行政院國家科學委員會專題研究計畫 成果報告

針對 3D 整合之電子設計自動化技術開發--子計畫一：三維度積體電路的隨機電熱模擬及其對功率最佳化的應用(2/2)

研究成果報告(完整版)

計畫類別：整合型
計畫編號：NSC 99-2220-E-009-035-
執行期間：99年08月01日至100年07月31日
執行單位：國立交通大學電信工程學系（所）

計畫主持人：李育民

計畫參與人員：碩士班研究生-兼任助理人員：吳宗恆
碩士班研究生-兼任助理人員：李亭蓉
碩士班研究生-兼任助理人員：王志升
博士班研究生-兼任助理人員：黃培育
博士班研究生-兼任助理人員：魏書含
博士班研究生-兼任助理人員：潘麒文

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫涉及專利或其他智慧財產權，2年後可公開查詢

中華民國 100年10月30日

針對3D整合之電子設計自動化技術開發

子計畫一：三維度積體電路的隨機電熱模擬及其對功率最佳化的應用（2/2）

Stochastic Electro-Thermal Simulation for 3-D ICs and Its Application to 3-D IC Power Optimization

計畫編號：NSC 99-2220-E-009-035-

執行期間：99 年 8 月 1 日 至 100 年 7 月 31 日

計畫主持人：李育民

一、中文摘要

目前針對三維度積體電路功耗優化技術鮮少討論到多重電壓供應技巧。本計畫利用多重電壓供應技巧以降低三維度積體電路功耗，發展的方法包括三大部分。（1)三維度積體電路的電源電壓分配:分配的方法考慮三個方面的因素–靈敏度、鄰近效應和電壓位準移位器的預算;（2）三維度積體電路電熱分析:得到三維度積體電路之溫度分佈;（3）考慮熱感知的靜態時序分析: 以分析三維度積體電路的延遲。實驗結果驗證了此多重電壓供應技術的有效性。

關鍵詞：三維度積體電路；電熱模擬；功率最佳化；低功率設計；多重電壓設計

二、英文摘要

Few of existing works on power reduction in 3D ICs discuss the ability of supply voltage scaling techniques for power optimization. In this work, a supply voltage assignment based power reduction method for minimizing the power consumption of 3D ICs is presented. The proposed approach includes three major headings: (1) *3D IC Voltage Assignment* for power reduction with including three factors--sensitivity, proximity effect and level shifter (LS) budget; (2) *3D Electro-Thermal Analysis* for getting the temperature distribution of 3D IC; (3) *Thermal Aware Static Timing Analysis* for obtaining thermal-related delay values of functional gates. The experimental results demonstrate the effectiveness of the developed voltage assignment method in power reduction.

Keywords：3-D IC, Electro-Thermal Analysis, Power Optimization, Low Power Design, Multiple Supply Voltage Design

三、研究計畫之背景及目的

In recent years, many researchers have shown that 3D ICs can provide the powerful enhancement of system integration. However, the heat removal is a great challenge in 3D IC design due to the high power density and the low thermal conductivities of inter-layer dielectrics. Moreover, the high temperature induces serious impacts on the timing, power and reliability of circuit design [1]. Therefore, it is necessary to reduce the power consumption of circuits for mitigating the thermal problem of circuit.

Among the existing power reduction techniques, the multiple supply voltage (MSV) [2-5] is an effective technique to reduce dynamic power and leakage power. Intuitively, any MSV techniques of 2D ICs can be extended to 3D ICs and the voltage scaling can be performed tier by tier. However, without simultaneously considering every tier in voltage assignment procedure might lead the power consumption distribution of each layer to be off-balance and cause the thermal problem.

In this work, based on [6], a grid-based post-placement MSV method will be developed for reducing the 3D IC power consumption and will be extended to explore more possibilities. This developed technique will simultaneously consider the level shifter (LS) budget and the thermal effect. Compared with previous works ignoring the LS issue and using the thermal-unrelated models, the proposed approach is more flexible and practical.

The report is organized as follows. First, the proposed power optimization methodology is presented in section 、四. After that, experimental results are given in section 、五. Finally, some conclusions are drawn in section 、六.

四、研究方法

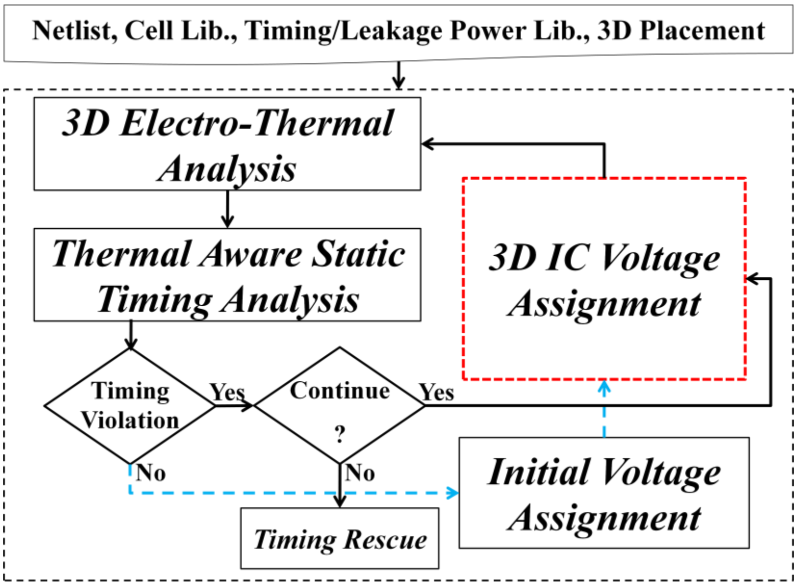

Netlist, Cell Lib., Timing/Leakage Power Lib., 3D Placement
3D Electro-Thermal Analysis
Thermal Aware Static Timing Analysis
3D IC Voltage Assignment
Timing Violation
Yes
Continue ?
Yes
No
No
Initial Voltage Assignment
Timing Rescue

Fig. 1. Flowchart of the proposed power optimization for 3D ICs

The proposed post-placement MSV based method for power reduction during the 3D IC design flow is shown in Fig. 1. Given a known 3D IC design placement, netlist, cell library and timing/leakage power cell library, for the grid-based procedure, each tier is partitioned into *n* grids as illustrated in Fig. 2.(a). First, the initial supply voltage of all gates is set to be the high supply voltage V_{DDH}, and the initial temperature of chip is obtained by *3D Electro-Thermal Analysis* (section、四. A). Then, the *Thermal Aware Static Timing Analysis* (section、四. B) is performed with the temperature-related delay of gate got from the initial temperature in *3D Electro-Thermal Analysis* step. After that, the *Initial Voltage Assignment* (section、四. C) is executed, and the circuit timing might be violated due to aggressively assigning low supply voltage V_{DDL} to all gates. Then, a grid-based procedure is developed for the *3D IC Voltage Assignment* (section、四. D) that assigns an appropriate V_{DD} for trading off the power consumption penalty and the timing saving advantage. After executing the voltage assignment procedure once, the power consumption and the delay of each gate are changed; hence, the thermal and timing analysis should be done again to update the temperature-related delay of gate and leakage power. *3D IC Voltage Assignment* is executed until no grid can be selected or timing violation is rescued (section、四. E). The voltage assignment result got by the proposed method can be applied to any voltage island generators. Moreover, *3D IC Voltage Assignment* method could be more beneficial for the voltage island generation because of considering the proximity effect.

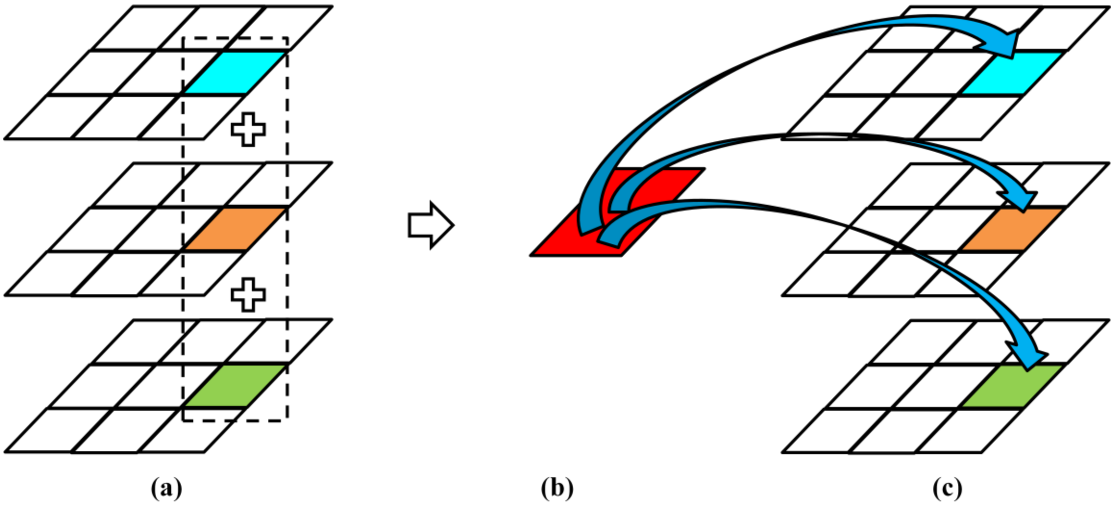


Fig. 2. A three-tier design example of the grid-based procedure for generating the voltage assignment

A. *3D Electro-Thermal Analysis*

Generally, the dynamic power is independent of temperature but the leakage power is significantly affected by temperature. Based on the empirical model [7], the gate leakage current I_{gate} is related to the oxide thickness, and the subthreshold current I_{sub} is related to the channel length and temperature. Fixing the oxide thickness and the channel length, the subthreshold current model of gate can be built by utilizing the least square fitting method to the estimated results of HSPICE under the 90nm technology as follow.

$$I_{gate} = \text{constant} \tag{1}$$

$$I_{sub} = s_0 \exp{(s_1 T)} \tag{2}$$

Here, s_0 and s_1 are fitting constants, and T is the operating temperature of gate/cell.

Since the fitting constants and the supply voltage are dependent, a look-up table is set up to store them for different supply voltages, V_{DDL} and V_{DDH}. With (1) and (2), the gate tunneling leakage power and the subthreshold leakage power are

$$P_{gate} = V_{DD} I_{gate} \tag{3}$$

$$P_{sub} = V_{DD} I_{sub} \tag{4}$$

The *3D Electro-Thermal Analysis* is performed with an electro-thermal iterative updating loop which is built by integrating (3)-(4), the 3D IC statistically thermal simulator [6] and the 3D IC analytical thermal simulator [8].

B. *Thermal Aware Static Timing Analysis*

The thermal-aware static timing analysis (STA) is a conventional block-based STA. Each gate delay is thermal dependent and is built as a canonical first-order form by applying the least square fitting method to fit the Monte Carlo results of HSPICE under 90nm technology. The general delay expression form of a specific gate is

$$Delay = (a_0 + a_1 T)(a_2 + a_3 C_L)\exp{(a_4 C_L)} \tag{5}$$

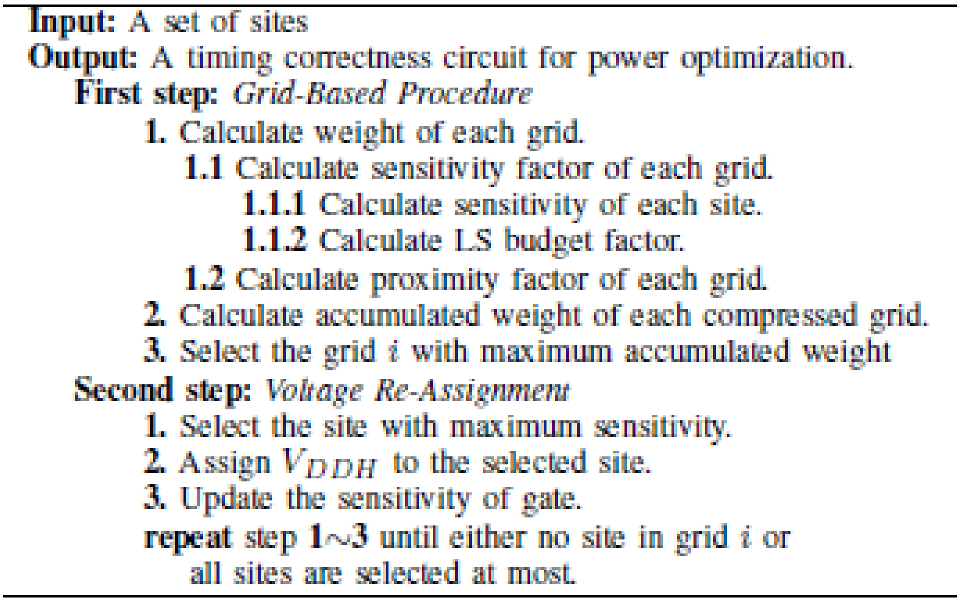
Here, a_0-a_4 are V_{DD} dependent fitting constants, and a look-up table is constructed to store these coefficients for V_{DDH} and V_{DDL}. C_L is the load capacitance.

C. *Initial Voltage Assignment*

Given a timing satisfied circuit with the initial supply voltage being V_{DDH}, an *Initial Voltage Assignment* procedure is performed to save the most power without considering the timing constraints. All gates are assigned to operate at V_{DDL}'s, and this step might cause the timing violation. After the assignment, a *Thermal Aware STA* step is executed to calculate the slack of each gate. With the updated operating temperature of each gate from *3D Electro-Thermal Analysis* and (5), the updated delay of each gate can be obtained. Then, the updated arrival time (AT), the updated required arrival time (RT) and the slack of each gate are calculated. Finally, a gate is referred to as a site if the slack of that gate is found to be negative, and a set of sites is output to the next step.

D. 3D IC Voltage Assignment

The algorithm of *3D IC Voltage Assignment* is shown in Fig. 3. Given a set of sites from *Initial Voltage Assignment*, the *Grid-Based Procedure* is executed to decide which grid should be firstly picked, and then sites in this chosen grid are assigned to operate at V_{DDH} to rescue timing. Then, *Voltage Re-Assignment* is performed by selecting several sites in the selected grid to operate at V_{DDH} for timing rescue. The above procedure is repeat until none of sites in the selected grid can be re-assigned to operate at V_{DDH}, or all sites in this grid have been rescued successfully.



```
Input: A set of sites
Output: A timing correctness circuit for power optimization.
  First step: Grid-Based Procedure
      1. Calculate weight of each grid.
         1.1 Calculate sensitivity factor of each grid.
             1.1.1 Calculate sensitivity of each site.
             1.1.2 Calculate LS budget factor.
         1.2 Calculate proximity factor of each grid.
      2. Calculate accumulated weight of each compressed grid.
      3. Select the grid $i$ with maximum accumulated weight
  Second step: Voltage Re-Assignment
      1. Select the site with maximum sensitivity.
      2. Assign $V_{DDH}$ to the selected site.
      3. Update the sensitivity of gate.
      repeat step 1~3 until either no site in grid $i$ or
         all sites are selected at most.
```

Fig. 3. Algorithm of 3D IC voltage assignment

E.1、 Grid-Based Procedure

We improve the idea of [6] to perform a grid-based procedure to decide which grid is firstly picked, and then sites in this grid are assigned to operate at V_{DDH} to meet timing constraints under the complex three dimensional structure. The goal of this decision is to effectively trade off the power consumption penalty and the timing saving advantage.

A three-tier design example of the grid-based procedure is illustrated in Fig.

2. Firstly, the weight of each grid is determined by considering three factors: 1) sensitivity factor, 2) proximity factor, 3) LS budget factor. Then, the three dimensional structure is vertically compressed into a two dimensional planar illustrated in Fig. 2.(b), and the weight of each compressed grid is obtained by accumulating the weight of each grid of *z*-axis. After that, the compressed grid with the maximum accumulated weight is selected, and this grid-based decision step is finished. Finally, as shown in Fig. 2.(c), the selected compressed grid is restored back to the original three layer structure. This procedure helps us to decide which grid is firstly assigned for the next stage *Voltage Re-Assignment*. The weight of a grid *i* is defined as

$$W_i \overset{\text{def}}{=} c_1\alpha_i + c_2\beta_i, \tag{6}$$

where α_i is the sensitivity factor and LS budget factor, and β_i is the proximity factor. There are two main concerns for the definition of α_i. The first concern (S_{site}) is how to make an assignment decision can obtain the most rescue of timing and the least penalty of power saving. The second concern λ_i is how to make a assignment decision can lead to fewer LS overheads (# level shifters). The β_i factor considers clustering. We hope more gates in clusters in a single grid, and all gates in a cluster operate at the same V_{DD}. The sensitivity factor α_i and proximity factor β_i are defined as follows.

$$\alpha_i \overset{\text{def}}{=} \frac{S_{grid_i}}{\max\left\{S_{grid_k}, k=1\sim n\right\}}\lambda_i \tag{7}$$

$$\beta_i \overset{\text{def}}{=} \frac{N_i - N_i^H}{N_i} = \frac{N_i^L}{N_i}, \tag{8}$$

where S_{grid_i} is the sum of site sensitivities S_{site}'s in grid *i*, N_i is the number of all gates in grid *i*, N_i^H is the number of gates with V_{DDH}, N_i^L is the number of gates with V_{DDL}, and λ_i is the LS budget factor that indicates whether the estimated S_{grid_i} is appropriate.

