Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs

Sungjun Im and Kaustav Banerjee

Center for Integrated Systems, Stanford University, Stanford, CA 94305
{sjim, kaustav}@stanford.edu

ABSTRACT

This work presents a full chip thermal analysis of 2-D high performance ICs based on technological, structural, and material data from ITRS '99 [1]. It is shown that interconnect Joule heating in advanced technology nodes can strongly impact the magnitude of the maximum temperature within 2-D chips despite negligible change in the chip power density, as per the ITRS. This result has been shown to have significant implications for interconnect reliability and performance not foreseeable by the ITRS. Furthermore, detailed thermal analysis of vertically integrated (3-D) ICs has been carried out using analytical modeling and numerical simulations. Additionally, comparison between the thermal design of two alternative 3-D technologies has been presented for the first time using ITRS data.

INTRODUCTION

Management of thermal issues is central to the development of future generation microprocessors, integrated networks, and other highly integrated systems. It is rapidly becoming one of the most challenging issues in high-performance chip design due to ever increasing device count and clock speed. Thermal problems have important implications for performance and reliability. Additionally, design and performance of emerging circuit architectures, such as vertically integrated (3-D) ICs will be significantly affected by thermal effects arising due to increased power density in these architectures. The 3-D architecture offers unique advantages both in terms of circuit performance (lower RC delay), and on-chip integration of digital, analog, and mixed signal circuits [2,3]. With the growing menace of RC delay in 2-D circuits, the 3-D architecture is being viewed as a potential alternative that can not only maintain chip performance [2-5] but also become a vehicle for System-on-a-Chip design in the near future [2,3]. Hence, careful thermal design of 3-D ICs is central to their development. At present, there is very limited information available on the thermal issues in 3-D ICs [2,3,6].

METHODOLOGY

In this work, the actual full chip models for 2-D ICs as per ITRS data at each technology node have been adopted and analyzed. Three-dimensional Finite Element (FE) thermal simulation (ANSYS) has been employed to account for Joule heating of orthogonal multilevel interconnects. Under steady state conditions, with uniform heat generation in a homogeneous medium with constant properties, the general heat equation can be expressed as,

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}+\frac{\dot{q}}{k}=0 \qquad (1)$$

where $\dot{q}$ is the uniform internal heat generation in W/m^3 and k is the thermal conductivity of the material. For the interconnects, $\dot{q}$ represents the volumetric Joule heating given by $j_{rms}^2\,\rho(T_{Die})$, where T_{Die} is the die temperature. For the other materials $\dot{q}=0$. It should be noted that the temperature dependence of the metal resistivity $\rho(T)$ must be included for most rigorous analysis. However, simulations based on both $\rho(T)$ and fixed resistivity at the die temperature $\rho(T_{Die})$ was carried out for less complex structures and the difference in temperature rise was found to be small. Hence, ρ was assumed to be constant in order to avoid convergence problems. Inter-layer and inter-metal dielectric materials (ILD and IMD) were assumed to be identical. Adiabatic boundary conditions have been applied on four sidewalls where all the metal wires and insulators satisfy symmetric boundary conditions. Assuming no heat removal from the top surface of the chip to ambient air, adiabatic boundary condition is applied on that surface as well. The substrate layer plus the package thermal resistance (R_n) has been determined by using a one-dimensional heat equation given by,

$$T_{Die}=T_0+R_n\left(\frac{P}{A}\right) \qquad (2)$$

where T_0 (=25 ^{0}C) is the ambient temperature, P is power dissipation, A is the chip area. Using (2) R_n was found to be 4.75 cm^2 °C/W, based on the operating chip temperature (T_{Die}=120 °C) for the present technology node (180 nm). Assuming the same value for R_n, the die temperatures at other technology nodes were obtained using (2). These die temperatures were applied to the bottom surface of the Si substrate to maintain same R_n. The power density ($\Phi=P/A$) at a given technology node was applied as surface loads in W/m^2 at the upper surface of the Si substrate. All interconnect layers were assumed to carry the maximum RMS current density (based on Table 1) for worst case analysis.

