Rethinking Thermal Via Planning with Timing-Power-Temperature Dependence for 3D ICs*

Kan Wang[1],Yuchun Ma[1], Sheqin Dong[1], Yu Wang[2], Xianlong Hong[1], Jason Cong[3]
[1]Dept. of Computer Science and Technology, Tsinghua University, TNList
[2]Dept. of Electronic Engineering, Tsinghua University, Beijing, China, 100084
[3]Dept. of Computer Science, UCLA, Los Angeles, CA, 90095

Abstract -Due to the increased power density and lower thermal conductivity, 3D is faced with heat dissipation and temperature problem seriously. Previous researches show that leakage power and delay are both relevant to temperature. The timing-power-temperature dependence will potentially negate the performance improvement of 3D designs. TSV (Through-Silicon-Vias) has been shown as an effective way to help heat removal, but they create routing congestions. Therefore, how to reach the trade-off between temperature, via number and delay is required to be solved. Different from previous works on TSV planning which ignored the effects of leakage power, in this paper, we integrate temperature-leakage-timing dependence into thermal via planning of 3D ICs. A weighted via insertion approach, considering both performance and heat dissipation with resource constraint, is proposed to achieve the best balance among delay, via number and temperature. Experiment results show that, with leakage power and resource constraint considered the temperature and via number required can be quite different, and weighted TSV insertion approach can improve thermal via number, by about 5.6%.

I. INTRODUCTION

In CMOS circuits, power dissipation consists of dynamic and static components. Leakage power is the main composition of static power. Currently, with continuous shrinking of minimal feature size, leakage power has become more and more and is definitely non-negligible [2]. Furthermore, leakage power and temperature are interacting with each other. To get an accurate value of the on-chip temperature caused by both dynamic and leakage power, an iterative computation process is needed which usually requires a few iterations. Previous researches showed that the temperature also influences the chip performance and increase the delay on each module. [3] investigated the compromises among timing, power and temperature using iterative optimization and it showed that temperature, performance and power tend to converge with about about 4 iterations.

In 3D ICs, high temperature will greatly reduce the performance of circuits. It is urgent to optimize thermal in 3D design. During the past few years, several works on thermal optimization have been proposed including thermal-driven floorplanning [2,7,10,13] and routing [5,6]. Unfortunately, even with complicated thermal-aware approach to improve heat dissipation, the maximum on-chip temperature is still too high for the circuit to operate properly [13].

On the other hand, many researches about thermal control and management are proposed, which provide efficient methods to solve thermal problem in 3D ICs. Introducing thermal vias into the circuit, as shown in Fig.1, in fact, is one of the efficient ways to reduce the chip temperature to a satisfactory level [13]. However, thermal vias bring in additional cost. First, it is expensive to manufacture thermal vias and the common thermal via pitch is very large compared to regular metal wires. Besides, thermal vias take the dead-space between blocks, which will create routing congestion and also lead to longer interconnects and larger delays. As a result, via number must be minimized to meet the target temperature. [13] formulated and solved the thermal via problem as a post-floorplan procedure to improve heat resource distribution while the works in [9] incorporated thermal via planning into the thermal-driven floorplanning and placement. Many works paid attention to the TSV planning and proposed methods to optimize the results. The works in [15,16] redistributed dead-space between blocks to favor thermal via insertion, considering the spatially variant power. [17] proposed the method to allocate thermal vias using Lagrangian relaxation instead of steady-state thermal analysis in 3D IC which can avoid the over-estimation caused by steady-state analysis.

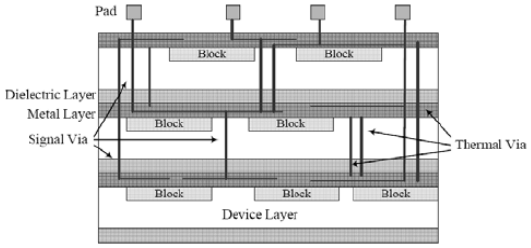


Figure 1. Thermal via in 3D IC stack

However, most of the previous thermal via planning approaches have not taken leakage power into account. But without considering the impact of leakage power on temperature, via number needed in the design will be under-estimated. Fig. 2 shows the interaction between temperature, power and performance. Higher temperature leads to larger delay and leakage power, and potentially degrades the performance. At the same time, increased leakage power will in turn raise the temperature which requires more thermal vias. On the other hand, thermal vias inserted will reduce the on-chip temperature, which will also influence the leakage power dissipation. These design factors affect with each other and it is necessary to consider the leakage-delay-temperature dependence. Without considering of these factors together, the thermal vias planning might be over-estimated or under-estimated.

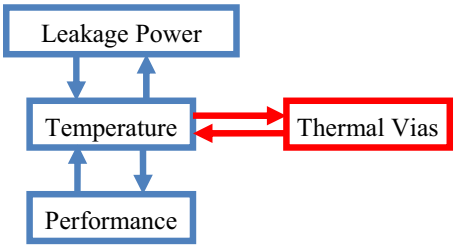


Figure 2. Relationship between leakage, via number, temperature and performance

Furthermore, most of the previous TSV insertion made use of the dead-space as much as possible and didn't consider the resource constraint in dead-space. Actually, the space between blocks is not only for thermal vias, but also for certain wires, signal vias, buffers and so on. [3] proposed that the via space is only 26% of the total black space area. In this case, the space left around the hotspots may not be enough for thermal vias, which may limit the effects of thermal vias to remove the hotspots.

