Category: Circuit and Device Interaction (CDI) (Student Paper)

Cooling Three-Dimensional Integrated Circuits using Power Delivery Networks

Hai Wei*[+], Tony F. Wu[+], Deepak Sekar[&], Brian Cronquist[#], Roger Fabian Pease[+], Subhasish Mitra[+^]
Department of Electrical Engineering[+] and Department of Computer Science^, Stanford University, Stanford, CA,
Monolithic 3D Inc.[#], San Jose, CA, Rambus[&], Sunnyvale, CA, Email*: haiwei@stanford.edu

Introduction

All technology options for three-dimensional integrated circuits (*3D ICs*) [Batude 11, Naito 10, Topol 05, Van Olmen 08, Wei 09, Wong 07] face the challenge of how to remove the heat dissipated on the upper layers (Fig. 1) [Kleiner 95]. Most existing techniques rely on an array of Through-Silicon Vias (*TSVs*) for this purpose. Here we report how on-chip power delivery networks (*PDNs*) designed to deliver noise-free power can significantly contribute to heat removal in 3D ICs.

Existing publications on heat removal in 3D ICs using TSVs and interconnects [Banerjee 01, Cong 07, Cong 11, Lau 09, Yu 09, Zhang 06] typically assume the silicon to be the primary heat conduit to the TSVs. Unfortunately, advanced 3D technologies, such as sequential or monolithic 3D (Fig. 1), with thin upper layers of silicon have significantly reduced heat conduction through the silicon. Hence, other conduits such as PDNs are essential for effective heat removal. We use the term inter-layer vias (*ILVs*) to refer to vias connecting various components (e.g., metal wires, transistors) belonging to different layers of a 3D IC, as opposed to conventional vias that connect various components within the same layer of a 3D IC. Depending on the 3D technology, ILVs may be implemented using TSVs (e.g., in parallel 3D) or conventional vias (e.g., in monolithic 3D).

Heat Conduction Challenge in 3D ICs

Heat can escape laterally along silicon layers and vertically through ILVs and inter-layer dielectric (*ILD*) (Fig. 1). A major challenge is that silicon layers thinner than 1µm exhibit thermal conductivity significantly lower (e.g., 2-fold lower) than the bulk value [Ju 99]. Hence, advanced 3D technologies such as monolithic 3D (Fig. 1 with T_{Si} of 100 nm) must employ effective ways of achieving lateral heat conduction. This is where PDNs significantly improve heat removal through their thermal conduction. As a result, we achieve lower maximum chip temperature and significantly reduced area consumed by ILVs (that may otherwise be required for heat removal).

Computational Approach

Conventional methods for simulating PDNs [COMSOL, FLO] based on finite element analysis (*FEA*) become computationally expensive, if not infeasible, for large designs such as the OpenSPARC T2 processor core [OpenSPARC] analyzed later in this paper. Similar observation was also made in [Yang 07] for temperature distributions in 2D ICs. To overcome this computation challenge for 3D ICs, we created a full-chip thermal analysis methodology based on the Power Blurring technique [Kemper 06] using abstracted models for PDNs. This methodology (Fig. 2) comprises two steps:

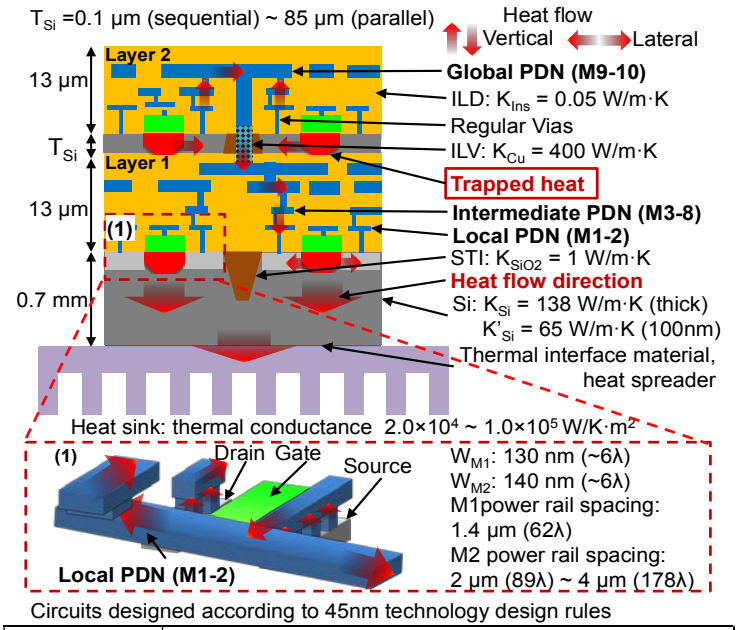
Step 1: For each layer of a 3D IC, abstracted models (3 in total) with anisotropic effective thermal conductivities [Incropera] are constructed for the local PDN and ILD (M1, ILD1, M2, ILD2), the intermediate PDN and ILD (M3, ILD3 - M8, ILD8), and the thin silicon layer with Shallow Trench Isolation (STI) (for monolithic 3D integration). The global PDN in each layer is not abstracted because the dimensions are large enough for the FEA tool, e.g., COMSOL, to properly mesh and simulate. With these abstracted models, the abstracted chip thermal model is created (Fig. 2b). Comparison between the abstracted chip thermal model and a detailed FEA (with no abstraction) shows less than 1% error (defined in Fig. 2c) for a two-layer 3D IC (Fig. 2b).