Thermal conductivity values of the ILD and IMD materials ($k_{Dielectric}$) used in the FE simulations were estimated as follows. Values of $k_{Dielectric}$ were plotted against the dielectric constant (ε) for four materials with known thermal conductivities as shown in Fig. 1. Using average values of ε at various technology nodes (based on ITRS), corresponding values of $k_{Dielectric}$ were interpolated based on Fig. 1.

THERMAL ANALYSIS OF 2-D ICs

Temperature contours within the ICs were obtained using fully coupled three-dimensional FE thermal simulation accounting for both power dissipation in the devices and Joule heating in the wires. The simulations at various technology nodes were based on parameters shown in Table. 1, which have been determined from the ITRS data. The thickness of the ILD layers was extracted from the aspect ratio of the vias assuming via size to be equal to the mimimum line widths at any given tier (local, semi-global, or global). For estimating the interconnect Joule heating, RMS values of the current density were calculated from values of J_{max} (=J_{peak}) given in Table 1 assuming a duty factor of 0.12 [7].

Fig. 2 shows a plot of the die temperature (T_{Die}) and the maximum temperature in the chip (T_{max}) as a function of the technology node for 2-D ICs. It can be observed that the power density Φ_{2D}, remains almost constant, and T_{Die}, which is directly proportional to Φ_{2D}, remains more or less within 133 ^{0}C ± 15 ^{0}C. However, T_{max}, increases rapidly with scaling due to increased Joule heating of the interconnects.

The spatial temperature distributions along the vertical direction from the Si substrate have been plotted from simulations as shown in Fig. 3. It can be observed that as technology scales down, T_{max} increases and the temperature gradient between the top metal lines and the Si substrate becomes larger. For the 35 nm node, T_{max} and temperature gradient are smaller than that for the 50 nm node due to the larger fraction of Cu in the (Cu+ILD) layers. It should be noted that the total height of the (Cu+ILD) layers decreases as scaling continues, due to the smaller vertical dimensions of wires and insulators despite increasing number of metal levels. It can also be observed that the maximum chip temperature, T_{max}, occurs around the long global wires, which are most prone to electromigration failures and also give rise to highest RC delays [7]. This has important implication for both reliability and performance.

Fig. 4 plots the percentage decrease in electromigration (EM) time to failure (TTF) (%D_TTF) of global wires at various technology nodes with respect to the TTFs at the corresponding die temperatures. It can be observed that the %D_TTF increases rapidly beyond the 130 nm node. Fig. 5 shows the performance implication of increased temperature for the global lines. It can be observed that the percentage increase in RC delay (%I_RC) due to temperature rise of interconnects above the corresponding die temperatures increases with scaling. These results indicate that for future 2-D ICs chip level thermal issues must be

0-7803-6438-4/00/$10.00 ©2000 IEEE

carefully considered at an early design (both layout and process) phase and advancement in chip cooling technologies will become imperative.

THERMAL ANALYSIS OF 3-D ICs

Analytical Die Temperature Model

A simple analytical model is proposed to estimate the temperature rise in each active layer of 3-D chips. The temperature rise (above the ambient temperature) of the j^{th} active layer in an n-layer 3-D chip (Fig. 6) can be expressed as,

$$\Delta T_j = \sum_{i=1}^{j}\left[R_i\left(\sum_{k=i}^{n}\frac{P_k}{A}\right)\right] \qquad (3)$$

where n is the total number of active layers, R_i represents the thermal resistance between the i^{th} and the $(i\text{-}1)^{th}$ layers and P_k is the power dissipation in the k^{th} layer. Note that this model does not take into account interconnect Joule heating. Assuming identical power dissipation (P) in each layer and identical thermal resistances (R) between layers, the temperature rise of the uppermost (n^{th}) layer in an n-layer 3-D chip can be expressed as,

$$\Delta T_n = \left(\frac{P}{A}\right)\left[\frac{R}{2}n^2 + \left(R_1 - \frac{R}{2}\right)n\right] \qquad (4)$$