What' more, as a result of effect of leakage power on temperature and delay, thermal vias should be inserted not only around the hot spots to minimize the maximal on-chip temperature, but also on the critical path to satisfy performance demand (e.g. delay). In this paper, the impact of these factors will be taken into TSV planning in the form of weights. We refer to it as weighted via insertion.

*This work is supported by NSFC 60606007, 60870001 and 61076035, 863 project (No. 2009AA01Z130) and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation

978-1-4244-7516-2/11/$26.00 ©2011 IEEE

In this paper, we propose an exploration approach to implement the TSV planning considering the power-temperature-timing dependence with some design constraints such as via density, timing constraints, and temperature threshold. In general, the contributions of this paper include:

1. **Iterative TSV planning considering leakage power-delay-temperature dependence**: The relationship between leakage power and thermal via insertion is a chicken-egg problem. They are interacting with each other. Besides, delay is also relevant to temperature. In this paper, we propose an iterative TSV planning process to obtain the converged results with leakage-power-delay-temperature dependence considered. Experiment results show that, with leakage power considered, the temperature increases by 35%, total via number increases by 19%, and delay is enlarged by 14.2%.
2. **Performance aware TSV planning with resource constraints**: The temperature will increase with leakage power considered, which leads to more thermal vias needed to reach the target temperature. But the limited space for vias may degrade the heat removal effects. In this paper, we analyze the impact of resource constraint to TSV planning and take it into account while TSV planning.
3. **Weighted via planning with power-temperature-timing evaluation**: In this paper, a weighted via insertion approach considering both performance and heat dissipation and the balanced evaluation method is proposed to obtain the trade-off between different design factors. The weighted approach can improve thermal via number by about 5.6%.

The rest of the paper is organized as follows. In Section II, thermal model and thermal via insertion technology are introduced. In Section III, we describe the problems in this paper. In Section IV, we investigate leakage-temperature-delay dependence, with which we analyze the impact between thermal via planning and leakage power In Section V, temperature and time aware TSV insertion method is proposed In Section VI, and experiment results are shown in Section VII. The conclusions are provided in Section VIII.

II. THERMAL MODEL AND THERMAL VIA INSERTION

In this section, we describe the thermal model and introduce thermal via insertion approach used in this paper.

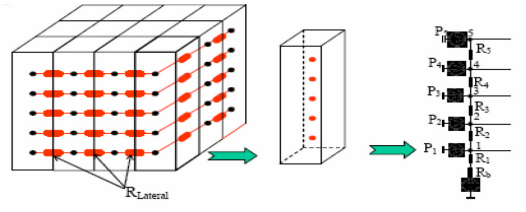
2.1 Thermal Model

The 3D circuit stacking is divided by a two-dimensional array of tile stacks, as shown in Fig. 3(a). Each tile stack is composed of several vertically-stacked tiles, as shown in Fig. 3(b). These tile stacks are connected by lateral thermal resistances, $R_{lateral}$. Within each tile stack, a thermal resistor R_i is modeled for the i-th device layer, while thermal resistance of the bottom layer and silicon substrate is modeled as R_b as shown in Fig.3(c).

Similar to [13], a tile stack is modeled as a resistive network. The isothermal bases of room temperature are modeled as a voltage source. A current source is present at every node in the network to represent the heat sources. The tile stacks are connected by lateral resistances. The system can be spatially discretized and be solved using the following equation to determine the steady-state thermal profile as a function of power profile:

$$T = PA^{-1} \tag{1}$$

where A is an $N \times N$ sparse thermal conductivity matrix. T and $P(T)$ are $N \times 1$ temperature and power vectors.



(a) *Tiles Stack Array* (b) *Single Tile Stack* (c)*Tile Stack Analysis*
Figure 3. Resistive thermal model for a 3D IC [13]

2.2 Thermal via insertion

As mentioned in Section I, the use of TSVs for inter-layer communication can efficiently remove heat and reduce the chip temperature to a satisfactory level. Cong, et al [13] formulated the thermal via planning as a nonlinear programming (NLP) problem and solved it by solving a sequence of simplified via planning sub-problems in alternating directions. In this paper, to further control the thermal effects with leakage power, we use the same via planning model to estimate the requirement of via number in each tile and fulfill thermal via insertion.

III. PROBLEM DEFINITION

Given a packing with certain number of blocks $\{b_1, b_2, \ldots b_n\}$ on L layers, our approach try to insert the thermal vias properly with leakage power and performance considered. The object is to minimize the total via number with certain constraints.

In this paper, we take thermal resource into consideration and constrain the packing ratio of thermal vias in each grid as *via_ratio*. *via_ratio* =0.5 means that available total area for vias in a grid is less than half of dead-space area in the grid.

