MANCHESTER 1824
The University of Manchester

The University of Manchester Research

An Overview of Thermal Challenges and Opportunities for Monolithic 3D ICs

DOI:
10.1145/3299874.3319485

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Shukla, P., Coskun, A., Pavlidis, V., & Salman, E. (2019). An Overview of Thermal Challenges and Opportunities for Monolithic 3D ICs. In *Great Lakes Symposium on VLSI* Association for Computing Machinery. https://doi.org/10.1145/3299874.3319485

Published in:
Great Lakes Symposium on VLSI

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:28. Aug. 2022

An Overview of Thermal Challenges and Opportunities for Monolithic 3D ICs

Prachi Shukla
Boston University
Boston, USA
prachis@bu.edu

Ayse K. Coskun
Boston Univeristy
Boston, USA
acoskun@bu.edu

Vasilis F. Pavlidis
University of Manchester
Manchester, UK
pavlidis@cs.man.ac.uk

Emre Salman
Stony Brook University (SUNY)
Stony Brook, USA
emre.salman@stonybrook.edu

ABSTRACT

Monolithic 3D (Mono3D) is a three-dimensional integration technology that can overcome some of the fundamental limitations faced by traditional, two-dimensional scaling. This paper analyzes the unique thermal characteristics of Mono3D ICs by simulating a two-tier flip-chip Mono3D IC and highlights the primary differences in comparison to a similarly-sized flip-chip TSV-based 3D IC. Specifically, we perform architectural-level thermal simulations for both technologies and demonstrate that vertical thermal coupling is stronger in Mono3D ICs, leading to lower upper tier temperatures. We also investigate the significance of lateral versus vertical flow of heat in Mono3D ICs. We simulate different hot spot scenarios in a two-tier Mono3D IC and show that although the lateral heat flow is limited as compared to TSV-based 3D ICs, ignoring this mechanism can cause nonnegligible error (~ 4 °C) in temperature estimation, particularly for layers farther from the heat sink. In addition, we show that with increasing interconnect utilization (due to the contribution of Joule heating to overall temperature), the on-chip temperatures and the significance of lateral heat flow within the two-tier Mono3D IC also increase. Finally, we discuss potential opportunities in Mono3D ICs to enhance their thermal integrity.

KEYWORDS

Monolithic 3D, thermal integrity, vertical thermal coupling, lateral heat spreading

ACM Reference Format:
Prachi Shukla, Ayse K. Coskun, Vasilis F. Pavlidis, and Emre Salman. 2019. An Overview of Thermal Challenges and Opportunities for Monolithic 3D ICs. In *Great Lakes Symposium on VLSI 2019 (GLSVLSI '19), May 9–11, 2019, Tysons Corner, VA, USA*, 6 pages. https://doi.org/10.1145/3299874.3319485

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GLSVLSI '19, May 9–11, 2019, Tysons Corner, VA, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6252-8/19/05…$15.00
https://doi.org/10.1145/3299874.3319485

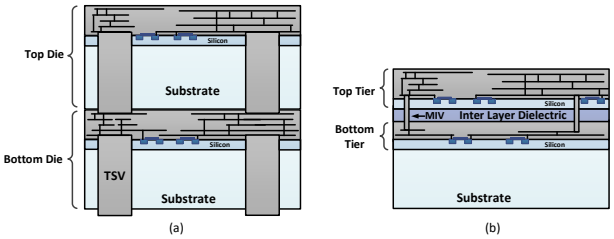

Figure 1: Two 3D integration technologies: (a) TSV-based 3D integration, and (b) Mono3D integration with MIVs

1 INTRODUCTION

Two-dimensional scaling of transistors is reaching its limits due to factors such as (i) lithographic challenges, (ii) increasing power densities and related thermal challenges, (iii) on-chip communication overhead as metal wires typically scale less than scaling of transistors, and (iv) increasing manufacturing cost [8]. Three-dimensional integrated circuits (3D ICs) are highly promising as they can mitigate the interconnect related challenges and provide a significant boost in performance and power, while also reducing the chip footprint as compared to two-dimensional ICs.