As the operating voltages of all sites in the selected grid are determined, the needed number of level shifters is determined, and λ_i can be defined as

$$\lambda_i \overset{\text{def}}{=} \min\left\{1, \frac{N_i^{LSa}}{N_i^{LSn}}\right\} \tag{9}$$

Here, N_i^{LSn} is the needed number of level shifters if all sites are assigned to operate at V_{DDH} in grid *i*, and N_i^{LSa} is the available number of level shifters in grid *i* that is estimated by the attainable white space of grid *i*.

Finally, the grid sensitivity S_{grid_i} and the site sensitivity S_{site_j} are defined as follows.

$$S_{grid_i} \stackrel{\text{def}}{=} \sum_{site_j \in grid_i} S_{site_j} \tag{10}$$

$$S_{site_j} \stackrel{\text{def}}{=} \frac{\Delta D}{\Delta P} \left| Slack_{site_j} \right| \tag{11}$$

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

E.2、 Voltage Re-Assignment

Based on the decision of grid-based procedure, a *Voltage Re-Assignment* procedure is performed to obtain the best timing rescue and the least power saving penalty within the selected grid *i*. Two factors, sensitivity and LS budget, are considered for selecting the site to operate at V_{DDH}. To start with, the site with the maximum sensitivity is selected. Then, the number of usage level shifters is checked if the site is assigned to operate at V_{DDH}. If the number of usage level shifters is larger than the LS budget of the selected grid, the selected site is not assigned to operate at V_{DDH}, and a new site with the second largest sensitivity is selected until the site meets these two constraints at the same time. After that, the selected site in grid *i* is assigned to operate at V_{DDH} for rescuing the timing. When the site is assigned to operate at V_{DDH}, the timing and sensitivity information of gates are affected and should be updated. The number of sites in $grid_i$ can be reduced during the assignment procedure. After the updating procedure, the next site with the maximum sensitivity is selected, and the assignment step is executed repeatedly. In this step, the selection and updating procedure are executed repeatedly until no site in grid *i*, or all sites in grid *i* have been selected and assigned. Therefore, the maximum possible times for executing the re-assignment procedure are equal to the initial number of sites in grid *i*, and the computation load of updating sensitivity is reduced.

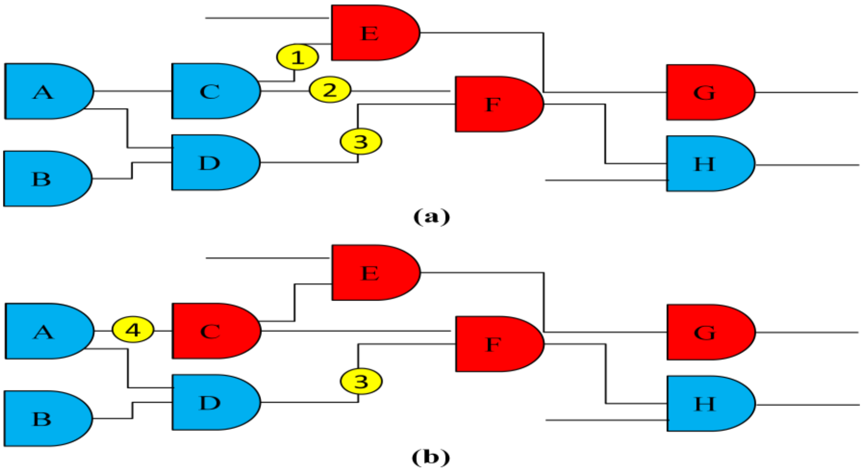


Fig. 4. Schematic of LS budget rescue

E. Timing and LS Budget Rescue

E.1、 Timing Rescue

The *3D IC Voltage Assignment* result might lead the circuit timing to be violated because the LS budget is limited. Hence, we can not arbitrarily assign V_{DDH} to the site. Therefore, a timing rescue is executed to rescue the timing of circuit. For example, the gates with red color are sites and the gates with blue color are operating at V_{DDL} as shown in the circuit schematic of Fig. 4.(a). First, we compute the gain of each site and sort them. The gain is the delay difference between V_{DDH} and V_{DDL}. The site with the maximum gain is selected firstly and assigned to operate at V_{DDH}. Then, the fan-in gates of the site are selected. For example, we firstly select the fan-in gate *C* with the maximum AT to operate at V_{DDH}, and update its timing information. If the maximum AT is changed from gate *C* to gate *D* after gate *C* is assigned to operate at V_{DDH}, we must assign V_{DDH} to gate *D* again. In this way, all gates of fan-in of the site are checked until the dominant gate is found. Although the timing is rescued, the number of level shifters might be over the LS budget due to executing the timing rescue procedure without considering LS budget. Therefore, we have to sacrifice the power consumption by assigning more gates with V_{DDL} to be V_{DDH} for reducing the number of level shifters. In the following, we are going to perform our method for rescuing the level shifter usage to meet the LS budget.

E.2、 LS Budget Rescue

Intuitively, we can start with the fan-in gate of the LS, and assign V_{DDH} to it. This way not only reduces the usage of level shifters but also maintains the correctness of circuit. In Fig. 4.(a), the gates with yellow color are level shifters, the red color gates operate at V_{DDH}, and the blue color gates operate at V_{DDL}. The LS budget rescue step finds out all level shifters, and checks whether their fan-in gates with V_{DDL} can be changed to V_{DDH} for reducing the number of level shifters. For example, Fig. 4.(a) shows that gate *C* is the fan-in of LS (No. 1 and No. 2), and gate *D* is the fan-in of LS (No. 3). The gain of gate *C* is equal to $-2+1=-1$, and the gain of gate *D* is equal to $-1+2=1$. Obviously, gate *C* reduces one LS but gate *D* increases one LS. Therefore, we should select gate *C* to operate at V_{DDH} for reducing the usage of level shifters as shown in Fig. 4.(b). Similarly, the gains of gate *A* and *D* in Fig. 4.(b) are checked again. The rescue procedure stops until no gate can be assigned to operate at V_{DDH} for reducing the usage of level shifters.

五、實驗方法與結果

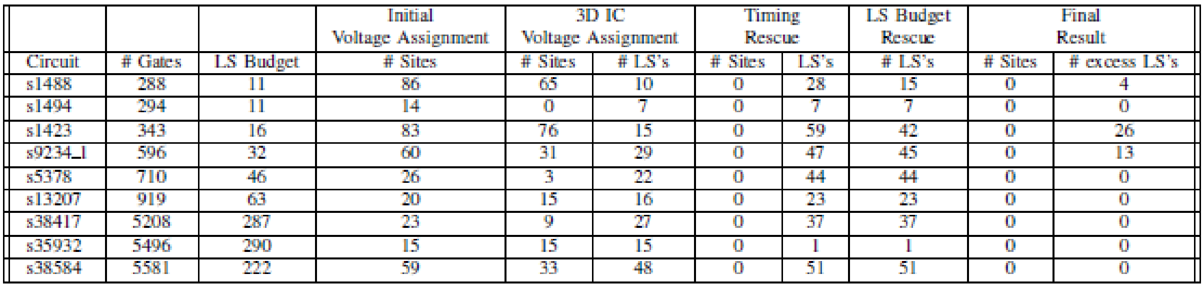
We implement our proposed method in C++ language and apply the algorithm to a set of ISCAS89 benchmark circuits and private designs. First, Design Compiler is used to synthesize the benchmark circuits with the UMC 90nm standard cell library. Next, the initial 2D placement of each test circuit is generated by the SOC Encounter, and its related 3D placement is obtained by transforming the 2D placement with Z—Place [9]. The timing/leakage power cell library with temperature effect is generated by evaluating the average leakage current and the delay of gate based on H-SPICE simulation for various types of logic gates. After getting the H-SPICE simulation results, based on the least square fitting method, the fitting constants of the leakage current and delay models are obtained. For simplicity, a single type of LS is used in the experimental results.

A. Comparison of Voltage Assignment Results

TABLE I lists the results of voltage assignment in different phases. The number of sites after *Initial Voltage Assignment* is listed in column 4. Columns 5-6 show the number of sites and the usage of level shifters after *3D IC Voltage Assignment*. The results of *Timing Rescue* and *LS Budget Rescue* are listed in columns 7-8 and column 9, respectively. Finally, the summarization is listed in columns 10-11.

First, the results of *3D IC Voltage Assignment* show that the timing rescue after initial voltage assignment is limited significantly by the LS budget. Most circuits cannot meet timing constraints after *3D IC Voltage Assignment* procedure under the limited LS budget constraint. Next, to deal with this problem, we try to rescue the sites of circuit by *Timing Rescue*. After executing *Timing Rescue* procedure, all circuits meet timing constraints finally. However, the usage of level shifters is more than LS budget because the timing rescue procedure only considers the influence of voltage assignment on timing and does not take LS budget into consideration. Finally, we try to rescue the circuit again by *LS Budget Rescue*. The results of column 11 show that it still has three circuits that are rescued unsuccessfully.

For reducing the problem size, *3D IC Voltage Assignment* method limits the voltage assignment decision to the sites, and obviously the number of sites should not be large in a circuit. Moreover, we think that the site is the most important gate for rescuing the timing. However, based on sites, it is not enough for the timing rescue because of the LS budget constraint.



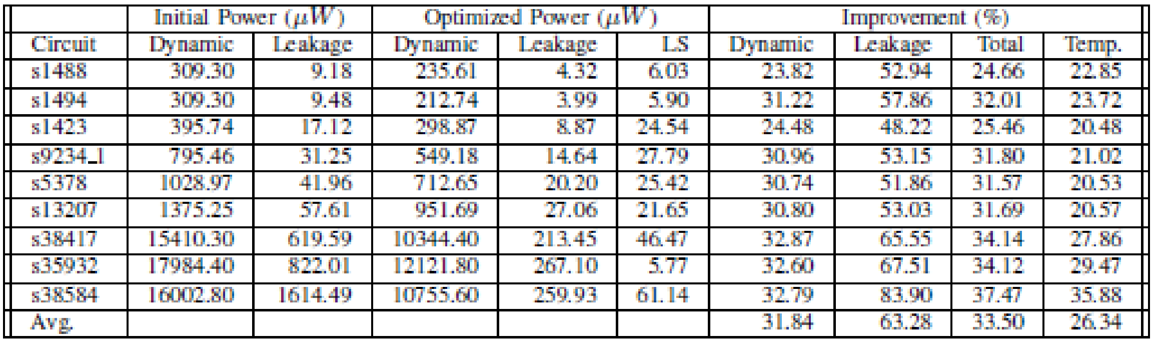
			Initial Voltage Assignment	3D IC Voltage Assignment		Timing Rescue		LS Budget Rescue	Final Result	
Circuit	# Gates	LS Budget	# Sites	# Sites	# LS's	# Sites	LS's	# LS's	# Sites	# excess LS's
s1488	288	11	86	65	10	0	28	15	0	4
s1494	294	11	14	0	7	0	7	7	0	0
s1423	343	16	83	76	15	0	59	42	0	26
s9234_1	596	32	60	31	29	0	47	45	0	13
s5378	710	46	26	3	22	0	44	44	0	0
s13207	919	63	20	15	16	0	23	23	0	0
s38417	5208	287	23	9	27	0	37	37	0	0
s35932	5496	290	15	15	15	0	1	1	0	0
s38584	5581	222	59	33	48	0	51	51	0	0

TABLE I. Results of the Proposed Voltage Assignment Method

B. Power Reduction

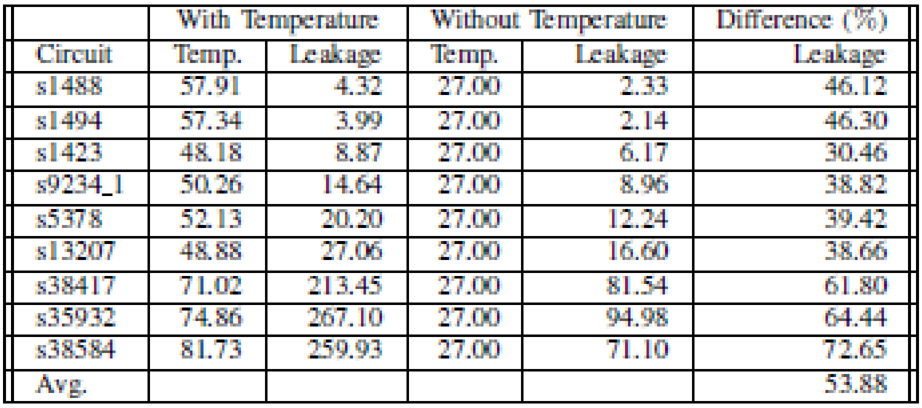
TABLE II summarizes the optimization results. TABLE I shows that three circuits are failed, and six circuits are successful after executing the proposed power reduction method. Therefore, the average of improvement is the average of the six circuit results. The initial power and average temperature of the circuit are listed in columns 2-4, and the power and average temperature of the circuits and the power of level shifters after optimization are listed in columns 5-8. The improvement percentages of the dynamic power, the leakage power and the total power are listed in columns 9-11, respectively. The temperature decrement of each circuit is listed in column 11. The improvement columns indicate that the proposed *3D IC Voltage Assignment* method can provide almost *33.50%* total power saving and *26.34* degree decrement of temperature in average. It can be observed that the leakage power reduction is greatly improved because of the temperature decrement.

TABLE III shows the leakage power and temperature estimation with the simulated temperature in columns 2-3 and without the simulated temperature in columns 4-5, and the percentage differences of leakage power in column 6. As shown in TABLE III, the full chip leakage power analysis without accurate temperature can lead to *53.88%* error in average. If the leakage power is underestimated, the power reduction can be dominated by the dynamic power, which is quite impractical.



	Initial Power (μW)		Optimized Power (μW)			Improvement (%)			
Circuit	Dynamic	Leakage	Dynamic	Leakage	LS	Dynamic	Leakage	Total	Temp.
s1488	309.30	9.18	235.61	4.32	6.03	23.82	52.94	24.66	22.85
s1494	309.30	9.48	212.74	3.99	5.90	31.22	57.86	32.01	23.72
s1423	395.74	17.12	298.87	8.87	24.54	24.48	48.22	25.46	20.48
s9234_1	795.46	31.25	549.18	14.64	27.79	30.96	53.15	31.80	21.02
s5378	1028.97	41.96	712.65	20.20	25.42	30.74	51.86	31.57	20.53
s13207	1375.25	57.61	951.69	27.06	21.65	30.80	53.03	31.69	20.57
s38417	15410.30	619.59	10344.40	213.45	46.47	32.87	65.55	34.14	27.86
s35932	17984.40	822.01	12121.80	267.10	5.77	32.60	67.51	34.12	29.47
s38584	16002.80	1614.49	10755.60	259.93	61.14	32.79	83.90	37.47	35.88
Avg.						31.84	63.28	33.50	26.34

TABLE II. Results of Power Optimization



	With Temperature		Without Temperature		Difference (%)
Circuit	Temp.	Leakage	Temp.	Leakage	Leakage
s1488	57.91	4.32	27.00	2.33	46.12
s1494	57.34	3.99	27.00	2.14	46.30
s1423	48.18	8.87	27.00	6.17	30.46
s9234_1	50.26	14.64	27.00	8.96	38.82
s5378	52.13	20.20	27.00	12.24	39.42
s13207	48.88	27.06	27.00	16.60	38.66
s38417	71.02	213.45	27.00	81.54	61.80
s35932	74.86	267.10	27.00	94.98	64.44
s38584	81.73	259.93	27.00	71.10	72.65
Avg.					53.88

TABLE III. Leakage Power Estimation

六、結論與討論

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

七、成果

[1] Huai-Chung Chang, Pei-Yu Huang, Ting-Jung Li, and Yu-Min Lee, *"Statistical Electro-Thermal Analysis with High Compatibility of Leakage Power Models,"* International SoC Conference (SOCC), 2010.

[2] Shu-Han Whi and Yu-Min Lee, *"Dual Supply Voltage Assignment in 3D ICs Considering Thermal Effects,"* The 16th Workshop on Synthesis And System Integration of Mixed Information Technologies (SASIMI), 2010.

[3] Yu-Min Lee and Chi-Wen Pan, *"*Redundant Via Insertion with Wire Spreading Capability*,"* International Journal of Electrical Engineering (IJEE), vol. 17, no. 6, pp. 383-398, December 2010.

[4] Chi-Wen Pan, Yu-Min Lee and Chih-Sheng Wang, *"Redundant Via Insertion under Timing Constraints,"* International Symposium on Quality Electronic Design (ISQED), 2011.

[5] Shu-Han Whi and Yu-Min Lee, *"Supply Voltage Assignment for Power Reduction in 3D ICs Considering Thermal Effect and Level Shifter Budget,"* International Symposium on VLSI Design, Automation and Test (VLSI-DAT), 2011.

[6] Pei-Yu Huang and Yu-Min Lee, *"Statistical Hot-Spot Identification Using On-Chip Thermal Yield Profile,"* VLSI Design/CAD Symposium (VLSI/CAD), 2011.

[7] Pei-Yu Huang and Yu-Min Lee, *"On-Chip Statistical Hot-Spot Estimation Using Mixed-Mesh Statistical Polynomial Expression Generating and Skew-Normal Based Moment Matching Techniques,"* Accepted by Asia South Pacific Design Automation Conference (ASPDAC), 2012.

八、參考文獻

[1] V. Reddy and A. T. Krishnan. Impact of negative bias temperature instability on digital circuit reliability. Proceedings of IRPS, pages 248-254, 2002.

[2] S. H. Kulkarni and A. N. Srivastava. A new algorithm for improved VDD assignment in low power dual VDD systems. Proceedings of ISLPED, pages 200-205, 2004.