The notions of the thermal via planning problem include:
$VN_{i,j,k}$: number of vias inserted in $\text{grid}_{i,j,k}$
A_{via} : area of one thermal via;
$Area^{ds}_{i,j,k}$: area of dead-space in $\text{grid}_{i,j,k}$;
via_ratio: packing ratio of thermal vias comparing to dead-space
$T_{i,j,k}$: temperature on $\text{grid}_{i,j,k}$ with leakage power considered;
required_T: target maximal on-chip temperature;
$Delay_{path}$: delay on a path;
Delay_threshold: the delay constraint set for the design.
With the above notions, we can define the problem as:

Objective:
$$Min \sum_{g_{i,j,k}} VN_{i,j,k} \tag{2}$$

a. Temperature constraints: $Max\{T_{i,j,k}\} \le required\_T$ (3)

b. Delay constraints: $Max\{Delay_{path}\} \le Delay\_threshold$ (4)

c. Dead-space resource constraints:
$$A_{via} \times VN_{i,j,k} \le Area^{ds}_{i,j,k} \times via\_ratio \tag{5}$$

Thermal via planning is to minimize (2) subject to equation (3) (4) and (5). With leakage-performance-temperature dependence considered, not only the number of thermal vias will be influenced, but also the optimization approach needs to be extended.

IV. LEAKAGE-TEMPERATURE-DELAY DEPENDENT MODEL

In this section, we introduce the two models which describe the Leakage-Temperature-Timing dependence.

4.1 Leakage-temperature dependent model

Leakage current is a super-linear function of its supply voltage, threshold voltage and temperature. To express the relationship more concisely, [3] modeled the leakage current with polynomial function. A third-order polynomial can describe the dependencies very well, with a maximum error of 5%. The model is of the form:

$$\frac{I_{leakage}(T)}{I_{leakage}(T_0)} = 1 + \alpha_1 \cdot (T - T_0) + \alpha_2 \cdot (T - T_0)^2 + \alpha_3 \cdot (T - T_0)^3 \tag{6}$$

where $I_{leakage}(T)$ is the leakage power under the current temperature T. α_1, α_2 and α_3 are empirical coefficient that have different values for different technologies. Typically, α_1=0.0226, α_2=0.00033, α_3=1.77e-6, and in this paper, we define $I_{leakage}(T_0)$=0.01 when T_0=0°C.

The thermal profile can be obtained by iteratively conducting thermal analysis and leakage power estimation until convergence, as shown in Fig.4. This usually requires only a few iterations.

4.2 Delay-temperature dependent model

As mentioned in [3], the delay-temperature dependence can be expressed as (7).

$$delay(T) = \frac{delay(T_0)(V_{DD} - V_{TH}(T_0))^{\alpha} T^{\beta}}{T_0^{\beta}(V_{DD} - V_{TH}(T_0) + k(T - T_0))^{\alpha}} \tag{7}$$

Where $k=k_0+\gamma(T-T_0)$. T is the absolute temperature in Kelvin, α is the velocity saturation index, and μ is the mobility.

Here we use the values of α, β, γ, k_0 in [3]. The maximum error of these values is just 6.1% and corresponding temperature is just 0°C.

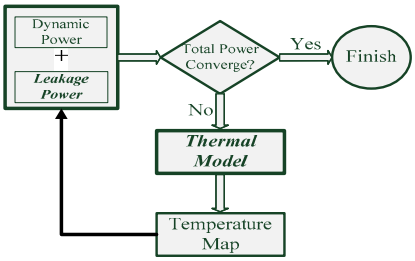


Figure 4. Loop for accurate leakage power calculation

In this paper, we calculate the delay on paths between the blocks and the performance of the chip is evaluated by the delay on the longest critical path. Therefore, to meet the timing constraints in (4), the delay on the longest delay should not exceed the maximum delay allowed in circuits.

V. Impact between Thermal Via Planning and Leakage

As mentioned in Section I, temperature, power and performance are interacting with each other and there should be a balance between them. Without considering the leakage power contribution to the temperature, via number needed in the design will be under-estimated. On the other hand, thermal via insertion will reduce the on-chip temperature which in turn influences the leakage power dissipation. We take two packings of benchmarks Ami33 and N100 for example and analyze the effects on thermal via planning by comparing the cases with or without leakage power considered. The thermal via insertion approach in [13] is used to fulfill the thermal via planning. The required temperature with thermal via inserted is set to 77°C which is 350K.

For a better comparison, we think of three schemes: neglecting leakage power (*NLP*), considering leakage power (*CLP*) and considering leakage power with via insertion (*VLP*). Three different TS-via planning schemes are as following: 1) *CLP* computes the leakage power based on initial packing, updates the temperature, and then plan the thermal via insertion with the updated temperature. 2) *NLP* performs the thermal via insertion without leakage power. 3) *VLP* initially inserts the thermal via and computes the leakage power based on the reduced temperature, and then implements the final thermal via insertion.

5.1 Impact of leakage power on via number

In order to get better comparison between *CLP* and *VLP* on impact of leakage, we show the ratio of leakage power to dynamic power after via insertion, divided into maximum value (*Max L/D*) and average value (*Avg L/D*), as shown in Table I.

