{N_{KL}}$ as the union of $\{\varsigma_n(\varpi)\}_{n=1}^{N_{t_{ox}}}$ and $\{\zeta_n(\theta)\}_{n=1}^{N_{L_{eff}}}$, N_{KL} = $N_{t_{ox}}$ + $N_{L_{eff}}$ and ξ_n, ς_n and ζ_n are used to represent $\xi_n(\varpi,\theta)$, $\varsigma_n(\varpi)$ and $\zeta_n(\theta)$ for simplicity in the following content.

2.3 Statistical Leakage Current Modeling

In this section, we introduce the empirical models for subthreshold and gate leakage currents with the uncertainty in physical parameters such as channel length and oxide thickness. Actually, the leakage currents depend on the input pattern and logic topology. We evaluate the average leakage currents based on H-SPICE simulation for various types of logic gate with input pattern considered. From the H-SPICE simulation results, we obtain the fitting constants of the empirical current models based on the least square method. The maximum errors of fitting models are no more than 2% in comparison with the H-SPICE simulation results. Because the leakage current is supply voltage dependent, we fit two pairs of coefficients corresponding to the high/low supply voltages.

2.3.1 Gate Tunneling Leakage Current

According to quantum mechanics, there is a finite probability that carriers can tunnel through the gate oxide. The result is so-called that the gate tunneling leakage current flows into the gate. The finite probability is an exponential function of oxide thickness. The gate tunneling leakage current increases exponentially as oxide thickness decreases. When the oxide thickness is thicker than $20\mathring{A}$, the gate tunneling leakage current is relatively small in comparison with other leakage currents such as the subthreshold leakage current. For the oxide thickness is thinner than $15-20\mathring{A}$, the tunneling current becomes an important factor and may become comparable with the subthreshold leakage current in advanced process. To put it briefly, the dependence of gate leakage current on oxide thickness is given by the following formula [46]:

$$I_{gate} = (A \cdot C)(W \cdot L)e^{-B \cdot \frac{t_{ox}}{V_{gs}} \alpha},$$

where $A = q_3/8\pi h\phi_b$, $B = 8\pi\sqrt{2m_{ox}}\phi_b^{3/2}/3hq$, $C = (V_{gs}/t_{ox})^2$, α is a parameter which is ranged from 0.1 to 1 depending on the voltage drop across the oxide, H is the Plancks constant, and ϕ_b is the barrier height for electronics/holes in the conduction/valance band. Note that the parameter variations are in general around 10-20% [47]. Hence, we make use of a first-order Taylor expansion at the nominal value of parameter oxide thickness and utilize the following gate tunneling leakage current model derived in [48].

$$I_{gate} = b_0 \exp(b_1 t_{ox}), \tag{2.8}$$

where b_0 and b_1 are fitting constants, t_{ox} is the oxide thickness. Because the parameter V_{gs} is related to the supply voltages, we fit two pairs of coefficients for the high/low supply voltage, respectively.

2.3.2 Subthreshold Leakage Current

The subthreshold leakage current is defined as the conduction current between source and drain in an "off" state CMOS transistor. The subthreshold leakge current for a MOSFET can be modeled as [49]

$$\begin{aligned} I_{OFF} &= I_{ds} \\ &= \mu_{eff} C_{ox} \frac{W_{eff}}{L_{eff}}(m-1)V_T^2(1-exp(-\frac{V_{ds}}{V_T}))exp(\frac{V_{gs}-V_{th}}{mV_T}), \qquad (2.9) \\ m &= 1+\frac{\sqrt{\varepsilon_{si} q N_a/4\Psi_B}}{C_{ox}}, \end{aligned}$$

where μ_{eff} is the effective mobility, C_{ox} is the gate-oxide capacitance, L_{eff} is the effective channel length, W_{eff} is the effective width, V_T is the thermal voltage, N_a is the channel doping concentration , q is the charge of electron, ε_{si} is the permittivity of silicon and Ψ_B is the difference between Fermi potential and intrinsic potential. Here, we still use the first-order formulation and consider the temperature impact [29], so the subthreshold leakage current model is

$$I_{sub} = c_0 \exp\left(c_1 L_{eff} + c_2 T\right), \tag{2.10}$$

where c_0, c_1 and c_2 are fitting constants, L_{eff} is the channel length and T is the temperature. As the gate leakage current, we fit two pairs of coefficients due to the relation between V_{gs} and supply voltage.

2.4 Timing Analysis

2.4.1 Static Timing Analysis (STA)

Before introducing the SSTA, we first review the traditional STA and several basic knowledge. In timing analysis, each gate has three types of timing information: 1) arrival time (AT) 2) required time (RT) 3) slack. The AT means the real time that the signal arrives; the RT represents the designer's constraint that when the signal should arrive. The definition of slack is the difference between AT and RT. When the slack is positive, which means that the signal arrives earlier than the designer's request, there is no timing violation. The gate with positive slack can be optimized to reduce the circuit's area or power dissipation.

Now we use the example shown in Fig. 2.2 to describe the details of STA. The method that is commonly referred to as PERT (Program Evaluation and Review Technique) is popularly used in STA [50]. In this figure, each block could be as simple as a logic gate or a more complex combinational block, and is characterized by the delay from each input pin to each output pin. *A* and *B* are the primary inputs; *Y* is the primary output. Initially, a depth first search (DFS) is performed and the blocks are put into a queue according to their sequence. Generally, it is assumed that all primary inputs are available at time zero. Hence, the AT of the output of *A* is simply the delay of *A*, so does *B*. Then, for *C*, it has two different inputs *A* and *B*. The AT from *A* is computed as 2+3=5; while the AT from *B* is computed as 5+1=6. Since the timing analysis considers the worst case, the AT of the output of *C* should choose the latest one, which is the maximum of 5 and 6. Doing a forward traversal, we can get the arrival time of all blocks in the circuit. On the other hand, the RT is calculated by a backward traversal from

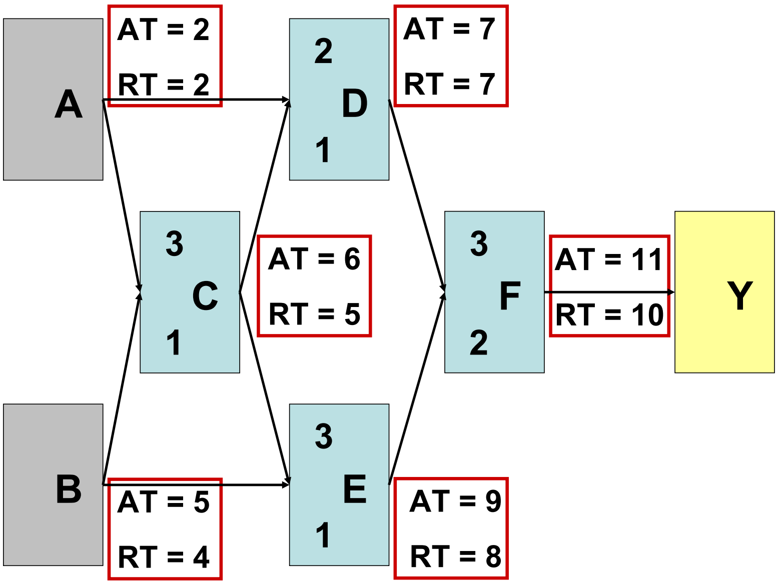


Fig. 2.2: An example illustrating the application of the STA. The numbers within the block correspond to the delay of the block. The primary inputs are assumed to be available at time zero.

the primary output. The RT at the output of *F* is 10, which is set by the designer. Then the RT of *D* and *E* are computed as 10-3=7 and 10-2=8, respectively. After the computation of AT and RT, we can get the slack of each gate by subtracting the AT from RT. The critical path, which is defined as the path between an input and an output with the maximum delay, is the path *B*-*C*-*E*-*F*-*Y*. In the above example, the delay time of each gate can be searched in the cell library. The conventional method builds the library according to several process parameters and tries to make the design work under the defined worst situation. Unfortunately, in presence of process variations, the traditional STA is pessimistic and not effective. The increasing variation parameters make it complex and impractical to build such a huge library. Moreover, from the viewpoint of probability, the possibility of worst case is extremely small.

2.4.2 Statistical Static Timing Analysis (SSTA)

The alternative approach in timing analysis is the SSTA which treats delays as random variables and propagates the random variables in the circuit. Existing SSTA methods can be categorized into two approaches: the path based SSTA and the block based SSTA. The path based SSTA tries to find the statistical critical paths. However, the task of selecting several timing critical paths statistically has high complexity that grows exponentially with respect to the circuit size. On the other hand, the block based SSTA treats each gate/wire as a timing block and performs the timing analysis block by block in the circuit timing graph without looking back to the path history. Hence, the computation complexity would only grow linearly with losing part of accuracy.

Numerous literatures have investigated the SSTA in various directions. Many researchers suggest the variations to be Gaussian random variables and use the canonical first order formulation to represent the delay [51, 52, 53, 54]. In contrast, various studies assume the fluctuations to be non-Gaussian distributions [55, 56, 57] or probabilistic interval variables [58] and use non-linear delay functions [59]. How to reduce the error caused by the MAX operation is proposed by [60], and how to take the spatial delay correlations into account is introduced in [61]. However, authors in [62] indicated that considering the parameters as Gaussian random variables and the use of linear delay model can provide sufficiently accurate simulation. Hence, we follow this assumption in our experiment.

2.5 Power Optimization

In this section, we review several previous publications about power optimization and introduce our utilized method. Power dissipation in CMOS digital circuits consists of dynamic power, short circuit power and leakage power. The short circuit power is usually negligible compared to dynamic power and leakage power. Therefore, most of the optimization works focus on the latter two sources of power consumption. Generally, the dynamic power is insensitive to process variations and can be assumed to be deterministic [63]. The leakage power is greatly affected by physical parameters with uncertainties because of manufactured process variations, and needs to be treated as random processes. Existing useful power reduction methods at the circuit-level are the supply voltage scaling, threshold voltage scaling, gate-oxide scaling, gate-sizing, retiming, and any combination of these methods.

2.5.1 Statistical Leakage Power Optimization and Deterministic Dynamic Power Optimization

A performance optimization based on the criticality is proposed in [64]. By modeling the statistics of leakage and delay as posynomial functions, authors in [65] formulate a geometric programming problem and solve it by the convex optimization method. In [66], it is formulated as an unconstrained nonlinear optimization problem and solved based on the efficient power and delay gradient computation. A statistical power optimization algorithm under the timing yield constraint is presented in [67], where the second order cone programming is employed. Some sensitivity-based heuristic methods are proposed to reduce the leakage power [68, 69]. The above works utilize the techniques of gate-sizing and dual-threshold voltage to reduce the leakage power statistically. However, most of them neglect the importance of dynamic power.

On the other hand, many researches use the multiple supply voltages to do the deterministic dynamic power optimization [70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84]; nevertheless, the leakage power is ignored in their experiments. Although several studies consider the dynamic and leakage power at the same time, they still have their limitations. Authors in [85] propose an algorithm based on the linear programming. The genetic algorithm is employed in [86] to do the power optimization. A two-phase flow is presented in [87] to minimize the power consumption. By the use of retiming and Vdd/Vth scaling, the authors [88] formulate the problem by using the integer linear programming approach. These studies are all deterministic methods. Although the work in [67] considers the total power reduction statistically, it does not take the temperature influence into account.

2.5.2 Multiple Supply Voltage (MSV) Technique

The concept of MSV method is to assign lower supply voltages to gates on the non-critical path for power saving and assign higher supply voltages to gates on the critical path for satisfying the timing constraint. It provides the premium result with less penalty [89]. The two constraints introduced by the MSV method are the electrical constraint and the physical constraint. In a voltage-scaled circuit, if a low supply voltage gate drives a high supply voltage gate, a level converter (LC) must be inserted to eliminate the undesirable static current. The additional level converters would increase the cost in area, delay and power; hence, the number of level converters must be controlled. Moreover, cells operating at different supply voltages should be placed carefully to facilitate the power network design and reduce the routing complexity. Previous efforts toward reducing the level-shifting overhead include: 1) clustered voltage scaling (CVS) and 2) extended CVS (ECVS). The CVS partitions a circuit into two clusters - one having only cells operating at high supply voltage and the other having only cells operating at low supply voltage. The scenario in which a cell driven by low supply voltage directly feeds a cell driven by high supply voltage is clearly precluded in this partition. The ECVS relaxes this topological constraint and allows a cell with low supply voltage to feed a cell with high supply voltage after its output has undergone level conversion. Thus, ECVS has more freedom in finding parts of the circuit that can be operated at the lower supply voltage and can potentially lead to higher power saving. However, the delay penalty tends to be larger too. An effective solution is grouping cells of different supply voltages into a small number of “voltage islands”, where each voltage island occupies a contiguous physical space and operates at a single supply voltage and meets the performance requirement.

Logic boundaries are largely used in this grouping process mainly because they are the boundaries that designers are most familiar with. Nevertheless, these natural boundaries in a design are almost always nonoptimal boundaries for supply voltages. Fig. 2.3 illustrates why sticking to logic boundaries is limiting the solution space in producing optimal MSV. In the example, there are three modules, each of them contains only leaf cells, and both modules A and B contain some timing-critical cells that require high voltage Fig. 2.3(a). Fig. 2.3(b) and (c) are the designs based on logic boundaries. While Fig. 2.3(b) guarantees the performance using high power, Fig. 2.3(c) reduces the power consumption without meeting the timing requirement. None of them are optimal MSV. By using placement proximity (instead of logic) information, the optimal MSV meets power and timing requirements at the same time while keeping the number of power domains small as shown in Fig. 2.3(d). This idea is called post-placement voltage island generation [72]. Due to the advantages of the post-placement voltage island, we employ this methodology in our experiment. Another reason is that the location of each gate is provided in the post-placement stage so the spatial correlation can be considered in the SSTA.

