On-chip thermal gradient analysis and temperature flattening for SoC design*

Takashi Sato[1], Junji Ichimiya[2], Nobuto Ono[3], Kotaro Hachiya[4], and Masanori Hashimoto[5]
[1]Renesas Technology, [2]Ricoh, [3]Jedat Innovation, [4]NEC Electronics, [5]Osaka Univ.

ABSTRACT

This paper quantitatively analyzes thermal gradient of SoC and proposes a thermal flattening procedure. First, the impact of dominant parameters, such as area occupancy of memory/logic, power density, and floorplan on thermal gradient and clock skew are studied. Important results obtained here are 1) the maximum temperature difference increases with higher memory area occupancy and 2) the difference is very floorplan sensitive. Then, we propose a procedure to amend thermal gradient. A slight floorplan modification using the proposed procedure improves on-chip thermal gradient significantly.

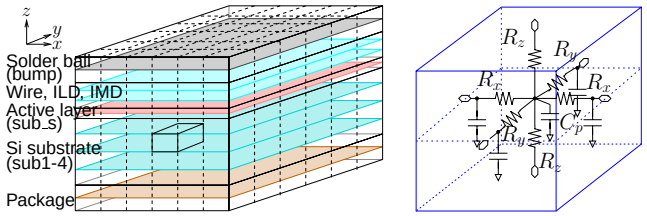
I. INTRODUCTION

As device density increases with scaling, higher power consumption and temperature have been rapidly becoming crucial for the system on a chip (SoC) design. Thermal management techniques as well as simulation techniques are proposed for high-end processors [1, 2, 3], but thermal impacts on SoC design have not been studied throughly yet. Since SoC and processors are very different in cost, especially for packaging, the techniques tailored for processors may not be adequate for SoC designs. Studying thermal impact considering appropriate packaging environment for SoC is required.

Temperature variation on a chip affects gate and wire delay characteristic, thus causes the fluctuation of signal timing. Specifically, long wires used in a global clock tree suffer increased clock skew and performance degradation [4]. Calculating temperature distribution enables instance-based temperature-dependent timing simulation. But obtaining a block placement or power consumption distribution that yields best or worst thermal gradient is not a trivial task. From timing design standpoint, it is much useful to exploit constant temperature distribution through intended placement or architectural techniques [3].

Thermal flattening will become a key technology for more advanced process nodes. It not just achieves smaller timing variation but it also improves reliability since flattening eventually reduces the maximum temperature. Higher temperature exponentially shortens EM lifetime [5, 6]. Moreover, flattened temperature contributes to accurate performance predictability because it matches conventional worst case simulation scenario which applies one worst temperature to all instances.

As far as the authors know of, there is no paper which presents quantitative analysis targeted for SoC varying key parameters for floorplan design, for example where to or not to place 'hot' logic circuits. Prior researches analyze local thermal distribution accurately but considering a limited number of wires [4]. Efficient calculation for thermal distribution is proposed in [7] but it assumes homogeneity of material. More general methods for faster thermal simulation are proposed in [8, 9] for final verification purposes. Papers [10, 11] modified force-directed placement [12] considering temperature as additional force to achieve flat thermal distribution in the cell placement.



(a) Division of a chip (b) A cell construction

Fig. 1. Thermal simulation model used in this analysis.

Novel contributions of this paper are summarized as follows:

- analyzes thermal impacts and clarifies the necessity of considering temperature effects in SoC design.
- proposes temperature flattening procedure which incrementally modifies floorplan to improve performance such as timing and reliability.