Table I Impact of leakage power and vias on temperature and via number with target temperature 77°C

Test cases		*W/O leakage (NLP)*	*With Leakage (CLP)*	*Leakage With Via (VLP)*
Ami33	***Initial T***(°C)	273.47	435.48	345.32
	T with TSV(°C)	76.82	76.83	76.99
	Via number	1143	6033	1328
	Max L/D	0	0.0658	0.0660
	Avg L/D	0	0.0444	0.0449
	Runtime(s)	3.25	3.97	6.33
N100	***Initial T***(°C)	191.36	261.11	228.34
	T with TSV(°C)	77.11	76.86	76.80
	Via number	12738	18311	14641
	Max L/D	0	0.0659	0.0658
	Avg L/D	0	0.0421	0.0421
	Runtime(s)	7.76	8.05	11.56

Results in Table I show that, considering leakage power, the temperature of the chip is generally higher than that without leakage considered, by nearly 60% for ami33 and 36% for n100. With the increased temperature, more thermal vias are needed to reduce on-chip temperature to the required temprature. In ami33, *NLP* need 1143 vias, while *CLP* needs 6033 vias to meet the temperature threshold with leakage power considered. In n100, *NLP* need 12738 vias, while *CLP* needs 18311 vias to meet the temperature threshold with leakage power considered.

By rethinking the situation, we find that both *NLP* and *CLP* are far away from the final chip results. In *NLP*, no leakage involved will result in under-estimation of the on-chip temperature along with thermal vias. But in *CLP*, the working on-chip temperature could not be that high so that the over-estimation of leakage power will lead to redundant thermal vias. After the thermal vias are inserted, *VLP* seems to provide practical results since the leakage power is computed with the final working temperature. But still 1328 thermal vias are needed in ami33 and 14641 in n100. Therefore, the leakage effect greatly affects the thermal via planning.

5.2. Impact of leakage power on via distribution

Power is distributed differently on each layer of the chip and power density is not the same everywhere, which leads to the non-uniform thermal distribution. Therefore, there are some regions which have higher temperature than others. We refer to these places as hot spots. In order to achieve better cooling effect, vias are inserted around hot spots as much as possible. Thus, via density meets certain distribution, and is consistent with the power density to some extent.

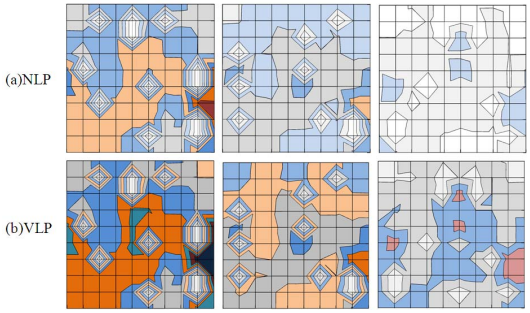


Figure 5. Thermal via distribution for n100

Figure 5 shows the thermal via distribution for n100 after via insertion without consideration of leakage power (a) and with consideration of leakage power (b). It shows that with leakage power consideration, even the maximal on chip temperature has be controlled to 77oC, the via demand increase greatly on each layer, averagely 14% increase.

VI. Temperature and Performance Aware TSV Insertion

6.1 Iterative Thermal Via Planning with Timing, Power and Temperature

As mentioned before, leakage aware thermal via planning may need an iterative process. If the space for thermal vias is enough, the temperature can directly reach the target temperature, so we could calculate the leakage power from the final (target) temperature. In most of the cases, the space for thermal vias is not enough to reach the target temperature; therefore a few iterations are needed to obtain the convergence, as shown in Fig.6. In this case, there are two methods to deal with:

1) Start with the initial temperature and calculate leakage according to temperature. Then reach a stable value after a few iterations.

2) Start with the target temperature and calculate leakage according to target temperature. And then also converge after iterations.

During the first iteration process, leakage and delay are calculated according to the initial temperature before TSV insertion. After TSV insertion, the temperature decreases, with values of leakage and delay also decreasing. As a result, the actual temperature is lower than that we expect to arrive. On this occasion, we over-estimate the leakage and the final temperature which will lead to more vias than actually needed.

During the second iteration, the actual temperature after thermal vias insertion is indeed higher than the expected one, which makes

us under-estimate the number of vias required.

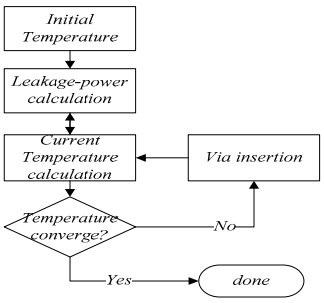


Figure 6. Flow of the iteration of leakage and via insertion

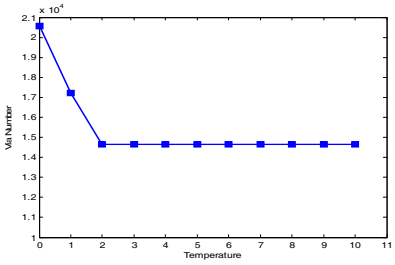


Figure 7. Iteration for n100 with *required_T*=350K

Either of the two iterations may lead to slow converge, which contributes to the iteration times of calculation. We use the first iteration process and the experiments show that the temperature, delay, and power tend to converge with less than 1% error in about 2 iterations as shown in Fig.7.

6.2 Weighted thermal via insertion with performance optimization

As mentioned before, temperature distribution impacts the performance of the design. In previous thermal via planning, though the maximal temperature can be reduced to the threshold by inserted vias, the temperature is still not even so that the delay on critical path might be increased due to the relatively high temperature. As Fig.8 shows, block 1 is the hottest block, around which some thermal vias might be allocated. But the temperature also influences the delay on critical path. In Fig.8, Block 3 might be the hottest block along the critical path, and we may want to make block 3 cooler so that the overall delay can be optimized. Therefore, considering both heat dissipation and the delay constraint, both block 1 and block 3 needs some thermal vias to be inserted around on the same layer or even on different layers.