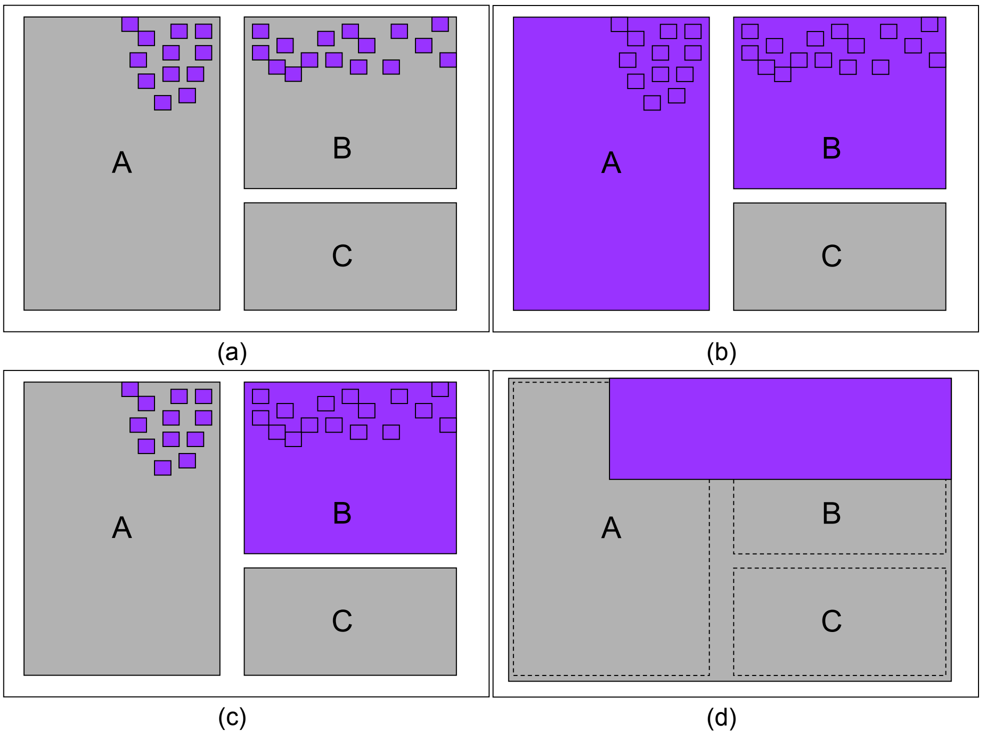


Fig. 2.3: (a) Design with timing-critical cells (small purple cells). (b) Power consumption too high. (c) Timing requirement not met for small cells in module A. (d) Placement-proximity-based solution with nonlogical boundary.

2.5.3 Previous Works of Post-Placement Voltage Island

The post-placement voltage island can be performed in two stages: supply voltage assignment [73] and voltage island generation [72]. Using the concept of the zero slack algorithm and the Voronoi diagram, authors in [73] propose a proximity-driven-voltage-assignment algorithm. Based on the placement and the voltage requirement of each cell, they continue implementing an efficient algorithm to find the voltage islands for the best tradeoff between the total power and the number of islands [72]. The method developed in [74] allows the generated voltage islands to be any shape instead of only rectangular [72].

Chapter 3

Statistical Power Optimization in 3D ICs

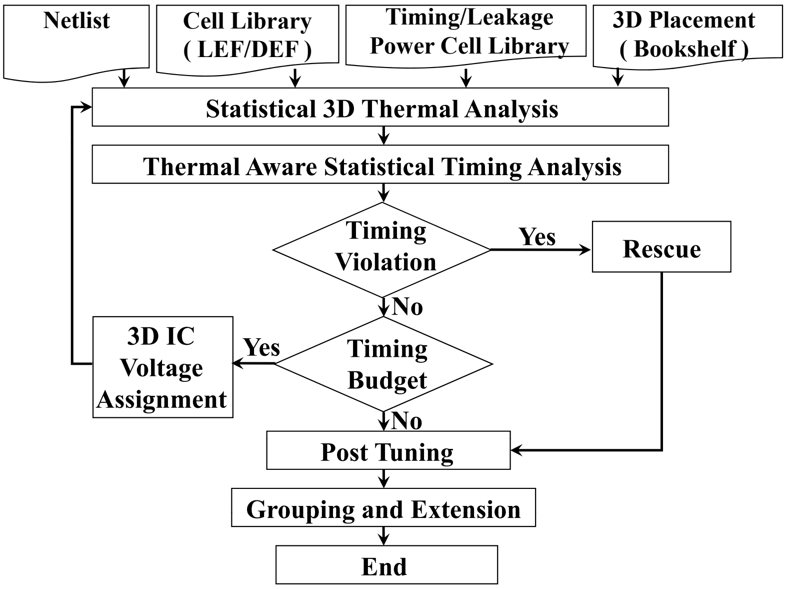


Fig. 3.1: Flowchart of the proposed statistical power optimization for 3D ICs

3.1 Problem Formulation and Flowchart

The proposed power reduction design flow for 3D ICs is shown in Fig. 3.1. Given a known placement, design netlist and a standard cell library, the flow first executes a statistical thermal simulation to obtain the statistical temperature distribution for the specified 3D IC. After that, a thermal aware SSTA is performed with the statistical temperature distribution got from the previous step. The slack data provided by SSTA is used to compute the power-delay sensitivity; next, a grid-based procedure is developed for the voltage assignment. After the assignment procedure, the power consumption and delay of gates are changed so we do the thermal and

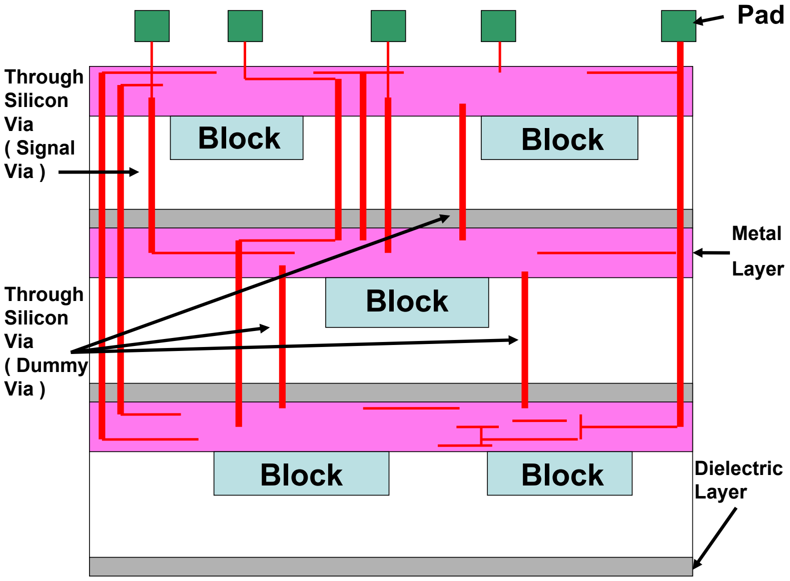


Fig. 3.2: The schematic diagram of a 3D IC with 3 chip layers

timing analysis again, and the timing of the circuit is verified. If there is any timing violation, a rescue procedure is enforced to assure the timing correctness and finish the iteration. When the circuit satisfies the timing constraint, the program starts the assignment process again. The program also terminates the loop when the iteration can not provide more improvement. Then, a post-tuning step is employed to further lower the power consumption. In the last process, the proposed method uses grouping and extension to implement the voltage island generation. Each executing step will be described in the following sections.

3.2 Statistical 3D Thermal Analysis

The utilized statistical electro-thermal aware 3D IC thermal simulator extends the Hermite polynomial chaoses (H-PCs) based 2D IC statistical thermal simulator [90, 91] to the 3D IC, and combines the developed 3D IC statistical thermal simulator with the electro-thermal iterative updating loop. The details of the algorithm are presented in the Appendix A. The entire thermal simulation flow is summarized in Fig. 3.4. The simulation can be divided into two stages; we introduce the contents of the method in the following subsections.

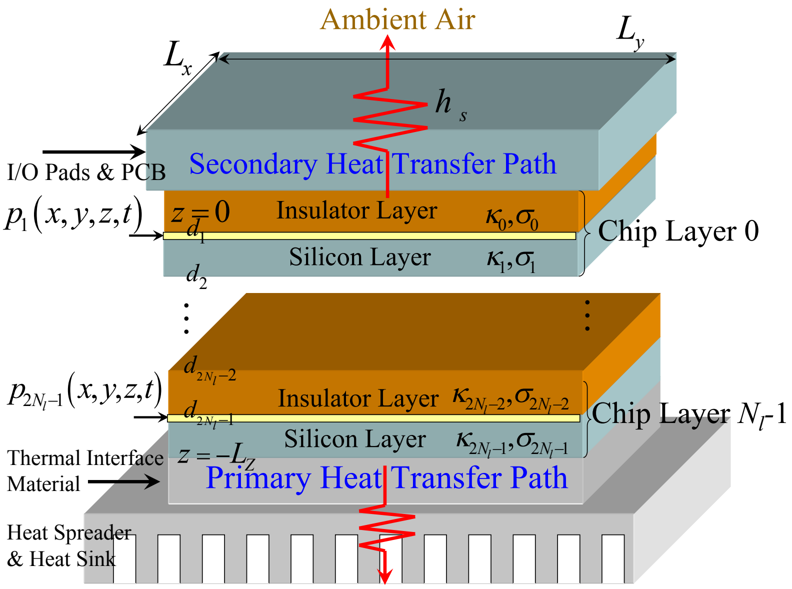


Fig. 3.3: The schematic diagram of a 3D IC with N_l chip layers

3.2.1 Thermal Model of 3D IC

Fig. 3.2 is the schematic representation of 3D integration, and Fig. 3.3 is its compact thermal model. As shown in Fig. 3.3, the structure of 3D IC is a multi-layer structure with stacking silicon and insulator layers one by one [22, 29]. This model consists of three portions [92]: the primary heat flow path, the secondary heat flow path, and the heat transfer characteristic of each macro/block on the silicon die. The primary heat flow path is composed of thermal interface material, heat spreader and heat sink. The secondary heat flow path contains I/O pads and the print circuit board (PCB). The functional blocks are modeled as many power generating sources distributed in a thin layer close to the top surface of each active silicon layer in the z-direction, and each insulator layer consists of Cu, ILD and glue materials.

3.2.2 First Stage

In the first stage, given an initial temperature, we substitute it into (2.10) and multiply the leakage currents shown in (2.8) and (2.10) with supply voltages to get the leakage power P_{sub} = $V_{DD}I_{sub}$ and P_{gate} = $V_{DD}I_{gate}$. Then, we build the projected leakage power cell library for each type of gate in every KL grid of each layer, and we use the 1-D thermal model and the mean of the power to set the thermal parameters and do the thermal simulation. After getting the new temperature, we rebuild the subthershold leakage power cell library by the average mean temperature. Note that the gate leakage power is temperature independent so we only calculate it once in the whole flow. The above steps are performed iteratively until the average mean temperature converges; thus we get a more accurate initial temperature.

3.2.3 Second Stage

In the second stage, we first compute the power with the more accurate initial temperature got in the first stage and obtain the mean temperature by using an efficient 3D deterministic GIT thermal simulatior [91]. Then, we obtain the first order PC expansion of the multi-layer temperature distribution.

$$T(\mathbf{r}, \varpi, \theta) \simeq \sum_{k=0}^{N_{KL}} T_k(\mathbf{r})\xi_k, \tag{3.1}$$

where each $T_k(\mathbf{r})$ is the coefficient function of the temperature projection onto the k-th H-PC, $\mathbf{r}$ is the position. The spatial mean and variance distributions of the full-chip temperature distribution can be obtained as

$$\begin{aligned} E\{T(\mathbf{r}, \varpi, \theta) + T_a\} &\approx T_0(\mathbf{r}) + T_a, && (3.2) \\ Var\{T(\mathbf{r}, \varpi, \theta) + T_a\} &\approx \sum_{k=1}^{N_{KL}} T_k^2(\mathbf{r})E\{\xi^2\}. && (3.3) \end{aligned}$$

The temperature is treated as a random process and is substituted into (2.10) to rebuild the subthreshold leakage power cell library. Those new leakage values are used to analyze the temperature distribution on the chip statistically. The above steps are performed iteratively until the mean and variance of the temperature converge.