where R_1 is mostly due to the package thermal resistance between the first layer and the heat sink (separated by the package layer of thickness t_{pkg}) and R is the thermal resistance between the i^{th} and the $(i\text{-}1)^{th}$ layers for $i \neq 1$. $R_1=(t_{Si\_1}/k_{Si})+(t_{pkg}/k_{pkg})$, and $R=(t_{Si\_i}/k_{Si})+(t_{glue,i}/k_{glue,i})+(t_{ins,i}/k_{ins,i})$ respectively. Here, t_{Si_i} is the thickness of the i^{th} Si layer, and $t_{glue,i}$ and $t_{ins,i}$ are the thickness of the i^{th} glue and insulator layers respectively. From (4) the temperature rise can be expected to increase linearly with power density and the square of the number of active layers, n. However, for all practical 3-D ICs, $R_1 >> R$, which gives rise to an approximately linear relationship between ΔT_n and n as shown in Fig. 7. Equation (4) also suggests that for most 3-D ICs with $n \leq 5$, R_1 will dominate the temperature rise of any layer, if interconnect Joule heating is ignored.

Full Chip Thermal Analysis

For full chip thermal analysis which included interconnect Joule heating (excluding the vias), two technologies employing processed wafer bonding have been analyzed in detail. As a practical case study, two-layer 3-D structures with each layer representing a 100 nm node chip as per ITRS have been analyzed as shown in Fig. 8. The thermal design of two wafer bonding techniques, (a) Case I: 3-D ICs fabricated by wafer bonding using polymer adhesives (glue layer) [8] and (b) Case II: 3-D ICs fabricated by wafer bonding using a thermocompression method [9], have been compared in this study. It should be noted that 3-D ICs fabricated by solid-phase crystallization (SPC) [3] can be thought of as a special example of Case I with $t_{glue}=0$.

Since the power dissipation in 3-D circuits (P_{3D}) has a strong design dependence [2], we chose Φ_{3D} as an independent variable within the range of $\Phi_{2D}<\Phi_{3D}<2\Phi_{2D}$ (with $\Phi_{2D} \approx 0.3$ W/mm^2) for the purpose of thermal analysis. For each value of Φ_{3D}, the temperature of the bottom silicon substrate, T_{Si_1}, was determined using (2), assuming that R_n remains constant for both 2-D and 3-D ICs if same packaging and chip cooling technologies are employed at all technology nodes. Φ_{3D} was applied as surface loads in W/m^2 at the upper surface of both Si_1 and Si_2. Interconnect Joule heating estimated from $j_{rms}^2\rho(T_{Si\_1})$ was used as input loads for all the wires to represent a worst case analysis. Heat sinks were attached to only Si_1 for both cases, which resulted in the T_{max} to occur near the top of the chip, similar to 2-D ICs.

Fig. 9 shows T_{max}, T_{Si_1} and the temperature of the upper layer (T_{Si_2}) as a function of Φ_{3D} for a given interconnect current density based on the ITRS at the 100 nm node. It can be observed that T_{Si_2} is higher than T_{Si_1}. This is due to the higher thermal impedance for the second layer and due to thermal coupling from adjacent interconnects. The lines of T_{max} and T_{Si_2} can be observed to gradually diverge from T_{Si_1} as Φ_{3D} increases. This is due to the increased $\rho(T_{Si\_1})$ at higher power densities, which gives rise to more interconnect Joule heating. It can also be observed that T_{max} and T_{Si_2} for Case I are slightly higher than those for Case II due to the lower thermal conductivity of the glue layer compared to that of the Cu bonding pads. Also, the temperature of Si_1 attached to the heat sink remains constant for both cases due to the heat sink.

Fig. 10 shows the temperature distributions within the 3-D chip. It can be observed that the temperatures are nearly identical for both Case I and Case II within the first chip. However, for the second chip, temperature is higher for Case I, which is consistent with Fig. 9. Additionally, the effect of changing the thickness of Si_2 and Cu pads were found to be negligible due to their high thermal conductivities. Furthermore, the thickness of the glue layer in Case I with low thermal conductivity affects T_{max} and T_{Si_2} as shown in Fig. 11. Finally, the impact of the area of Cu pads on the chip temperature in Case II has been observed to be negligible above 50% Cu pad area as shown in Fig.12.