In Fig. 2a, without loss of generality, we describe the abstraction process for effective thermal conductivity in the z direction, k_z, for the local PDN highlighted in Fig. 2b. We first stack two copper blocks (for modeling purposes only) along the z direction on the top and the bottom of the local PDN to serve as heat source and sink. This forms a copper-PDN-copper stack. The length, breadth, and thickness of the two copper blocks are identical to the local PDN. (If Δz was the thickness of the local PDN, the thickness of this stack would be $3\Delta z$). Heat flux, q (e.g. $5\times10^5 W/m^2$ for OpenSPARC T2 core power density), is applied along the z direction on the top surface of the heat source. The bottom surface of the heat sink is set to 300K (ambient). All other boundaries of the stack (i.e., 4 surfaces perpendicular to x and y axes) are set to be adiabatic. Through FEA [COMSOL], temperature values at the interfaces between the local PDN and the top block ($T_{top}(x,y)$) and the bottom block ($T_{bot}(x,y)$) are computed. Finally, k_z is calculated using equation (1) in Fig. 2a [Incropera]. This process can be repeated for the x and y directions to calculate effective thermal conductivities k_x and k_y, respectively. One needs to stack copper blocks and apply heat flux along the x and y directions (instead of the z direction) and use the length (Δx) and breadth (Δy), instead of the local PDN thickness (Δz) in equation (1), to calculate k_x and k_y.

To quantify the accuracy of the abstracted model for various power density distributions, we ran Monte Carlo simulations of a two-layer 3D IC (25µm×25µm) using randomly-generated power density distributions (Fig. 2c) and compared the temperature distribution using our abstracted model (Fig. 2b) vs. the model using detailed PDNs. To generate power density distributions, each layer was divided into a 5×5 matrix of 5µm×5µm blocks. For each block, a randomly sampled power density value (uniformly distributed from 0 to $300 W/cm^2$) was assigned. The specific error statistic (Fig. 2c) is chosen to make the error independent of the ambient temperature used for the simulations (300K). The mean and standard deviation of the error are 0.034% and 0.01%, respectively.

Step 2: Similar to the Power Blurring technique in [Kemper 06], we compute steady-state temperature distributions by treating a 3D IC (described using the abstracted models in Step 1) as a linear system (with power density distributions as input signals and temperature distributions as responses). Hence, the temperature distributions can be calculated through convolution of (thermal) impulse responses with power density distributions [Oppenheim] of 3D ICs. To calculate the steady-state impulse responses, *H(x,y)*, each layer of the 3D IC was first divided into a matrix of 5µm×5µm

blocks (for ICs with dimension greater than 25µm×25µm, the matrix will be bigger than 5×5). Next, power density of 300 W/cm^2 (maximum possible power in the 3D IC simulated) was assigned to each 5µm×5µm block. The resulting temperature distributions (obtained through FEA) were normalized by the power (300 W/cm^2) to produce the impulse response at each block. With the steady-state impulse responses, the temperature distribution *T(x,y)* can be obtained by convolution of the impulse responses with the power density distribution *P(x,y)* (Fig. 2d). Power density distribution can be estimated using tools such as McPAT [McPAT]. The maximum temperature increase, obtained using our technique, agrees within 5% of previously published data [Etessam-Yazdani 06] (Fig. 2e), thus establishing common ground with accepted approaches. Our approach allows us to explore a large variety of 3D configurations because we can quickly calculate temperature distributions for 3D ICs integrated using a range of integration methods (i.e., different T_{Si} values).