3.3 Temperature Aware Statistical Timing Analysis

In this work, our SSTA is based on the widely used block based algorithm [51]. The delay is expressed in the canonical first-order form below:

$$Delay = a_0 + a_1 L_{eff} + a_2 t_{ox} + a_3 T, \tag{3.4}$$

where a_0 is the nominal value, and a_i's which are V_{DD} dependent are the sensitivities to the variations. As the leakage current, we evaluate the average delay by H-SPICE simulation and

```
Input: Fitted cell leakage power, Dynamic power, Placement, Spatial
       correlation function
Output: The spatial mean and variances for the full–chip temperature
        distribution
    First stage
        1. Parameter modeling by Karhunen–Loeve expansion
        2. Build projected power cell library of each layer with
           the given initial temperature
        3. Use 1–D thermal model and average mean temperature to
           set the thermal parameter
        4. Re–build the subthreshold leakage power library by average
           mean temperature
        repeat steps 2∼4 until the mean temperature converges
    Second stage
        1. Obtain mean temperature by using 3D deterministic GIT
           thermal simulator
        2. Obtain the first order PC expansion of the multi–layer
           temperature distribution
        3. Re–build the subthreshold leakage power library
        repeat steps 1∼3 until means and variances of temperature
```

Fig. 3.4: Algorithm of the statistical thermal simulation.

fit the coefficients based on the least square method. The KL expansion is employed to manipulate the complicated spatial correlation of L_{eff} and t_{ox}. Combined with the approximated temperature obtained in the previous thermal analysis stage, the delay formulation for a specific type of functional gate located at $\mathbf{r}^* = (x^*, y^*, z^*)$ becomes

$$\begin{aligned}
Delay &= a_0 + a_1 L_{eff} + a_2 t_{ox} + a_3 T \\
&= a_0 + a_1 L_{eff}(x^*, y^*, z^*, \theta) + a_2 t_{ox}(x^*, y^*, z^*, \varpi) + a_3 T \\
&= a_0^{'} + a_1 \sum_{k=0}^{N_{KL}} q_k^* \xi_k + a_2 \sum_{k=0}^{N_{KL}} f_k^* \xi_k + a_3 \sum_{k=0}^{N_{KL}} T_k(\mathbf{r}^*) \xi_k \\
&= a_0^{'} + \sum_{k=0}^{N_{KL}} \left(a_1 q_k^* + a_2 f_k^* + a_3 T_k(\mathbf{r}^*) \right) \xi_k \\
&= a_0^{'} + \sum_{k=0}^{N_{KL}} a_k^{'} \xi_k, \qquad (3.5)
\end{aligned}$$

where $q_k^* = \sqrt{\chi_m} q_m(x^*, y^*, z^*)$ and $f_k^* = \sqrt{\beta_n} f_n(x^*, y^*, z^*)$, respectively. Equation (3.5) is consistent with the canonical first-order delay form in [51]. As a result, the thermal effect can be easily considered in the SSTA without any extra complexity.

After setting up the canonical form of the delay, we explain how to execute the *SUM* and *MAX* operations in the statistical timing analysis. Before showing the computation of the *MAX* operation, we introduce the concept of *tightness probability*. Given any two random variables

X and *Y*, the *tightness probability* T_X of *X* is the probability that it is larger than (or dominates) *Y*. The *tightness probability* T_Y of *Y* is (1-T_X). Below we show how to compute the maximum of two timing quantities in the canonical first-order form and how to determine their tightness probabilities. Given two timing quantities

$$\begin{aligned} A &= a_0 + \sum_{i=i}^{n} a_i \Delta X_i, \qquad (3.6) \\ B &= b_0 + \sum_{i=i}^{n} b_i \Delta X_i, \qquad (3.7) \end{aligned}$$

and the definitions

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}), \qquad (3.8) \\ \Phi(y) &= \int_{-\infty}^{y} \phi(x) dx, \qquad (3.9) \\ \theta &= (\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B)^{1/2}, \qquad (3.10) \end{aligned}$$

where ρ is the correlation coefficient, σ_A and σ_B are the standard deviation of *A* and *B* respectively, the probability that *A* is larger than B is

$$\begin{aligned} T_A &= \int_{-\infty}^{\infty} \frac{1}{\sigma_A} \phi(\frac{x - a_0}{\sigma_A}) \Phi(\frac{(\frac{x-b_0}{\sigma_B}) - \rho(\frac{x-a_0}{\sigma_A})}{\sqrt{1-\rho^2}}) dx \\ &= \Phi(\frac{a_0 - b_0}{\theta}), \qquad (3.11) \end{aligned}$$

The mean and variance of *MAX(A,B)* can also be analytically expressed as

$$\begin{aligned} E\{MAX(A,B)\} &= a_0 T_A + b_0(1 - T_A) + \theta\phi(\frac{a_0 - b_0}{\theta}), \qquad (3.12) \\ Var\{MAX(A,B)\} &= (\sigma_A^2 + a_0^2)T_A + (\sigma_B^2 + b_0^2)(1 - T_A) \qquad (3.13) \\ &\quad + (a_0 + b_0)\theta\phi(\frac{a_0 - b_0}{\theta}) - (E\{MAX(A,B)\})^2, \qquad (3.14) \end{aligned}$$

Hence we get the mean and variance of *C=MAX(A,B)*. Mathematically,

$$c_i = T_A a_i + (1 - T_A) b_i, \qquad (3.15)$$

where T_A is the tightness probability of *A*. The maximum of two Gaussians is not a Gaussian but we re-express it in the canonical Gaussian form so we can propagate the delay in the analysis. The method proposed in [60] is used to reduce the error induced by this approximation.

On the other hand, when a new timing quantity *C* is the summation of *A* and *B*, it is computed as follows

$$\begin{aligned} C &= A + B \\ &= c_0 + \sum_{i=i}^{n} c_i \Delta X_i \\ &= (a_0 + b_0) + \sum_{i=i}^{n} (a_i + b_i) \Delta X_i \end{aligned} \tag{3.16}$$

3.4 Voltage Island Generation in 3D ICs

A significant concept of three dimensional voltage island generation is that we have to consider all the tiers at the same time instead of considering each tier sequentially. An obvious reason is that the timing budget is limited. If we perform the voltage island generation tier by tier, the available timing budget will become less and less. This will result in an ill circuit which its power consumption distribution on the chip is extremely unbalanced. On the other hand, the structure of the 3D IC is different with the 2D IC because of the vertical counterpart. For example, the power consumption may be acceptable for a region in a specific tier, but its upper or lower counterpart has high power consumption, which causes the thermal problem in the vertical space. If the upper or lower counterpart can not operate at the low supply voltage, we can do the power saving in the central region.

3.4.1 Voltage Assignment

In the beginning, the timing budget of the circuit is checked to guarantee that the power reduction is available. The program stops the execution when the timing budget is not sufficient; otherwise, the assignment procedure is performed. After the verification, each tier is partitioned into many grids, and the sensitivity of each grid i ($grid_i$) is defined as the summation of each gate's sensitivity in $grid_i$.

$$Sen_{grid_i} = \sum_{gate_j \in grid_i} Sen_{gate_j}. \tag{3.17}$$

The number of *sites* in each grid is also recorded.

The next step is to vertically compress the three dimensional structure into a two dimensional planar, and several criteria are accumulated for each compressed grid. According to the importance, these criteria include: 1) the power-delay sensitivity, 2) the power density induced weights and 3) the total number of *sites*.

The power-delay sensitivity is the most important criterion since the power dissipation reduction is the primary target. A low (high) supply voltage reduces (increases) the power consumption but slows (accelerates) the speed of the gate. To utilize the timing budget effectively and achieve the most power saving, the power-delay sensitivity metric [93] is used as our optimization criterion. For each gate i, the power-delay sensitivity [93] is

$$Sen_{gate_i} = \begin{cases} \frac{\Delta P_i}{\Delta D_i} Slack_i, & Slack_i \geq \Delta D_i; \\ 0, & \text{else;} \end{cases} \tag{3.18}$$

where $Slack_i$ is the timing slack of gate i, ΔP_i and ΔD_i are the power dissipation difference and the delay difference for providing the low supply voltage for gate i instead of the high supply voltage, respectively. The gate which does not have enough slack for using the low supply voltage is called a *site*. For the purpose of the statistical power-delay sensitivity computation, each ΔD_i is fitted by the least square method as the first-order form with respect to L_{eff}, t_{ox} and T which is similar to equation (3.4), and each ΔP_i is fitted by the least square method as the exponential form with respect to L_{eff}, t_{ox} and T which is similar to equations (2.8) and (2.10). With equations (2.4)(2.5)(3.1) and a similar derivation to equation (3.5), for each gate i, the expressions of ΔP_i and ΔD_i are

$$\begin{aligned} \Delta P_i &= \Delta P_{dynamic_i} + \Delta P_{sub_i} + \Delta P_{gate_i} \\ &= p_{i_0} + c_{i_0}^{'} \exp\left(\sum_{k=1}^{N_{KL}} c_{i_k}^{'} \xi_k\right) + b_{i_0}^{'} \exp\left(\sum_{k=1}^{N_{KL}} b_{i_k}^{'} \xi_k\right) \end{aligned} \tag{3.19}$$

$$\Delta D_i = d_{i_0} + \sum_{k=1}^{N_{KL}} d_{i_k} \xi_k. \tag{3.20}$$

After executing our thermal aware SSTA, $Slack_i$ can also be obtained as the following canonical first-order form

$$Slack_i = s_{i_0} + \sum_{k=1}^{N_{KL}} s_{i_k} \xi_k. \tag{3.21}$$

With the definition of the power-delay sensitivity metric in equation (3.18), the following statistical power-delay sensitivity metric is applied in our power reduction flow and Sen_{gate_i} is re-deifined as

$$Sen_{gate_i} = \begin{cases} \frac{E\{\Delta P_i\}}{E\{\Delta D_i\}} \times E\{Slack_i\} & , \mathbf{prob}_i \geq \eta; \\ 0 & , \text{else.} \end{cases} \tag{3.22}$$

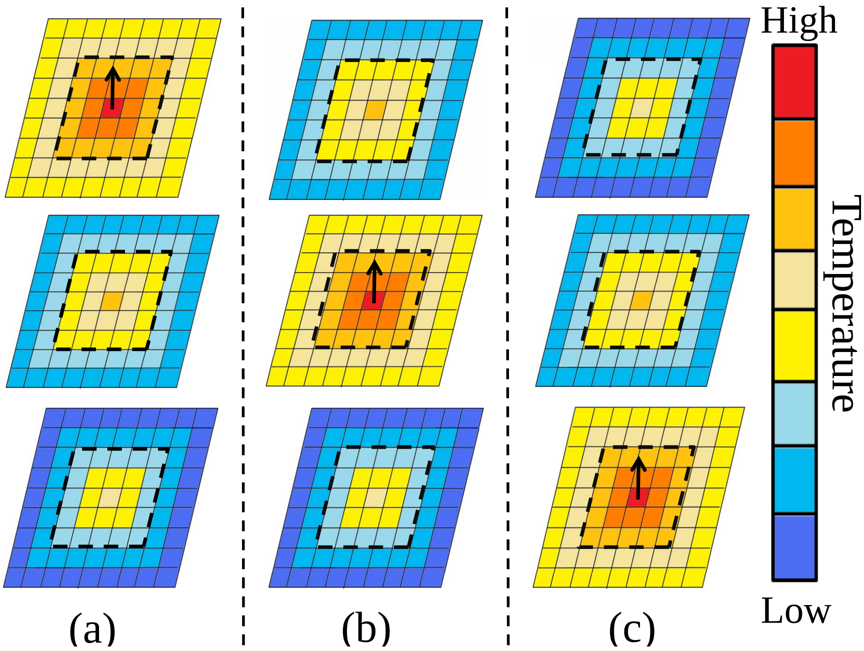


Fig. 3.5: A three-tier chip example for constructing the experimental model of power density induced weights. (a) An 1W power is inserted into the grid (5, 5) on the tier 1 and the corresponding temperature profile on each tier. (b) An 1W power is inserted into the grid (5, 5) on the tier 2 and the corresponding temperature profile on each tier. (c) An 1W power is inserted into the grid (5, 5) on the tier 3 and the corresponding temperature profile on each tier. The color bar shows the related levels of the temperature with respect to the colors on tiers. The arrows indicate the 1W power sources. The rectangles with dotted line margins are the truncation regions.

Here, $\mathbf{prob}_i = \mathbf{prob}\left(Slack_i \geq \Delta D_i\right)$, which is the *tightness probability* of $Slack_i$ and can be obtained in constant time by performing table look-up, and η is a user-specified threshold value.

The second important criterion is the power density induced weights among the compressed grids because the hot-spot issue must be avoided during executing our power reduction method. The power density induced weights of grids are obtained as follows. First, we make an experimental model for the power density induced weights. As shown in Fig. 3.5, an 1W power source is orederly inserted to the central gird of each tier and its correponding temperature profiles of tiers are obtained. The above step is to approximate the spatial impulse response of the heat transfer equation for a specified 3D IC. Then, these correponding temperature profiles are normalized in the range from 0 to 1 by using the maximum temperature among all of these correponding temperature profiles. Finally, the truncation regions of these normalized temperature profiles are obtained by a specified threshold value.

The above truncation regions can be pre-characterized before the placement because the spatial impulse responses for the heat transfer equation only depends on the package structure, chip dimension and the thermal parameters of a specified 3D circuit. Therefore, it can be re-used during our power reduction method being executed. Although the spatial impulse responses of the heat transfer equation for a specified 3D circuit are various in different locations, the weights in the truncation regions of grids are slightly different in our experimental results. For the efficiency, all of grids are set to share the same weights in their truncation regions in our implementation[1]. With these truncation regions, the power density induced weight of grid j induced by grid i can be obtained as the product of the power density of grid i with the weight of the truncation regions which corresponds to grid j. Consequently, the accumulated value of each grid is the summation of the power density induced weights from itself and from the neighboring grids among tiers.

Here, power density of grid i is obtained by using $\mu_{P_{T_i}} + 3\sigma_{P_{T_i}}$ to ensure the thermal safety. P_{T_i} is the power consumption in grid i. $\mu_{P_{T_i}}$ and $\sigma_{P_{T_i}}$ can be computed by using the means and variances of powers of various types of gates in grid i. For a specific type of gate in grid i, mean and variance of the power can be obtained by substituting equations (2.4)(2.5)(3.1) into equations (2.8)(2.10) and then obtain the means and variances of equations (2.8)(2.10). For example, after substituting equations (2.5)(3.1) into equation (2.10), we obtain the following computational formulas for mean and variance of the subthreshold leakage power for a type of gate in grid i.

$$E\{P_{sub}\} = \hat{c}_0 \exp\left(\frac{1}{2}\sum_{k=1}^{N_{KL}} \beta_k^2\right), \tag{3.23}$$

$$Var\{P_{sub}\} = \hat{c}_0^2 \exp\left(\sum_{k=1}^{N_{KL}} \beta_k^2\right), \tag{3.24}$$

where $\hat{c}_0$ and each β_k are known values consisted of supply voltage and the leading coefficients of ξ_k of L_{eff} and $T(\mathbf{r}, \theta, \omega)$ in equations (2.5)(3.1).