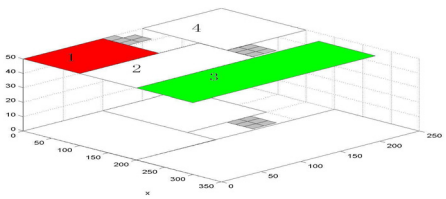


Figure 8. Thermal vias around hotspots and blocks on critical path

We propose a weighted via insertion approach considering both performance and heat dissipation. Normally, via number can be ideally budgeted according to the heat dissipation. The via number in each grid can be proportionally computed according to the heat flow between grids.

Thermal via insertion are implemented in the following two steps:

(1) ***Vertical thermal via planning***. First the vertical thermal via planning distributes the vias to different layers by assuming the temperature distribution inside each layer is uniform. Suppose that the total number of vias on layer m is VN_m ($2 \le m \le L$), where L is the maximal layer number, the total power density on layer m is P_m. Then via number between layers follows:

$$VN_L : VN_{L-1} \ldots\ldots : VN_2 = Q_L : Q_{L-1} \ldots\ldots : Q_2 \qquad (8)$$

where $Q_n = \sum_{i=n}^{L} P_i \quad (n = 2,3,.....L)$.

(2) ***Horizontal thermal via planning.*** Normally, horizontal thermal via planning will assign thermal vias within one device layer to different grids according to heat propogation. Beside the heat dissipation, we also consider the heat removal demands caused by the leakage power and delay optimization.

We set two kinds of weights for blocks: $W_{Leakage}$ to evaluate leakage criticality and W_{delay} to evaluate delay criticality. According to leakage power model, the leakage power density depends on the temperature and dynamic power density. We define the weight for leakage power of block b_i as:

$$W_{leakage}(b_i) = \alpha \times \frac{T_{bi}}{T_{\max}} \times \frac{P_{bi}}{P_{\max}} \qquad (9)$$

where T_{max} is the maximal on-chip-temperature before via insertion. T_{bi} is the temperature on block b_i. P_{bi} is the dynamic power density of block b_i, and P_{max} is the maximal power density for all blocks. α is constant factor which is defined by users. Here, we set α=0.2. According to delay model, delay on block depends on the temperature and initial delay. Since the performance is decided by the longest path on the chip. Therefore, we only assign the delay weight for the blocks on critical path.

$$W_{delay}(b_i) = \beta \times \frac{T_{bi}}{T_{\max}} \times \frac{delay(T_0)}{Delay\_threshold} \qquad (10)$$

where $delay(T_0)$ is the delay of block b_i at the temperature of T_0, while *Delay_threshold* is the target maximal delay on whole chip. β is also constant factor and we set β=0.2. For the blocks on non-critical path, the delay weight is set as 0. Since both W_{delay} and $W_{leakage}$ are between (0, 1), we accordingly distribute the weights on blocks to the nearby grids which have available dead-space and then via insertion will be guided by the weight values along with the heat dissipation. Here we combine the weights with the heat flow for each grid as:

$$H_{i,j,k} = (1 + W_{delay} + W_{leakage}) \cdot I_{i,j,k} \qquad (11)$$

Where the heat flow $I_{i,j,k}$ for each grid is updated through path counting approach according to [13]. Then the via number initially assigned to each tile is proportional to $H_{i,j,k}$ instead of $I_{i,j,k}$. Ideally without the dead-space resources, for two tiles on a layer k, the number of thermal vias allocated follows:

$$VN'_{i1,j1,k} : VN'_{i2,j2,k} = H_{i1,j1,k} : H_{i2,j2,k} \qquad (12)$$

Considering the resource constraint, the via number in a grid is:

$$VN_{i,j,k} = \min\left(\frac{Area^{ds}_{i,j,k}}{A_{via}}, VN'_{i,j,k}\right) \qquad (13)$$

Therefore, by iteratively updating the number of vias in each grid to reach the target temperature or meet some stop criteria, the vias can be inserted between blocks.

6.3 Evaluation Process

The target temperature on chip really influences the chip performance as well as via budget, and the balance between thermal vias, delay and temperature needs a specific searching strategy. Instead of setting a fixed target temperature, we give a feasible range for required temperature with a lower bound as T_{low} and an upper bound as T_{up}. We first treat the *required_T* as our target temperature and insert vias until below it. The we range the temperature around this value between T_{low} and T_{up} to search all the solution space to get a temperature that best balance the dependence between via number, delay and temperature.

By changing value *required_T*, the total via number would change. When temperature changes, delay of the circuit will also differ. In this paper, we provide the function to value the results in terms of delay, temperature and via number. They are defined as follows:

A. Balance Function:

$$Bal(T,d,n) = \alpha \cdot (T - T_{low}) + \beta \cdot delay + \gamma \cdot via\_num \qquad (14)$$

Where $\alpha, \beta, \gamma > 0$ and $\alpha + \beta + \gamma = 1$. This function is used to balance the weight of temperature, delay and via number.

B. Punish Function:

$$M(T,d,n)=\begin{cases}M & T>T_{up}\\ 0 & otherwise\end{cases} \quad (15)$$

where M is value large enough as the a penalty item. Once current temperature is higher than T_{up}, we will add a large item to the cost function as a penalty.