Circuits designed according to 45nm technology design rules

3D integration technology options	**1. Sequential or Monolithic 3D:** T_{Si}<100nm, high ILV density [Batude 11] **2. Parallel 3D:** T_{Si} 1-85µm, low ILV density [Topol 05]
Application options	**1. Memory-on-logic** **2. Logic-on-logic**
Cooling technology options	**1. Conventional air cooling (heat sink & fan):** 2×10^4 W/m^{2}K [Etessam-Yazdani 06] **2. External liquid cooling (not microfluidic cooling using in-chip channels):** 1.0×10^5 W/m^{2}K [Tuckerman 81]

Figure 1. Simulated 3D IC structures [Banerjee 01] including PDNs. The thickness of Layer 2 silicon (T_{Si}) varies from 0.1 to 85µm for different 3D integration technologies. 45 nm technology design rules [FreePDK] are used for metal and transistor dimensions. Material thermal properties are obtained from [Im 05, Ju 99].

Results

To understand the impact of PDNs, consider a hypothetical 3D IC comprising 4 regions (A-D) with distinct power densities given in Fig. 3a for a 2mm×2mm IC. Figure 3b shows the maximum chip temperature (on Layer 2) for this 2-layer 3D IC (structure corresponds to Fig. 1) using conventional air cooling technique. PDNs for both Layer 1 (bottom layer) and Layer 2 (top layer) are analyzed using the abstracted PDN models in Fig. 2b. The cross section area of ILVs is 3µm×3µm in parallel 3D ICs [ITRS 11] and 300nm×300nm for monolithic 3D ICs [Naito 10]. Assuming the ILV keep-out-zone to be the same as the ILV width [Yu 11], the area taken by one ILV is 9µm×9µm for parallel 3D and 900nm×900nm for monolithic 3D. This implies for 3% chip-level area impact, up to 400 ILVs/mm^2 and 40,000 ILVs/mm^2 can be allowed for parallel and monolithic 3D ICs, respectively (Fig. 3b). The ILV density can be further increased if smaller ILVs are used and larger chip area is allocated for ILVs.