The gate called *site* will violate the timing constraint when it is assigned with the low supply voltage so we must avoid generating too many *sites* during the proposed assigning procedure. In fact, this criterion is related to the power-delay sensitivity criterion because the *site* contributes zero sensitivity, and the grid with the high sensitivity should not contain too many *sites*.

[1] If a more accurate result is required, truncation regions with different weights can be obtained for the grids in the regions of the center and the corners of a specified 3D circuit.

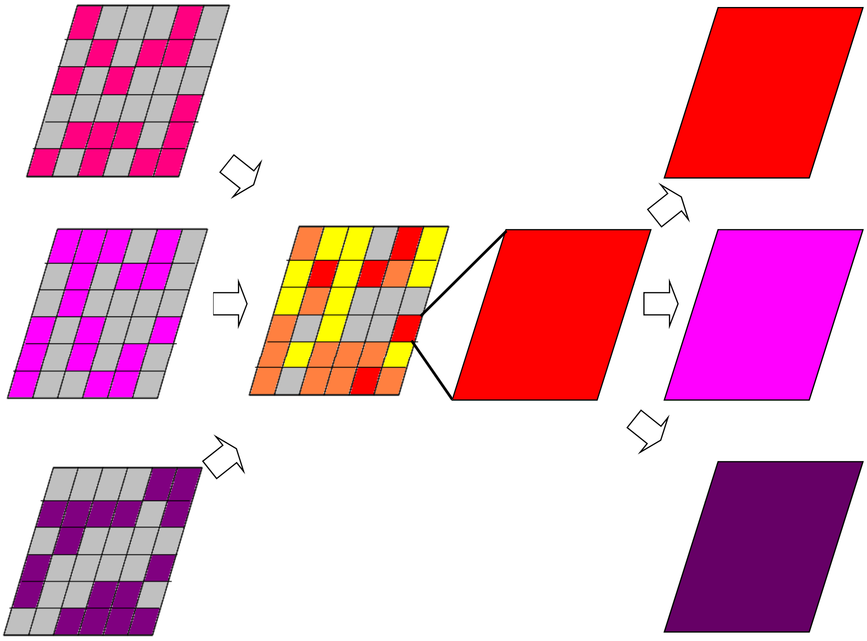


Fig. 3.6: A three-tier design example of the grid-based procedure for generating the voltage islands. Each tier is first divided into many grids, and the grid with the high sensitivity has dark color. After compressing, the criteria are accumulated, the higher priority the grid has, the darker the color 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.

After compressing, the priority of each compressed grid is decided by these three criteria, and the grid with the highest priority is selected. Then, the selected compressed grid is restored back to the multi-layer structure. This procedure is helpful for us to understand whether the accumulated sensitivity is contributed by a grid on a specific layer or by all the grids on every layer averagely and to decide the grid for using the low supply voltage. Finally, every gate in this grid is assigned with the low supply voltage. When there are more than two supply voltages, the designer can decide to lower the supply voltage to next level or to the smallest level. A three-tier design example of the above grid-based procedure for the voltage assignment is illustrated in Fig. 3.6.

3.4.2 Incremental Update

When a grid is assigned with the low supply voltage, the timing and power-delay sensitivity information of many gates are affected and should be updated. To get the exact slack and sensitivity information in the circuit, performing the SSTA after every low supply voltage assignment is an instinctive but costly idea. For reducing the computational load, we use an incremental approach to approximately update the sensitivity. The basic idea is borrowed from the STA. For example, the gate C in Fig. 3.7 has two input signals which are from the outputs of gate A and gate B. The AT of gate A and gate B are 2 and 5 by using the STA, respectively; the AT of C is determined by the latest arriving input B. If gate A is decided to operate at the low supply voltage and the AT of A increases to be 4 now, C is not affected because B is still the slowest input. Nevertheless, if the AT of A becomes 6, the AT of gate C is changed and the increment is 1. To conclude, the low supply voltage assignment will impact C only when the new AT exceeds the dominated one.

This similar concept can be used in the statistical manner by replacing the deterministic delay by the random variable delay and using the available mathematical method to do comparison between two random variables in probabilistic forms. In this example, the gate C records its latest arriving input $M = MAX(A, B)$ as shown in Fig. 3.7(d) after the initial timing analysis. When the low supply voltage is assigned to gate A, the AT of A is changed to A'. Then, we compare these two random variables. If $\mathbf{prob}(A' \geq M) \geq \alpha$, where α is a user-defined bonding parameter, it means that gate C is heavily affected by this assignment. Therefore, the new latest arriving input becomes to $M' = MAX(A', M)$ and the arrival time of C is updated. Similarly, the required time of gate can be updated by the smallest one. After the update of AT and RT, the slack and power-delay sensitivity information of gate are updated.

3.4.3 Site Rescue

The next issue is about the *sites* in the circuit. Although the grid with *sites* is not assigned with low supply voltage, timing violations may still happen because we reduce the supply voltage for the gates in one grid at the same time and update the timing approximately. In the timing-check stage, if any gate violates the timing constraint, the program will use several ways to recover the timing. The first method is the placement refinement. Since the *site* can not operate at the low supply voltage, we try to find a *replacer*, which is a gate with the largest sensitivity in the neighboring grids operating at the high supply voltage, to exchange its location with the *site* as shown in Fig. 3.4.5(a). If a *replacer* is not available, the *site* will be pushed to a neighboring grid operating at the high supply voltage as illustrated in Fig. 3.4.5(b). The second method is the gate-sizing. According to the cell library, the program selects a size for the *site* to satisfy

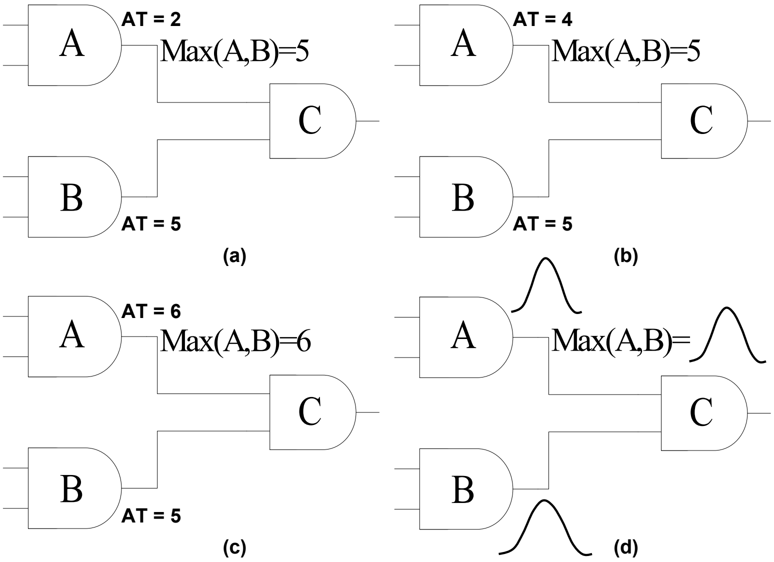


Fig. 3.7: (a) Initial state. (b) The low supply voltage is assigned to gate A, but the AT of gate C is not affected. (c) The low supply voltage is assigned to gate A, and the AT of gate C is affected. (d) Statistical form.

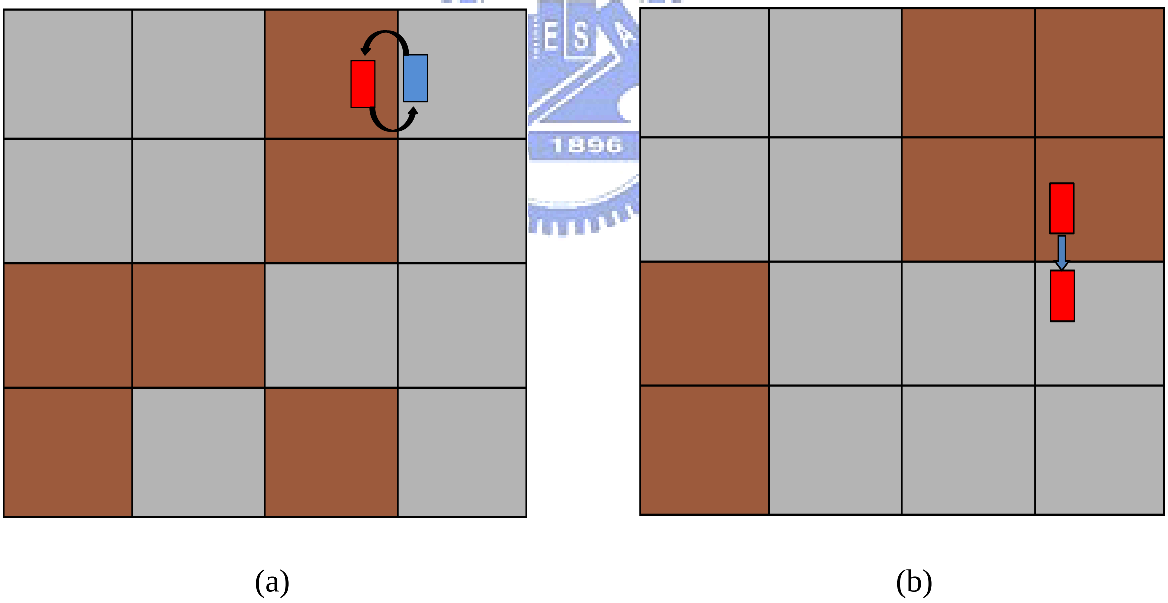


Fig. 3.8: The grids with dark color operate at low supply voltage and the grids with light colors operate at high supply voltage. The red block is the *site* and the blue block is the *replacer*.

the timing constraint with less power dissipation. Although sizing up the gate size will increase a little power consumption, it is worth because of the large power saving from other gates in the grids. Lastly, if these techniques fail to recover the timing, the design will return to the last timing-satisfied condition with the least power dissipation.

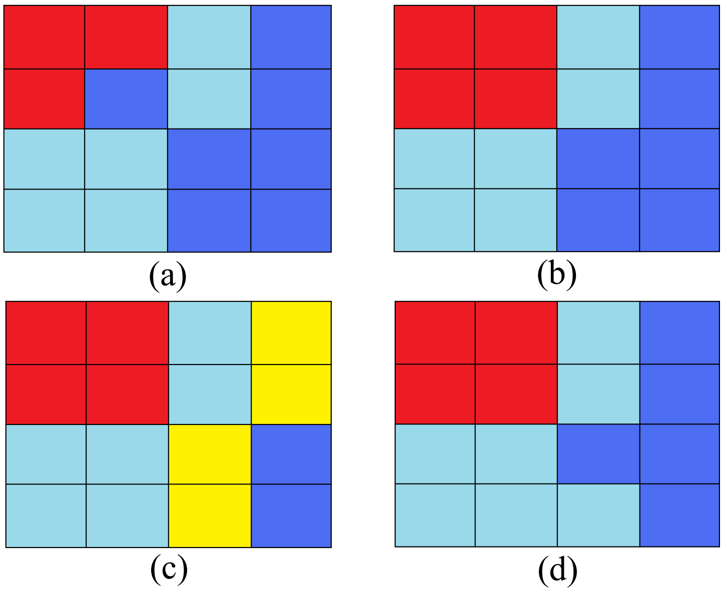
3.4.4 Post Tuning

The proposed voltage assignment and incremental update approach are very efficient in runtime. The avoidance of timing violations and the grid-based structure nevertheless limit the assignment to the grid without any *site* and sacrifice some gates that have potential power saving. Hence, we provide a slower but finer gate-based post tuning step to further improve the assignment result. This gate-based procedure uses the same statistical power-delay sensitivity as guidance. Once the gate with highest sensitivity is found, it is assigned to low supply voltage for power reduction. Through a SSTA run, the action is accepted unless the timing constraint is violated. This process repeats until no further power reduction can be made.

3.4.5 Grouping and Extension

After the voltage assignment and post tuning, methods proposed in [72, 74] can be utilized to complete the voltage island generation. Based on their idea, there are still several improvements which can be exploited. Fig. 3.9 illustrates this idea. In this figure, red color means the highest supply voltage, and dark blue and light blue represent the middle and the least supply voltage, respectively. Fig. 3.9(a) is the result of voltage assignment. Fig. 3.9(b) is the result of process in [72]. A dark blue grid is annexed to the red grids to reduce the number of islands. Since the timing resource belonged to the dark blue grid is released due to this mergence, the island with lower supply voltage can extend its range. When there are more than two supply voltages, the timing resource can be utilized by any supply voltage except the highest one. In this example, for the least voltage, the yellow grids are the possible regions for the extension as shown in Fig. 3.9(c). Fig. 3.9(d) is the result of the island extension, so the power consumption is further reduced.

Initially, every grid with low supply voltage is viewed as an single island. The proposed method first scans the topology and groups the single islands that are adjacent to each other into big islands. The action results many *base islands* which may be composed of only one grid or several grids and each *base island* has a size which is equal to the number of grids it contains. Second, some *base islands* are deleted when their size is smaller than an user-specified size bound. This is similar to [72] since these small islands can consume many power network


(a) (b)

(c) (d)

Fig. 3.9: Voltage Island Extension

resource and they would be merged like Fig. 3.9(a)(b). The process next finds the *neighbors*, which are the grids next to the *base islands*. Then, these *base islands* start to extend their region to the *neighbors* with the resource from the deletion as shown in Fig. 3.9(c)(d). The power-delay sensitivity is still used as guidance and the extension is allowed when there is no timing violation. Eventually, the voltage islands are generated.