Implications of T_{max} on the EM reliability of the 3-D chip were analyzed and the %D_TTF with respect to the TTFs at the die temperatures was found to be around 85% with negligible variation within the range of Φ_{3D} used in this study. Fig. 13 shows the %I_RC as a function of Φ_{3D}. It can be observed that the %I_RC increases with Φ_{3D} because the rate of change in T_{max} is slightly higher than that of T_{Die} as shown in Fig. 9. Additionally, the 3-D chip of Case I shows a higher %I_RC than that of Case II, due to higher value of T_{max} as shown in Fig. 10. Finally, Fig. 14 plots the effective thermal resistance (R_{eff}) values that would be required to maintain the 3-D chip temperatures to around present 2-D chip temperatures (~120 ^{0}C) for different power densities. It should be noted that most of the contribution in R_{eff} comes from the package thermal resistance. From this data it is evident that in order to extract maximum performance from 3-D ICs advancement in chip cooling solutions similar to [10] to lower the package thermal resistance will be necessary.

CONCLUSION

In conclusion, full chip thermal analysis of 2-D and 3-D high performance chips has been performed using analytical models and numerical simulations. It has been shown that thermal design issues are going to be critical for both 2-D and 3-D ICs, and must be considered during their early design phase. It has also been shown that interconnect Joule heating and low thermal conductivity of dielectric materials will strongly impact the magnitude of the maximum temperature within these chips, with significant implications for reliability and performance. Hence, advancement in thermal properties of low-k dielectrics and chip packaging/cooling technology will be required.

ACKNOWLEDGMENTS

This work was supported by the MARCO Interconnect Focus Center and by the SRC.

REFERENCES

[1] Int. Technology. Roadmap for Semiconductors (*ITRS*), 1999, {http://public.itrs.net/}

[2] S. J. Souri, K. Banerjee, A. Mehrotra, and K. C. Saraswat, "Multiple Si layer ICs: motivation, performance analysis, and design implications," *37th ACM Design Automation Conf.*, 2000, pp. 873-880.

[3] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep submicron interconnect performance and systems-on-chip integration," to appear in *Proceedings of the IEEE, Special Issue on Interconnects*, 2001.

[4] A. Rahman, A. Fan, and R. Reif, "Comparison of key performance metrics in two- and three-dimensional integrated circuits," *Proc. IITC*, 2000, pp. 18-20.

[5] J. W. Joyner et al., "A three-dimensional stochastic wire length distribution for variable separation of strata," *Proc. IITC*, 2000, pp. 126-128.

[6] M. B. Kleiner, S. A. Kuhn, P. Ramm, and W. Weber, "Thermal analysis of vertically integrated circuits," *Tech. Dig. IEDM*, 1995, pp. 487-490.

[7] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," *36th ACM Design Automation Conference*, 1999, pp.885-891.

[8] P. Ramm et al., "Three dimensional metallization for vertically integrated circuits," *Microelectronic Engineering*, 37/38, pp. 39-47, 1997.

[9] A. Fan, A. Rahman, and R. Reif, "Copper wafer bonding," *Electrochemical and Solid-State Letters*, vol. 2 (10), pp. 534-536, 1999.

[10] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, EDL-2, no. 5, pp. 126-129, 1981.

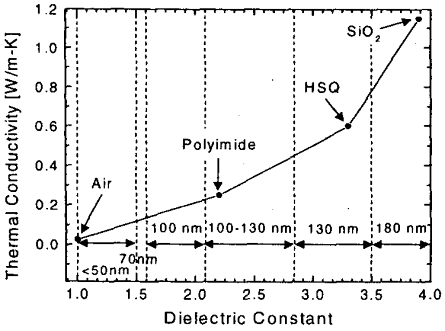

Fig. 1. Thermal conductivity ($k_{Dielectric}$) as a function of dielectric constant, ε. The expected ranges of dielectric constant of ILDs for different technology nodes are specified based on ITRS.