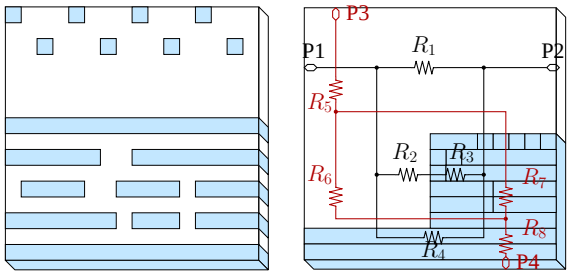
II. NUMERICAL ANALYSIS OF ON-CHIP THERMAL GRADIENT FOR SOC

A. Modeling for thermal simulation

In our analysis, we utilize finite-difference approach similar to [13] to solve the heat diffusion equation. Figure 1 shows the model structure including package. A die as well as packaging material is discretized by 3-D grids to form a cell which are connecting mutually in x, y, and z directions. Each cell consists of thermal resistance (R_x, R_y, R_z) and capacitance (C_p) as depicted in Fig. 1(b). The idea of the model calculation for x, y-direction is illustrated in Fig. 2. We first rearrange metal wires in a cell. Thermal resistance between ports P1 and P2 is calculated as combined resistance of $R_1, R_2 + R_3$, and R_4, where R_1 represents inter-metal dielectric (IMD), R_2+R_3 represents the IMD and the metal wire, and R_4 is metal wire only. The wires which run through the cell (R_4) and the wires which do not (R_3) are distinguished in the model to capture thermal resistance accurately. The equivalent resistances for y and z directions are calculated in the same manner. C_p consists of the sum of metal wire and insulation material thermal capacitance calculated from their volume and specific heat. Thermal resistance of the package is usually much larger than that of a silicon die therefore it dominantly determines chip temperature. In this paper, we use a package of 2.7 K/W thermal resistance as a typical package for high-end SoC.

*This work was conducted as part of activities in the physical design methodology (PDM) study group, an EDA technical committee of JEITA.

0-7803-8736-8/05/$20.00 ©2005 IEEE.



(a) Wire distribution (b) Simplification for modeling

Fig. 2. Equivalent thermal resistance calculation for x, y directions.

TABLE I
EXAMPLE LAYER DIVISION AND THERMAL PROPERTIES.

Layer	Thickness (μm)	R_x (K/W)	R_y (K/W)	R_z (K/W)	C_p (J/K)
package	200	2.7e4	2.7e4	2.72e3	9.81e-5
sub1∼4	125	63.49	63.5	2.54	8.00e-5
sub_s	2	9.0e3	1.8e4	0.037	1.56e-6
wire1∼2	3.1	7.8e3	7.8e3	0.93	5.39e-6
bump	200	5.0e4	5.0e4	1.0e3	1.69e-5

B. Example thermal property

Table I shows an example of layer stack-up and corresponding thermal characteristics. Here the size of the die is 10 mm square for x, y-direction. The dimensions are determined by referring 90-nm node process in the ITRS [14]. LSI is divided into 16 in x, y and 9 in z direction respectively creating 2304 cells in total. Wire material is copper and dielectrics (IMD and inter-layer dielectric (ILD)) are both SiO_2. Wire and ILD are modeled as 2 layers ('wire1' and 'wire2'). Both consist of 4 metal and 4 ILD layers.

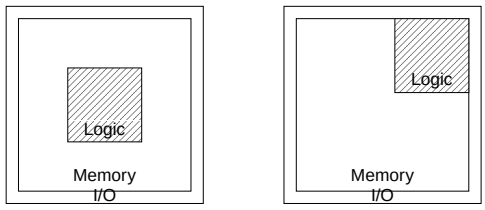
We assume that the allowed maximum junction temperature $T_{j_{max}}$ for this LSI is 120 ˚C and the environment temperature is 27 ˚C throughout the analysis. Total power consumption of the chip is $P_t = 32$ W which is the largest power allowed for the assumed package and $T_{j_{max}}$. P_t is obtained by considering the chip as a point (this is equivalent to using average temperature as the chip temperature ignoring the gradient).

C. Varying parameters

Circuit functionality dependency In recent SoC, memory area occupancy has increasing trend — ITRS expects it becomes 75% in year 2003 and 93% in 2012 [14]. Since there is usually substantial difference in the power density between memory and logic, their ratio should affect temperature distribution. Here we define memory area occupancy α_m by excluding I/O circuit area, i.e. α_m =(memory area)/(logic area + memory area).