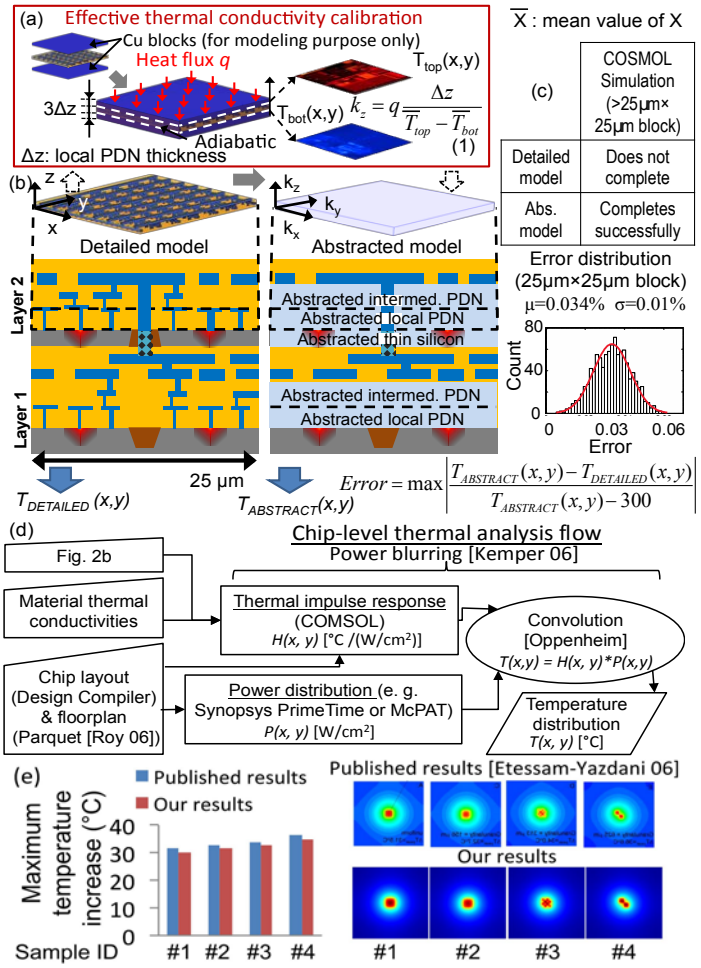


Figure 2. (a) Effective thermal conductivity abstraction. (b) The abstracted thermal model and the original thermal model with detailed structures for one example of the structure in Fig. 1 (the area is 25µm×25µm). The local PDN, intermediate PDN, and thin silicon layer are each abstracted to construct the abstracted thermal model. (c) Comparison made between the abstracted model and the model with detailed structures shows good agreement. (d) Flow to compute chip-level temperature distribution for a given power density distribution using the abstracted models in part (b) and the Power Blurring technique [Kemper 06]. (e) Temperature distributions produced using the method in (d) for various power density distributions in [Etessam-Yazdani 06]. The maximum temperature increase (max(T(x,y))–300K) of all samples (2D ICs) in [Etessam-Yazdani 06] show good agreement with published simulation results with error (defined in (c)) less than 5%.

As shown in Fig. 3b, PDNs are highly effective conduits for

lateral heat conduction for monolithic 3D ICs with 100 nm-thin Layer 2 silicon (T_{Si}): the maximum chip temperature can be reduced by 35 °C. When PDNs are considered by the thermal models, the X-axis represents the density of ILVs in the 2-layer 3D IC that are connected to PDNs to deliver power to the Layer 1 circuits. When PDNs are not taken into consideration (i.e. local, intermediate and global PDNs in both Layer 1 and Layer 2 are not represented in the thermal models), the corresponding ILVs (now electrically inactive) are used solely for heat conduction, transferring heat directly from Layer 2 silicon to Layer 1, and then the heat sink. While sweeping ILV density, the overall chip area is assumed to stay the same. (The error in area estimation can be up to 3% for monolithic 3D ICs and 15% for parallel 3D ICs.) Without PDNs, the monolithic 3D IC (with T_{Si}=100nm) cannot be cooled below the temperature constraint of 85°C [Weste]. Although parallel 3D ICs can be cooled below the temperature constraint of 85°C using ILVs alone (without taking PDNs into consideration), demands on high-density ILVs can improve (i.e., reduce) significantly by 18X (Fig. 3c) when PDNs are carefully incorporated in the chip-level thermal analysis model. This can result in area cost savings.

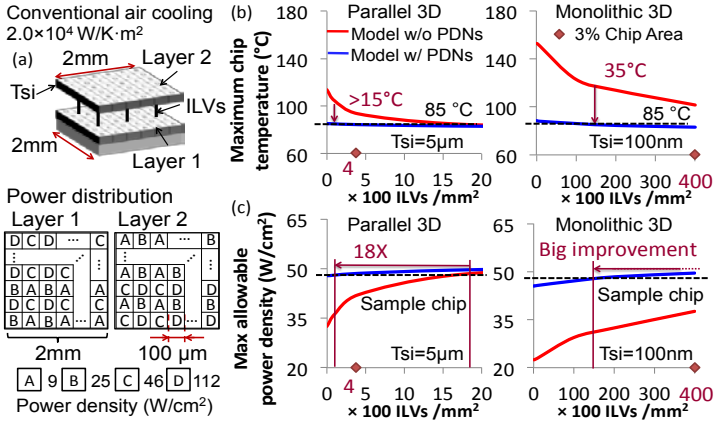


Figure 3. (a) Detailed power density distribution of a simulated 3D IC. (b) Maximum chip temperature vs. ILV density. Two 3D ICs with different T_{Si} are shown and compared. The dashed horizontal line marks the temperature constraint of 85°C [Weste]. (c) Given a thermal constraint, e.g., 85°C, the demands on high-density ILVs can be significantly reduced by 18X for the sample 3D IC shown in (a) using parallel 3D integration. Benefits for monolithic 3D ICs are even higher. This results in large area cost savings. ILD thickness of 13µm remains constant in all cases [Naito 11, Cong 11].

Figure 4 breaks down the maximum chip temperature of the 3D IC in Fig. 3a into subcomponents. For monolithic 3D ICs, the lateral heat resistance through Layer 2 silicon becomes a major challenge. PDNs can significantly improve lateral heat conduction, resulting in reduced maximum chip temperature. For parallel 3D ICs, the lateral heat conduction through silicon substrate lessens the need for PDNs. Effective heat sinks benefit both parallel and monolithic 3D ICs.