Chapter 4

Experimental Results

The proposed approach has been implemented in C++, and been applied to a set of ISCAS89 benchmark circuits and private designs. The benchmark circuits are synthesized by using Design Compiler with the UMC 90 nm standard cell library. After that, the SOC Encounter is used to generate an initial 2D placement. Then we transform the 2D placement to a 3D placement with Z–Place provided by professor Renato [94]. The ranges of the process variations (3σ) are: L_{eff}: 20 $\%$, t_{ox}: 20 $\%$. The Karhunen–Loeve transformation is employed to deal with the physical parameters with spatial correlation. The number of reference points is set to be 16 for the parameter modeling of the channel length and the oxide thickness. For the delay constraints, we consider the timing yield target of 99 $\%$. The timing/leakage power cell library with process variations is generated as follows. We evaluate the average leakage current and gate delay based on H-SPICE simulation for various types of logic gates. From the H-SPICE simulation results, we obtain the fitting constants of the leakage current and gate delay models based on the least square method.

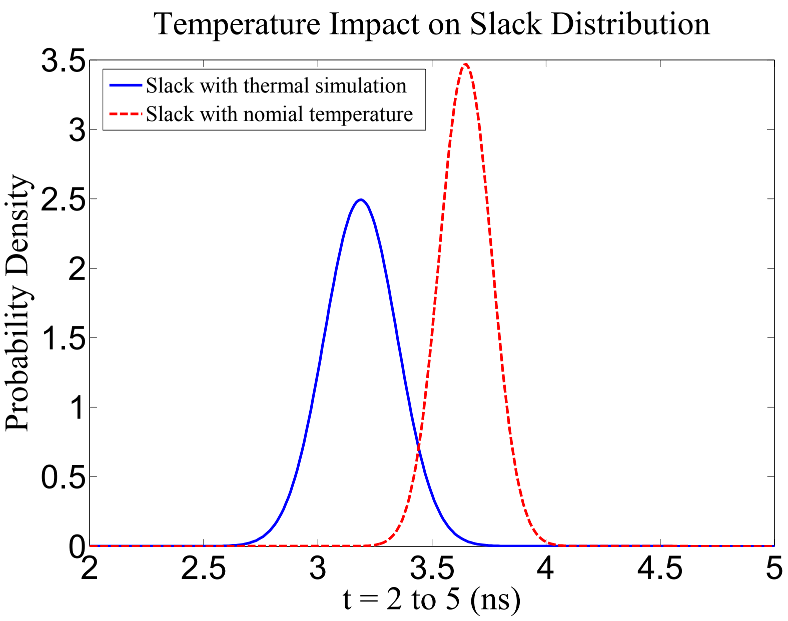


Fig. 4.1: Temperature impact on slack distribution.

- **Thermal Aware SSTA**

Our first comparison gives the impact of temperature on delay computation. Fig. 4.1 shows the slack distribution both with the statistical 3D thermal simulation result and the nominal temperature in [19]. As observed in this figure, for the slack distribution, the mean will decrease and the variance will increase when statistical thermal simulation is utilized. Since thermal problem is one of the critical challenges in 3D IC design, considering the temperature impact in circuit analysis is an essential work.

Table 4.1: Leakage Power Estimation

	With Statistical Temperature		With Nomial Temperature		Difference(%)	
Circuit	Leakage Power (μW)	Total Power (μW)	Leakage Power (μW)	Total Power (μW)	Leakage Power	Total Power
s1488	46.15	560.03	31.51	555.64	31.71	0.78
s1494	46.47	564.10	31.93	559.74	31.30	0.77
s5378	1144.90	2190.06	104.71	1878.00	90.85	14.25
s9234_1	490.82	1514.86	99.35	1397.42	79.76	7.75
s13207	1180.84	2461.10	185.81	2162.59	84.26	12.13
s35932	3381.22	6682.40	1122.90	6004.90	66.79	10.14
s38417	5025.33	10312.62	904.73	9076.44	82.00	11.99
s38584	3257.71	7149.88	847.62	6426.85	73.98	10.11
Circuit 1	1701.47	5281.69	1217.42	5136.48	28.45	2.75
Circuit 2	5884.27	9204.04	5161.80	8987.30	12.28	2.35
Circuit 3	3516.46	6085.30	2875.25	5892.94	18.23	3.16
Avg.					54.51	6.93

- **Temperature Impact ono Leakage Power Estimation**

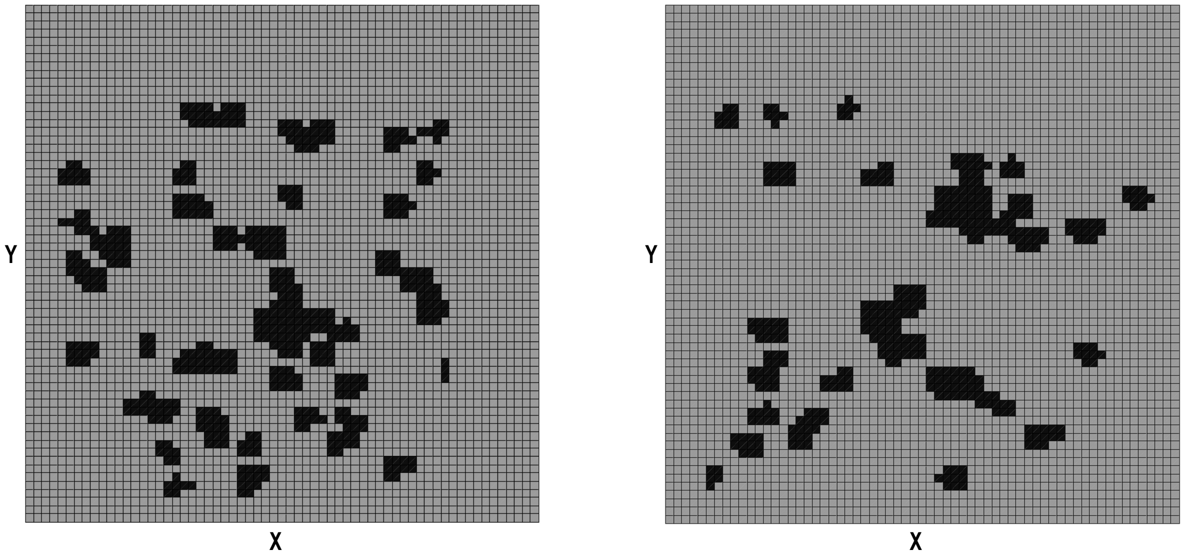
  Table 4.1 lists the leakage power and total power estimation with the simulated temperature (columns 2, 3) and the nominal temperature in [19] (columns 4, 5), the percentage differences of leakage power, and total power (columns 6, 7), respectively. Here, we suppose the ratio of the leakage power to the total power is 30 % [16]. As show in this table, the full chip leakage power analysis without accurate temperature can lead to 54 % error in average. When the leakage power is underestimated, the optimization work would be dominated by the dynamic power, which may result a non-ideal design.

Table 4.2: Optimization Result

		Initial Power(μW)		Optimized Power(μW)		Improvement			
Circuit	#Gates	Dynamic	Leakage	Dynamic	Leakage	Dynamic($\%$)	Leakage($\%$)	Total($\%$)	ΔT_{max}
s1488	288	780.26	46.15	619.22	29.84	20.64	35.34	21.00	11.523
s1494	294	785.95	46.47	593.30	27.10	24.51	41.69	24.94	13.459
s5378	710	2637.98	1144.90	2386.67	662.85	9.53	42.10	14.64	15.451
s9234_1	596	1953.73	490.82	1478.29	236.22	24.33	51.87	27.01	18.467
s13207	919	3009.78	1180.84	2442.97	612.02	18.83	48.17	23.06	17.252
s35932	5496	8097.19	3381.22	6570.67	2317.17	18.85	31.47	20.77	6.580
s38417	5208	12578.60	5025.33	12089.60	4577.34	3.89	8.91	4.62	3.218
s38584	5581	8817.95	3257.71	7197.77	1911.05	18.83	41.34	21.51	12.065
Circuit 1	8819	6816.07	1701.47	5875.38	1482.88	13.80	12.85	13.71	2.114
Circuit 2	16285	10626.80	5884.27	8273.69	4956.71	22.14	15.76	20.92	3.156
Circuit 3	11077	7186.23	3516.46	5633.85	3207.17	21.60	8.80	19.38	1.972
Avg.						18.43	31.30	19.23	9.569

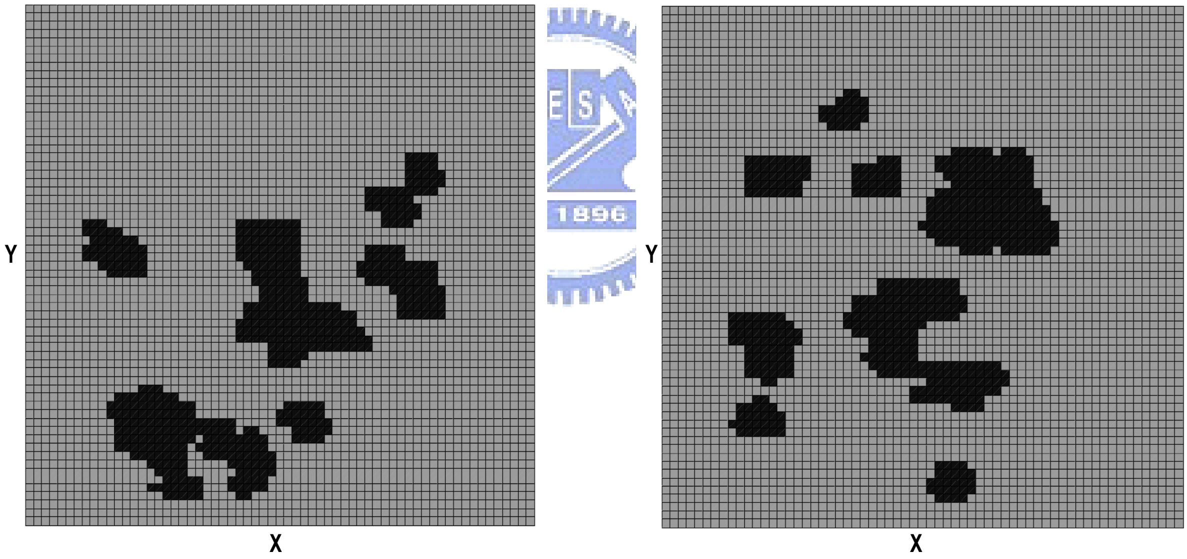
- **Power Reduction**

  Table 4.2 lists the optimization results of each circuit. In this table, we show the initial power (columns 3, 4), the power after optimization (columns 5, 6), and the percentage improvements of dynamic power, leakage power, total power and the decrement of temperature (columns 7–10). It indicates that this method can provide almost $20\,\%$ power saving and 9 degree decrement in temperature in average. Here, ΔT_{max} is the maximum μ + 3σ temperature among tiers of each test cases. Notice that the leakage power reduction is very well due to the temperature improvement.



(a) (b)

Fig. 4.2: (a) Voltage assignment result on layer 1. (b) Voltage assignment result on layer 2.

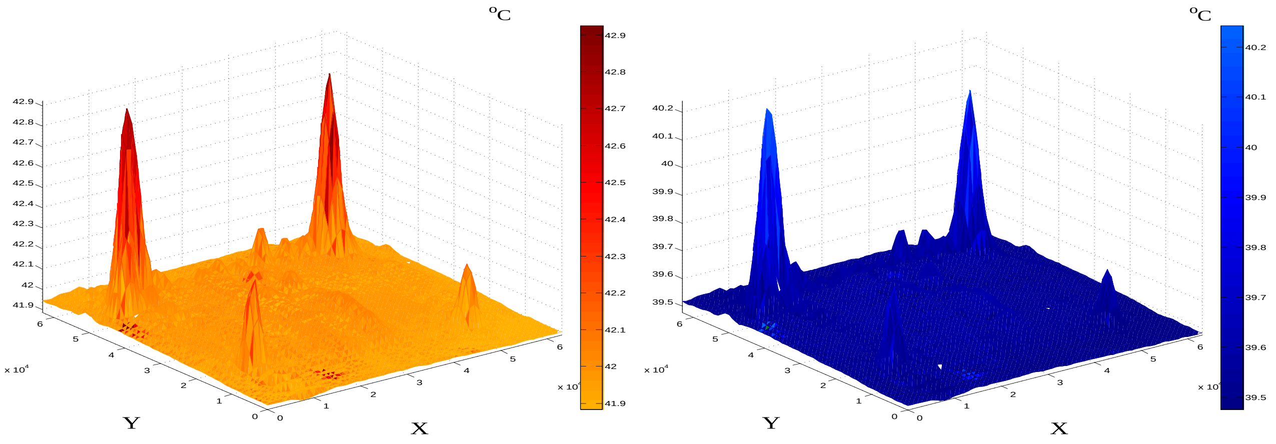


(a) (b)

Fig. 4.3: (a) Voltage islands on layer 1. (b) Voltage islands on layer 2.