Floorplan and power density dependency Floorplan is another parameter. Due to the difference in power density, placement of memory and logic may have a big influence on the thermal profile. Figure 3 shows two different floorplans: type-C places logic at the center of the die, and type-L places logic at the corner.

Block partitioning and placement dependency Starting from type-C or type-L as initial placement, we vary block partition and placement at the same time to find the optimal floorplan of the blocks.



(a) Logic center (type-C) (b) Logic corner (type-L)

Fig. 3. Two different floorplan strategies for logic circuit placement.

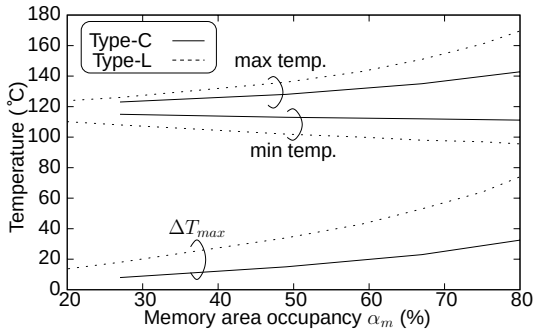


Fig. 4. Memory area occupancy dependency of on-chip temperature.

III. THERMAL SIMULATION RESULTS

In this section we present simulation results of key parameters to analyze qualitative importance of on-chip thermal profile on SoC design.

A. Temperature characteristics of gate and wire delay

Gate and wire delay change were calculated through SPICE simulations. Gate delay change is about 4% with 40 ˚C temperature difference in a 130-nm industrial process. The wire resistance increased about 12% for 40 ˚C around the nominal temperature. Delay change for the wire resulted in about 5% for 40 ˚C using the same process.

B. Circuit functionality and floorplan dependency

Figure 4 shows maximum and minimum temperature, and ΔT_{max}, which we define as maximum minus minimum temperature, as a function of the memory area occupancy α_m. Memory power density is assumed to be 0.25 W/mm^2, I/O circuit consumes 20 % of total power, and the rest of the power is consumed by logic circuit. Independent of α_m, type-L floorplan always shows larger ΔT_{max}. This is because of the adiabatic boundary condition at the chip periphery. As α_m increases, ΔT_{max} becomes larger since power concentrates on smaller logic area. We see that high power consumption SoC with small logic area tends to have steeper thermal gradient. When memory occupies more than 60% of the chip area, which is not a rare case in recent SoC designs, ΔT_{max} becomes larger than 40 ˚C in this example.

C. Block partitioning and placement dependency

We investigate how the temperature profile is affected by the placement and partitioning. Chip area organization and total

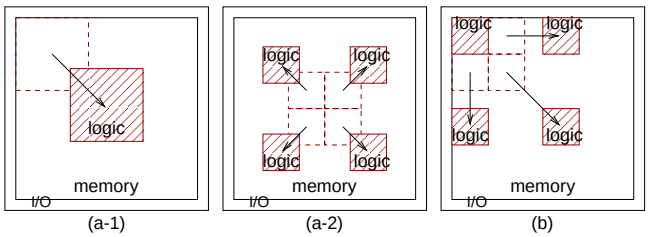


Fig. 5. Logic block movements and partitioning. Dotted blocks represent initial placement and arrows show directions of the block-shift.

chip power are both constant.

Pattern (a-1): Locate a logic block at the corner of the chip (same as type-L in Fig.3) first, then move it diagonally to the center (becomes type-C).

Pattern (a-2): Starting from type-C, divide logic circuit into four smaller blocks and shift each block to four different corners diagonally and synchronously.

Pattern (b): Starting from type-L, divide logic circuit into four smaller blocks and shift three blocks to x, y directions, and diagonally. One block at the corner stays unmoved due to a constraint.