For higher power densities (e.g., 250 W/cm^2), PDNs are highly beneficial when combined with appropriate cooling solutions, e.g., external liquid cooling [Tuckerman 81], as shown in Fig. 5. This minimizes demands on highly exotic cooling techniques, e.g., in-chip micro-fluidic channels [Sekar 08]. Similar to Fig. 4, PDNs incorporating high-density (e.g., 15,000 ILVs/mm^2) ILVs enable sufficient cooling of 3D ICs (for temperature constraint of 85 °C).

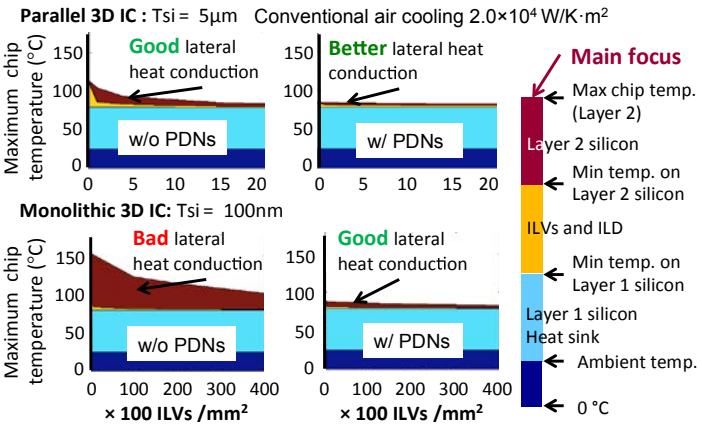


Figure 4. Breakdown of maximum chip temperature for the 3D IC in Fig. 3a with and without PDNs taken into consideration by thermal models. Simulations are run using the method described in Fig. 2 and temperature distributions are recorded in each layer. PDNs improve lateral heat conduction.

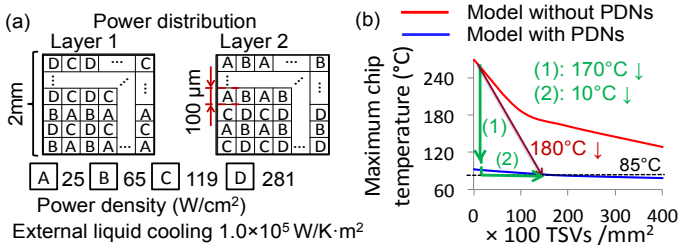


Figure 5. (a) Simulated 3D IC with local power density of over 250 W/cm^2 for a 3D IC with dimensions same as Fig. 4. Instead of conventional air cooling, external liquid cooling is adopted due to the increased power density. (b) Even with external liquid cooling, PDNs can greatly reduce the maximum chip temperature by over 170 °C (arrow 1, vertical). PDNs with high-density ILVs (15,000 ILVs/mm^2) enable effective cooling (arrow 2, horizontal) resulting in 180 °C reduction in maximum chip temperature, thereby satisfying the temperature constraint of 85 °C.

Case Study using OpenSPARC T2 Processor Cores

To demonstrate the application-level effectiveness of PDNs in cooling 3D ICs, we used the OpenSPARC T2 processor core design, which is part of a large open-source industrial multi-core design. We floorplanned various modules of the processor core using Parquet [Roy 06] (Fig. 6), and then obtained power density estimates using McPAT by running an 8-threaded program that solves the Black-Scholes application from the PARSEC benchmark suite [PARSEC].

Next, we constructed two 3D IC scenarios (Fig. 6): 1. An OpenSPARC T2 processor core on top of another OpenSPARC T2 processor core; and, 2. An L2 cache bank on top of an OpenSPARC T2 processor core. We used CACTI [CACTI] to obtain power consumption estimates for the L2 cache. The power density distribution is assumed to be uniform for each module inside the processor core. Note that, McPAT and CACTI estimates can be inaccurate. Figures 7 and 8 clearly demonstrate:

1. For monolithic 3D ICs, even with external liquid cooling, for high power densities (138 W/cm^2 for the EXU unit in Fig. 7), PDNs must be considered for effective heat removal. PDNs with

high-density (10,000 ILVs/mm^2) ILVs reduce the maximum chip temperature to 74°C. When PDNs are not taken into account, the maximum temperature can rise to 122°C even with 10,000 ILVs/mm^2, demonstrating the limitations of cooling techniques using ILVs alone.