[3] H. Wu and I. M. Liu. Post-placement voltage island generation under performance requirement. Proceedings of ICCAD, pages 309-316, 2005.

[4] R. L. S. Ching and E. F. Y. Young, E.F.Y. Post-placement voltage island generation. Proceeding of ICCAD, pages 641-646, 2006.

[5] H. Wu and M. D. F. Wong. Timing-constrained and voltage-island-aware voltage assignment. Proceedings of DAC, pages 432, 2006.

[6] S. A. Yu and P. Y. Huang and Y. M. Lee. A multiple supply voltage based power reduction method in 3-D ICs considering process variations and thermal effects. Proceedings of ASPDAC, pages 55-60, 2009.

[7] H. F. Dadgour and S. C. Lin. A statistical framework for estimation of full-chip leakage-power distribution under parameter variations. IEEE Transactions on Electron Devices, 54(11):2930-2945, 2007.

[8] P. Y. Huang and Y. M. Lee. Full-chip thermal analysis for the early design stage via generalized integral transforms. IEEE Transactions on Very Large Scale Integration Systems, 17(4):613-626, 2009.

[9] R. Hentschke and G. Flach. 3D-vias aware quadratic placement for 3D VLSI circuits. Proceedings of ISVLSI, pages 67-72, 2007.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

99 年 10 月 14 日

附件三

報告人姓名	李亭蓉	服務機構及職稱	國立交通大學電信工程系(所) 碩士
會議 時間 地點	99.9.27-99.9.29 美國、內華達州 (Nevada, US)	本會核定補助文號	NSC 98-2220-E-009 -058 - NSC 99-2220-E-009 -035 -
會議名稱	(中文)第 23 屆國際系統晶片會議 (英文) 23rd IEEE International SoC Conference (SOCC 2010)		
發表論文題目	(中文)對功率模型具有高度相容性的統計型電熱分析 (英文)Statistical Electro-Thermal Analysis with High Compatibility of Leakage Power Models		

報告內容應包括下列各項：

一、 參加會議經過

本次會議除了口頭論文發表、海報論文發表之外；主辦單位並且邀請不同領域的專家針對不同的系統晶片設計考量方向給與前瞻的演講。整個會議中聽取了功率考量專家剖析現在及未來能量的降低之於系統晶片、微處理器及計算系統的發展趨勢，並與數個發表論文的作者及相關業界的學者專家討論研究議題內容。

此研討會一直以來都是受研究單位矚目的國際會議之一，吾人非常榮幸有機會參加此次2010年在美國內華達州(Nevada, U.S.A)舉行的會議。會議分為口頭發表及壁報發表兩部分，此國際會議中發表之文章不只範圍廣且技術先進，因此參加此會議不僅可增進自己研究領域之知識，亦可了解現今系統晶片的趨勢。

我們在此次會議發表論文為對功率模型具有高度相容性的統計型電熱分析。論文全文請參見附件。

二、 與會心得

參加本屆國際系統晶片會議，令我獲益良多。不僅吸收到眾多設計系統晶片上考量的方針與趨勢；如，為節省能源消耗及手持(handheld)便利性綠系統("Green" System)。經由口頭論文發表會，接觸到不同子領域中解決相似問題的演算法，並且獲得其他專家們所提出的改進方向與建議。

此次會議中主要有幾部分：系統晶片的能源最佳化技術及電路、類比電路、系統設計方法、通訊電路系統和嵌入式記憶體系統。

三、 建議

與會人士除了歐美國家外，韓國、日本、及中國都是積極參與國際的會議，如果往後能多鼓勵參予類似活動，對於國際交流與合作上會有很大的幫助，也可以藉由接觸國外學者獲得更廣的國際觀，增加研究的能力。

四、 攜回資料名稱及內容

會議論文海報集光碟：集合發表於此研討會議中所有論文及海報內容。

會議手冊：所有演講、論文及海報的摘要還有會議議程。

國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：100年 3 月20日

計畫編號	NSC99－2220－E－009－035		
計畫名稱	針對 3D 整合之電子設計自動化技術開發－子計畫一：三維度積體電路的隨機電熱模擬及其對功率最佳化的應用(2/2)		
出國人員姓名	潘麒文	服務機構及職稱	國立交通大學電信工程研究所 SOC 組 博士班三年級
會議時間	100 年 3 月 14 日至 100 年 3 月 16 日	會議地點	美國 聖克拉拉(Santa Clara)
會議名稱	(中文) 電子設計品質會議 (英文)11th International Symposium on Quality Electronic Design		
發表論文題目	(中文)在時序限制下的冗餘接點安插 (英文) Redundant Via Insertion under Timing Constraints		

一、參加會議經過

本次 2011 年 ISQED 的舉辦地點為美國的聖克拉拉，我們所發表的論文：Redundant Via Insertion under Timing Constraints，很榮幸在 ISQED 的論文甄選中，成為被選中的論文之一。在這三天的會議中，將近有 14 個 workshops 在這段時間內進行，也因為時間緊迫，很多 workshop 都是同時進行。因此，在這段時間內可以自由選擇自己有興趣的題目去參加會議，瞭解到現在電子設計自動化上的發展趨勢以及一些相關應用成果，同時也可以知道其他學者所發表的論文成果。

到會場做完報到的手續之後，就是自行決定要去聽哪一個演說，雖然中間有一

些同時進行的 session，不過像是 demo 和 poster 這類的 session，就可以先去 demo 的會場看每一個的主題，在那邊聽作者的解釋、說明、示範，也有人提供實機讓你借出會場去測試使用。大致看完 demo 後，可以再去外面看 poster，幾乎每個海報都會有作者在旁邊解說。雖然這兩場是同時進行，可是時間安排剛剛好，不會讓人有一下子就看完的感覺。另外像是 Best papers 的 session，雖然他是在一個很大的會議廳，可是參加人數也非常的多，太晚到甚至有可能找不到座位。

本人是在會議的最後一天上台報告，當天有 4 個 session 同時進行，因此每間會議室的人都不像之前那麼多。此外，在會議進行時有發生一點小意外，就是其中一位演講者的投影片數據變成亂碼，當時該位教授還開玩笑說是因為水土不服嗎，也讓人感受到隨機應變的重要性，也很欽佩演講者沒有被投影片失常而影響後面的報告。在同一場 session 中碰到幾個同樣來自交大的學生，也算是一個很特別的經驗。

二、與會心得

這次出國參加 ISQED 2011，是本人第一次參加大型的國際研討會，也是第一次去美國，很高興也很榮幸有這個機會可以去參加研討會。在研討會的中途休息時間可以看到有部份學者、研究人員在討論剛剛聽得演講內容，不禁讓人覺得佩服，也期許自己可以像他們一樣。

由於是第一次參加大型研討會，很多事情都不知道該怎麼做，幸好同行者有這方面的經驗，很多事情都請他幫忙處理，在這過程中也瞭解到事先多做一點功課，之後到美國那邊就比較輕鬆了。美國地大物博，在聖克拉拉如果沒有交通工具，真的是行動不便，好加在還有一些鐵路系統，可以帶我們到比較繁榮的市區走走•在飯店裡，可以看到來自不同國家的人，我發現歐美的人都比較熱情，會主動向我們

打招呼，而華人通常都是一群一群的自成一個團體•

在參與別的 session 時，會去另外注意別人製作簡報的方式，研究如何使用簡單明瞭的方法讓聽眾很快的吸收、理解。也因為這樣，一直到上台報告的前幾天，我還是一直在調整自己的簡報，使用一些動畫的方式來強調重點，希望讓聽眾能夠較易瞭解。可惜的是有部份 session 是同時進行，無法聽完所有的演講，因此只能從演講主題下去挑選要參加的研討會。

住在旅館時，旅館有提供無線網路的服務，可是在房間內使用常常會覺得訊號差以及連線不穩定，無法從網路查資料，這也突顯了事先準備的重要性•這趟行程讓我見識到大型國際研討會的規模與水準，以及來自各地的研究人員，深感國際競爭的壓力，我們必須各加倍專注於我們的學術研究•

三、考察參觀活動(無是項活動者略)

四、建議

可以在國內多舉辦類似的大型國際研討會，並邀請一些國會知名學者出席演講，以提升會議的規模與水準，與世界接軌•

五、攜回資料名稱及內容

11th ISQED 光碟x1 : 內含本次會議的所有論文資料

六、其他

---------- Forwarded Message -----------
From: SOCC2010@ieee-socc.org
To: yumin@cm.nctu.edu.tw
Cc: SOCC2010@ieee-socc.org
Sent: Sat, 26 Jun 2010 06:14:30 -0700
Subject: Your SOCC 2010 Submission (Number 68)

Dear Prof. Yu-Min Lee:

On behalf of the SOCC 2010 Program Committee, we are pleased to inform you that the following submission has been accepted to appear at the conference as a regular paper:

Statistical Electro-Thermal Analysis with High
Compatibility of Leakage Power Models

Please revise your paper according to the the reviews. Your final manuscript will appear in the proceedings. The manuscript is limited to SIX pages. The deadline for submission is Friday July 9, 2010.

To upload your final manuscript, please visit the following
site:

https://www.softconf.com/b/socc2010/

and, on the left-hand side of the page, enter the passcode associated with your submission. Your passcode is as follows:

68X-F8B3B5H7B5

Alternatively, you can click on the following URL, which will take you directly to a form to submit your final paper:

https://www.softconf.com/b/socc2010/cgi-bin/scmd.cgi?
scmd=aLogin&passcode=68X-F8B3B5H7B5

The reviews and comments are attached below. Please try to follow the reviewers' advice when you revise your paper.

Congratulations on your fine work.   If you have any additional questions, please feel free to get in touch.

Best Regards,
Ramalingam Sridhar and Norbert Schuhmann SOCC 2010

---------------------------------- 原文 ----------------------------------

主旨:　Your ISQED 2011 Submission (Number 238)

寄件者: kamesh.gadepally@gmail.com

日期:　四, 十一月 25, 2010 7:01 am

收件者: wayne1234.cm97g@nctu.edu.tw

Dear Mr. Chi-Wen Pan:

On behalf of the ISQED 2011 Program Committee, I am delighted to inform you that the following submission has been accepted to appear at the conference:

Redundant Via Insertion under Timing Constraint

The Program Committee worked very hard to thoroughly review all the submitted papers.　Please repay their efforts, by following their suggestions when you revise your paper.

To upload your final manuscript, please visit the following
site:

https://www.softconf.com/b/isqed2011/

and, on the left-hand side of the page, enter the passcode associated with your submission.　Your passcode is as follows:

238X-F5G3P6H9C5

Alternatively, you can click on the following URL, which will take you directly to a form to submit your final paper:

https://www.softconf.com/b/isqed2011/cgi-bin/scmd.cgi?scmd=aLogin&passcode=238X-F5G3P6H9C5

The reviews and comments are attached below.　Again, try to follow their advice when you revise and improve the quality of your paper.

Congratulations on your fine work.　If you have any additional questions, please feel free to get in touch.

Best Regards,
Kamesh Gadepally - ISQED2011 TPC Chair
Keith Bowman , ISQED2011 TPC Co-Chair
ISQED 2011

STATISTICAL ELECTRO-THERMAL ANALYSIS WITH HIGH COMPATIBILITY OF LEAKAGE POWER MODELS

Huai-Chung Chang, Pei-Yu Huang, Ting-Jung Li and Yu-Min Lee
National Chiao Tung University, Hsinchu, Taiwan

Abstract— **In this work, a statistical electro-thermal analyzer with high compatibility of power model is developed. The developed analyzer takes both the easily implementing advantage of Monte Carlo method and the fast convergent advantage of stochastic analysis method to effectively solve the statistical electro-thermal problem. Experimental results indicate that the developed electro-thermal analyzer can be orders of magnitude faster than the Monte Carlo method under the same accuracy level. The computational time is only 1.16 seconds for a design with over one million gates, and the maximum errors are only 0.34% and 1.84%, compared with the Monte Carlo method, for estimating the mean and the standard deviation profiles of full-chip temperature distribution, respectively.**

I. INTRODUCTION

Power dissipation and thermal effect are important issues of VLSI design as the technology continuously scales down, and the power density rapidly increases. The chip-temperature profiles and gradients significantly influence on IC performance, reliability, and package cost. Because leakage power contributes a large portion of total power in the modern technology, it is necessary to model and estimate leakage power accurately. Furthermore, the leakage power of a circuit element exponentially depends on its operating temperature and process parameters. Hence, process variations and thermal impacts need to be cautiously considered.

In recent years, several thermal-power related analysis methods have been proposed. In the power analysis, [1]–[3] quantified process variations of leakage power. Nevertheless, none of them simultaneously considers the statistical power and the electro-thermal effect. In the thermal analysis, [4] proposed a deterministic electro-thermal analyzer considering the temperature dependence of leakage power.

To include process variation effects, the electro-thermal simulation needs to be considered as a statistical fashion to ensure design reliability. Hence, several statistical thermal analyzers were developed [5], [6]. However, [5] didn't consider electro-thermal coupling. Though [6] presented a statistical electro-thermal analysis, it needs to re-fit the leakage power model as the design or its geometry changes because its model fits the leakage power of each temperature grid rather than that of each gate. This limits its usage for early physical design stages. Moreover, both of them need *specified* leakage power models for the power projection [5] and the iteratively log-normal approximation [6].

Because the scaling technology can lead more complicated leakage power models for enhancing the accuracy, it is urgent to develop a statistical thermal analyzer with the high capability of accurate but complicated leakage power models. Compared with [5], [6], our developed statistical electro-thermal analyzer is more applicable since we take the advantage of sparse grid collocation technique [7] to avoid the convoluted statistical calculation algorithm. The sparse grid collocation technique has been adopted in thermal-power related researches such as building leakage power models [3] and analyzing statistical leakage power [2]. However, both of [2], [3] didn't consider and indicate how to treat temperature dependence issues in their power analysis methods.

In this work, we will present how to easily, accurately and efficiently solve the statistical electro-thermal problem with any temperature-dependent leakage power models. Moreover, rather than [6], the developed electro-thermal analyzer doesn't need to re-fit leakage power models during thermal-driven early physical design stages such as floorplanning or placement because the cell based leakage power models are adopted. Firstly, the Karhunen-Loève (KL) expansion is used to transform spatially correlated physical parameters to a set of uncorrelated random variables. Then, the Smolyak sparse grid formulation [7] is applied to obtain the sampling values of physical parameters for obtaining the deterministic power models in executing deterministic electro-thermal simulations. After a set of deterministic electro-thermal simulations being solved, the Newton interpolating formula is utilized to calculate the expression coefficients of temperature profile. Finally, the statistical characteristics of temperature distribution can be extracted.

Our major contributions are

1) This work presents an easily, accurately and efficiently statistical electro-thermal simulation, and it has the high compatibility to incorporate any power models.
2) The developed statistical electro-thermal analyzer can accurately and efficiently provide the mean and standard deviation profiles of full-chip temperature distribution.
3) Experimental results reveal that ignoring electro-thermal coupling in statistical thermal analysis can lead to significant errors of full-chip temperature distribution.

This paper is organized as follows. Firstly, the leakage power modeling and the problem formulation are described in section II. After that, the proposed statistical electro-thermal analyzer is detailed in section III. Finally, experimental results and conclusion are given in sections IV and V, respectively.

II. LEAKAGE POWER MODELING AND PROBLEM FORMULATION

A. Leakage Power Modeling

Many leakage power models were developed in [1], [2], [5], [6], [8]. However, none of them in [1], [2], [5] simultaneously considered temperature and process variation effects. Hence, their accuracy degrades as the technology scales down. For the authors' best knowledge, only [6], [8] simultaneously considered both effects. Nevertheless, the leakage current model in [8] was based on $90nm$ technology. Hence, as the technology advances, its accuracy deteriorates shown in TABLE I. A grid-based leakage power model was developed in [6]. Each fitted model was used to coarsely approximate the total leakage current in each grid, and this limits its usage after the floorplanning stage.

978-1-4244-6683-2/10/$26.00 ©2010 IEEE

TABLE I

ERROR COMPARISON OF I_s AND I_g WITH THE RESULTS OF HSPICE UNDER $65nm$ TECHNOLOGY FOR AN NAND GATE.

f_g		**Max Error**	**Avg. Error**	**Error $>$ 3%**
Without temperature	$T_{ox}, T_{ox}^2, L_{ch}, L_{ch}^2$ [1]	6.48%	2.70%	4.37%
With temperature	**Our adopted model**: a polynomial function constructed by L_{ch}, T, T_{ox} and T_{ox}^2	1.55%	0.29%	0.00%
f_s		**Max Error**	**Avg. Error**	**Error $>$ 3%**
Without temperature	$L_{ch}, L_{ch}^2, T_{ox}^{-1}, T_{ox}^2$ [1]	347.32%	70.65%	98.27%
	$L_{ch}, L_{ch}^2, T_{ox}^{-1}, T_{ox}, T_{ox}^2, T_{ox}/L_{ch}, L_{ch}/T_{ox}, T_{ox}L_{ch}$ [2]	314.13%	70.52%	100.00%
With temperature	L_{ch}, T, T_{ox} [8]	32.23%	8.73%	76.62%
	Our adopted model: a 3^{rd} order polynomial function completely expanded by L_{ch}, T_{ox} and T	1.31%	0.19%	0.00%

Here, a cell-based leakage power fitting model including the process variation effect and temperature dependence is presented. Firstly, for each cell, different input patterns, various physical process parameter values and operating temperatures are combined and put into HSPICE with industrial design kit under the BSIM4 model to generate its leakage current data. After that, the average leakage currents of input patterns are fitted by the least square fitting method. Finally, the fitted coefficients of different average leakage current models such as the average subthreshold leakage (I_s) and the average gate tunneling leakage (I_g) can be obtained.