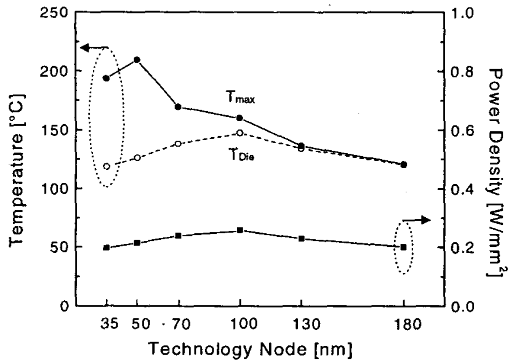

Fig. 2. Maximum temperature (T_{max}), die temperature (T_{Die}), and power density (Φ_{2D}) in 2-D ICs as a function of technology node.

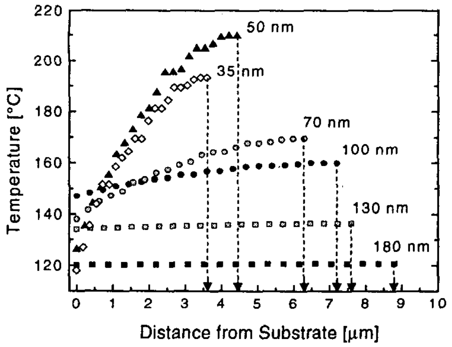

Fig. 3. Temperature distribution along vertical distance from upper surface of Si substrate to top metal level in 2-D ICs. The total height of (Cu+ILD) layers decreases as technology node scales down.

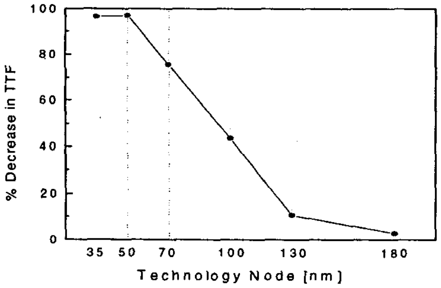

Fig. 4. Percentage decrease in EM time to failure (TTF) for global wires (at T_{max}) with respect to TTF based on T_{Die} at each node for 2-D ICs.

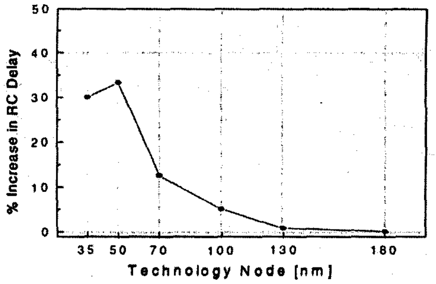

Fig. 5. Percentage increase in RC delay for global wires with respect to that at T_{Die} at each node for 2-D ICs. T_{Die} at each node has been chosen as the reference temperatures and temperature coefficient of resistance (TCR)=0.004 was used for Cu.

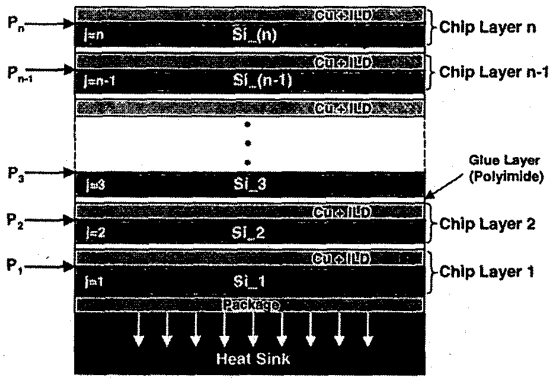

Fig. 6. Schematic of an n-layer 3-D IC with a heat sink at the bottom.

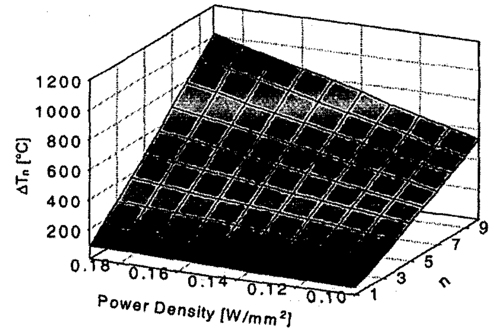

Fig. 7. Temperature increase as a function of the number of chip-layers (n) and the power density in each layer (Φ) calculated using the analytical equation (4).