C. Cost Function:

$$Cost(T,d,n)=Bal(T,d,n)+M(T,d,n) \quad (16)$$

This function helps to evaluate the result of Performance aware TSV insertion. In this paper, we use this cost function to evaluate the results by arrange the weight of α, β, γ. We consider a result to be a good one once it has a low cost. The overall optimization flow is shown in Fig. 9.

```
Input: T0, Required_T, Tup, Tlow
Output: Via number, Max T, Delay
Method:
CurT=LeakageModel (T0); //Calculate the leakage and update Current temperature
While(CurT> Required_T){
    ViaInsertion(CurT, Required_T);//Insert vias to reduce curT to below Required_T
}
Calculate via_number0, delay0 and maxT0;
MinCost=CostCalculation(via_number0, delay0, maxT0);
For (Temperature between Tup and Tlow){//Search all the temperatures between Tup and Tlow
  LeakageModel (T0); //Calculate the leakage and update Current temperature
  While(CurT> Required_T){
    ViaInsertion(CurT, Required_T);
   Calculate via_numberi, delayi and maxTi;
   Costi=CostCalculation(via_numberi, delayi, maxTi);
   If(Costi<MinCost){
     MinCost=Costi; n=i;
   }
}
Output via_numbern, delayn and maxTn;
```

Figure 9. Algorithm of temperature and time aware TSV insertion

VII. EXPERIMENTAL RESULTS

In this section, we show the dependence of leakage-temperature and delay-temperature and also the results of temperature and performance aware TSV insertion which is greatly influenced by required temperature and values of *via_ratio*.

All experiments were performed on a workstation with 3.0 GHz CPU and 4GB physical memory. We use five typical MCNC and GSRC benchmarks [7] in our experiments. The number in each benchmark's name indicates the number of blocks, i.e., these numbers correspond to problem instance size. A four-device layer configuration is assumed for all circuits. Each device layer is silicon based, and there are two metal routing layers on top of each device layer. A thermal TS-via will extend to the metal layers above and below its device layer. The parameters of thermal vias between layers are set as [13]. The floorplan layout is generated by a 3-D thermal-driven floorplanning tool [7].

7.1 Impacts between leakage-thermal vias

As mentioned before, leakage power greatly affects thermal vias insertion and increase the number of thermal vias to great extent. What's more, with leakage power considered, available dead-sapce resource may be not enough for required vias to meet the temperature threshold ***Required_T***. In this section, we provide five benchmark examples of experimental results with ***Required_T*** as 77°C (350K), through three times of iterations. Here we set *via_ratio*=1, which means the whole dead space can be used for vias.

Table II Impact of leakage power on thermal via number

Circuit	W/O leakage (NLP)			With Leakage (VLP)		
	Initial T	Max T with via	Via Number	Max T(°C)	Max T with via	Via Number
Ami33	152.9	76.83	483	184.2	77.09	548
Ami49	195.2	76.87	48065	245.3	88.74	57179
N100	191.4	77.11	12738	261.1	76.80	14641
N200	204.2	76.94	16243	267.7	76.86	19237
N300	258.7	76.89	26602	396.9	77.03	32130
Ratio	1	1	1	1.35	1.03	1.19

As shown in Table II, we use two TS-via planning schemes: one with leakage considered (*VLP*) and the other not (*NLP*). Compared to NLP, the temperature with leakage power is evidently 35% higher and thus *VLP* require 19% more vias to reach ***Required T*** than that of neglecting leakage power (*NLP*).

With leakage power considered, 77°C is an unreachable threshold for ami49. Even all of the dead-space is used for thermal via insertion; the final on-chip temperature is 88.74°C which is above ***Required_T***.

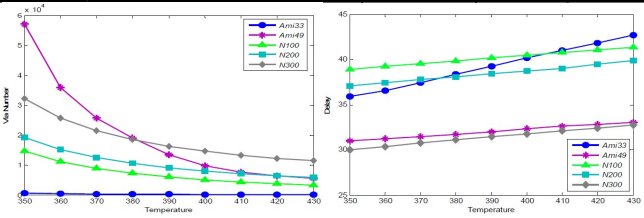
7.2 Impacts of Temprature-delay-power dependence

In this section, we analyze the delay-temperature dependence and the impact of temperature to delay, via number and power.

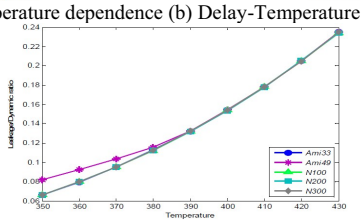
From Table III we can see that delay can be very different under different ***Requried_T***. Take ami33 for example, the normalized delay is about 35.9, with ***Required T*** of 350K. Delay with 410K is about 14.2% larger that of 350K. Ami33, N100, N200 and N300 also verify the dependence, increase about 7.54% on average.