2. Without PDNs, a forbidden range of allowable Layer 2 power densities exists for monolithic 3D ICs even with external liquid cooling. Figure 8 shows 3D technology (characterized by T_{Si}) vs. Layer 2 circuit power density, with temperature constraint contours. The Layer 1 consists of an OpenSPARC T2 processor core with power density of 45 W/cm^2 (up to 138 W/cm^2 in the EXU module). For Layer 2, we explore a range of power densities from those corresponding to an OpenSPARC T2 processor core all the way to a single-bank L2 cache of OpenSPARC T2. The reduction in maximum chip temperature is more significant for applications with high power (performance) components on Layer 2 for a two-layer monolithic 3D IC (Fig. 8).

The effectiveness of PDNs for heat removal in monolithic 3D ICs was also confirmed for low-power applications using conventional air cooling.

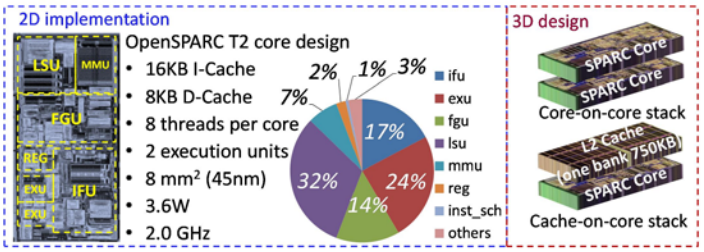


Figure 6. OpenSPARC T2 processor core. Power consumption is computed using McPAT by executing the 8-threaded Black-Scholes application [PARSEC]. Two 3D ICs are formed: 1. Two identical OpenSPARC T2 processor cores are stacked. Modules on Layer 2 overlap exactly with the corresponding modules on Layer 1; 2. One cache bank is stacked on top of one OpenSPARC T2 processor core.

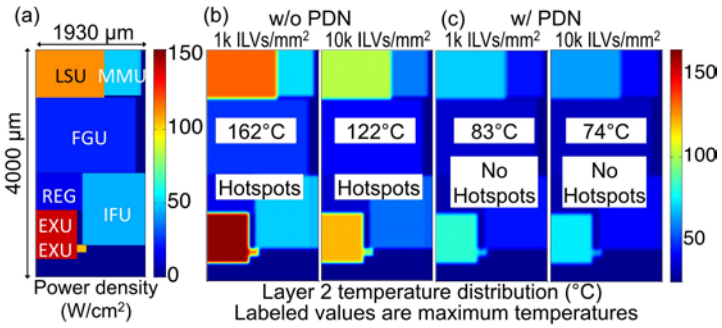
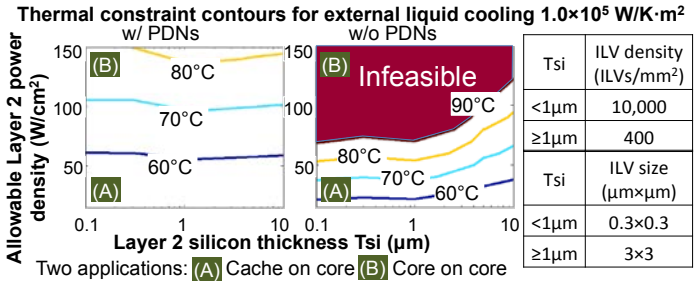


Figure 7. Temperature distributions for a 3D IC (T_{Si} = 100 nm) with one OpenSPARC T2 processor core on top of another core with external liquid cooling. (a) Power density distribution of the core for the Black-Scholes application. (b) Temperature distributions when PDNs are not considered. Thermal hotspots are visible. (c) Temperature distributions when PDNs are considered. No hotspots are observed and over 75 °C temperature reduction can be achieved. Maximum chip temperature is labeled for each situation.



Tsi	ILV density (ILVs/mm²)
<1µm	10,000
≥1µm	400

Tsi	ILV size (µm×µm)
<1µm	0.3×0.3
≥1µm	3×3

Figure 8. 3D technology (characterized by T_{Si}) vs. Layer 2 power density with temperature constraint contours to demonstrate major benefits of PDNs. Two applications are considered: (A) One L2 cache bank (30 W/cm^2) on Layer 2 and OpenSPARC T2 processor core in Layer 1. (B) One OpenSPARC T2 processor core in Layer 2 and another OpenSPARC T2 processor core on Layer 1.