Since I_s is the off-state leakage, and I_g occurs in both on and off states of transistor, the cell leakage power can be written as

$$P_{leakage} \quad = \quad V_{dd} \times (I_g + (1 - Sw)\, I_s)\,, \tag{1}$$

where

$$I_g \quad = \quad a_0 \cdot e^{f_g(T_{ox}, L_{ch}, T)}, \tag{2}$$

$$I_s \quad = \quad b_0 \cdot e^{f_s(T_{ox}, L_{ch}, T)}. \tag{3}$$

Here, a_0 and b_0 are fitted constants, L_{ch} and T_{ox} are the channel length and oxide thickness, respectively. T is the operating temperature, Sw is the switching activity, V_{dd} is the supply voltage, and $f_g(T_{ox}, L_{ch}, T)$ and $f_s(T_{ox}, L_{ch}, T)$ are specific fitting forms[1]. I_g is modeled as exponentially dependent on temperature since it is exponentially affected by the threshold voltage [9].

The accuracy of several existing leakage current models has been investigated [1], [2], [8]. Because they do not present enough accuracy, much more accurate leakage current models are proposed in this work. The error comparison of existing leakage current models and the proposed leakage current models for an two-input NAND gate is shown in TABLE I.

As shown in TABLE I, different f_g and f_s lead to different errors compared with the results of HSPICE. The drastic errors of [1], [2], [8] are because of the ignorance of either temperature or developing technology. The maximum error and average error of proposed models are less than 1.55% and 0.29%, respectively. Actually, for all cell types given by an industrial design kit, the maximum error and average error of our developed leakage current models are only 1.55% and 0.5%, respectively.

With the above demonstration, the leakage current (power) model might be very complicated for achieving acceptable accuracy. This fact indicates that the statistical power analyzer or the statistical thermal analyzer should have the ability to handle complicated leakage current (power) models.

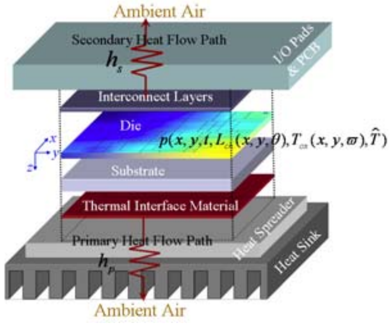


Fig. 1. Compact thermal model of physical design.

B. Problem Formulation

The compact thermal model of a chip for the physical design stage is shown in Fig. 1 [10]. The primary heat flow path is composed of thermal interface material, heat spreader and heat sink. The secondary heat flow path involves interconnect layers, I/O pads, and the print circuit board. The functional blocks are modeled as many power sources attached to the thin layer close to the top surface of die. The main heat sources consist of the dynamic and leakage power consumed by devices. Because the dynamic power is insensitive to process variations and operating temperature [1], it is viewed to be deterministic. However, the leakage power is strongly dependent on process parameters and operating temperature. Hence, the leakage power is viewed as random processes [1], and the thermal coupling effect needs to be considered for the full-chip temperature distribution analysis. By combining the compact thermal model and the statistical power consumption considering the thermal coupling effect, the steady state temperature distribution $T(\mathbf{r}, \theta, \varpi)$ of die is determined by the following statistical steady-state heat transfer equation.

$$\nabla \cdot (\kappa(\mathbf{r}, T) \nabla T(\mathbf{r}, \theta, \varpi)) = -p(\mathbf{r}, L_{ch}(x, y, \theta), T_{ox}(x, y, \varpi), T), \tag{4}$$

subject to the following boundary condition

$$\kappa(\mathbf{r}_{b_s}, T) \frac{\partial T(\mathbf{r}_{b_s}, \theta, \varpi)}{\partial n_{b_s}} + h_{b_s} T(\mathbf{r}_{b_s}, \theta, \varpi) = f_{b_s}(\mathbf{r}_{b_s}). \tag{5}$$

Here, ∇ is the diverge operator, and $\kappa(\mathbf{r}, T)$ is the thermal conductivity of die. The $p(\mathbf{r}, L_{ch}(x, y, \theta), T_{ox}(x, y, \varpi), T)$ is the random process of power density profile which consists of the dynamic power density profile $p_d(\mathbf{r})$, the sub-threshold leakage power density profile $p_s(\mathbf{r}, L_{ch}(x, y, \theta), T_{ox}(x, y, \varpi), T)$, and the gate leakage power density profile $p_g(\mathbf{r}, L_{ch}(x, y, \theta), T_{ox}(x, y, \varpi), T)$. The $\mathbf{r} = (x, y, z) \in D$, $D = (0, L_x) \times (0, L_y) \times (-L_z, 0)$ is the domain of die, L_x and L_y are lateral sizes of die, and L_z is the thickness of die. The θ and ϖ are sampling values of manufacturing outcomes $\Omega_{L_{ch}}$ and $\Omega_{T_{ox}}$ for the channel length and oxide thickness, respectively. The $L_{ch}(x, y, \theta)$ and $T_{ox}(x, y, \varpi)$ are the random processes of the device channel

[1] We consider the variations of the device channel length and the oxide thickness since the leakage power is very sensitive to them [1]. It should be noted that although only these two parameters are considered, our framework can be easily extended to include the effects of any process variation parameters such as the channel dopant variation, etc.

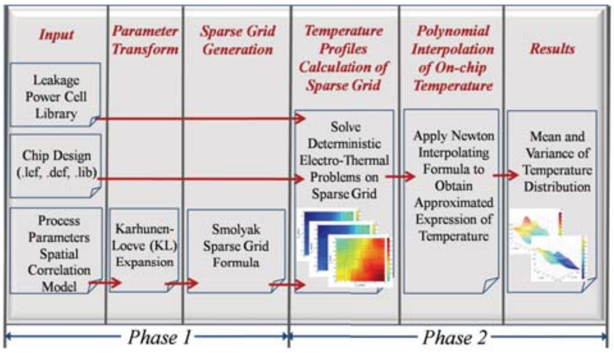

Input
Parameter Transform
Sparse Grid Generation
Temperature Profiles Calculation of Sparse Grid
Polynomial Interpolation of On-chip Temperature
Results
Leakage Power Cell Library
Chip Design (.lef, .def, .lib)
Process Parameters Spatial Correlation Model
Karhunen-Loeve (KL) Expansion
Smolyak Sparse Grid Formula
Solve Deterministic Electro-Thermal Problems on Sparse Grid
Apply Newton Interpolating Formula to Obtain Approximated Expression of Temperature
Mean and Variance of Temperature Distribution
Phase 1
Phase 2

Fig. 2. The flowchart of proposed statistical electro-thermal analyzer.

length and the oxide thickness, respectively. The b_s is any specific boundary surface of the die, and $\mathbf{r}_{b_s}$ is the position located on b_s. The h_{b_s} is the heat-transfer coefficient on b_s, $f_{b_s}(\mathbf{r}_{b_s})$ is the heat flux on b_s, and $\partial/\partial n_{b_s}$ is the differential operator along the outward direction normal to b_s. Since the major part of device current passes through the region close to the channel, the power density profile has its value only in that region which its thickness is equal to the junction depth for dynamic and sub-threshold leakage power and is equal to the Debye length for gate tunneling leakage power.

With equations (4)–(5), our goal is to evaluate the mean and variance profiles of steady-state full-chip temperature distribution.

III. Proposed Statistical Electro-Thermal Analyzer

The flowchart of proposed statistical electro-thermal analyzer is shown in Fig. 2. Each operation in *Phase 1* is only related with the technology node rather than design pattern, and operations of *Phase 2* are design dependent.

In *Phase 1*, given a spatial covariance function of physical parameters, the KL expansion is employed to decompose correlated parameters into a set of uncorrelated random variables. After that, the Smolyak sparse grid formula is used to generate sparse grids, which are a set of sampling random vectors of KL expanded and inter-die random variables. In *Phase 2*, with each sampled random vector on sparse grids, a deterministic electro-thermal simulation with the deterministic power profile obtained by this sampled random vector is performed. Then, with all thermal profiles corresponding to sampled random vectors on sparse grids, an approximated representation of stochastic full-chip temperature distribution is obtained by the Newton interpolating formula. Finally, the statistical characteristics, such as mean and variance profiles, of the full-chip temperature distribution are extracted.

Different from the existing statistical thermal/electro-thermal analyzers [5], [6], the proposed framework can easily, accurately and efficiently obtain an approximated expression of the full-chip temperature distribution without suffering from complicated statistically calculating algorithms such as the power projection [5] and the iteratively statistical temperature moment extraction [6]. This is because each power profile corresponding to each sampled random vector is deterministic during each deterministic electro-thermal analysis being performed. Hence, accurate but complicated leakage power models can be adopted in this framework. Each step in Fig. 2 is detailed in the rest subsections.

A. Parameter Transformation

Generally, process variations of one physical parameter P can be classified into intra-die $\triangle \mathrm{P}^{intra}$ and inter-die $\triangle \mathrm{P}^{inter}$ variations which both can be modeled as Gaussian random variables [1]. The physical parameter $\mathrm{P} \in \{T_{ox}, L_{ch}\}$ with its expected value $\overline{\mathrm{P}}$ at position $\mathbf{r}_{xy} = (x, y)$ can be written as

$$T_{ox}(\mathbf{r}_{xy},\varpi)=\overline{T}_{ox}(\mathbf{r}_{xy})+\Delta T_{ox}^{intra}(\mathbf{r}_{xy},\varpi_i)+\Delta T_{ox}^{inter}(\mathbf{r}_{xy},\varpi_j)\,, \quad (6)$$

$$L_{ch}(\mathbf{r}_{xy},\theta)=\overline{L}_{ch}(\mathbf{r}_{xy})+\Delta L_{ch}^{intra}(\mathbf{r}_{xy},\theta_i)+\Delta L_{ch}^{inter}(\mathbf{r}_{xy},\theta_j)\,. \quad (7)$$

The ϖ_i and ϖ_j are subsets of ϖ, and θ_i and θ_j are subsets of θ.

According to [1], $T_{ox}(\mathbf{r}_{xy}, \varpi) = T_{ox}(x, y, \varpi)$ is assumed to be spatially uncorrelated[2]. Because the spatial correlation of $\Delta L_{ch}^{intra}(\mathbf{r}_{xy}, \theta_i)$ might have different decreasing rates in x- and y-directions, the spatial covariance function proposed in [11] is adopted for $\Delta L_{ch}^{intra}(\mathbf{r}_{xy}, \theta_i)$[3]. Given σ as the standard deviation of target random process, and correlation lengths η_x and η_y in x- and y-directions, respectively, the spatial covariance function between two random variables at points $\mathbf{r}_{x_1y_1}$ and $\mathbf{r}_{x_2y_2}$ is

$$C(\mathbf{r}_{x_1y_1}, \mathbf{r}_{x_2y_2}) = \sigma^2 e^{-\frac{|x_1-x_2|}{\eta_x}} e^{-\frac{|y_1-y_2|}{\eta_y}}. \quad (8)$$

With applying the KL expansion, $\Delta L_{ch}^{intra}(\mathbf{r}_{xy}, \theta_i)$ based on (8) can be approximated as

$$\Delta L_{ch}^{intra}(\mathbf{r}_{xy}, \theta_i) \approx \sum_{m=1}^{N_{L_{ch}}} \sqrt{\chi_m} q_m(\mathbf{r}_{xy}) \zeta_m(\theta_i). \quad (9)$$

Here, χ_m's are eigenvalues of $C(\mathbf{r}_{x_1y_1}, \mathbf{r}_{x_2y_2})$, q_m's are related eigenvectors, and $N_{L_{ch}}$ is the expansion length. $\{\zeta_m(\theta_i)\}$ is the set of uncorrelated standard normal random variables.

Because of the KL expansion property, the expanded random variables are Gaussian random variables if the target random process is Gaussian, and the eigen-pair $(\chi_m, q_m(\mathbf{r}_{xy}))$ closed form can be derived [12]. In the rest of this paper, $\zeta = \{\zeta_m\}$ and $\varsigma = \{\varsigma_n\}$ are sets of random variables to represent L_{ch} and T_{ox}, respectively, $\tilde{\xi} = \zeta \cup \varsigma$, and θ and ϖ are dropped for the sake of notation simplicity.

B. Smolyak Sparse Grid Formulation

The basic idea of Smolyak sparse grid formulation is to build an interpolating approximation of a high dimensional multivariate-function $u \in C^r$ by much less sampling values of the desired function than the full tensor product interpolation formula but with an acceptable error bound in the order of $O(M^{-r} \log M^{(d-1)(r-1)})$ [13]. Here, M is the number of sampling points, and d is the number of variables.

For the Monte Carlo method, the random variable samples are randomly generated, and a large number of samples is required to achieve accurate mean and variance estimation. For the Smolyak sparse grid formulation, the random variable samples are generated by using roots of Hermite polynomial chaos (H-PCs) or extrema of the Chebyshev polynomial [14], and the desired solution is obtained by using interpolation with these samples.

[2] Although T_{ox} is assumed to be spatially uncorrelated, the proposed simulation mechanism still works for T_{ox} being spatially correlated.

[3] Although we choose this specific spatial covariance function (8), any valid spatial covariance functions can be adopted

The high order interpolating approximation can be achieved with a small number of samples [13].

According to the Smolyak sparse grid formulation [7], our desired full-chip statistical temperature distribution $T(\mathbf{r},\tilde{\xi})$ represented by a set of KL expanded random variables $\tilde{\xi}$ can be explicitly approximated as [15]

$$\tilde{T}_q^d(\mathbf{r},\tilde{\xi}) = \sum_{q-d+1\le|\mathbf{i}|\le q} (-1)^{q-|\mathbf{i}|}\binom{d-1}{q-|\mathbf{i}|}(Q^{i_1}(T)\otimes\cdots\otimes Q^{i_d}(T)), \qquad (10)$$

where d is the number of random variables in $\tilde{\xi}$, q is the level of desired solution, Q^{i_n} with the level $i_n \ge 1$ is the one-dimensional interpolating operator of $T(\mathbf{r},\tilde{\xi})$ with respect to the n-th random variable in $\tilde{\xi}$, $\otimes$ is the functional cross product, and $|\mathbf{i}| = i_1 + \cdots + i_n + \cdots + i_d$. The level i_n is the index to decide the number of sampling values for the interpolating polynomial Q^{i_n}. As suggested in [16], the relation between the number of sampling values m_{i_n} and the level i_n is $m_1 = 1$ and $m_{ij_n} = 2^{i_n-1}+1$ for $i_n > 1$.

From (10), we only need to know the temperature on the following small set of sampling values for $\tilde{\xi}$ [17]. The sparse grid, the set of sampling values of $\tilde{\xi}$, in (10) is derived as

$$H(q,d) = \bigcup_{q-d+1\le|\mathbf{i}|\le q} \left(\vartheta^{i_1}\times\cdots\times\vartheta^{i_n}\times\cdots\times\vartheta^{i_d}\right), \qquad (11)$$

where ϑ^{i_n} denotes the set of sampling points of $\tilde{\xi}_n$, and '$\times$' is the cross product of the points of set.

The number of sampling points from Smolyak sparse grid formulation increases as $O(\frac{d^{q-d}}{(q-d)!})$ that is less severe than that of full tensor product formulation. The runtime complexity of our proposed statistical electro-thermal analyzer can be analyzed to be $O(\mathrm{C}_{\det}\frac{d^{q-d}}{(q-d)!})$. The $\mathrm{C}_{\det}$ is the runtime complexity for executing a deterministic electro-thermal simulation.

The sampling values corresponding to ϑ^{i_n} must be properly decided. Adopting the roots of H-PCs with its order being corresponding to the level i_n can achieve the most accurate result if $\tilde{\xi}$ is a set of normal random variables [14]. On the other hand, adopting the extrema of the Chebyshev polynomial with its order being corresponding to the level i_n can achieve the nested sparse grid structure for any levels and acceptable accuracy [16]. In this paper, we adopt the roots of H-PCs in our experimental implementation because the results are shown to be very accurate by using the low level approximation, and the nested sparse grid structure is still preserved for $q = d + 1$[4].