Table III Impacts of Temperature-delay-power dependence

Test cases	Required T	Max T	Normalized Delay	Via Number	Leakage/ Dynamic	Run Time (s)
Ami33	350	350.09	35.90	548	0.0662	7.06
	380	380.14	38.34	243	0.113	7.97
	410	410.00	41.00	126	0.178	8.39
Ami49	350	361.74	31.02	57179	0.0821	4.43
	380	381.60	31.72	18967	0.115	4.70
	410	410.09	32.61	7553	0.178	5.66
N100	350	349.80	38.87	14641	0.0658	12.08
	380	379.85	39.85	7309	0.112	11.4
	410	410.02	40.76	4329	0.178	15.37
N200	350	349.86	37.10	19237	0.0659	16.61
	380	379.95	38.09	10545	0.112	17.72
	410	409.82	39.53	7174	0.178	17.20
N300	350	350.03	30.01	32130	0.0661	16.98
	380	379.94	31.09	18534	0.112	22.08
	410	410.13	32.10	13248	0.178	18.07



(a) Via-Temperature dependence (b) Delay-Temperature dependence



(c) Leakage-Temperature dependence

Figure.10 Temperature-delay-power dependence

Via number is also determined by ***Required_T***, which varies greatly between different temperatures. In Ami33, the total via number required for 350K is 548, which is 3.35 times more than that of 410K. Ami49 calls for 57179 to reduce as much thermal as possible with 361K, 6.57 times more than that of 410K, while n100, n200, n300 by 3.38, 2.68, 2.43 times as many as 410K. Since leakage power is proportional to temperature, the ratio of ***Leakage/Dynamic*** will increases with temperature. The value increases by 170.5% from 350K to 410K, as shown in Table III.

Fig.10 shows the dependence of temperature with via number, delay and leakage power. When temperature increases, leakage and delay will also increase, while via number decrease heavily. Different cases have different characters due to different packing ratio and distribution of critical path. Ami49 seems to be sensitive to via number since the dead-space resource in the packing of ami49 is relatively limited. But ami33 seems to be sensitive to delay instead. Normally, the thermal vias can be inserted successfully to meet the required temperature. The ratio between leakage power and dynamic power will be similar as shown in Fig.10(c).

7.3 Performance aware TSV insertion with resource constraints.

In this section, we analyze the impact of *via_ratio* to via number and delay. Besides, we use the evaluation method mentioned in

section VI to value the results of temperature and performance aware TSV insertion.

A. Impact of thermal resource to Via Number

Thermal vias are supposed to be inserted around hot spots and the critical path as many as possible. However, with thermal resource considered, vias cannot be simply inserted around hot spots as much as possible, due to the limited space. Here, we use *via_ratio* to stand for the packing ratio of thermal vias in the dead-space at each grid. Table IV shows the impacts of *via_raito* to via number and delay in ami33 with initial temperature 273°C. It can be obviously seen that *via_ratio* can greatly influence via number. Under the ***Required_T***=350K, the total via number needed when *via_ratio*=0.6 is 47.1% higher than that of *via_ratio*=1. What's more, when *via_ratio*=0.2, there is not enough space for vias to reach the ***Required_T.*** The final temperature is 25°C higher than required temperature. Under the ***Required_T***=380K and 410K, via number required when *via_ratio*=0.2 is 76.0% and 23.5% higher compared to *via_ratio*=1, which proves the impact of *via_ratio* to via number.

Table IV Impact of *via_ratio* in Ami33

via_ratio	*Required T(K)*	*Max T(K)*	*Delay*	*Via Number*	*Max L/D*	*Runtime* (s)
1	350	349.95	36.84	1328	0.0660	3.55
0.6	350	349.80	36.82	1953	0.0658	3.75
0.2	350	375.44	37.54	1409	0.0769	4.13
1	380	380.05	39.29	633	0.1126	4.63
0.6	380	380.05	39.27	672	0.1121	5.13
0.2	380	380.15	39.26	1114	0.1123	5.41
1	410	409.92	41.54	443	0.1779	4.39
0.6	410	409.92	41.55	436	0.1779	5.23
0.2	410	409.90	41.47	547	0.1782	5.35

B. Evaluation of Weighted Performance aware TSV insertion

In this section, we will show the evaluation results of weighted thermal via insertion with performance optimization. In order to compare the cases of with weight considered and without, we divide the results into two groups: **NP** (Normal Pattern) and **WP** (With Weight). Table V shows the effect of weighted TSV insertion on ami33 with initial temperature 152.9°C. Weighted approach can improve via number by 5.6% at most and 3.0% on average and the lower *required_T* is the more significant effect is created. When *required_T* is higher enough, weighted approach can result in more vias to some extent.

Table V Effect of weighted TSV insertion of Ami33

Required T	*Normal Pattern (NP)*			*Weighted Pattern (WP)*		
	Max T	*Delay*	*Via Number*	*Max T*	*Delay*	*Via Number*
350	350.09	36.84	1327	349.95	36.89	1253
370	369.99	38.46	758	369.81	38.68	723
390	389.99	40.07	549	389.89	40.05	536
410	410.00	41.55	436	411.93	41.69	442
430	430.05	42.96	370	430.10	42.94	382
Ratio	1	1	1	1	1.	0.97

Table VI Best results with different weight

Type	*Weight* (α, β, γ)	*Max T* (K)	*Normalized Delay*	*Via Number*
Ami33	**0,0,1**	430.13	42.67	84
	0,1,0	350.09	35.90	548
	0.5,0.3,0.2	430.13	42.67	84
Ami49	**0,0,1**	430.17	33.03	5464
	0,1,0	361.74	31.02	57179
	0.5,0.3,0.2	430.17	33.03	5464
N100	**0,0,1**	429.91	41.32	3226
	0,1,0	349.80	38.87	14641
	0.5,0.3,0.2	410.02	40.76	4329
N200	**0,0,1**	429.84	39.89	5954
	0,1,0	349.86	37.10	19237
	0.5,0.3,0.2	409.82	39.02	7174
N300	**0,0,1**	430.13	32.72	11436
	0,1,0	350.03	30.01	32130
	0.5,0.3,0.2	430.13	32.72	11436

Table VI shows the best results of the five benchmarks under different weights combination in (16). The result can be different due to the different values of parameters of α, β, γ in Section VI. Therefore, using the different cost functions, we can obtain the different trade-off between different factors.