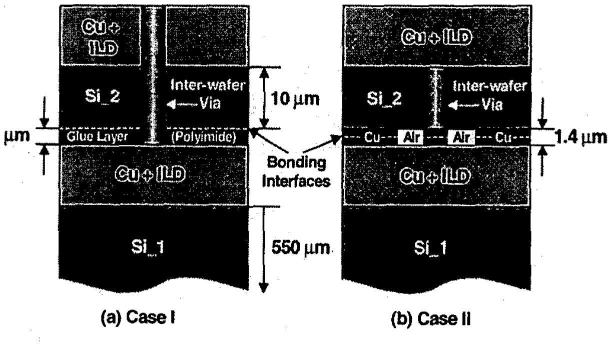

Fig. 8. Schematic of two 3-D structures (a) Case I: Fabricated by wafer bonding with a glue layer or solid-phase crystallization (SPC). (b) Case II: Wafer bonding using Cu pad thermocompression.

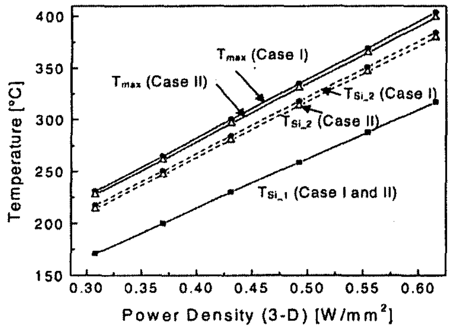

Fig. 9. Maximum chip temperature, temperature at the second Si layer (T_{Si_2}), and substrate temperature (T_{Si_1}) in Case I and Case II for different power density obtained using FE simulations. Each layer is based on a 100 nm node IC. t_{Si_2}=10 μm, $t_{Cu\ pad}$=1.4 μm, and t_{glue} = 2 μm, k_{ILD}=k_{IMD}=0.19 W/m-K, k_{glue}=0.25 W/m-K, k_{Cu} = 400 W/m-K, k_{Si} = 148 W/m-K, $Jrms$=4.85 x10^5 A/cm2.

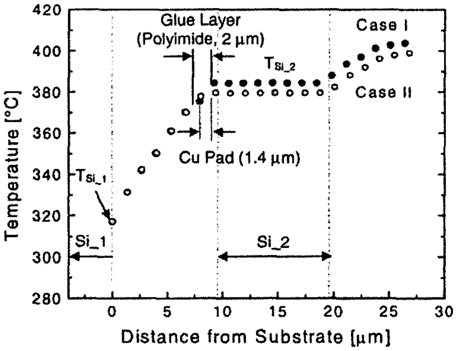

Fig. 10. Temperature distribution along the vertical distance from the upper surface of Si substrate (Si_1) to top metal level for the second Si layer (Si_2) for Case I and Case II for 100 nm node *with* Φ_{3D}=0.6154 W/mm2. Other parameters are same as in Fig. 9.

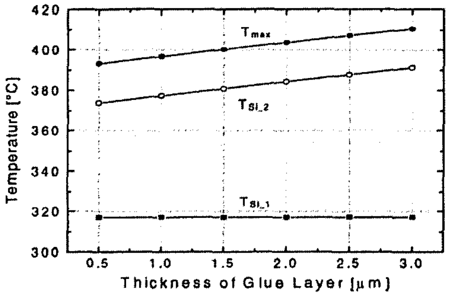

Fig. 11. $T_{max}$and T_{Die} in 3-D ICs (Case I) as a function of thickness of the glue layer (polyimide) for 100 nm node *with* Φ_{3D}=0.6154 W/mm2. Other parameters are same as in Fig. 9.