The ΔT_{max} is illustrated in Fig. 6(a). Horizontal axis represents shift distance from original location measured by a cell-grid regardless of the direction. The final placement of (a-1) and the initial placement of (a-2) are identical; the final placement of (a-2) and (b) are also the same. The graph shows that the closer the block is to the corner, the higher the ΔT_{max} becomes. ΔT_{max} is 75 ℃ for type-L and 34 ℃ for type-C. Dividing a block into smaller ones reduces ΔT_{max} further until 11 ℃ and there exists convex optimal position.

Clock skew due to thermal distribution is analyzed using SPICE simulation. Assumed global clock forms symmetrically balancing tree over the entire chip as depicted in Fig. 6(c). The clock tree has 6 buffer stages at the maximum from tree root to the starting points of local clock distribution. The longest wire length in this example is 1.25-mm and each wire is modeled as 4 stage π-model. Correspondent to Fig. 6(a), Fig. 6(b) shows the maximum skew normalized by the nominal clock delay measured at 36 buffer output points. The maximum skew reaches about 10% of clock latency when ΔT_{max} is at its maximum. The clock skew becomes smaller for floorplans with smaller ΔT_{max}. As for the reliability standpoint, 8x shorter EM lifetime is observed using Black's equation [5] since local temperature exceeds $T_{j_{max}}$ significantly in Fig. 6(a).

Above analysis shows that effects of on-chip thermal gradient are unavoidable for high-end SoC especially in the designs containing a lot of memory and concentrated logic power. From Fig. 6, we see there are two ways to reduce temperature-dependent clock skew. One is to minimize ΔT_{max}, and the other is to equalize temperature for clock distributing branches. The efficacy of matching branch temperatures can be understood by pattern (a-2). Because the clock tree is constructed as both x- and y-axis symmetric, temperature distribution of each branch becomes the same, resulting in zero skew although the ΔT_{max} is non-zero. However, this situation is practically hard to realize. On the other hand, reducing ΔT_{max} is block placement dependent and is worth considering. Thermal flattening to minimize ΔT_{max} will be discussed in the next section.

IV. Temperature flattening procedure in floorplan design

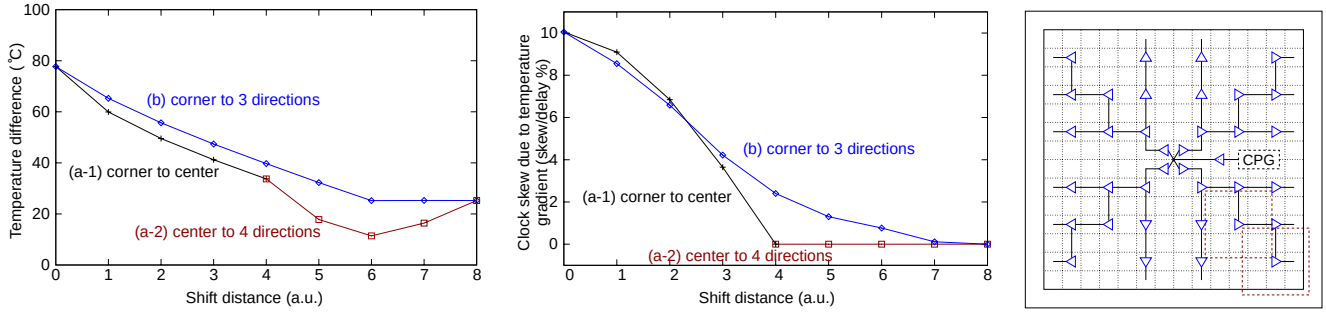
Above simulation results suggest earlier design stages are more suitable for improving thermal gradient since small change made in floorplan makes large difference in temperature distribution. We take incremental optimization approach after initial floorplan based on timing. The reason is as follows. We found that the temperature effect in delay is approximately 10 % at the largest for recent SoC. It may be a rare case when clock and data are routed using paths that are significantly different in temperature but converge in the same FF because the thermal gradient on an SoC cannot become sharp for typical combination of the chip power and package thermal resistance in our experiments. Thermal impact on delay becomes progressively important but it is, for a while, a secondary problem.