C. Calculation of Temperature Profiles on Sparse Grids

After the sparse grid $H(q,d)$ being obtained, the samples of channel length and oxide thickness corresponding to the m-th sampling grid $\tilde{\xi}^m$ of $H(q,d)$ can be obtained by using the parameter modeling technique stated in section III-A. Hence, the deterministic power density profile corresponds to $\tilde{\xi}^m$ can be obtained. With the deterministic power density profile, we have the following deterministic steady heat transfer equation.

$$\nabla\cdot\left(\kappa(\mathbf{r},T)\nabla T(\mathbf{r},\tilde{\xi}^m)\right) = -p(\mathbf{r},\tilde{\xi}^m,T), \qquad (12)$$

subject to the following boundary condition

$$\kappa(\mathbf{r}_{b_s},T)\frac{\partial T(\mathbf{r}_{b_s},\tilde{\xi}^m)}{\partial n_{b_s}} + h_{b_s}T(\mathbf{r}_{b_s},\tilde{\xi}^m) = f_{b_s}(\mathbf{r}_{b_s}). \qquad (13)$$

Here, $p(\mathbf{r},\tilde{\xi}^m,T)$ and $T(\mathbf{r},\tilde{\xi}^m)$ are deterministic power density and temperature profiles with respect to $\tilde{\xi}^m$, respectively. Since the power density profile is temperature dependent in equation (12), the deterministic electro-thermal analysis is used to get each $T(\mathbf{r},\tilde{\xi}^m)$ and is summarized in Fig. 3.

```
Algorithm Calculation of Temperature Profiles on Sparse Grid
Input:Sampling point $\tilde{\xi}^i$, initial temperature $T^{ini}$ and $p_{dyn}(\mathbf{r})$
Output: Stable temperature profiles $T(\mathbf{r},\tilde{\xi}^i)$ of $\tilde{\xi}^i$
Begin
  Obtain $T_{ox}(\mathbf{r}_{xy},\tilde{\xi}^i)$ and $L_{ch}(\mathbf{r}_{xy},\tilde{\xi}^i)$ according to $\tilde{\xi}^i$;
  $T(\mathbf{r},\tilde{\xi}^i) \leftarrow T^{ini}$;
  $T'(\mathbf{r},\tilde{\xi}^i) \leftarrow 0$;
  While ($T(\mathbf{r},\tilde{\xi}^i) - T'(\mathbf{r},\tilde{\xi}^i) \le$ Converging criterion)
      $T'(\mathbf{r},\tilde{\xi}^i) \leftarrow T(\mathbf{r},\tilde{\xi}^i)$;
      Update $p_{leakage}(\mathbf{r},\tilde{\xi}^i,T)$ by $T(\mathbf{r},\tilde{\xi}^i)$;
      $p_{total}(\mathbf{r},\tilde{\xi}^i,T) \leftarrow p_{leakage}(\mathbf{r},\tilde{\xi}^i,T) + p_{dyn}(\mathbf{r})$;
      † Solve deterministic thermal equations (12) and (13)
        with $p_{total}(\mathbf{r},\tilde{\xi}^i,T)$ to obtain a new $T(\mathbf{r},\tilde{\xi}^i)$;
      if ($T(\mathbf{r},\tilde{\xi}^i)$ = Infinite) then Thermal runaway;
  Return $T(\mathbf{r},\tilde{\xi}^i)$
End
† The deterministic thermal analyzer [18] is used to obtain $T^*$.
  Any deterministic thermal analyzer can be used here.
```

Fig. 3. Deterministic electro-thermal analysis for each sampling point in sparse grid. $p_{leakage}$, p_{dyn} and p_{total} are the leakage, dynamic and total power density profiles for each sampling point of sparse grid, respectively.

D. Polynomial Interpolation of Temperature Distribution

Instead of directly using equation (10) which requires to obtain different $Q^{i_1}(T)\otimes\cdots\otimes Q^{i_d}(T)$ for each different $|\mathbf{i}| = i_1+\cdots+i_d$, we take the advantage of nested sparse grid structure and then perform one time of Newton interpolating method [14] to globally interpolate $T(\mathbf{r},\tilde{\xi})$ by the deterministic temperature profiles of all sampling values in sparse grid. For the sparse grid that can not preserve nested structure, the Newton interpolating method can be applied to obtain each different $Q^{i_1}(T)\otimes\cdots\otimes Q^{i_d}(T)$.

Based on the Newton interpolating formula, the temperature at the specified die position $\mathbf{r}^*$ can be approximated as

$$T(\mathbf{r}^*,\tilde{\xi}) \approx \sum_{m=0}^{m=N} \hat{a}_m(\mathbf{r}^*)\phi_m(\tilde{\xi}), \qquad (14)$$

where $\phi_m(\tilde{\xi})$ is an interpolating polynomial with respect to the m-th sampling value $\tilde{\xi}^m$, and its form can be found in [14]. The $N = |H(q,d)| - 1$, $|H(q,d)|$ is the number of sampling values in sparse grid, and $\hat{a}_m(\mathbf{r}^*)'s$ need to be determined.

Based on the basic idea of interpolation that the approximated function must match each known data, the interpolated polynomial in (14) must satisfy equation (15) for each $\tilde{\xi}^k$.

$$\sum_{m=0}^{m=N} \hat{a}_m(\mathbf{r}^*)\phi_m(\tilde{\xi}^k) = T(\mathbf{r}^*,\tilde{\xi}^k). \qquad (15)$$

Based on the property of $\phi_n(\tilde{\xi})$ [14], the matrix equation (16)

[4] If a highly order approximation is required for the accuracy, we suggest the extrema of the Chebyshev polynomial because its nested sparse grid structure is preserved for any levels; hence, it needs much less sampling points than that of choosing the roots of the H-PCs for highly order approximation.

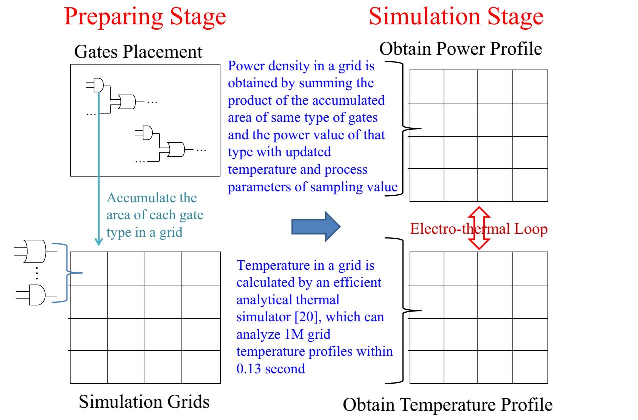

Preparing Stage
Simulation Stage
Gates Placement
Obtain Power Profile
Power density in a grid is obtained by summing the product of the accumulated area of same type of gates and the power value of that type with updated temperature and process parameters of sampling value
Accumulate the area of each gate type in a grid
Electro-thermal Loop
Temperature in a grid is calculated by an efficient analytical thermal simulator [20], which can analyze 1M grid temperature profiles within 0.13 second
Simulation Grids
Obtain Temperature Profile

Fig. 4. The implementation of proposed electro-thermal simulation for each sampling grid.

can be obtained to find each $\hat{a}_m(\mathbf{r}^*)$ at chip position $\mathbf{r}^*$.

$$\begin{bmatrix} \phi_0\left(\tilde{\xi}^0\right) & 0 & \cdots & 0 \\ \phi_0\left(\tilde{\xi}^1\right) & \phi_1\left(\tilde{\xi}^1\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0\left(\tilde{\xi}^N\right) & \phi_1\left(\tilde{\xi}^N\right) & \cdots & \phi_N\left(\tilde{\xi}^N\right) \end{bmatrix} \begin{bmatrix} \hat{a}_0\left(\mathbf{r}^*\right) \\ \hat{a}_1\left(\mathbf{r}^*\right) \\ \vdots \\ \hat{a}_N\left(\mathbf{r}^*\right) \end{bmatrix} = \begin{bmatrix} T\left(\mathbf{r}^*, \tilde{\xi}^0\right) \\ T\left(\mathbf{r}^*, \tilde{\xi}^1\right) \\ \vdots \\ T\left(\mathbf{r}^*, \tilde{\xi}^N\right) \end{bmatrix} \tag{16}$$

Each $\hat{a}_m(\mathbf{r}^*)$ can be calculated in linear time since the system matrix of (16) is a lower triangular matrix. After each $\hat{a}_m(\mathbf{r}^*)$ has been calculated, the mean and variance profiles of the full-chip temperature distribution can be estimated as

$$E\{T(\mathbf{r}^*, \tilde{\xi})\} = E\left\{\sum_{m=0}^{m=N} \hat{a}_m(\mathbf{r}^*)\phi_m(\tilde{\xi})\right\}, \tag{17}$$

$$Var\{T(\mathbf{r}^*, \tilde{\xi})\} = Var\left\{\sum_{m=0}^{m=N} \hat{a}_m(\mathbf{r}^*)\phi_m(\tilde{\xi})\right\}. \tag{18}$$

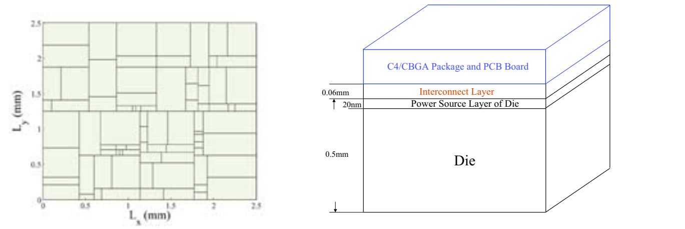
E. *Implementation of the Deterministic Thermal Simulation*

Fig. 4 shows the implementation of electro-thermal simulation for each sampling point. In the preparing stage, the accumulated area of each same type gate in each temperature grid is obtained. With the accumulated area of each different gate type in a temperature grid, the power density in that grid is got by summing each product of the accumulated area of a specific gate type and its power value (dynamic power plus its leakage power calculated at the sampled process parameters and operating temperature). Hence, the complexity to get the power profile for each sampling point is $O(P \times Q \times N_{type})$. P and Q are numbers of divisions of grids along $x-$ and $y-$directions, respectively. N_{type} is the number of gate types and is determined by the cell library, and it is far less than the number of simulation grids $P \times Q$.

For the deterministic electro-thermal simulation with respect to a sampling point a deterministic thermal solver [18], which can get the temperature profile for one million grids in 0.13 seconds, is adopted. In the experimental setting, the number of simulation grid is 128×128, and the runtime is 0.018 seconds for executing one time of [18] in the electro-thermal loop shown in Fig. 3.

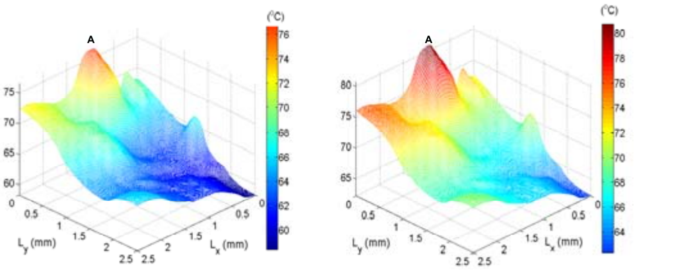
IV. Experimental Results

The developed analyzer is implemented in C++ language and tested on a Linux system with Intel Xeon 3.0-GHz CPU and 32 GB memory. The die size is $2.5mm \times 2.5mm \times 0.5mm$. The junction depth is $20nm$ that is the nominal value for the $65nm$ technology [19], and the Debye length is $2nm$ [20]. The floorplanning of test chip has 1.2 million functional gates as shown in Fig. 5(a), and the geometries of chip are shown in Fig. 5(b). By applying the modeling skill of thermal parameter and iterative 1-D thermal computation scheme [10], the equivalent heat transfer coefficients of primary and secondary heat flow paths, and the average thermal conductivity of chip material are $12000\mathrm{W}/(\mathrm{m}\cdot{}^\circ\mathrm{C})$, $2017\mathrm{W}/(\mathrm{m}\cdot{}^\circ\mathrm{C})$, and $148.13\mathrm{W}/(\mathrm{m}\cdot{}^\circ\mathrm{C})$, respectively. The boundary condition of each vertical surface is isothermal [18]. The number of simulation grids for executing [18] is 128×128 at the top surface of die.



(a) (b)

Fig. 5. (a) The floorplanning of test chip. (b) The geometry of test chip.



(a) (b)

Fig. 6. Temperature profile at the top surface of die. (a) The mean temperature distribution without including the electro-thermal coupling. (b) The mean temperature distribution with including the electro-thermal coupling.

The nominal values of channel length and oxide thickness are $65nm$ and $1.5nm$, respectively. The $3\sigma_{L_{ch}}$ and $3\sigma_{T_{ox}}$ are set to 12% and 5% of nominal values, respectively. Both η_y/L_y and η_x/L_x are set to 0.98 that means that the correlation between two devices located half of the chip dimension away in the x-direction or the y-direction is 0.6.

A. *Accuracy and Efficiency*

To verify the analyzer, the Monte Carlo method is also implemented with 10^5 samples as reference solutions considering the same issues such as the electro-thermal coupling effect, spatially intra-die variations and inter-die variations. The proposed electro-thermal analyzer takes 10 random variables to expand process variations and uses Smolyak sparse grid formula with $q = 11$. Hence, the stochastic thermal profile of test chip is interpolated by 21 individual sampling points. The results with three different ratios of inter-die variations and intra-die variations to the total variations in a reasonable region are shown in TABLE II.

Compared with the golden solution, the proposed analyzer is extremely accurate and can be finished in seconds for the test chip. For example, as the inter-die variation is 50% of total variations, the maximum errors of the calculated spatial mean distribution and spatial standard deviation distribution for the

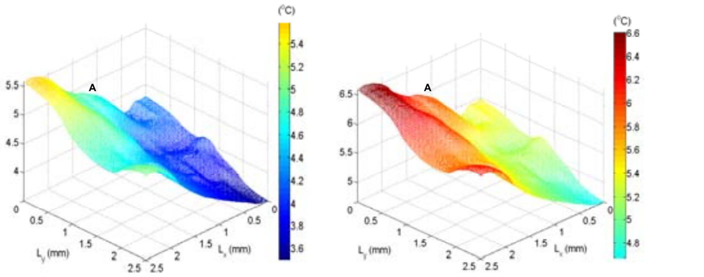
TABLE II

ACCURACY AND EFFICIENCY COMPARED WITH THE MONTE CARLO METHOD.

Inter-die / Total Variations	Intra-die / Total Variations	Our Proposed Method†				Monte Carlo‡		Speedup (X)
		maximum mean error	maximum std. error	runtime (s)		sampling knots	runtime (s)‡	
				Phase 1	Phase 2			
40%	60%	0.31%	1.68%	3.23	1.16	6736	357.88	308.51
50%	50%	0.32%	1.84%	3.27	1.16	6465	347.82	299.84
60%	40%	0.34%	1.81%	3.40	1.16	6422	341.72	294.59

† Our proposed method is compared with the golden solution constructed by the Monte Carlo method using 10^5 samples.

‡ To show the efficiency, the Monte Carlo method is simulated till the standard deviation achieves the same accuracy as the proposed method. The runtime does not include the time of input parser which is only performed once in the Monte Carlo simulation.



(a) (b)

Fig. 7. Temperature profile at the top surface of die. (a) The spatial standard deviation distribution without including electro-thermal coupling. (b) The spatial standard deviation distribution with including electro-thermal coupling.

full-chip temperature distribution are only 0.34% and 1.84%, respectively. The execution time is only 3.27 seconds and 1.164 seconds in *Phase 1* and *Phase 2*, respectively.

Since operations in *Phase 1* are irrelevant to design patterns, they only need to be pre-performed once while applying the analyzer to thermal-aware design procedures. Hence, to show the efficiency of proposed analyzer, the runtime of *Phase 2* is compared with the runtime through the Monte Carlo simulation fulfilling the same accuracy of standard deviation as ours. TABLE II shows that the developed analyzer is orders of magnitude faster than the Monte Carlo method. Since each sampling point is independent, the parallel programming technique can be used to further enhance the speedup.

B. Without v.s. With Including Electro-Thermal Coupling Effect

Fig. 6 and Fig. 7 show the spatial mean distribution and spatial standard deviation distribution of temperature distribution at the top surface of test chip, respectively. Fig. 6(a) and Fig. 7(a) are the results without considering the electro-thermal coupling effect. Fig. 6(b) and Fig. 7(b) are the results with considering the electro-thermal coupling effect. These two figures reveal the dramatic differences of these two results. As we can see, the difference of spatial mean profile can reach 6.54%, and the difference of spatial standard deviation profile is over 25.01%. This impact shows that it is necessary to consider the electro-thermal coupling effect for the statistical thermal analysis, and the developed analyzer can accurately and efficiently achieve this goal.

V. CONCLUSION

An efficient statistical electro-thermal analyzer with high compatibility of power models has been presented. The analyzer can efficiently provide the accuracy results and is high capability with any leakage power models and the spatial correlation function. The experimental results also indicated that it is not allowable to ignore the electro-thermal coupling effect when considering process variations in statistical thermal simulation.

ACKNOWLEDGMENT

This work was supported in part by the National Science Council of Taiwan under Grant NSC 98-2220-E-009-058.

REFERENCES

[1] H. Chang and S. S. Sapatnekar. Prediction of leakage power under process uncertainties. *ACM TODAES*, 12:1–27, April 2007.

[2] R. Shen, N. Mi., and S. Tan. Statistical modeling and analysis of chip-level leakage power by spectral stochastic method. *Proc. ASP-DAC*, pages 31–6, June 2009.

[3] S. Bhardwaj, S. Vrudhula, and A. Goel. A unified approach for full chip statistical timing and leakage analysis of nanoscale circuits considering intra-die process variations. *IEEE TCAD*, 27(10):1812–1825, 2008.

[4] Y. K. Cheng, P. Raha, C. C. Teng, E. Rosenbaum, and S. M. Kang. ILLIADS-T: an electrothermal timing simulator fortemperature-sensitive reliability diagnosis of CMOS VLSI chips. *IEEE TCAD*, 17(8):668–81, 1998.

[5] P. Y. Huang, J. H. Wu, and Y. M. Lee. Stochastic thermal simulation considering spatial correlated within-die process variations. *Proc. ASP-DAC*, pages 55–60, June 2009.

[6] J. Jaffari and M. Anis. Statistical thermal profile considering process variation: Analysis and appllications. *IEEE TCAD*, 27:1027–40, June 2008.

[7] S. A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR*, pages 240–3, 1963.

[8] S. A. Yu, P. Y. Huang, and Y. M. Lee. A multiple supply voltage based power reduction method in 3-d ics considering process variations and thermal effects. *Proc. ASP-DAC*, pages 55–60, June 2009.