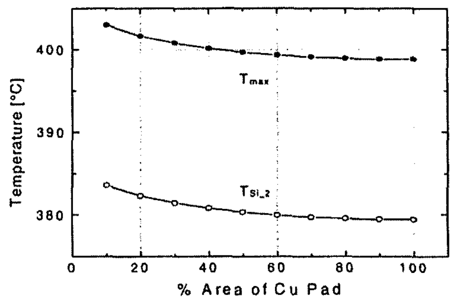

Fig. 12. $T_{max}$and T_{Die} in 3-D ICs (Case II) as a function of percent area of Cu pads for 100 nm node with Φ_{3D} =0.6154 W/mm2. Other parameters are same as in Fig. 9. T_{Si_1} remains constant at 317 °C.

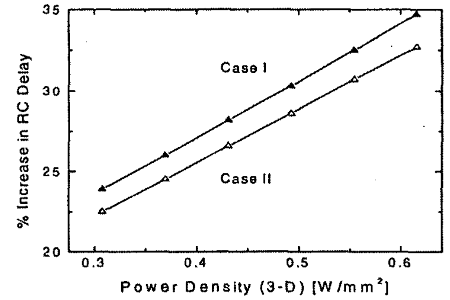

Fig. 13. Percent increase in RC delay for wires at $Tmax$ for various power densities in 3-D ICs at the 100 technology node. The die temperatures have been chosen as the reference temperature and TCR=0.004 was used for Cu. $Jrms$=4.85x10^5 A/cm2 has been applied to all metal wires.

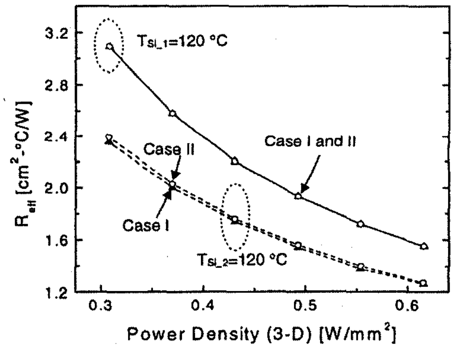

Fig. 14. The effective thermal resistance, R_{eff}, for 3-D ICs required to have the substrate (Si_1) or the second layer (Si_2) at 120 °C as a function of power density for 100 nm technology node. The R_{eff} values for Si_2 have been estimated from FE simulations and equation (2) in the text. Note that smaller values of R_{eff} will be required to operate the upper layer devices at 120 °C due to interconnect Joule heating adjacent to the second chip.

Node	P_{max} *(W)*	*Chip Area (mm^2)*	*# of Metal Layers (L/S/G)*	*Wire Width (μm) (L/S/G)*	*Wire Height (μm) (L/S/G)*	*ILD Thickness (μm) (L/S/G)*	ε	k_{ILD} *(W/m-K)*	ρ_{Cu}*(μΩ-cm) @100 °C*	J_{max}=J_{peak} *(A/cm^2)*
180 nm	90	450	6 (2/2/2)	(0.2/ 0.3/ 0.6)	(0.3/ 0.6/ 1.2)	(0.3/ 0.7/ 1.3)	3.75	1.02	2.2	5.8E5
130 nm	130	567	7 (2/3/2)	(0.2/ 0.2/ 0.4)	(0.3/ 0.4/ 1.0)	(0.3/ 0.4/ 1.0)	3.1	0.54	2.2	9.6E5
100 nm	160	622	8 (2/4/2)	(0.1/ 0.2/ 0.3)	(0.2/ 0.4/ 0.8)	(0.2/ 0.4/ 0.8)	1.9	0.19	2.2	1.4E6
70 nm	170	713	9 (3/3/3)	(0.1/ 0.1/ 0.2)	(0.2/ 0.3/ 0.5)	(0.2/ 0.3/ 0.6)	1.5	0.12	2.2	2.1E6
50 nm	174	817	9 (3/3/3)	(0.07/ 0.07/ 0.14)	(0.14/ 0.2/ 0.4)	(0.14/ 0.2/ 0.4)	1.25	0.07	2.2	3.7E6
35 nm	183	937	10 (3/4/3)	(0.05/ 0.05/ 0.1)	(0.1/ 0.2/ 0.3)	(0.1/ 0.1/ 0.3)	1.25	0.07	2.2	4.6E6

Table 1. ITRS based interconnect parameters for local (L), semi-global (S), and global (G) tier wires used in this study.