In the following, we limit grid-cells in discussion are at the substrate surface, layer 'sub_s'. Let T_i be temperature of i-th grid-cell. Here $i \in S_k$ and S_k is a set of cell number of continuous grid-cells that belongs to a logic block k. The procedure to decrease ΔT_{max} is summarized as follows.

1. Generate an initial floorplan based on timing optimization without considering temperature.
2. Calculate temperature distribution. Stop if ΔT_{max} is less than predefined threshold or iteration count exceeds predefined limit. Otherwise, go to step-3.
3. For each block, calculate sum of thermal gradient vector $\boldsymbol{v_d}$ projected on x-y plane. Move a block j that has the largest $\|\boldsymbol{v_d}\|$ to the direction that $\boldsymbol{v_d}$ points to improve logic block placement, then go to step-2. Here, $\boldsymbol{v_d} = \sum_i \mathrm{grad}\, T_i$ and $\|\boldsymbol{a}\|$ is Euclidean norm of a vector $\boldsymbol{a}$.

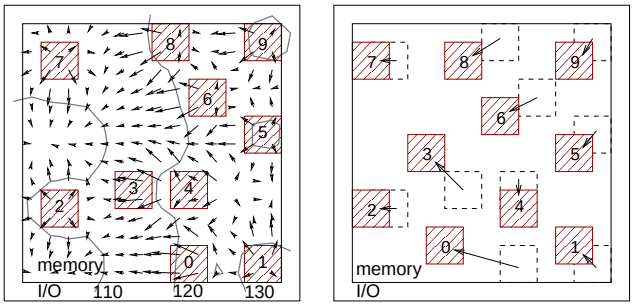
Results in Sec. III suggest that 1) partitioning a logic block into smaller pieces and 2) distributing them apart are effective to flatten chip temperature. Because effective thermal resistivity of a chip is determined mainly by a package used, projection of $\boldsymbol{v_d}$ on x-y plane can usually be well approximated by ignoring a z coordinate of the gradient vector $\partial T_i/\partial z$. In the following experiments, the direction of the movement is 8-ways neighbor and its distance is limited to one cell-grid.

Figure 7(a) shows logic block placement after initial floorplan. Ten blocks each of 2x2 grid size are numbered for reference. Initial isothermal line and thermal gradient vector projection are also over-wrapping. The calculated ΔT_{max} for initial floorplan is 26.1 ℃. Due to power concentration on the right-half of the chip, we notice thermal gradient as many left-pointing arrows. The gradient can also be clearly observed in temperature map in Fig. 8(a). This gradient pushes blocks mainly to the left. After 23 iterations, ΔT_{max} has been reduced to 7.4 ℃ (Fig. 8(b)). Here, we see that relative position of the blocks does not change much during the flattening process, which is a good property for timing convergence. The procedure not only improves quality of timing design by reduced timing difference due to thermal gradient, but signifi-



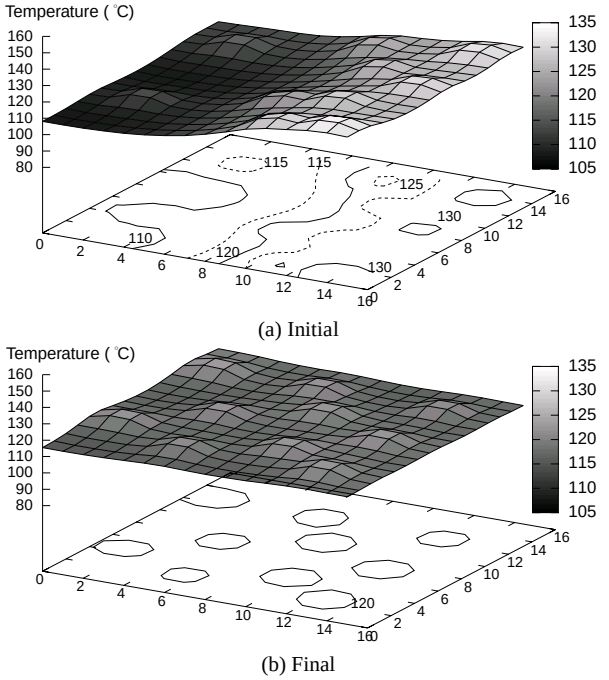
(a) ΔT_{max} (b) Maximum clock skew (c) Clock tree topology