Two major types of 3D integration are sequential Mono3D process and through silicon vias (TSV)-based 3D process, as illustrated in Fig. 1 [15]. Mono3D is an emerging 3D integration technology where the tiers are fabricated sequentially and interconnected using nanoscale monolithic inter-tier vias (MIVs). Sequential integration in Mono3D results in thin tiers (~ 500 nm) that are separated by an inter-layer dielectric (ILD) (SiO_2) with a thickness in the range of tens of nanometers [18].

Alternatively, in TSV-based 3D ICs, the tiers originate from separately fabricated wafers. As such, each tier is substantially thicker compared to Mono3D ICs. These tiers are interconnected using micrometer scale TSVs and are integrated with a thick insulating bonding layer [18]. This significant difference between the size of a TSV (several micrometers) and an MIV (nanometers) not only affects the integration density, but also the chip-level power and performance characteristics. For example, a practical TSV with several micrometers of diameter exhibits a capacitance in the range of tens of femtofarads, which is equivalent to approximately 100 gates (with fanout of two) in a relatively old 45 nm technology

node [20, 25]. In the 7 nm technology node, a TSV is equivalent to approximately several thousand gates, thereby consuming significant dynamic power and causing *RC* delay. Furthermore, the keep-out zone around a TSV (i.e., region that should be free of transistors due to undesirable TSV effects such as stress and threshold voltage variation) exacerbates these issues.

Thus, Mono3D ICs with MIVs not only reduce the overall footprint and increase integration density, but also significantly enhance the power and performance characteristics. For example, a two-tier Mono3D IC has been shown to potentially achieve up to a 50% reduction in die size and ~50% reduction in power compared to its 2D counterpart at the same feature size, providing advantages equivalent to a generation of dimensional scaling [1, 12, 26, 28]. The key advantages of Mono3D ICs can be summarized as: (i) smaller chip footprint, (ii) significant performance enhancements and power savings due to the increased bandwidth and reduced wire length, and (iii) reduced on-chip communication overhead as a result of shorter interconnects.

A primary and well-known limitation facing 3D ICs is effective heat dissipation, particularly from the upper tiers (see Figure 2a for upper/bottom tiers). In Mono3D ICs, additional thermal issues can arise due to very dense device integration, routing congestion (which may contribute to Joule heating), and strong thermal coupling among tiers, which exacerbates the effects of thermal hot spots. These factors differentiate Mono3D technology from TSV-based 3D ICs. Furthermore, Mono3D process requires low temperature fabrication steps to protect the devices within the bottom tier. To withstand higher temperatures on the bottom tier, tungsten-based interconnects have been proposed [2]. However, such methods also increase the vertical thermal resistance and lead to performance degradation and higher on-chip temperatures [14, 19].

Overall, Mono3D ICs face unique thermal issues during both fabrication and design stages. This paper discusses these thermal challenges and also the opportunities to mitigate them. The main contributions of this work are as follows:

(1) We model and compare Mono3D and TSV-based two-tier 3D ICs in terms of inter-tier thermal coupling. We further investigate the effects of thermal coupling on circuit temperatures for these technologies.
(2) We present an architectural-level thermal analysis on the lateral and vertical heat flow in a two-tier Mono3D IC. We show that lateral heat flow plays an important role with increasing distance from the heat sink and in scenarios with high-density hot spots on the upper tier.
(3) We model different utilization levels of the interconnects to represent various application behaviors (slightly, moderately, and highly parallel), and calculate the effect that such application scenarios have on circuit temperature and lateral versus vertical heat flow.

The rest of the paper is organized as follows. Section 2 briefly discusses the differences between Mono3D and TSV-based 3D integration technologies, followed by a summary of recent work on thermal management in Mono3D ICs. In Section 3, we investigate the unique thermal issues faced by Mono3D ICs. We then discuss a few opportunities to enhance thermal integrity of Mono3D ICs in Section 4. Finally, we conclude in Section 5.

2 BACKGROUND

In the following subsections, we discuss how Mono3D differs from TSV-based 3D ICs. We also describe the recent work in Mono3D ICs with a focus on thermal issues.