VIII. CONCLUSION

Leakage power and delay are both relevant to temperature, and increase as the temperature increases so that the leakage effects cannot be ignored in 3D design. Though TSV (Through-silicon-vias) can help to remove heat which will in turn reduce the leakage power, but they create routing congestion, which also leads to longer interconnects and also delays. Therefore, how to reach the trade-off between temperature, via number and delay, required to be solved. In this paper, we rethink of thermal via planning of 3D ICs with Timing-Power-Temperature Dependence aware. We try to get the most appropriate via number with time, area and temperature constraints, which make the best balance of delay, via number and temperature. The approaches proposed for the evaluation of TSV planning could be very helpful for a new leakage and time aware TSV-Driven floorplanning.

REFERENCES

[1] B. Black, D. W. Nelson, C. Webb, and N. Samra., "3D processing technology and its impact on IA32 microprocessors," in Proc. Int. Conf. Computer Design, Oct. 2004, pp. 316–318.

[2] Pingqiang Zhou, Yuchun Ma, Qiang Zhou, Xianlong Hong, "Thermal Effects with Leakage Power Considered in 2D/3D Floorplanning." In Procceedings of CADCG,2007, 338-343, USA.

[3] Hao Hua, Chris Mineo, Kory Schoenfliess, Ambarish Sule, Samson Melamed, Ravi Jenkal, and W. Rhett Davis, "Exploring Compromises among Timing, Power and Temperature in Three-Dimensional Integrated Circuits", DAC, 2006, Pages:997 – 1002, San Francisco.

[4] Y. J Lee, R. Goel, S. K. Lim, "Multi-functional Interconnect Co-optimization for Fast and Reliable 3D Stacked ICs", in Procceedings of ICCAD, 2009, Pages: 645-651, California.

[5] JL Ayala, A Sridhar, V Pangracious, D Atienza, YLeblebici, "Through Silicon Via-Based Grid for Thermal Control in 3D Chips", Nano-Net 4th International ICST Conference, 2009, Pages: 90-98, Lucerne, Switzerland.

[6] J. Cong and Y. Zhang. Thermal-Driven Multilevel Routing for 3-D ICs. *ASPDAC*, 2005, pages 121–126, Jan.

[7] J.Cong, J. Wei and Y. Zhang, "A Thermal-Driven Floorplanning Algorithm for 3D ICs", in Procceedings of ICCAD, 2004

[8] W. L. Huang, G.M. Link, Y. Xie, N. Vijaykrishnan and M.J Irwin, "Interconnect and Thermal-Driven floorplanning for 3D microprocessors", in Procceedings of ISQED, Mar. 2006

[9] Z.Li, X.Hong, et. al, "Efficient thermal via planning approach and its application in 3D floorplanning," IEEE Trans. Computer-Aided Design, 2007

[10] P. Zhou, Y. Ma, Z. Li, R.P. Dick, L. Shang, H. zhou, X.L. Hong and Q. Zhou, "3D-STAF: Scalable Temperature and Leakage Aware Floorplanning for Three Dimensional Integrated Circuits", In procceedings of ICCAD, 2007

[11] C.H. Tsai and S.M.S Kang, "Standard cell placement for even on-chip thermal distribution", in Procceedings of ISPD, 1999

[12] K. Skadron, M.R Stan, W. Huang, S. Velusamy, K. Sankaranarayanan D. Tarjan, "Temperature-aware Microarchitecture", in Procceedings of ISCA, 2003.

[13] J. Cong and Y. Zhang, "Thermal Via Planning for 3D ICs", Proc. Of IEEE/ACM ICCAD, Nov.2005

[14] T.D. Richardson and Y. Xie, "Evaluation of Thermal-aware design Techniques for Microprocessors", in Proceedings of ASICON, 2005.

[15] Eric Wong, Sung Kyu Lim: Whitespace redistribution for thermal via insertion in 3D stacked ICs. ICCD 2007: 267-272

[16] Xin Li, Yuchun Ma, Xianlong Hong, Sheqin Dong, Jason Cong: LP based white space redistribution for thermal via planning and performance optimization in 3D ICs. ASP-DAC 2008: 209-212

[17] Hao Yu, Yiyu Shi, Lei He, Tanay Karnik: Thermal Via Allocation for 3-D ICs Considering Temporally and Spatially Variant Thermal Power. IEEE Trans. VLSI Syst. 16(12): 1609-1619 (2008)

[18] B. Goplen and S. Sapatnekar. "Thermal Via Placement in 3D ICs," Proc. of ISPD, pp 167-174, Apr. 2005