[9] K. M. Cao, W. C. Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu. BSIM4 gate leakage model including source-drain partition. pages 815–818, 2000.

[10] 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 TVLSI*, 14:501–13, May 2006.

[11] S. Bhardwaj, S. Vrudhula, P. Ghanta, and Y. Cao. Modeling of intra-die process variations for accurate analysis and optimization of nanoscale circuits. *Proc. DAC*, pages 791–6, 2006.

[12] B. Cline, K. Chopra, D. Blaauw, and Y. Cao. Analysis and modeling of cd variation for statistical static timing. *Proc. ICCAD*, pages 60–6, 2006.

[13] J. Taylor and F. Hover. High dimensional stochastic simulation and electric ship models. *Proc. ESTS*, 21-3:402–7, May 2007.

[14] G. M. Phillips. *Interpolation and Approximation by Polynomials*. Springer, 2003.

[15] G. W. Wasilkowski and H. Wozniakowski. Explicit cost bounds of algorithms for multivariate tensor product problems. *J. Complexity*, 11(1):1–56, 1995.

[16] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Advan. Comput. Math.*, 12(4):273–88, 2000.

[17] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SINUM*, pages 2309–45, May 2008.

[18] P. Y. Huang and Y. M. Lee. Full-chip thermal analysis for the early design stage via generalized integral transforms. *IEEE TVLSI*, 17:613–26, May 2009.

[19] F. Lallement, B. Duriee, A. Grouillet, F. Amaud, B. Tavel, F. Wacquant, P. Stalk, M. Woo, Y. Erokhin, J. Scheuer, L. Gadet, J. Weeman, D. Distaso, and D. Lenoble. Ultra-low cost and high performance 65nm cmos device fabricated with plasma doping. *Symp. VLSl Technol. Dig. Tech. Papers*, pages 178–9, 2004.

[20] J. Bienacel, D. Barge, M. Bidaud, N. Emonet, D. Roy, L. Vishnubhotla, I. Pouilloux, and K. Barla. Anticipation of nitrided oxides electrical thickness based on XPS measurement. *Materials Science in Semiconductor Processing*, 7(4-6):181–3, 2004.

Redundant Via Insertion under Timing Constraints

Chi-Wen Pan, Yu-Min Lee
National Chiao Tung University, Hsinchu, Taiwan
wayne1234.cm97g@nctu.edu.tw, yumin@cm.nctu.edu.tw

Abstract— **Redundant via insertion is a useful technique to alleviate the yield loss and elevate the reliability of VLSI designs. While extra vias are inserted into the design, the electronic properties of designed circuit might be altered, and the circuit timing might be changed and needs to be efficiently re-analyzed. Therefore, a fast timing (incremental timing) analyzer is required to assistant the redundant via insertion procedure.**

This work develops an efficient redundant via insertion method under timing constraints. Firstly, an effectively incremental circuit timing analysis method is developed, and the redundant via insertion task is transformed into a mixed bipartite-conflict graph matching problem. Then, the insertion problem is solved by a timing-driven minimum weighted matching algorithm.

The experimental results show that the developed algorithm can achieve 3.2% extra insertion rates over the method without considering timing effects, which all redundant vias would be removed if the timing of that net does not meet the timing requirements, in average. In addition, the developed incremental timing analysis mechanism can speed up the runtime of redundant via insertion procedure under timing constraints by over 34 times in average.

I. Introduction

As the feature size continuously scales down, yield becomes an important issue for the current modern design. Yield-loss comes from many physical factors, and via failure is one major factor caused by electromigration, thermal stress and random defects. Partial via failure increases circuit resistances that result in the unexpected timing delay, and complete via failure can break the net connection and lead to the inaccurate signal that makes the circuit fail. Without violating any design rules, inserting a redundant via (RV) adjacent to a single via as a safeguard is a widely recommended method for improving the circuit yield and reliability [1], [2]. With the double vias, the via failure rate can be dramatically reduced as reported in [3].

Many researches performed the RV insertion procedure during the routing stage [4]–[6]. Xu et al. [4] developed a maze routing with considering RV insertion for the yield enhancement. Yao et al. [5] minimized the via usages and inserted redundant vias in the routing stage to improve the multilevel routing framework. Chen et al. [6] proposed a full-chip gridless routing method with considering double-via insertion based on the bipartite matching graph algorithm. However, they stacked the vertical vias as one stack via that the solution space was reduced. Besides, many researches worked on the post-routing stage [7], [8]. Lee et al. [7] proposed a zero-one integer linear program to solve double-cut via insertion problem and handle the via density constraint. Lei et al. [8] transformed the RV insertion task into a mixed bipartite matching graph and presented a heuristic minimum weighted matching algorithm to solve the problem. They also developed the wire spreading method to insert redundant vias for dead vias.

Inserting RV into the circuit might vary the timing characteristic of design. Luo et al. [9] created a set of untouched nets such as the timing critical nets or the nets specified by customers etc., and utilized the geotopological technology to insert redundant vias. Although they considered the timing issue, the insertion algorithm cannot guarantee whether the circuit timing is degraded or not after the insertion. Chiang et al. [10] developed a two-phase insertion approach method with considering timing constraints. However, they only checked the timing behavior after executing the insertion procedure and simply got rid of inserted redundant vias violating timing constraints. In other words, they didn't immediately update the timing behavior during the RV insertion procedure. Therefore, the insertion rate is degraded.

Recently, Lin et al. [11] pointed out that "How to tackle the timing issue more accurately during double-via insertion is still worthy further study.". The experimental results illustrated in TABLE IV also show that the ratio range of net sink node delay differences after performing the conventional (without considering the timing effects) redundant via insertion method [8] on test circuits can be $35\% \sim 58\%$. Therefore, it is necessary to develop timing-driven redundant via insertion methods.

Because the timing information should be frequently predicted and updated during the RV insertion process, a fast timing analysis method is required. In this work, firstly, an incremental analysis method is developed to effectively perform the timing analysis and predict the timing behavior for the RV insertion process. The developed incremental analysis method does not need to recalculate the entire net timing to predict the timing effect of adding an extra via. It only needs to analyze the timing influence induced by the modified circuit parts. Therefore, the computational load of timing analysis can be dramatically reduced. After that, the proposed incremental timing analysis method is incorporated with the mixed bipartite-conflict (MBC) graph [8] for developing a timing-driven minimum weighted matching (*t*-MWM) algorithm to solve the RV insertion task.

To the best of our knowledge, this is the *first* redundant via insertion method that *truly* considers the timing effects of inserted vias during the insertion procedure.

The paper is organized as follows. The redundant via insertion problem under timing constraints is formulated in section II. Then, an efficient incremental timing analysis method for the redundant via insertion problem is detailed in section III, and the MBC graph [8] is briefly presented in section IV. After that, the developed *t*-MWM algorithm and the experimental results are presented and discussed in section V and section VI, respectively. Finally, the conclusion is given in section VII.

978-1-61284-914-0/11/$26.00 ©2011 IEEE

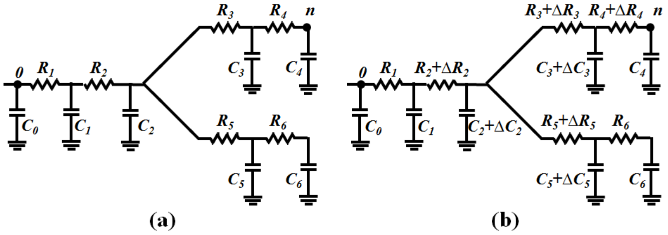


Fig. 1. A simple RC tree. (a) An original RC tree. (b) A locally modified RC tree.

II. Problem Formulation

Given a post-routing design, its routed netlist and each net timing constraint[1], the RV insertion problem under timing constraints is to simultaneously insert extra vias as many as possible to the design and satisfy the given timing constraints. The problem can be formulated as follows.

$$\max \sum_{i}^{N} \sum_{k}^{M_i} Rv_{ik},$$

subject to

$$D_j^i \leq \overline{D_j^i}, \qquad \forall\,(\text{net}\, i)\ \&\ (\text{sink}\, j\, \text{of net}\, i)$$

Here,

$$Rv_{ik} \quad = \quad \begin{cases} 1 & \text{if an RV is inserted to a single via } k \text{ of net } i \\ 0 & \text{otherwise.} \end{cases}$$

N is the number of nets in the given design, M_i is the number of single vias on net i, D_j^i is the propagation delay from the source of net i to its sink j, and $\overline{D_j^i}$ is the timing constraint of sink j on net i.

III. Incremental Timing Analysis

A. Incremental Timing Formula

A simple RC tree shown in Fig. 1 is used to describe the closed form of the incremental timing formula. Given an original RC tree shown in Fig. 1.(a) and its locally modified RC tree shown in Fig. 1.(b), their Elmore[2] delay values from node 0 to node n can be calculated as

$$\begin{aligned} t_n^{(\mathrm{a})} &= R_1 \sum_{i=1}^{6} C_i + R_2 \sum_{j=2}^{6} C_2 + R_3 \sum_{k=3}^{4} C_k + R_4 C_4 \\ t_n^{(\mathrm{b})} &= t_n^{(\mathrm{a})} + \triangle t_n. \end{aligned}$$

Here, their timing difference $\triangle t_n$ can be derived as

$$\begin{aligned} \triangle t_n &= \triangle C_2 \sum_{i=1}^{2} R_i + \triangle C_3 \sum_{j=1}^{3} R_j + \triangle C_5 \sum_{k=1}^{2} R_k \\ &\quad + \triangle R_2 \sum_{i=2}^{6} C_i + \triangle R_3 \sum_{j=3}^{4} C_j \\ &\quad + \triangle R_2 \sum_{i=2,3,5} \triangle C_i + \triangle R_3 \triangle C_3. \end{aligned} \tag{1}$$

Here, each ΔC_i is the difference value of nodal capacitance at node i after and before via insertion, and each ΔR_j is the difference value of branch resistance on branch j after and before via insertion. The expression of (1) consists of three types. Each term in the first line is the product of a delta nodal capacitance and its original upstream common path equivalent resistance. The original upstream common path equivalent resistance is the sum of original overlapped segment resistances between the paths from source node 0 to the nodal-capacitance changed node and from source node 0 to node n in Fig. 1.(a). Each term in the second line is the product of a delta branch-resistance (on the path from source node 0 to node n) and its original downstream equivalent capacitance. Each term in the last line is an interactive term that is the product of a delta branch-resistance (on the path from source node 0 to node n) and its delta downstream equivalent capacitance.

With (1), the timing difference $\triangle t_n$ for each node n after a RC tree is locally modified can be generally formulated as follows.

$$\Delta t_n = \sum_i R_{ni} \Delta C_i + \sum_{j \in P_n} \Delta R_j C_{j*} + \sum_{j \in P_n} \Delta R_j \Delta C_{j*} \tag{2}$$

Here, R_{ni} is the original upstream common path equivalent resistance between P_n and P_i, and P_n/P_i is the routed path from source node to node n/i. C_{j*} is the original equivalent downstream capacitance seen from branch j, and ΔC_{j*} is the difference value of equivalent downstream capacitance seen from branch j after and before via insertion.

By utilizing the timing difference formula of node n shown in (2), it can be very efficient to predict each specific net sink node delay without recalculating all node delays of a specific net.

B. Redundant Via Shapes

According to the connection between wires and the single via, the structures of redundant vias can be categorized into two different shapes. If a single via only connects two wire segments, it is called an *L-shape* since this structure looks like an alphabet letter 'L'. If a single via connects more than two wire segments, it is called a *T-shape* since it looks like an alphabet letter 'T'. Moreover, according to the characteristic of redundant vias, they can be divided into two types, on-track RV and off-track RV. The on-track RV is placed on the original net and needs only one extra wire segment, and an off-track RV is placed out of the original net and needs to add two extra wire segments.

Since *L-shape* RVs and *T-shape* RVs are cyclic circuits, in order to effectively calculate their Elmore delay values, several fundamental circuit transformation techniques are utilized to make them acyclic. The equivalent acyclic RC circuits of each *L-shape* RV and each *T-shape* RV are derived as follows to incrementally update the circuit timing efficiently.

B.1. L-shape

The *L-shape* RVs contain three templates that two types are on-track RVs, and one type is off-track RV. These templates and their corresponding RC trees are shown in Fig. 2. Fig. 3 shows the equivalent acyclic RC circuits of above three *L-shape* RV templates, and their related ΔR values and ΔC values are summarized in TABLE I.

B.2. T-shape

According to the transformation methods, *T-shape* RVs can be divided into *simple T-shape* RVs and *complicated T-shape*

[1] In this work, the main issue is how to consider the timing effects during the redundant via insertion procedure. Hence, we simplify the problem with the net timing constraint; however, our developed method also works with the circuit timing constraint.

[2] Although the Elmore delay [12] is used as the delay metric in this work, the proposed incremental delay analysis method can be easily extended to any higher order delay metrics.

TABLE I

ΔR VALUES AND ΔC VALUES OF THE EQUIVALENT ACYCLIC RC CIRCUIT OF L-SHAPE RV AND SIMPLE T-SHAPE RV.

L-shape RV	ΔR			ΔC			
	ΔR_1	ΔR_2	ΔR_3	ΔC_0	ΔC_1	ΔC_2	ΔC_3
on-track type I	$-r_{w1}x$	$\frac{(r_{w1}x+r_v l_v)(r_{w2}x+r_v l_v)}{(r_{w1}+r_{w2})x+2r_v l_v} - r_v l_v$	0	$-\frac{c_{w1}x}{2}$	$\frac{c_{w2}x+c_v l_v}{2}$	$\frac{c_{w1}x+c_{w2}x+c_v l_v}{2}$	0
on-track type II	0	$\frac{(r_{w1}x+r_v l_v)(r_{w2}x+r_v l_v)}{(r_{w1}+r_{w2})x+2r_v l_v} - r_v l_v$	$-r_{w2}x$	0	$\frac{c_{w1}x+c_{w2}x+c_v l_v}{2}$	$\frac{c_{w1}x+c_v l_v}{2}$	$-\frac{c_{w2}x}{2}$
off-track type	0	$\frac{(r_{w1}x+r_v l_v)(r_{w2}x+r_v l_v)}{(r_{w1}+r_{w2})x+2r_v l_v} - r_v l_v$	0	0	$\frac{c_{w1}x+c_{w2}x+c_v l_v}{2}$	$\frac{c_{w1}x+c_{w2}x+c_v l_v}{2}$	0

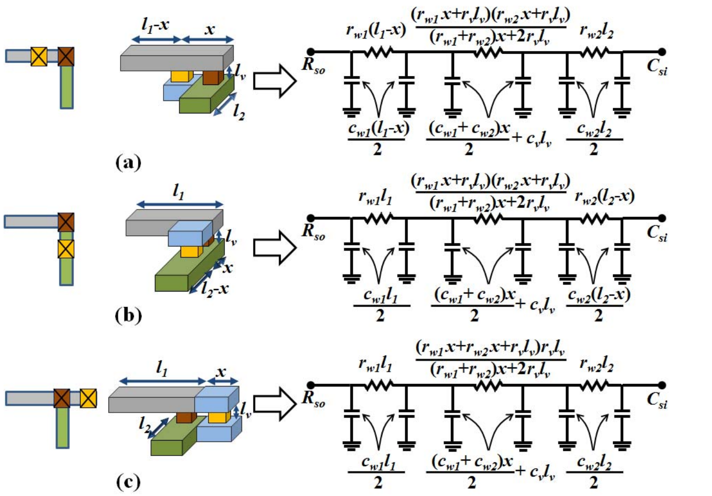

(a)
(b)
(c)
l_1-x
x
l_v
l_2
l_2-x
l_1
R_{so}
C_{si}
$r_{w1}(l_1-x)$
$r_{w1}l_1$
$r_{w2}l_2$
$r_{w2}(l_2-x)$
$\frac{(r_{w1}x+r_v l_v)(r_{w2}x+r_v l_v)}{(r_{w1}+r_{w2})x+2r_v l_v}$
$\frac{(r_{w1}x+r_{w2}x+r_v l_v)r_v l_v}{(r_{w1}+r_{w2})x+2r_v l_v}$
$\frac{c_{w1}(l_1-x)}{2}$
$\frac{c_{w1}l_1}{2}$
$\frac{(c_{w1}+c_{w2})x}{2}+c_v l_v$
$\frac{c_{w2}l_2}{2}$
$\frac{c_{w2}(l_2-x)}{2}$

Fig. 2. L-shape RV. (a) on-track type I. (b) on-track type II. (c) off-track type.

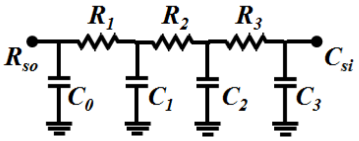


Fig. 3. The equivalent acyclic RC circuit of L-shape RV.

RVs precisely. The *simple T-shape* RVs contain three templates that two are on-track RVs, and one is off-track RV. Utilizing the similar circuit transformation techniques as *L-shape* RV templates, they can be transformed to corresponding equivalent RC tree structures. The *simple T-shape* RV templates and their equivalent RC trees are shown in Fig. 4. Fig. 5 represents equivalent acyclic RC circuits of *simple T-shape* RV templates. Although the equivalent circuit structure of *simple T-shape* RV is different with that of *L-shape* RV, they have the same delta RC values as shown in Fig. I because the extra wire segment in each *simple T-shape* RV has no impact on the equivalent circuit transformation.