Fig. 6. Floorplan dependency of the maximum temperature difference and clock skew.



(a) Initial (b) Final

Fig. 7. Logic block placement before and after thermal flattening. Non-hatched areas are memory and I/O circuits. (a) also shows thermal gradient vector and isothermal map, (b) shows block movements from (a). Only slight changes in floorplan improve temperature difference significantly from $\Delta T_{max} = 26.1$ ℃ to 7.4 ℃ after 23 iterations.



(a) Initial

(b) Final

Fig. 8. Die surface temperature distribution of Fig. 7.

cantly improves reliability because the process reduce maximum temperature on a chip.

V. Conclusion

We have proposed a thermal modeling method which is effective for chip-level analysis and optimization in SoC design. Effects of dominant factors for determining the thermal gradient such as memory and logic area, power consumption, floorplan, etc. are quantitatively analyzed. Further, the global clock skew is simulated for various floorplans using industrial device models, which confirmed the correlation between maximum temperature difference and skew. We also showed that modifying floorplan effectively reduces EM risk. Finally, a practical temperature flattening procedure is presented. It was found that even a small shift of logic blocks is able to improve circuit performance substantially.

References

[1] J. Clabes et al., "Design and implementation of the POWER5™ microprocessor," in *Proc. ISSCC*, 2004, pp. 56–57.

[2] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. ISLPED*, August 2003, pp. 78–83.

[3] K. Skadron et al., "Temperature-aware computer systems: Opportunities and challenges," *IEEE Micro*, vol. 23, no. 6, pp. 52–61, Nov.-Dec. 2003.

[4] A. H. Ajami, M. Pedram, and K. Banerjee, "Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs," in *Proc. CICC*, 2001, pp. 233–236.

[5] J. Black, "Electromigration — a brief survey and some recent results," *IEEE Trans. Elec. Dev.*, vol. ED-16, no. 4, pp. 338–347, April 1969.

[6] K. Banerjee et al., "Analysis and optimization of thermal issues in high-performance VLSI," in *Proc. ISPD*, 2001, pp. 230–237.

[7] Y.-K. Cheng and S.-M. Kang, "Fast thermal analysis for CMOS VLSIC reliability," in *Proc. CICC*, 1996, pp. 479–482.

[8] Z. Yu, D. Yergeau, and R. Dutton, "Full chip thermal simulation," in *Proc. ISQED*, 2000, pp. 145–149.

[9] T.-Y. Wang and C. C.-P. Chen, "3-D thermal-ADI: a linear-time chip level transient thermal simulator," *IEEE Trans. on CAD*, vol. 21, no. 12, pp. 1434–1445, Dec. 2002.

[10] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," in *Proc. ICCAD*, 2003, pp. 86–89.

[11] B. Obermeier and F. M. Johannes, "Temperature-aware global placement," in *Proc. ASP-DAC*, 2004, pp. 143–148.

[12] H. Eisenmann and F. Johannes, "Generic global placement and floorplanning," in *Proc. DAC*, 1998, pp. 269–274.

[13] C.-H. Tsai and S.-M. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE Trans. on CAD*, vol. 19, no. 2, pp. 253–266, February 2000.

[14] SIA, *International Technology Roadmap for Semiconductors*, 2003 Edition.