Conclusion

Our comprehensive analysis, over a range of 3D integration methods and application power density characteristics, quantifies major benefits of PDNs on the temperature distribution of 3D ICs. For example, PDNs can reduce the maximum steady-state temperature by over 35 °C for a 2-layer monolithic 3D IC. Our OpenSPARC T2 case study also demonstrates that the cooling benefits of PDNs are essential to achieve monolithic 3D integration. Our analysis framework can be adopted for exploring technology-circuit-application interactions for a wide variety of 3D technologies, cooling options, PDN designs, or even software -level task scheduling approaches. Of course, it is essential to experimentally validate the simulation results presented in this paper.

Acknowledgement

This work was partially supported by FCRP C2S2 and NSF. We thank Prof. H.-S.P. Wong, and M. Shavezipur of Stanford, E. Cheng, J.-Y. Choi (Samsung), Y. Li, D. Lin, and F. Toufexis of the Stanford Robust Systems Group, Z. Or-Bach of Monolithic 3D Inc., and TM Mak of Intel for helpful suggestions.

References

[Banerjee 01] Banerjee, K., *et al., Proc. IEEE,* Vol. 89, pp. 602, 2001. [Batude 11] Batude, P., *et al., IEDM,* pp. 151, 2011. [CACTI] http://hpl.hp.com/research/cacti [COMSOL] http://comsol.com [Cong 07] Cong, J., *et al., ASP-DAC*, pp. 780, 2007. [Cong 11] Cong, J., *et al., DAC,* pp. 670, 2011. [Etessam-Yazdani 06] Etessam-Yazdani, K., *et al., Intersociety Conference on Phenomena in Electronics Systems*, pp. 666, 2006. [FLO] http://flomerics.com [FreePDK] http://eda.ncsu.edu/wiki/FreePDK [Im 05] Im, S., *et al., Trans. Electron Dev.*, Vol. 52, pp. 2710, 2005. [Incropera] Incropera, F. P., *et al., Introduction to Heat Transfer, 5th edition, John Wiley & Sons*, pp. 96, 2007. [ITRS 11] http://itrs.net [Ju 99] Ju, Y. S., *et al., Applied Physics Lett.*, pp. 3005, 1999. [Kemper 06] Kemper, T., *et al., THERMINIC*, pp. 133, 2006. [Kleiner 95] Kleiner, M.B., *et al., IEDM,* pp. 487, 1995. [Lau 09] Lau, J. H., *et al., Electronic Components & Technology Conf.,* pp. 635, 2009. [McPAT] http://hpl.hp.com/research/mcpat [Naito 10] Naito, T., *et al., Symp. VLSI Tech.,* pp. 219, 2010. [OpenSPARC] http://opensparc.net [Oppenheim] Oppenheim, A., *et al., Signals and Systems, Prentice Hall,* pp. 23, 1996. [PARSEC] http://parsec.cs.princeton.edu [Roy 06] Roy, J. A., *et al., IEEE Trans. CAD*, Vol. 25, pp.1313, 2006. [Sekar 08] Sekar, D., *et al., IITC*, pp. 13, 2008. [Topol 05] Topol, A. W., *et al., IEDM*, pp. 352, 2005. [Tuckerman 81] Tuckerman, D. B., *et al., IEEE Electron Dev. Lett.* Vol. 2, pp.126, 1981. [Van Olmen 08] Van Olmen, J., *et al., IEDM,* pp. 1, 2008. [Weste] Weste, N., *et al., CMOS VLSI Design, 3rd edition, Pearson Education,* pp. 233, 2005. [Wei 09] Wei, H., *et al., IEDM,* pp. 577, 2009. [Wong 07] Wong, S., *et al., VLSI-TSA*, pp. 1, 2007. [Yang 07] Yang, Y., *et al., IEEE Trans. CAD*, Vol. 26, pp. 86, 2007. [Yu 09] Yu, H., *et al., ACM Trans. DAES*, Vol. 4, pp. 41, 2009. [Yu 11] Yu, C.L., *et al., Symp. VLSI Tech.,* pp. 138, 2011. [Zhang 06] Zhang, T., *et al., ASP-DAC*, pp. 309, 2006.