The delta-wye transformation is needed for transforming the rest T-shape RV templates to be acyclic circuits with extra resistance and capacitance parameters. Their equivalent RC tree structures are more complicated than those of *simple T-shape* RV templates. We call them *complicated T-shape* RVs. The *complicated T-shape* RV templates and their equivalent RC trees are shown in Fig. 6. Fig. 7 is the equivalent acyclic RC circuit of *complicated T-shape* RV templates, and their related ΔR values and ΔC values are summarized in TABLE. II.

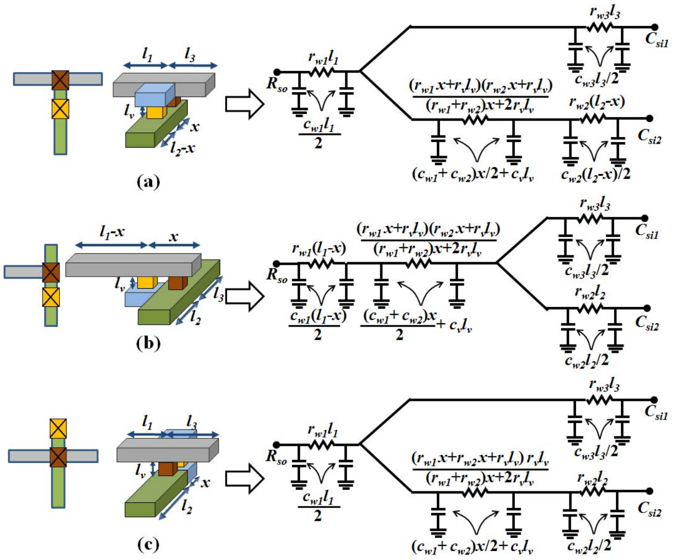

(a)
(b)
(c)
l_1
l_3
l_1-x
x
l_v
l_2
l_2-x
R_{so}
C_{si1}
C_{si2}
$r_{w1}l_1$
$r_{w1}(l_1-x)$
$r_{w3}l_3$
$r_{w2}l_2$
$r_{w2}(l_2-x)$
$\frac{(r_{w1}x+r_v l_v)(r_{w2}x+r_v l_v)}{(r_{w1}+r_{w2})x+2r_v l_v}$
$\frac{(r_{w1}x+r_{w2}x+r_v l_v)r_v l_v}{(r_{w1}+r_{w2})x+2r_v l_v}$
$\frac{c_{w1}l_1}{2}$
$\frac{c_{w1}(l_1-x)}{2}$
$c_{w3}l_3/2$
$(c_{w1}+c_{w2})x/2+c_v l_v$
$\frac{(c_{w1}+c_{w2})x}{2}+c_v l_v$
$c_{w2}(l_2-x)/2$
$c_{w2}l_2/2$

Fig. 4. Simple T-shape RV. (a) on-track type I. (b) on-track type II. (c) off-track type.

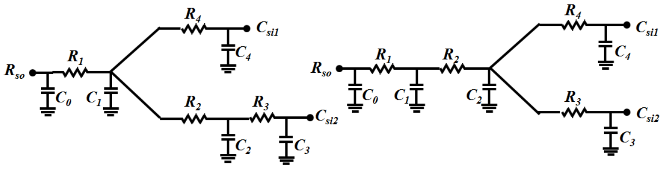


Fig. 5. Two equivalent acyclic RC circuits of simple T-shape RV.

C. Fast Sink Node Timing Check

To efficiently check whether the timing of each related sink node will violate its timing constraint or not if a related RV is inserted, the timing difference formula $\triangle t_n$ shown in (2) and the related ΔR values and ΔC values for *L-shape* RVs and *T-shape* RVs presented in TABLE. I and TABLE. II, respectively, are integrated to build a fast sink node timing check algorithm shown in Fig. 9. By executing the algorithm shown in Fig. 9, the timing of sink node as if adding an RV candidate can be fast checked.

Remark: The developed fast sink node timing check algorithm can be easily extended to perform the path delay check by using the static timing analysis method to calculate each gate's require arrival time (RT), arrival time (AT), and slack. By setting each gate's slack to be non-negative as a timing constraint during the RVI algorithm, one inserted redundant via might change the gate's RT or AT on the net that can be verified by the fast sink node timing check method. After that, the

TABLE II

ΔR VALUES AND ΔC VALUES OF THE EQUIVALENT ACYCLIC RC CIRCUIT OF COMPLICATED T-SHAPE RV.

Complicated *T-Shape* RV	ΔR					
	ΔR_1	ΔR_2	ΔR_3	ΔR_4	ΔR_5	ΔR_6
on-track type I	$-r_{w1}x$	$\frac{(r_{w2}x+r_vl_v)r_{w1}x}{(r_{w1}+r_{w2})x+2r_vl_v}$	$\frac{r_vl_vr_{w1}x}{(r_{w1}+r_{w2})x+2r_vl_v}$	$\frac{(r_{w2}x+r_vl_v)r_{w1}x}{(r_{w1}+r_{w2})x+2r_vl_v} - r_vl_v$	0	0
on-track type II	0	$\frac{r_vl_vr_{w3}x}{(r_{w2}+r_{w3})x+2r_vl_v}$	$\frac{(r_{w2}x+r_vl_v)r_{w3}x}{(r_{w2}+r_{w3})x+2r_vl_v}$	$\frac{(r_{w2}x+r_vl_v)r_{w1}x}{(r_{w2}+r_{w3})x+2r_vl_v} - r_vl_v$	$-r_{w3}x$	0
on-track type III	0	$\frac{(r_{w1}x+r_vl_v)r_vl_v}{(r_{w1}+r_{w2})x+2r_vl_v} - r_vl_v$	$\frac{r_vl_vr_{w2}x}{(r_{w1}+r_{w2})x+2r_vl_v}$	$\frac{(r_{w1}x+r_vl_v)r_{w2}x}{(r_{w1}+r_{w2})x+2r_vl_v}$	0	$-r_{w2}x$

Complicated *T-Shape* RV	ΔC						
	ΔC_0	ΔC_1	ΔC_2	ΔC_3	ΔC_4	ΔC_5	ΔC_6
on-track type I	$\frac{-c_{w1}x}{2}$	$\frac{c_{w2}x+c_vl_v}{2}$	$\frac{-c_vl_v}{2}$	$\frac{c_{w1}x+c_vl_v}{2}$	$\frac{c_{w2}x+c_vl_v}{2}$	0	0
on-track type II	0	$\frac{c_{w3}x+c_vl_v}{2}$	$\frac{-c_vl_v}{2}$	$\frac{c_{w2}x+c_vl_v}{2}$	$\frac{c_{w2}x+c_vl_v}{2}$	$\frac{-c_{w3}x}{2}$	0
on-track type III	0	$\frac{c_{w1}x+c_vl_v}{2}$	$\frac{-c_vl_v}{2}$	$\frac{c_{w2}x+c_vl_v}{2}$	$\frac{c_{w1}x+c_vl_v}{2}$	0	$\frac{-c_{w2}x}{2}$

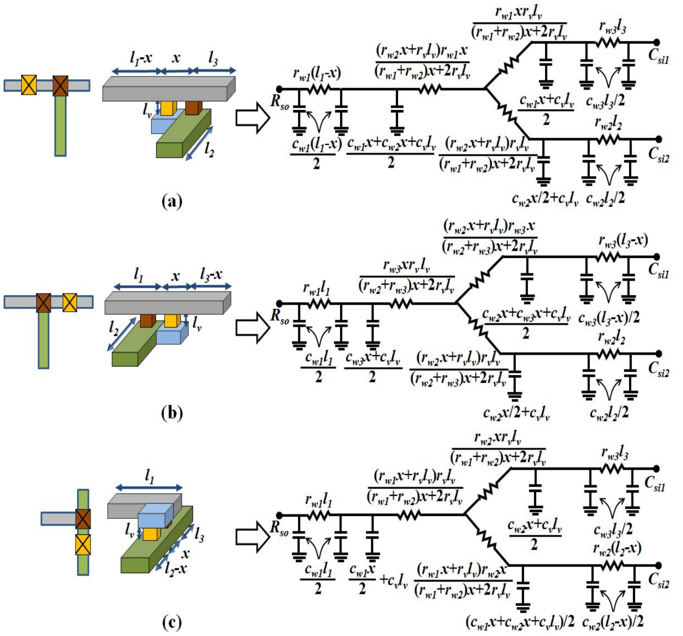

l_1-x
x
l_3
l_v
l_2
R_{so}
$r_{w1}(l_1$-$x)$
$\frac{(r_{w2}x+r_vl_v)r_{w1}x}{(r_{w1}+r_{w2})x+2r_vl_v}$
$\frac{r_{w1}xr_vl_v}{(r_{w1}+r_{w2})x+2r_vl_v}$
$r_{w3}l_3$
C_{si1}
$\frac{c_{w1}x+c_vl_v}{2}$
$c_{w3}l_3/2$
$r_{w2}l_2$
C_{si2}
$\frac{c_{w1}(l_1\text{-}x)}{2}$
$\frac{c_{w1}x+c_{w2}x+c_vl_v}{2}$
$\frac{(r_{w2}x+r_vl_v)r_vl_v}{(r_{w1}+r_{w2})x+2r_vl_v}$
$c_{w2}x/2+c_vl_v$
$c_{w2}l_2/2$
(a)
l_1
x
l_3-x
l_2
l_v
R_{so}
$r_{w1}l_1$
$\frac{r_{w3}xr_vl_v}{(r_{w2}+r_{w3})x+2r_vl_v}$
$\frac{(r_{w2}x+r_vl_v)r_{w3}x}{(r_{w2}+r_{w3})x+2r_vl_v}$
$r_{w3}(l_3$-$x)$
C_{si1}
$\frac{c_{w2}x+c_{w3}x+c_vl_v}{2}$
$c_{w3}(l_3$-$x)/2$
$r_{w2}l_2$
C_{si2}
$\frac{c_{w1}l_1}{2}$
$\frac{c_{w3}x+c_vl_v}{2}$
$\frac{(r_{w2}x+r_vl_v)r_vl_v}{(r_{w2}+r_{w3})x+2r_vl_v}$
$c_{w2}x/2+c_vl_v$
$c_{w2}l_2/2$
(b)
l_1
l_v
l_3
x
l_2-x
R_{so}
$r_{w1}l_1$
$\frac{(r_{w1}x+r_vl_v)r_vl_v}{(r_{w1}+r_{w2})x+2r_vl_v}$
$\frac{r_{w2}xr_vl_v}{(r_{w1}+r_{w2})x+2r_vl_v}$
$r_{w3}l_3$
C_{si1}
$\frac{c_{w2}x+c_vl_v}{2}$
$c_{w3}l_3/2$
$r_{w2}(l_2$-$x)$
C_{si2}
$\frac{c_{w1}l_1}{2}$
$\frac{c_{w1}x}{2}+c_vl_v$
$\frac{(r_{w1}x+r_vl_v)r_{w2}x}{(r_{w1}+r_{w2})x+2r_vl_v}$
$(c_{w1}x+c_{w2}x+c_vl_v)/2$
$c_{w2}(l_2$-$x)/2$
(c)

Fig. 6. Complicated T-shape RV. (a) on-track type I. (b) on-track type II. (c) on-track type III.

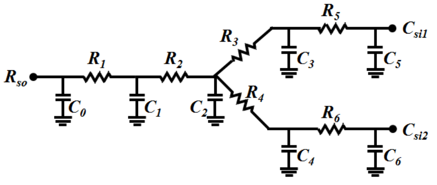
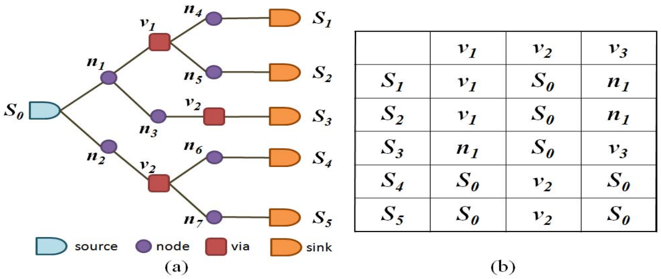


Fig. 7. The equivalent acyclic RC circuit of complicated T-shape RV.

delay effect is passed to the upstream and downstream path, and the circuit delay is calculated and checked. With this method, we can effectively update the circuit timing, keep the slack being non-negative, and has no any timing violation.

D. Runtime Complexity Analysis

The incremental timing analysis method can analyze the timing difference in $O(n_{sink})$ time if a relation topology table between vias and sinks can be constructed in advance. Here, n_{sink} is the number of sink nodes on the net. Once the developed algorithm decides a redundant via to be inserted, the timing influence can be easily analyzed by using the look-up table. For example, after a redundant via has been inserted to via v_1 shown in Fig. 8, the timing influence on each sink can be analyzed by finding the common path between the via v_1 and each sink node.



	v_1	v_2	v_3
S_1	v_1	S_0	n_1
S_2	v_1	S_0	n_1
S_3	n_1	S_0	v_3
S_4	S_0	v_2	S_0
S_5	S_0	v_2	S_0

Fig. 8. An example of look-up table. (a) A RC tree. (b) The related look-up table.

$$n_{cp}(v_1, S_1) = v_1 \tag{3}$$
$$n_{cp}(v_1, S_2) = v_1 \tag{4}$$
$$n_{cp}(v_1, S_3) = n_1 \tag{5}$$
$$n_{cp}(v_1, S_4) = S_0 \tag{6}$$
$$n_{cp}(v_1, S_5) = S_0 \tag{7}$$

Here, $n_{cp}(v, S)$ is the node that its upstream path is the common path of the via v and sink S. With the formulas (2) and (3)–(7), the timing information can be efficiently predicted and updated. The conventional timing analysis method reanalyzes the timing values of all nodes on the net and needs to recalculate the downstream capacitance of each node. The runtime complexity of the conventional method is $O(n)$, and n is the number of nodes o the net. Therefore, compared with the conventional method, the developed method is more efficiency.

IV. MBC GRAPH MATCHING PROBLEM

Lei et. al [8] proposed and defined the MBC graph matching problem with three essential definitions. The first definition constructs a via-candidate bipartite graph that includes the set of single vias on one side and the set of corresponding RV candidates on the other side. The second definition constructs a candidate relative graph that shows the relationship between RV candidates. The connection in graph means the conflict of two candidates because of design rules; in other words, they fight for one RV position. The third definition integrates two graphs

```
Algorithm: Sink Node Timing Check for Adding RV Candidate
Input:
  CRV: a redundant via candidate on net $i$
  $S_1, \cdots, S_{n_i}$: sink nodes of net $i$
  $t_{S_1}, \cdots, t_{S_{n_i}}$: original sink delays at $S_1, \cdots, S_{n_i}$, respectively
Output:
  Satisfaction or Violation

Find the delta RC values of equivalent acyclic RC circuit of
CRV by utilizing Fig. I or Fig. II
For sink nodes from $S_1$ to $S_{n_i}$
  Calculate $\triangle t_{S_j}$ by (2).
  If $t_{S_j} + \triangle t_{S_j} >$ the timing constraint of sink node $S_j$
    Return Violation
End
Return Satisfaction
```

Fig. 9. Sink Node Timing Check for Adding RV Candidate Algorithm

constructed by definition one and definition two to become a MBC graph.

Fig. 10 illustrates the above three stages for constructing a MBC graph. Given a postrouting design with RV candidates shown in Fig. 10.(a), the via-candidate bipartite graph is shown in Fig. 10.(b), and Fig. 10.(c) displays the conflicts between RV candidates. Combining the graphs shown in Fig. 10.(b) and Fig. 10.(c), the MBC graph is shown in Fig. 10.(d).

The goal of MBC graph matching problem is to find the maximum matching RVs in a given MBC graph.

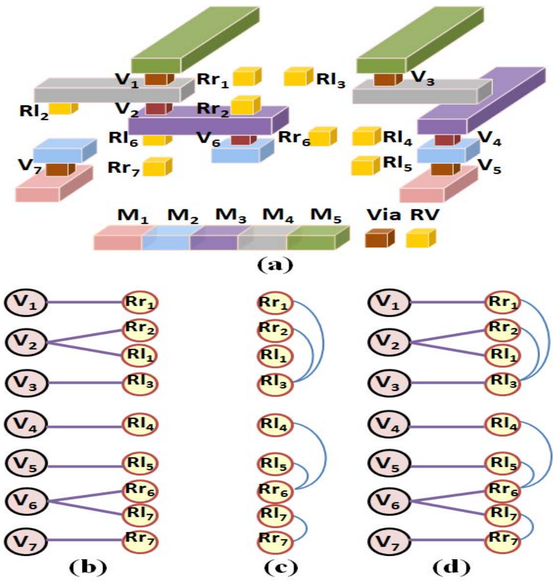


Fig. 10. An example of MBC graph. (a) A postrouting design with RV candidates. (b) Via-candidate bipartite graph. (c) Candidate relative graph. (d) MBC graph.

V. Timing Driven Redundant Via insertion

Based on the MBC graph, the relationship between the single via set and RV candidate set are well connected. Here, we are going to present the developed timing-driven minimum weighted matching (*t*-MWM) algorithm to solve the MBC graph matching problem and use the timing check algorithm shown in Fig. 9 to efficiently check the sink node timing as if an RV candidate is inserted on a specific net.