2.1 Mono3D versus TSV-based 3D Integration

Mono3D and TSV-based 3D integration technologies differ in many aspects. In TSV-based 3D technology, the device layers are thicker (tens of μm) because each device layer (tier) has its own substrate. Although the thicker substrates improve the lateral heat spreading, they also result in high vertical thermal resistance. In addition, the TSVs exhibit large physical dimensions (5-100 μm) and facilitate the flow of heat vertically. However, the presence of a thick bonding layer (~ 2.5 μm) between the device layers impedes the flow of heat both vertically and laterally from the upper tier to the bottom tier, thereby, degrading the overall conductivity of the 3D IC [18].

On the other hand, the device layers in Mono3D ICs are much thinner (100-500 *nm*) due to which the lateral heat spreading capability reduces, thereby resulting in higher-density hot spots. In addition, the size of the MIVs is similar to conventional metal vias (*nm*) and therefore facilitate significantly denser device integration in Mono3D ICs. This dense integration, however, may cause routing congestion and exacerbated Joule heating [3, 27].

2.2 Thermal Management in Mono3D ICs

Recent works in Mono3D ICs have focused on alleviating the thermal challenges stated in the previous section. Samal *et al.* show that optimizing PDN design styles can increase thermal conductivity in Mono3D ICs and further lead to reduction in on-chip temperature by up to 5% [17]. Wei *et al.* further show that ignoring PDNs in thermal models can result in overestimation of steady-state temperature in Mono3D ICs [24]. Iqbal *et al.* propose the use of nano pillars for extracting heat from selected hot spot regions without interfering with the vias [9]. Such techniques improve thermal conductivity, thus, coupling, which allows heat to flow faster. In another work, Samal *et al.* show that the lateral flow of heat is negligible in Mono3D ICs and then build a non-linear regression model for temperature estimation of the tiers [18]. However, their thermal analysis lacks a wide coverage of different power profiles.

Tier partitioning and floorplanning are also important design optimization decisions for Mono3D ICs with implications on temperature. Kim *et al.* propose four-tier partitioning methods, in which a 2D placement policy is applied to an initial 2D design to partition cells recursively and build a four tier Mono3D design with MIVs [10]. Another 3D placement technique minimizes wire length while reducing routing congestion [13]. Pletea *et al.* provide a logic and memory partitioning scheme for Mono3D ICs with large memories using 2D EDA tools [16]. Another recent approach uses simulated annealing to iteratively partition the cells into different tiers while minimizing wire length [6]. Guler *et al.* utilize simulated annealing to floorplan circuits with Mono3D technology at diverse integration granularities (block, gate, and transistor levels) [7]. Their floorplanner creates a hybrid Mono3D design that minimizes the area, wire length, and power consumption.

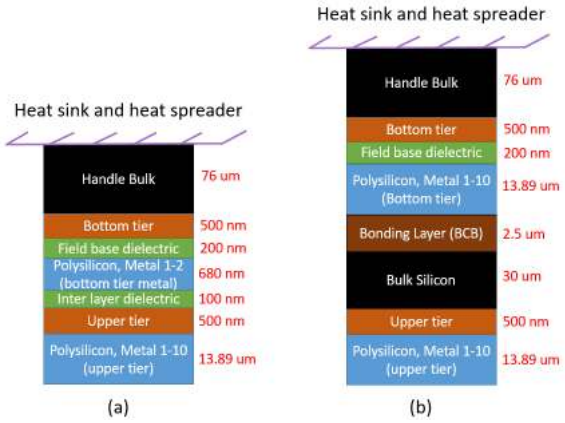

Figure 2: Cross-section of a flip-chip stack for (a) two-tier Mono3D IC, (b) two-tier TSV-based 3D IC. These structures are input into the thermal simulator for thermal analysis.

3 INSIGHTS ON MONO3D THERMAL ANALYSIS AND THERMAL INTEGRITY

In this section, we present architectural-level insights on thermal challenges faced by the Mono3D integration technology. We begin by comparing a two-tier Mono3D IC to a TSV-based 3D IC to investigate and quantify the distinct thermal characteristics of these technologies. We then analyze further the lateral and vertical heat flow in Mono3D ICs under different hot spot scenarios. In addition, we investigate the effect of different interconnect utilization levels on lateral versus vertical heat flow and on-chip temperatures. All of the analyses in this paper are performed using the architectural-level HotSpot-6.0 thermal simulator [22].