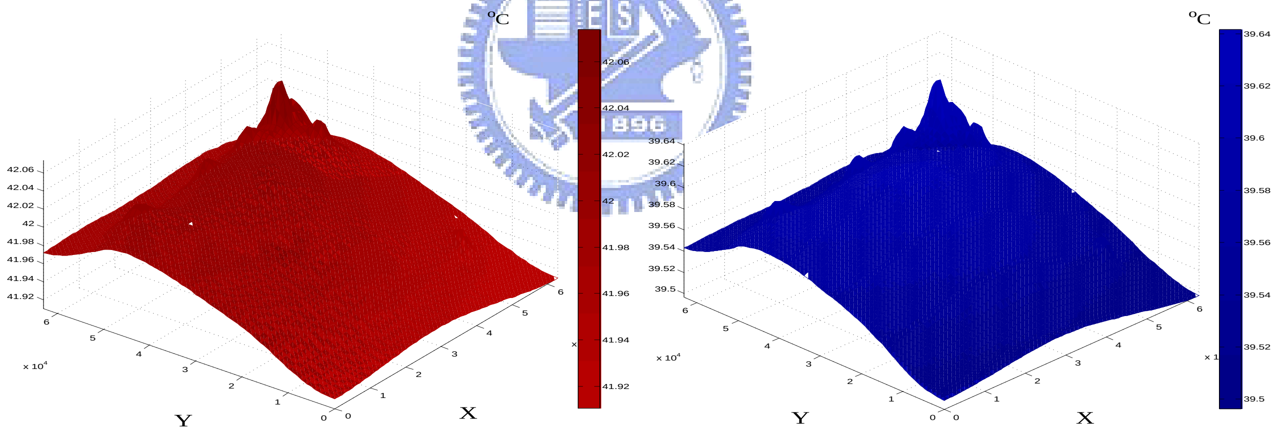
- **Voltage Island**

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. 4.2 illustrates the voltage assignment result of Circuit 3; Fig. 4.3 shows the corresponding voltage islands generated from these initial voltages assignments.



(a) (b)

Fig. 4.4: (a) Initial temperature distribution on layer 1. (b) Optimized temperature distribution on layer 1.



(a) (b)

Fig. 4.5: (a) Initial temperature distribution on layer 2. (b) Optimized temperature distribution on layer 2.

- **Temperature Improvement**

Fig. 4.4 and 4.5 illustrate the change of temperature distribution on each layer of Circuit 1. These figures show obvious improvement in average temperature and suggest that the thermal gradient issue should be controlled in early design stage.

Chapter 5

Conclusion

In this thesis, we have presented a novel grid–based statistical total power optimization approach for 3D ICs. The use of temperature related timing and leakage power models and the thermal aware statistical timing analysis method provides more accurate estimation of the circuit performance. The experimental results show that simulation without considering thermal effect can underestimate the power consumption and overestimate the timing information of the design. The experimental results also indicate that the proposed methodology can provide significant power saving.

Bibliography

[1] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *IEEE Des. and Test of Comput.*, vol. 22, no 6, pp. 498–510, Nov.–Dec. 2005.

[2] R. Kumar, and V. Kursun, "Impact of Temperature Fluctuations on Circuit Characteristics in 180nm and 65nm CMOS Technologies," in *Proc. Int. Symp. Circuit and Sys.*, 2006, pp. 3858–3861.

[3] B. Lasbouygues, R. Wilson, N. Azemard, and P. Maurine, "Timing Analysis in Presence of Supply Voltage and Temperature Variations," in *Int. Symp, Phys. Des.*, 2006, pp. 10–16.

[4] J. Daga, E. Ottaviano, and D. Auvergen, "Temperature Effect on Delay for Low Voltage Applications," in *Proc. Des. Autom. Test Europe*, 1998, pp. 680–685.

[5] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full Chip Leakage Estimation Considering Power Supply and Temperature Variations," in *Proc. Int. Symp. on Low Power Electron. Des.*, 2003, pp. 78–83.

[6] J. Cong, J. Wei, and Y. Zhang, "A Thermal–Driven Floorplanning Algorithm for 3D ICs," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2004, pp. 306–313.

[7] P. Zhou, Y. Ma, and Z. Li, "3D–STAF: Scalable Temperature and Leakage Aware Floorplanning for Three–Dimensional Integrated Circuits," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 590–597.

[8] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal–Aware 3D IC Placement Via Transformation," in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2007, pp. 780–785.

[9] J. Cong, and Y. Zhang, “Thermal–Driven Multilevel Routing for 3–D ICs,” in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2005, pp. 121–126.

[10] T. Zhang, Y. Zhan, and S. S. Sapatnekar, “Temperature–Aware Routing in 3D ICs,” in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2006, pp. 309–314.

[11] M. Pathak, and S. Lim, “Thermal–aware Steiner Routing for 3D Stacked ICs,” in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 205–211.

[12] H. Yu, Y. Shi, L. He, and T. Karnik, “Thermal Via Allocation for 3D ICs Considering Temporally and Spatially Variant Thermal Power,” in *Proc. Int. Symp. on Low Power Electron. Des.*, 2006, pp. 156–161.

[13] Z. Li, X. Hong, Q. Zhou, S. Zeng, J. Bian, H. Yang, V. Pitchumani, and C. Cheng, “Integrating Dynamic Thermal Via Planning with 3D Floorplanning Algorithm,” in *Int. Symp, Phys. Des.*, 2006, pp. 178–185.

[14] B. Goplen, and S. S. Sapatnekar, “Placement of 3D ICs with Thermal and Interlayer Via Considerations,” in *Proc. Des. Autom. Conf.*, 2007, pp. 626–631.

[15] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, “Parameter Variations and Impact on Circuits and Microarchitecture,” in *Proc. Des. Autom. Conf.*, 2003, pp. 338–342.

[16] Semiconductor Industry Association, International Technology Roadmap for Semiconductors, 2005.

[17] K. Roy, S. Mukhopadhyay, and H. Mahmoodi–meimand, “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep–Submicrometer CMOS Circuits,” *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[18] http://www-device.eecs.berkeley.edu/ ptm/

[19] D. Sinha, N. Shenoy, and H. Zhou, “Statistical Gate Sizing for Timing Yield Optimization,” in *Proc. Int. Conf. on Comput.-Aided Des.*, 2005, pp. 1037–1041.

[20] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, “Interconnect–Power Dissipation in a Microprocessor,” in *Proc. Int. Workshop on Syst. level Intercon. Predic.*, 2004, pp. 7–13.

[21] A. Rahman, A. Fan, and R. Reif, “Comparison of Key Performance Metrics in Two– and Three–Dimensional Integrated Circuits,” in *Proc. Int. Intercon. Tech. Conf.*, 2000, pp. 18–20.

[22] D. Banerjee, S. Souri, P. Kapur, and K. Saraswat, “3–D ICs: A Novel Chip Design for Improving Deep–Submicrometer Interconnect Performance and Systems–on–Chip Integration,” *Proc. IEEE.*, vol. 89, no. 5, pp. 602–633, May 2001.

[23] S. Das, A. Chandrakasan, and R. Reif “Timing, Energy, and Thermal Performance of Three–Dimensional Integrated Circuits,” *Proc. Great Lakes Symp. on VLSI*, 2004, pp. 338–343.

[24] C. C. Liu, H. Ganusov, M. Burtscher, and S. Tiwari, “Bridging the Processor–Memory Performance Gap with 3D IC Technology,” *IEEE Des. and Test of Comput.*, vol. 22, no 6, pp. 556–564, Nov.–Dec. 2005.

[25] P. Jacob, O. Erdogan, A. Zia, P. M. Belemjian, R. P. Kraft, and J. F. McDonald, “Predicting the Performance of a 3D Processor–Memory Chip Stack,” *IEEE Des. and Test of Comput.*, vol. 22, no 6, pp. 540–547, Nov.–Dec. 2005.

[26] A. Zeng, J. Li, K. Rose, and R. J. Gutmann, “First–Order Performance Prediction of Cache Memory with Wafer–Level 3D Integratin,” *IEEE Des. and Test of Comput.*, vol. 22, no 6, pp. 548–555, Nov.–Dec. 2005.

[27] K. Bernstein, P. Andry, J. Cann, P. Emma, D. Greenberg, W. Haensch, M. Ignatowski, S. Koester, J. Magerlein, R.Puri, and A. Young, “Interconnects in the Third Dimension: Design Challenges for 3D ICs,” in *Proc. Des. Autom. Conf.*, 2007, pp. 562–567.

[28] P. Leduc, F. Crecy, M. Fayolle, B. Charlet, T. Enot, M. Zussy, B. Jones, J. Barbe, N. Kernevez, N. Sillon, S. Maitrejean, and D. Louis, “Challenges for 3D IC integration: bonding quality and thermal management,” in *Proc. Int. Intercon. Tech. Conf.*, 2007, pp. 210–212.

[29] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, and W. Davis, "Exploring Compromises among Timing, Power and Temperature in Three–Dimensional Integrated Circuits,"

in *Proc. Des. Autom. Conf.*, 2006, pp. 997–1002.

[30] C. Ferri, S. Reda, and R. Bahar, "Strategies for Improving the Parametric Yield and Profits of 3D ICs," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 220–226.

[31] A. W. Topol, J. D. C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three–dimensional Integrated Circuits," *IBM Journal of Res. and Dev.*, 2006, pp.491–506.

[32] F. Niklaus, J. Q. Lu, J. J. Mcmahon, J. Yu, S. H. Lee, T. S. Cale, and R. J. Gutmann "Wafer–Level 3D Integration Technology Platforms for ICs and MEMs,"

[33] S. Das, A. Chandrakasan, and R. Reif, "Design Tools for 3–D Integrated Circuits," in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2003, pp. 53–56.

[34] P. Franzon, W. Davis, M. Steer, H. Hao, S. Lipa, S. Luniya, C. Mineo, J. Oh, A. Sule, and T. Thorolfsson, "Design for 3D Integration and Applications," in *Int. Symp. on Signal, Syst. and Electron.*, 2007, pp. 263–266.

[35] R. Weerasekera, L. R. Zheng, D. Pamunuwa, and H. Tenhunen "Extending Systems–on–Chip to the Third Dimension: Performance, Cost and Technological Tradeoffs," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 212–219.

[36] V. Suntharalingam, R. Berger, J. A. Burns, C. K. Chen, C. L. Keast, J. M. Knecht, R. D. Lambert, K. L. Newcomb, D. M. O'Mara, D. D. Rathman, D. C. Shaver, A. M. Soares, C. N. Stevenson, B. M. Tyrrell, K. Warner, B. D. Wheeler, D. W. Yost, and D. J. Young, "Megapixel CMOS Image Sensor Fabricated in Three–Dimensional Integrated Circuit Technology," in *IEEE Int. Solid–State Circuits Conf.*, 2005, pp. 356–357.

[37] Q. Gu, Z. Xu, J. Ko, and M. C. F. Chang, "Two 10Gb/s/pin Low–Power Interconnect Methods for 3D ICs," in *IEEE Int. Solid–State Circuits Conf.*, 2007, pp. 448–614.

[38] J. A. Burns, B. F. Aull, C. K. Chen, C. L. Chen, C. L. Keast, J. M. Knecht, V. Suntharalingam, K. Warner, P. W. Wyatt, and D. R. W. Yost, “A Wafer–Scale 3–D Circuit Integration Technology,” *IEEE Trans. on Electron Devices*, vol. 53, no. 10, pp. 2507-2516, Oct. 2006.

[39] M. Loeve *Probability Theory*, D. Van Nostrand Company Inc., 1960.

[40] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao, “Modeling of intra-die process variations for accurate analysis and optimization of nanoscale circuits,” in *Proc. Des. Autom. Conf.*, 2006, pp. 791–6.

[41] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, and C. C. -P. Chen, “Correlation-preserved statistical timing with a quadratic form of Gaussian variables,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 11, pp. 2437–49, Nov. 2006.

[42] F. Liu, “A general framework for spatial correlation modeling in VLSI design,” in *Proc. Des. Autom. Conf.*, 2007, pp. 817–22.

[43] J. Xiong, V. Zolotov, and L. He, “Robust extraction of spatial correlation,” in *Int. Symp. Phys. Des.*, 2005, pp. 2–9.

[44] B. Cline, K. Chopra, D. Blaauw, and Y. Cao, “Analysis and modeling of CD variation for statistical static timing,” in *Proc. Int. Conf. on Comput.- Aided Des.*, 2006, pp. 60–66.

[45] B. Liu, “Gate level statistical simulation based on parameterized models for process and signal variations,” in *Proc. Int. Symp. on Quality Electron. Des.*, 2007, pp. 257–262.

[46] Kyung Ki Kim, et al, “Accurate Macro-modeling for Leakage Current for IDDQ Test”, IMTC, 2007.

[47] Nassif, “Delay Variability: Source, Impacts and Trends”, ISSCC,pp. 368, Feb. 2000,

[48] S. Mukhopadhyay, and K. Roy “Modeling and Estimation of Total Leakage Current in Nano-scaled CMOS Devices Considering the Effect of parameter Variation” in *Proc. Int. Symp. on Low Power Electron. Des.*, 2003, pp. 172–175.

[49] H. Dadgour, S. C. Lin, and K. Banerjee, “A Statistical Framework for Estimation of Full–Chip Leakage–Power Distribution Under Parameter Variations,” *IEEE Trans. on Electron Devices* vol. 54, no. 11, pp. 2930–2945, Nov. 2007.

[50] S. S. Sapatnekar, *Timing*,

[51] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan, “First–Order Incremental Block–Based Statistical Timing Analysis,” in *Proc. Des. Autom. Conf.*, 2004, pp. 331–336.

[52] J. Le, X. Li, and L. T. Pileggi “STAC:Statistical Timing Analysis with Correlation,” in *Proc. Des. Autom. Conf.*, 2004, pp. 343–348.

[53] L. Zhang, W. Chen, Y. Hu, and C. C. P. Chen, “Statistical Timing Analysis with Extended Pseudo–Canonical Timing Model,” in *Proc. Des. Autom. Test Europe*, 2005, pp. 952–957.