A. Edge Weight Assignment for the MBC Graph

Given a constructed MBC graph, an edge weight value $\omega(e)$ between a single via and its specific RV candidate needs to be assigned. This edge weight is determined by several properties between this single via and its relative RV candidates as follows.

$$\omega(e) = \rho \cdot (\alpha \times F.N. + \beta \times C.D. + \gamma \times C.T.) \qquad (8)$$

Here, α, β and γ are user-defined constants. The ρ is a timing violation indicator which is equal to infinity (i.e., this RV cannot be a candidate) if inserting the RV candidate will violate the timing-constraint, and it is equal to 1 if inserting the RV candidate won't violate the timing-constraint. The key factors, $F.N.$, $C.D.$ and $C.T.$ are

1) *Feasible number* ($F.N.$): It is the number of feasible RV candidates that a single via has, and the maximum number is 4.
2) *Conflict degree* ($C.D.$): It is the number of conflicts between one RV candidate and the rest RV candidates.
3) *Candidate type* ($C.T.$): For the manufacturing reason, we prefer to insert on-track RVs rather than off-track RVs. The value of $C.T.$ is equal to 0 if it is an on-track RV candidate; otherwise, the value of $C.T.$ is assigned to be 1.

B. Two Phase t-MWM Algorithm for a MBC Graph

The HMWM algorithm was used for solving the RV insertion problem in [8] but it is not suitable for the timing-driven redundant via insertion. To deal with timing constraints, we propose a two phase *t*-MWM algorithm that utilizes the properties of HMWM algorithm and considers timing issues. Fig. 11 shows the individual phase flowchart of the proposed two phase *t*-MWM algorithm. For each phase, similarly with HMWM algorithm [8], firstly, *t*-MWM algorithm assigns a suitable weight for each edge in the MBC graph and sorts the edge weights in the increasing order. Then, *t*-MWM algorithm picks up the RV candidate connected with the minimum edge weight, and the MBC graph is updated. Finally, above steps are repeated until there is no more RV candidate.

Compared with HMWM algorithm, the primary differences are that the edge weight assignment step considers the timing effect in *t*-MWM algorithm, and *t*-MWM algorithm re-analyzes the timing of modified circuit while performing the "*edge weight update*" step. Owing to inserting an RV will alter the net timing behavior, the timing effect on the same net should be recalculated. However, estimating the timing effects frequently is time-consuming and exhausted. To alleviate the computation load of updating timing effects, *t*-MWM algorithm utilizes the proposed timing check algorithm shown in Fig. 9 to check and update the sink-node timing for adding an RV candidate.

Phase 1 of *t*-MWM algorithm simplifies the problem and handles RV candidates that have no conflicts. In this phase, the edge weight is modified as (9) to accommodate this simple problem.

$$\omega(e) = \begin{cases} \max\left\{\frac{t_{S_j}+\triangle t_{S_j}}{t_{S_j}}\right\}, & \text{if } \triangle t_{S_j} < 0,\ \forall \text{sink node } j \\ & \quad \text{of the related net,} \\ \infty, & \text{otherwise.} \end{cases} \qquad (9)$$

Here, t_{S_j} is the original sink delay at sink node j of the related net, and $\triangle t_{S_j}$ is the timing difference at sink node j of the related net after inserting a related RV.

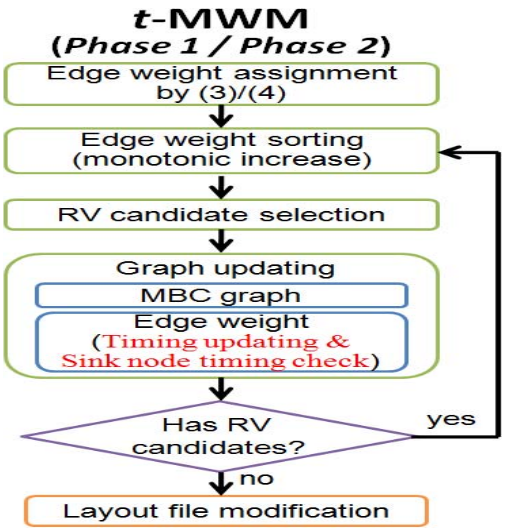

t-MWM
(Phase 1 / Phase 2)
Edge weight assignment by (3)/(4)
Edge weight sorting (monotonic increase)
RV candidate selection
Graph updating
MBC graph
Edge weight (Timing updating & Sink node timing check)
Has RV candidates?
yes
no
Layout file modification

Fig. 11. The individual phase flowchart of the two phase *t*-MWM algorithm.

TABLE III
PARASITIC PARAMETER

	Resistance	Capacitance
	Ω/□	$pF/\mu m^2$
Metal 1	0.101	1.12e-4
Metal 2	0.101	6.59e-5
Metal 3	0.101	5.97e-5
Metal 4	0.101	5.59e-5

Phase 1 is a pre-stage to increase the timing budget for other candidates in *Phase 2*. Without any conflicts of RV candidates, the candidates inserted in *Phase 1* will not decrease the RV candidate number in *Phase 2*. Furthermore, according to (9), each picked RV can increase the timing-budget.

Phase 2 deals with the rest RV candidates excluding those inserted in *Phase 1* and utilizes (8) for the edge weight assignment. The rest steps in *Phase 2* are the same as those of *Phase 1* shown in Fig. 11.

Remark: The concept of pre-increasing the timing budget in *Phase 1* has a significant influence on the insertion number of *Phase 2*. For example, the single via V_7 in Fig. 10 has only one RV candidate *Rr*$_7$ that conflicts to *Rl*$_6$. The better choice is to pick up *Rr*$_6$ and *Rr*$_7$ to match single vias V_6 and V_7 individually.

However, if we execute *Phase 2* directly instead of performing *Phase 1* firstly, the $\omega(e)$ related to *Rr*$_7$ will be infinity if the timing constraint is violated after *Rr*$_7$ is chosen to be inserted. In other words, *Rr*$_7$ cannot be a candidate under timing constraints unless other inserted RVs at the same net can provide extra enough timing budget for *Rr*$_7$.

As a result, if *Phase 1* can enhance the timing budget, it is helpful for improving the insertion rate in *Phase 2*.

VI. EXPERIMENTAL RESULTS

The developed RV insertion method under timing constraints has been implemented in C++ programming language and tested on an Intel Xeon 5160 quad core 3.0-GHz Linux based machine with 32GB memory. The test designs are the routed MCNC benchmarks that were built by the authors of [6] with $0.18\mu m$ technology. The RC parasitic parameters are listed in TABLE III, and the resistance of via is equal to 6.4Ω. The first two cases contain 4 metal layers and others have 3 metal layers. For the edge weight assignment constants, we choose $\alpha = 3$, $\beta = 1$ and $\gamma = 2$ that are the same as those in [8] for the trade-off between the RV insertion rate and the on-track RV insertion rate.

TABLE IV
NET SINK NODE TIMING DIFFERENCE RATIO AFTER RVI WITHOUT CONSIDERING TIMING CONSTRAINTS

Circuit	Timing Difference Ratio Interval (%)
Mcc1	$-43.87 \sim 7.40$
Mcc2	$-28.33 \sim 7.13$
Struct	$-44.51 \sim 0.01$
Primary1	$-48.51 \sim 9.81$
Primary2	$-48.45 \sim 9.86$
S5378	$-48.64 \sim 5.67$
S9234	$-48.61 \sim 3.51$
S13207	$-48.62 \sim 5.64$
S15850	$-48.62 \sim 4.04$
S38417	$-48.61 \sim 3.89$
S38584	$-48.62 \sim 5.63$

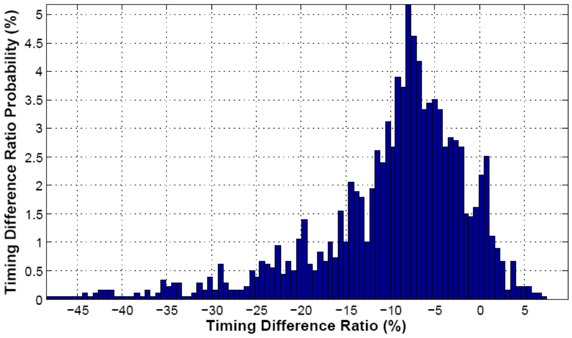


Fig. 12. Net sink node timing difference ratio probability graph of Primary1.

To investigate the timing impact of inserted redundant vias, we calculate the net sink node timing difference ratios after performing the conventional (without considering timing constraints) redundant via insertion method on each test circuit [8]. TABLE IV shows the difference ratio interval for each circuit. The 'Timing Difference Ratio Interval' is the timing disturbance ratio interval of net sink node delay differences after finishing the redundant via insertion procedure. It can be found that the range of timing difference ratio can be $35\% \sim 58\%$. Fig. 12 draws the sink node timing difference ratio probability graph for Primary1, and it reveals that the timing difference ratio interval can be very wide with non-ignorable probabilities. Therefore, it is meaningful to consider the timing effect while inserting redundant vias[3].

The results of redundant via insertion are presented in TABLE V. The "#Single Vias" is the total number of single vias in the given design, "#Alive Vias" is the total number of alive vias that a alive via is a single via having at least one RV candidate, "#Ins. RVs" is the number of RVs inserted by executing the insertion algorithms, and "#Ins. Rate" is the RV insertion rate. The "#Ins. RVs w/o vio." is equal to the total insertion number of RVs minus the number of inserted RVs causing timing constraint

[3] The high ratio of positive timing difference is because the nets are short in test circuit [8]; therefore, the insertion of redundant via may have high effect on timing delay.

TABLE V

COMPARISON OF RV INSERTION RESULTS WITH AND WITHOUT CONSIDERING TIMING CONSTRAINTS.

Circuit	Via Information		RVI without Considering Timing Constraints					RVI under Timing Constraints			
	#Single Vias	#Alive Vias	#Ins. RVs	Ins. Rate (%)	#Ins. RVs w/o viol.	Ins. Rate w/o viol. (%)	Runtime of HMWM [8] (s)	#Ins. RVs	Ins. Rate (%)	Runtime of *t*-MWM with Increment. (s)	Runtime of *t*-MWM with non-Increment. (s)
Mcc1	5948	5686	5661	99.56	5316	93.49	0.02	5350	94.09	1.26	25.33
Mcc2	34376	32228	32070	99.51	31799	98.67	0.34	31874	98.90	0.92	1.25
Struct	7598	7577	7577	100.00	6879	90.79	0.10	7088	93.55	0.42	1.87
Primary1	5536	5485	5483	99.96	4169	76.01	0.05	4616	84.16	0.26	0.97
Primary2	23154	22789	22787	99.99	20109	88.24	0.32	21294	93.44	2.46	13.54
S5378	6739	6532	6433	98.48	5869	89.85	0.05	6130	93.85	0.29	1.12
S9234	5365	5215	5212	99.94	4372	83.84	0.03	4606	88.32	0.24	1.17
S13207	13972	13583	13554	99.79	12574	92.57	0.10	12923	95.14	1.79	12.88
S15850	16922	16389	16349	99.76	14889	90.85	0.43	15386	93.88	1.66	16.25
S38417	40942	39840	39752	99.78	37197	93.37	0.49	37792	94.86	45.80	2002.50
S38584	55381	53700	53562	99.74	50793	94.59	0.98	51839	96.53	1807.13	494769.00
Comparison					**1.000**	**1.000**		**1.025**	**1.035**	**1.000**	**34.398**

violation, "#Ins. Rate w/o vio." is the RV insertion rate without violating timing constraints, and "Runtime of HMWM" is the runtime of HMWM algorithm [8]. The "Runtime of *t*-MWM with Increment." is the runtime of the proposed *t*-MWM algorithm with utilizing the developed incremental timing analysis mechanism to update the required timing information, and "Runtime of *t*-MWM with non-Increment." is the runtime of *t*-MWM algorithm with utilizing the conventionally non-incremental delay analysis method to update the timing information.

From TABLE V, the insertion rate can achieve to 99.68% in average without considering timing constraints. However, to meet the timing constraint, those inserted RVs located on the timing violation nets need to be removed. After removing those RVs, the insertion rate declines extremely to 90.20% in average. On the other hand, the insertion rate of the proposed RV insertion algorithm under timing constraints can achieve to 93.34% in average and guarantee the result satisfying timing constraints.

The last two columns in TABLE V also demonstrate that the developed incremental timing analysis technique can dramatically save the runtime of RV insertion procedure under timing constraints. Compared with the utilization of non-incremental timing analysis, it can speed up the runtime by over 34.398 times in average.

VII. CONCLUSION AND FUTURE WORK

In this paper, the RV insertion problem under timing constraints has been transformed to a MBC graph matching problem, and a developed two phase *t*-MWM algorithm utilizing the proposed incremental timing analysis method has been successfully used to solve the matching problem. The experimental results have showed that the insertion rate is improved, and how to fast and accurately estimate the timing behavior is crucial for the RV insertion under timing constraints.

Generally, some nets may not be the critical nets; hence, their timing delays after inserting redundant vias can be allowed to be worser than their original delays. To obtain the better and reasonable insertion rate, we are going to take the gate delay into consideration to fulfill the entire timing constraints in the future.

ACKNOWLEDGMENT

This work was supported in part by the National Science Council of Taiwan under Grant NSC 99-2220-E-009 -035.

REFERENCES

[1] Taiwan semiconductor manufacturing company (tsmc) reference flow 9.0.

[2] G. A. Allan. Targeted layout modifications for semiconductor yield/reliability enhancement. *IEEE Trans. Semiconduct. Manuf.*, pages 573–81, June 2004.

[3] Joe G. Xi. Improving yield in rtl-to-gdsii flows. *CMP Media LLC. EE Times: Design News.*, 2005.

[4] G. Xu, L.-D. Huang, D. Z. Pan, and M. D. F. Wong. Redundant-via enhanced maze routing for yield improvement. *in Proc. Conf. Asia South Pacific Des. Autom*, pages 1148–1151, 2005.

[5] H. Yao, Y. Cai, X. Hong, and Q. Zhou. Improved multilevel routing with redundant via placement for yield and reliability. *in Proc. Great Lakes Symp.*, pages 143–146, 2005.

[6] H. Y. Chen, M. F. Chiang, Y. W. Chang, L. Chen, and B. Han. Full-chip routing considering double-via insertion. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 27:844–57, May 2008.

[7] K. Y. Lee, C. K. Koh, T. C. Wang, and K. Y. Chao. Fast and optimal redundant via insertion. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, pages 2197–2208, December 2008.

[8] B. Y. Chiang, C. K. Lei, and Y. M. Lee. Redundant via insertion with wire spreading capability. *in Proc. Conf. Asia South Pacific Des. Autom.*, pages 468–473, 2009.

[9] F. Luo, Y. Jia, and W. M. Dai. Yield-preferred via insertion based on novel geotopological technology. *in Proc. Conf. Asia South Pacific Des. Autom.*, pages 730–735, 2006.

[10] J. T. Yan, B. Y. Chiang, and Z. W. Chen. Timing-constrained redundant via insertion for yield optimization. *in Northeast Workshop on Circuits and Syst.*, pages 1126–1129, 2007.

[11] C. W. Lin, M. C. Tsai, K. Y. Lee, T. C. Chen, T. C. Wang, and Y. W. Chang. Recent research and emerging challenges in physical design for manufacturability/reliability. *in Proc. Conf. Asia South Pacific Des. Autom.*, pages 238–243, 2007.

[12] W. C. Elmore. The transient response of damped linear networks. *Journal of Applied Physics*, 19:55–63, January 1948.

國科會補助計畫衍生研發成果推廣資料表

日期:2011/10/30

國科會補助計畫	計畫名稱: 子計畫一：三維度積體電路的隨機電熱模擬及其對功率最佳化的應用(2/2)	
	計畫主持人: 李育民	
	計畫編號: 99-2220-E-009-035-	學門領域: 晶片科技計畫--整合型學術研究計畫

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：李育民	**計畫編號：**99-2220-E-009-035-

計畫名稱：針對 3D 整合之電子設計自動化技術開發--子計畫一：三維度積體電路的隨機電熱模擬及其對功率最佳化的應用(2/2)

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	1	1	100%	篇	
		研究報告/技術報告	3	3	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	3	3	90%	人次	
		博士生	3	3	60%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	5	3	70%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	參加 2010 大學院校奈米元件電腦輔助模擬與設計軟體製作競賽榮獲奈米 CMOS 元件組佳作(頒獎單位:國家實驗研究院國家奈米元件實驗室)

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本計畫利用多重電壓供應技術以降低三維度積體電路功耗，在發展該技術時亦發展了一個利用混合網格來分析在考慮製程變異下晶片上具有隨機特性的溫度分佈。當元件參數統計特性的資訊以及設計電路的資訊被給定之後，我們首先利用混合網格來產生用來近似溫度的隨機多項式。此混合網格分析方法主要是利用一個較密的網格分析晶片上溫度分佈的期望值。由於要晶片溫度分布的二階以及三階統計特性可藉由執行許多次無隨機變異的溫度分析來近似之。因此，為了要達到效能上的提升，我們利用了較為稀疏的網格來執行分析二階以及三階統計特性所需的多次無隨機變異的溫度分析。在得到晶片溫度的期望值、標準差以及歪斜係數之後，我們將在晶片上任意點的隨機溫度近似成歪斜常態(skew-normal)的隨機變數。然後再利用此歪斜常態隨機變數的尾端機率求得晶片上的熱良率分布輪廓(thermal yield profile)。 利用此分析流程，我們能在可接受的準確率之下，使得具有製程變異之晶片溫度分析的效率達到與不考慮製程變異之溫度分析之效率相當的程度。此熱良率分析技術，可有效地提供晶片之相關熱資訊予任何熱導向之積體電路設計流程。