3.1 Inter-Tier Thermal Coupling in Mono3D versus TSV-based 3D

In this section, we study the differences in the thermal characteristics of TSV-based and Mono3D ICs. For this study, we model a flip-chip two-tier Mono3D and TSV-based 3D IC in HotSpot-6.0, with a heat sink mounted on top of the bottom tier (see Figure 2). We model all of the metal layers of both tiers in the Mono3D IC since each layer has comparable thickness. For a fair comparison between the two ICs, we also model the metal layers in the TSV-based 3D IC. In addition, HotSpot-6.0 supports features for heterogeneous modeling within a layer in a chip stack [11]. We use this functionality to model the TSVs along with silicon in the upper tier. However, we do not model the MIVs in the Mono3D IC since the size of the MIVs in the nanometer range is not practical for architecture-level analysis. Furthermore, ignoring MIVs underestimates the vertical heat flow in Mono3D technology, ensuring that our model presents a pessimistic analysis. Investigating the impact of MIV on temperature characteristics remains as future work that focuses on layout-level temperature estimation.

To study the thermal behavior of the two ICs, we model a 10 mm × 10 mm chip. Further, we model the TSVs as 4 TSV arrays, each of dimension 5 mm x 200 μm × 33 μm (length × width × height), passing through the upper tier, bulk silicon, and bonding layer. We

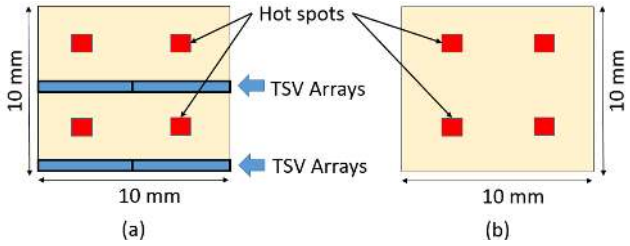

Figure 3: Top view of the upper tiers with four hot spots in (a) TSV-based 3D IC, and (b) Mono3D IC.

then place a hot spot at the center of the bottom tier and 4 hot spots in the upper tier. All of the hot spots are of size 400 μm × 400 μm. The bottom and the upper tiers consume a background power density of 25 and 30 W/cm^2, respectively. In addition, we assign 5% of the total power consumed by the tiers to the interconnects. Of that 5%, we assign 20% to the TSVs. The power densities are selected to ensure that the on-chip temperatures do not exceed 110 °C [23]. Since hot spot power densities in future processors are expected to reach up to 2 kW/cm^2, we assign each hot spot a power density of 750 W/cm^2 (to keep maximum on-chip temperature below 110 °C) [21]. The steady-state thermal maps obtained from the HotSpot thermal simulations are shown in Figure 4. The figures also illustrate the lateral heat spreading between the hot spots and cooler regions of the chip.

As shown in Figures 4a and 4b, hot spots in the upper tier reach very high temperatures (~ 109 °C) in the TSV-based 3D IC. The same locations on the bottom tier are approximately 38 °C cooler, thus, indicating weak vertical thermal coupling between the tiers. There are two main reasons for the observed weak vertical thermal coupling: (i) distance of the upper tier from the heat sink, and (ii) presence of an insulating bonding layer between the tiers. Note that the hot spot at the center of the bottom tier has a negligible effect on the center of the upper tier. This is because the vertical heat flow path from the bottom tier to the heat sink exhibits a lower thermal resistance due to the shorter distance and absence of an insulating layer.

On the other hand, Figures 4c and 4d illustrate that hot spots on the upper tier of Mono3D IC lead to similar hot spots on the bottom tier, reaching a temperature of ~ 77 °C. Overall, the temperature distribution among the two tiers varies up to only ~ 4 °C, while the temperature difference in TSV-based 3D IC reaches up to 38 °C. Note that the hot spots in Mono3D IC are more localized than those in TSV-based IC, thus, demonstrating that Mono3D technology exhibits lower lateral thermal coupling within a tier. Despite the reduced lateral thermal coupling in Mono3D IC, the maximum on-chip temperature is lower due to the stronger vertical inter-tier coupling as compared to TSV-based 3D IC. This, as a result, indicates enhanced overall thermal coupling in Mono3D ICs.