[54] H. Chang, and S. S. Sapatnekar, “Statistical Timing Analysis Under Spatial Correlations,” in *IEEE Trans. Comput.–Aided Design Integr. Circuits Syst.*, vol. 24, no. 9, pp. 1467–1482, Sept. 2005.

[55] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Vewmark, and M. Sharma, “Correlation–Aware Statistical Timing Analysis with Non–Gaussian Delay Distributions,” in *Proc. Des. Autom. Conf.*, 2005, pp. 77–82.

[56] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, and C. C. P. Chen, “Correlation–Preserved Non–Gaussian Statistical Timing Analysis with Quadratic Timing Model,” in *Proc. Des. Autom. Conf.*, 2005, pp. 83–88.

[57] H. Chand, V. Zolotov, S. Narayan, and C. Visweswariah, “Parameterized Block–Based Statistical Timing Analysis with Non–Gaussian Parameters, Nonlinear Delay Functions,” in *Proc. Des. Autom. Conf.*, 2005, pp. 71–76.

[58] W. S. Wang, V. Kreinovich, and M. Orshansky, “Statistical Timing Based on Incomplete Probabilistic Descriptions of Parameter Uncertainty,” in *Proc. Des. Autom. Conf.*, 2006, pp. 161–166.

[59] Z. Feng, P. Li, and Y. Zhan, “Fast Second–Order Statistical Static Timing Analysis Using Parameter Dimension Reduction,” in *Proc. Des. Autom. Conf.*, 2007, pp. 244–249.

[60] D. Sinha, H. Zhou, and N. Shenoy, “Advances in Computation of the Maximum of a Set of Random Variables,” in *Proc. Int. Symp. Quality Electron. Des.*, 2006, pp.

[61] B. N. Lee, Li-C. Wang, and M. S. Abadir “Refined Statistical Static Timing Analysis Through Learning Spatial Delay Correlations.” in *Proc. Des. Autom. Conf.*, 2006, pp. 149–154.

[62] B. Cline, K. Chopra, D. Blaauw, and Y. Cao, “Analysis and Modeling of CD Variation for Statistical Static Timing,” in *Proc. Int. Conf. on Comput.-Aided Des.,* 2006, pp. 60–66.

[63] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for VLSI: Timing and Power,* Springer-Verlag, 2004.

[64] M. Hashimoto, and H. Onodera, “A Performance Optimization Method by Gate Sizing using Statistical Static Timing Analysis,” in *Int. Symp, Phys. Des.,* 2000, pp. 111–116.

[65] S. Bhardwaj, and S. Vrudhula, “Leakage Minimization of Nano–scale Circuits in the presence of Systematic and Random Variations,” in *Proc. Des. Autom. Conf.,* 2005, pp. 541–546.

[66] K. Chopra, S. Shah, A. Srivastave, D. Blaauw, and D. Sylvester, “Parametric Yield Maximization using Gate Sizing based on Efficient Statistical Power and Delay Gradient Computation,” in *Proc. Int. Conf. on Comput.-Aided Des.,* 2005, pp. 1023–1028.

[67] M. Mani, A. Devgan, and M. Orshansky, “An Efficient Algorithm for Statistical Minimization of Total Power under Timing Yield Constraints,” in *Proc. Des. Autom. Conf.,* 2005, pp. 309–314.

[68] A. Srivastava, D. Sylvester, and D. Blaauw, “Statistical Optimization of Leakage Power Considering Process Variations using Dual–Vth and Sizing,” in *Proc. Des. Autom. Conf.,* 2004, pp. 773–778.

[69] X. Ye, Y. Zhan, and P. Li, "Statistical Leakage Power Minimization Using Fast Equi–Slack Shell Based Optimization," in *Proc. Des. Autom. Conf.*, 2007, pp. 853–858.

[70] K. Usami, and M. Horowitz, "Clustered voltage Scaling technique for low–power design," in *Proc. Int. Symp. on Low Power Electron. Des.*, 1995, pp. 3–8.

[71] K. Usami, M. Igarashi, F. Minami, M. Ishikawa, M. Ichida, and K. Nogami, "Automated low–power technique exploiting multiple supply voltages applied to a media processor," *IEEE J. Solid–State Circuits*, vol. 33, no. 3, pp. 463–472, March 1998.

[72] H. Wu, I. Liu, M. Wong, and Y. Wang, "Post–Placement Voltage Island Generation under Performance Requirement," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2005, pp. 309–316.

[73] H. Wu, M. Wong, and I. Liu, "Timing–Constrained and Voltage–Island–Aware Voltage Assignment," in *Proc. Des. Autom. Conf.*, 2006, pp. 429–432.

[74] R. Ching, E. Young, K. Leung, and C. Chu, "Post–Placement Voltage Island Generation," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2006, pp. 641–646.

[75] W. Lee, H. Liu, and Y. Chang, "Voltage Island Aware Floorplanning for Power and Timing Optimization," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2006, pp. 389–394.

[76] W. Lee, H. Liu, and Y. Chang, "An ILP Algorithm for Post–Floorplanning Voltage–Island Generation Considering Power–Network Planning," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 650–655.

[77] Q. Ma, and E. Young, "Voltage Island–Driven Floorplanning," in *Proc. Int. Conf. on Comput.-Aided Des.*, 2007, pp. 644–649.

[78] B. Liu, Y. Cai, Q. Zhou, and X. Hong, "Power Driven Placement with Layout Aware Supply Voltage Assignment for Voltage Island Generation in Dual–Vdd Designs," in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2006, pp. 582–587.

[79] L. Guo, Y. Cai, Q. Zhou, and X. Hong, "Logic and Layout Aware Voltage Island Generation for Low Power Design," in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2007, pp. 666–671.

[80] W. Mak, and J. Chen, “Voltage Island Generation under Performance Requirement for SoC Designs,” in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2007, pp. 798–803.

[81] W. Hung, G. Link, Y. Xie, N. Vijaykrishnan, N. Dhanwada, and J. Conner, “Temperature–Aware Voltage Island Architecting in System–on–Chip Design,” in *Proc. Int. Conf. on Comput. Des.*, 2005, pp. 689–694.

[82] J. Hu, Y. Shin, N. Dhanwada, and R. Marculescu, “Architecting Voltage Islands in Core–based System–on–a–Chip Designs,” in *Proc. Int. Symp. on Low Power Electron. Des.*, 2004, pp. 180–185.

[83] S. Kulkarni, A. Srivastava, and D. Sylvester, “A New Algorithm for Improved VDD Assignment in Low Power Dual VDD Systems,” in *Proc. Int. Symp. on Low Power Electron. Des.*, 2004, pp. 200–205.

[84] 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 Trans. Very Large Scale Integr. Syst.*, vol. 15, no. 6, pp. 637–648, June 2007.

[85] 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 *Proc. Int. Symp. on Low Power Electron. Des.*, 2003, pp. 158–163.

[86] W. Hung, Y. Xie, N. Vijaykrishnan, M. Kandemir, M. J. Irwin, and Y. Tsai, “Total Power Optimization through Simultaneously Multiple–V$_{DD}$ Multiple–V$_{TH}$ Assignment and Device Sizing with Stack Forcing,” in *Proc. Int. Symp. on Low Power Electron. Des.*, 2004, pp. 144–149.

[87] A. Srivastava, D. Sylvester, and D. Blaauw, “Power Minimization using Simultaneous Gate Sizing, Dual–Vdd and Dual–Vth Assignment,” in *Proc. Des. Autom. Conf.*, 2004, pp. 783–787.

[88] M. Ekpanyapong, and S. Lim, “Integrated Retiming and Simultaneous Vdd/Vth Scaling For Total Power Minimization,” in *Int. Symp. Phys. Des.*, 2006, pp. 142–148.

[89] "Architecting, Designing, Implementing, and Verifying Low-Power Digital Integrated Circuits," Cadence 2007.

[90] J. H. Wu, and Y. M. Lee, "Stochastic Thermal Simulator Considering Within–die Spatial Correlation under Process Variations," Master thesis, NCTU, 2007.

[91] P. Y. Huang, C. K. Lin, and Y. M. Lee, "Full-chip Thermal Analysis for the Early Design Stage via Generalized Integral Transforms," in *Proc. Asia and South Pacific Des. Autom. Conf.*, 2008, pp. 462–467.

[92] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early–stage VLSI design," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 14, no. 5, pp. 501–513, May 2006.

[93] D. Markovic, V. Stojanovic, B. Nikolic, M. A. Horowitz, and R. W. Brodersen, "Methods for true energy-performance optimization," in *IEEE J. of Solid-State Circuits*, vol. 39, no. 8, pp. 1282–93, 2006.

[94] R. Hentschke, G. Flach, F. Pinto, and R. Reis, "3D–Vias Aware Quadratic Placement for 3D VLSI Circuits," in *IEEE Comput. Society Ann. Symp. on VLSI*, 2007, pp. 67–72.

[95] P. Y. Huang, and Y. M. Lee, "A full–chip electro–thermal analysis for 3D ICs," personal discussion.

[96] X. Ye, P. Li, and F. Liu, "Practical Variation-Aware Interconnect Delay and Slew Analysis for Statistical Timing Verification," in *Proc. Int. Conf. on Comput. -Aided Des.* , pp. 54-59, Nov. 2006.

[97] P. Li, "Statistical Sampling-Based Parametric Analysis of Power Grids", in *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 12, pp. 2852-2867, Dec. 2006.

[98] H. Chang and S. S. Sapatankar, "Statistical timing analysis under spatial correction," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 9, pp. 1467-1482, Sept. 2005.

[99] N. Mi, J. Fan, S. X.-D. Tan, Y. Cai, and X. Hong, “Statistical Analysis of On-Chip Power Delivery Networks Considering Lognormal Leakage Current Variations with Spatial Correlation,” *IEEE Trans. on Circuits and Syst.*, accepted for future publication.

[100] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, revised edition, Springer-Verlag, 2003.

[101] R. H. Cameron, W. T. Martin, “The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals,” Ann. of Math., 1947.

Appendix A

Statistical Thermal Simulation

The utilized thermal simulation technique is presented in [95], and the contents of the methodology are described in the following sections.

A.1 Polynomial Chaos

The Taylor expansion method has been widely used to solve a linear system with random input. Its basic idea is to approximate the desire solution and the random input by taking Taylor expansion with respect to the random variables, and to match the leading coefficients of the approximating polynomials for the desire solution with the expanded coefficients of the random input. This method has been widely used in the statistical timing analysis and the circuit performance analysis [96, 97, 98, 41]. However, the small variation assumption of the desire solution with respect to the random variables is not appropriate for approximating the leakage power, because the leakage power depends on the physical parameters exponentially. Since the temperature is directly affected by the leakage power, the Taylor expansion method is not very suitable for the thermal analysis. An example for the power grid analysis with large variations of the random variable for the log-normal leakage current [99] was indicated for demonstrating the inaccuracy of the second order Taylor expansion method.

On the contrary, the polynomial chaos (PC) based method [100] is chosen because it can handle the desired solution with large variations of the random variables and achieve a minimal mean square error approximation. Moreover, the projected deterministic heat transfer equations in this work are un-coupled for each PC. Hence, the efficiency is equal to the Taylor expansion method for solving the stochastic heat transfer equations.

A.1.1 The Bases for Random Space

The generalized polynomial chaos, which is called the Askey-Chaos, utilizes the orthogonal polynomials as the trial basis in the random space to expand the stochastic process. The original polynomial chaos which is termed as the Hermite chaos was first introduced by Wiener [100]. Ghanem and Spanos are the pioneers that employ the Hermite orthogonal polynomials in terms of Gaussian random variable to deal with various problems in mechanics [100]. The theorem of Cameron and Martin [101] guarantees that a general second-order random process $u(\theta)$ can be represented in the following form:

$$\begin{aligned} u(\theta) &= c_0\Gamma_0 + \sum_{i_1=1}^{\infty} c_{i_1}\Gamma_1(\xi_{i_1}) \\ &+ \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} c_{i_1 i_2}\Gamma_2(\xi_{i_1},\xi_{i_2}) \\ &+ \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} c_{i_1 i_2 i_3}\Gamma_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + ... \end{aligned}$$

where $\Gamma_r(\xi_{i_1},...,\xi_{i_n})$ represents the polynomial chaos of order r in terms of the N-dimensional random variables $\vec{\xi} = (\xi_{i_1},...,\xi_{i_n})$. The polynomial chaos was so-called Hermite polynomial chaos for the Gaussian random variables. For the Hermite polynomials with multi-dimension $\Gamma_r(\xi_{i_1},...,\xi_{i_N})$, the general expression form can be obtained as

$$\Gamma_r(\xi_{i_1},...,\xi_{i_N}) = (-1)^r \frac{\partial^n}{\partial\xi_{i_1}...\xi_{i_n}} e^{-\frac{1}{2}\vec{\xi}^T\vec{\xi}}$$

The zero, first, and second-order Hermite polynomial chaos can be given by:

$$\Gamma_0 = 1; \;\; \Gamma_1(\xi_i) = \xi_i; \;\; \Gamma_2(\xi_i) = \xi_i\xi_j - \delta_{ij}$$

where δ_{ij} is the Kronecker delta. For charity, the above general second-order random process $u(\theta)$ can be expressed as more simplified form

$$u(\theta) = \sum_{j=1}^{\infty} \hat{a}_j \Phi_j(\vec{\xi}) \qquad \text{(A.1)}$$

where there is a one-to-one mapping between the polynomial chaos $\Gamma[.]$ and $\Phi[.]$, and also between the coefficients $\hat{a}_j$ and $c_{i_1...i_n}$. The polynomial chaos of the same order with different argument list are orthogonal to each other, so are ones of the different order. For notation, the polynomial chaos satisfy the following orthogonality property:

$$< \Phi_i\Phi_j >=< \Phi_i^2 > \delta_{ij}$$

where $< . >$ denotes the inter product defined in the following:

$$< f(\vec{\xi})g(\vec{\xi}) >= \frac{1}{\sqrt{(2\pi)^n}} \int f(\vec{\xi})g(\vec{\xi})e^{-\frac{1}{2}\vec{\xi}^T\vec{\xi}}d\vec{\xi}$$

A.2 Stochastic Galerkin Procedure

According to section A.1.1, the temperature $T(\mathbf{r}, t, \varpi, \theta)$ can be approximated as

$$T(\mathbf{r}, t, \varpi, \theta) \simeq \sum_{k=0}^{N_{KL}} T_k(\mathbf{r}, t)\xi_k, \tag{A.2}$$

where each $T_k(\mathbf{r}, t)$ is the coefficient function of the temperature projection onto the k-th H-PC, and N_{KL} is the truncation number.

Substituting equation (A.2) into the thermal equation and approximating the power density as a function of ξ, the residual of thermal equation is

$$R(\mathbf{r}, t, \xi) \quad \equiv \quad \sum_{k=0}^{N_{PC}} \left(\rho c_p \frac{\partial T_k(\mathbf{r}, t)}{\partial t} - \kappa \nabla^2 T_k(\mathbf{r}, t) \right) \Phi_k(\xi) - p(\mathbf{r}, t, \xi). \tag{A.3}$$

Based on the principle of stochastic Galerkin projection [100], the residuals of equations are enforced to be orthogonal to each H-PC, *i.e.*, $< R(\mathbf{r}, t, \xi)\Phi_k(\xi) >= 0$ for each k. The decoupled deterministic heat transfer equation is obtained to solve $T_k(\mathbf{r}, t)$ for each different k.

$$\rho c_p \frac{\partial T_k(\mathbf{r}, t)}{\partial t} = \kappa \nabla^2 T_k(\mathbf{r}, t) + \frac{p_k(\mathbf{r}, t)}{< \Phi_k^2(\xi) >}, \tag{A.4}$$

subject to the boundary condition

$$\kappa \frac{\partial T_k(\mathbf{r}_{b_s}, t)}{\partial n_{b_s}} + h_{b_s} T_k(\mathbf{r}_{b_s}, t) = \hat{f}_{b_s}(\mathbf{r}_{b_s}, t)\delta_{0k} \tag{A.5}$$

for each b_s.

The $p_k(\mathbf{r}, t) =< p(\mathbf{r}, t, \xi)\Phi_k(\xi) >$ in equation (A.4) is equal to

$$p_k(\mathbf{r}, t) = p_d(\mathbf{r}, t)\delta_{0k} + p_{g_k}(\mathbf{r}, t) + p_{s_k}(\mathbf{r}, t), \tag{A.6}$$

where $p_{g_k}(\mathbf{r}, t) =< p_g(\mathbf{r}, t, \varsigma)\Phi_k(\xi) >$ and $p_{s_k}(\mathbf{r}, t) =< p_s(\mathbf{r}, t, \zeta)\Phi_k(\xi) >$ are the projected gate-leakage and subthreshold-leakage power density profiles of the k-th H-PC, respectively. The term δ_{0k} in both equations (A.5) and (A.6) is because $< \Phi_k(\xi) >= \delta_{0k}$ [100]. Each δ_{0k} attends in right hand sides of equations (A.5) and (A.6) is due to $< \Phi_k(\xi) >= \delta_{0k}$ [100]. After

$p_k(\mathbf{r}, t)$ is calculated, any existing deterministic thermal simulator can be utilized to obtain each $T_k(\mathbf{r}, t)$.

The above un-coupled deterministic heat transfer equations have an advantage for both numerical and analytical thermal simulators. In this work, an efficient early-stage thermal simulator proposed in [91] is utilized to serve as the deterministic thermal simulator. The spatial mean and variance distributions of the full-chip temperature distribution can be obtained as

$$\begin{aligned} E\{T(\mathbf{r}, t, \varpi, \theta) + T_a\} &\approx T_0(\mathbf{r}, t) + T_a, \qquad \text{(A.7)} \\ Var\{T(\mathbf{r}, t, \varpi, \theta) + T_a\} &\approx \sum_{k=1}^{N_{PC}} T_k^2(\mathbf{r}, t) < \Phi_k^2(\xi) > . \qquad \text{(A.8)} \end{aligned}$$

We should note that only one deterministic heat transfer equation is needed to be solved for obtaining the spatial mean temperature distribution.

A.3 Gate Leakage Power Projection

By substituting equations (2.4) and (2.6) into equation (2.8) and multiplying it by the supply voltage V_{dd}, the stochastic gate tunneling leakage power for a specific type of functional gate located at (x^*, y^*, z^*) can be expressed as

$$\begin{aligned} I_{gate} &= b_0 \exp\left(b_1(\bar{t}_{ox} + \triangle t_{ox})\right), \\ P_g(x^*, y^*, z^*, \varsigma) &= \overline{P}_g \exp\left(\bar{b}_1 \varsigma^T \mathbf{f}^*\right), \qquad \text{(A.9)} \end{aligned}$$

where $\overline{P}_g = \bar{b}_0 V_{dd}$, $\bar{b}_0$ and $\bar{b}_1$ are known values, $\varsigma = [\varsigma_1, \varsigma_2, \cdots, \varsigma_{N_{tox}}]^T$, $\mathbf{f}^* = [f_1^*, \cdots, f_n^* \cdots, f_{N_{tox}}^*]^T$, and each $f_n^* = \sqrt{\beta_n} f_n(x^*, y^*, z^*)$.

By using equation (A.9), the gate tunneling leakage power projection onto the k-th H-PC for a specific type of functional gate located at a reference position (x^*, y^*, z^*) is

$$< P_g(x^*, y^*, z^*, \varsigma)\Phi_k(\xi) >= \overline{P}_g < \exp\left(\bar{a}_1 \varsigma^T \mathbf{f}^*\right) \Phi_k(\xi) > . \qquad \text{(A.10)}$$

Fig. A.1 shows an algorithm for calculating equation (A.10). Steps $4 \sim 5$ are owing to the independence of $\{\varsigma_i\}$ and $\{\zeta_i\}$. The rest steps of Fig. A.1 can be derived by utilizing the zeroth, first and second derivatives of the moment generating function of independent standard normal random variables. Although the algorithm in Fig. A.1 only calculates the gate leakage power projection up to the second order of H-PCs, it can be easily extended to the higher order of H-PCs.

```
Algorithm  Gate Tunneling Leakage Power Projection
Input:  Constants $\overline{P}_g$ and $\overline{a}_1$, vector $\mathbf{f}^*$, and the $k$-th H-PC $\{\Phi_k(\xi)\}$
Output:  Set $\{B_k^g = \overline{P}_g < \exp\left(\overline{a}_1 \varsigma^T \mathbf{f}^*\right) \Phi_k(\xi) >\}$
Begin
    $D_g^* \leftarrow \overline{P}_g \prod_{i=1}^{N_{T_{ox}}} \exp\left((\overline{a}_1 f_i^*)^2/2\right)$
    for $k \leftarrow 0$ to $N_{PC}$
      if $\Phi_k(\xi)$ is a function of $\{\zeta_i\}$
          $B_k^g \leftarrow 0$
      elseif $\Phi_k(\xi) = 1$
          $B_k^g \leftarrow D_g^*$
      elseif $\Phi_k(\xi) = \varsigma_i;\ i \in G$
          $B_k^g \leftarrow D_g^* \overline{a}_1 f_i^*$
      elseif $\Phi_k(\xi) = \varsigma_i \varsigma_j - \delta_{ij};\ i \in G,\ j \in G$
          $B_k^g \leftarrow D_g^* (\overline{a}_1)^2 f_i^* f_j^*$
 End
$* G = \{1, 2, 3, \cdots, N_{T_{ox}}\}$
```

Fig. A.1: Gate tunneling leakage power projection algorithm

A.4 Subthreshold Leakage Power Projection

Also, substituting equations (2.5) and (2.7) into equation (2.10) and multiplying it by the supply voltage V_{dd}, the subthreshold leakage power for a specified type of functional gate located at (x^*, y^*, z^*) can be given as

$$\begin{aligned} I_{sub} &= c_0 \exp\Big(c_1(\overline{L}_{eff} + \triangle L_{eff}) + c_2 T\Big) \\ P_s(x^*, y^*, z^*, \zeta) &= \overline{P}_s \exp\Big(\bar{c}_1 \zeta^T \mathbf{q}^* + \bar{c}_2 \xi^T \mathbf{T}(\mathbf{r}^*)\Big), \end{aligned} \tag{A.11}$$

where $\overline{P}_s = \bar{c}_0 V_{dd}$, $\bar{c}_0$, $\bar{c}_1$ and $\bar{c}_2$ are known values, $\zeta = [\zeta_1, \zeta_2, \cdots, \zeta_{N_{L_{eff}}}]^T$, $\mathbf{q}^* = [q_1^*, \cdots, q_n^* \cdots, q_{N_{L_{eff}}}^*]^T$, $\xi = [\xi_1, \xi_2, \cdots, \xi_{N_{KL}}]^T$, $\mathbf{T}(\mathbf{r}^*) = [T_1(\mathbf{r}^*), \cdots, T_n(\mathbf{r}^*) \cdots, T_{N_{KL}}(\mathbf{r}^*)]^T$, and each $q_n^* = \sqrt{\chi_n} q_n(x^*, y^*, z^*)$.

The leakage power have to project to both ζ and ς because of its temperature dependent characteristic. In addition, the temperature on one tier is affected by the power on the other tier, so the we have to do projection to different tier. Before deriving the subthreshold leakage projection, we first formulate the subthreshold leakage power on a specific layer l

$$\begin{aligned} E[P_s^l] &= E[\bar{c}_0 \exp(\bar{c}_1 L_i + \bar{c}_2 T_i^l)] \\ &= E[\bar{c}_0 \exp(\bar{c}_1(\overline{L}_i^l + \sum_j q_{ij}^{l*} \zeta_j^l) \end{aligned}$$

$$
\begin{aligned}
& +\overline{c}_2(\overline{T}_i^l + \sum_j T_{ij}^{l\rightarrow l}\zeta_j^l + \sum_{l'\neq l}\sum_j T_{ij}^{l'\rightarrow l}\zeta_j^{l'} + \sum_k T_{ik}^{l\rightarrow l}\varsigma_k^l + \sum_{l'\neq l}\sum_k T_{ik}^{l'\rightarrow l}\varsigma_k^{l'}))] \\
= \; & \overline{c}_0 \exp(\overline{c}_1\overline{L}_i^l + \overline{c}_2\overline{T}_i^l) E[\exp(\overline{c}_1 \sum_j q_{ij}^{l*}\zeta_j^l)] E[\exp(\overline{c}_2 \sum_j T_{ij}^{l\rightarrow l}\zeta_j^l)] \\
& E[\exp(\overline{c}_2 \sum_{l'\neq l}\sum_j T_{ij}^{l'\rightarrow l}\zeta_j^{l'})] E[\exp(\overline{c}_2 \sum_k T_{ik}^{l\rightarrow l}\varsigma_k^l)] E[\exp(\overline{c}_2 \sum_{l'\neq l}\sum_k T_{ik}^{l'\rightarrow l}\varsigma_k^{l'})] \\
= \; & \overline{c}_0 \exp(\overline{c}_1\overline{L}_i^l) \exp(\frac{1}{2}\sum_j(\overline{c}_1 q_{ij}^l)^2) \exp(\overline{c}_2\overline{T}_i^l) \exp(\frac{1}{2}\overline{c}_2^2 \sum_j (T_{ij}^{l\rightarrow l})^2) \\
& (\prod_{l'\neq l} \exp(\frac{1}{2}\overline{c}_2^2 \sum_j T_{ij}^2))(\exp(\frac{1}{2}\overline{c}_2^2 \sum_k (T_{ik}^{l\rightarrow l})^2))(\prod_{l'\neq l} \exp(\frac{1}{2}\overline{c}_2^2 \sum_k (T_{ik}^{l'\rightarrow l})^2)) \\
= \; & \overline{c}_0 \exp(\overline{c}_1\overline{L}_i^l) \exp(\frac{1}{2}\overline{c}_1^2 \sum_j (q_{ij}^l)^2) \exp(\overline{c}_2\overline{T}_i^l) \exp(\frac{1}{2}\overline{c}_2^2\sigma_{T_i^l}^2) \qquad \text{(A.12)}
\end{aligned}
$$

,where l' is the layer besides l, $i = (x^*, y^*)$ is the location. By using equation (A.11) and (A.12), the subthreshold leakage power projection onto the k-th H-PC for a specific type of functional gate located at a reference position (x^*, y^*) are

$$
\begin{aligned}
E[\varsigma_k^l P_s^l] &= \overline{c}_2 T_{ik}^{l\rightarrow l} E[P_s] && \text{(A.13)} \\
E[\varsigma_k^l P_s^{l'}] &= \overline{c}_2 T_{ik}^{l\rightarrow l'} E[P_s^{l'}] && \text{(A.14)} \\
E[\zeta_k^l P_s^l] &= \overline{c}_2 T_{iq}^{l\rightarrow l} E[P_s] && \text{(A.15)} \\
E[\zeta_q^l P_s^{l'}] &= \overline{c}_2 T_{iq}^{l\rightarrow l'} E[P_s^{l'}] && \text{(A.16)}
\end{aligned}
$$