3.2 Significance of Lateral versus Vertical Heat Flow in Mono3D ICs

In this section, we explore the significance of lateral versus vertical heat flow in Mono3D ICs. The chip stack used for this analysis is the same as shown in Figure 2a. We use the HotSpot simulator to

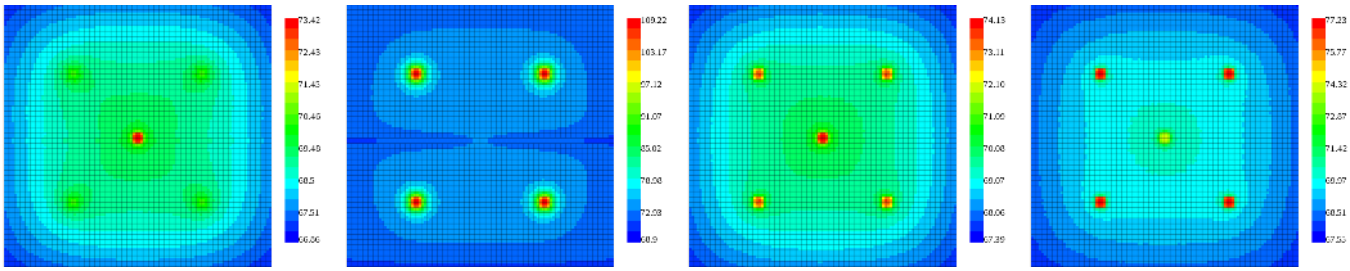

Figure 4: Thermal maps of (a) bottom tier and (b) upper tier of the TSV-based 3D IC, respectively, and (c) bottom tier and (d) upper tier of the Mono3D IC, respectively.

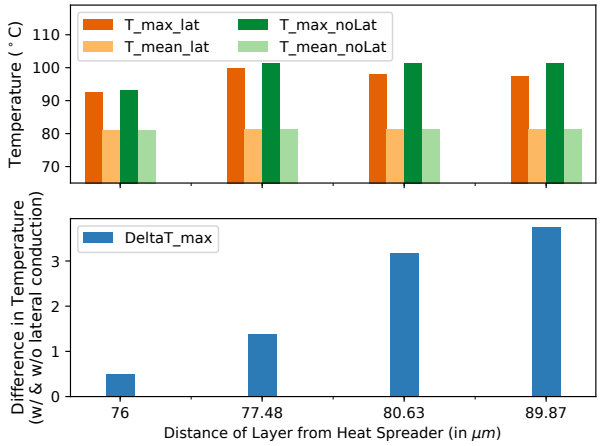

Figure 5: Effect on chip temperature and lateral heat flow modeling when a single hot spot is at the center of each tier.

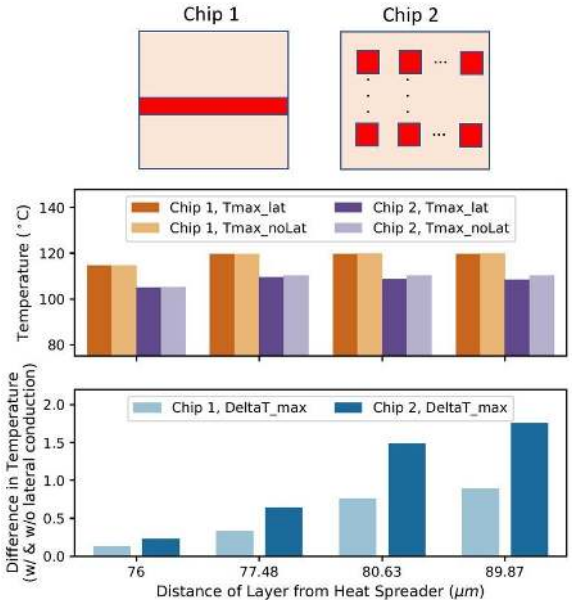

Figure 6: Top view of the upper tier with 10% hot spot area. Chip 1 contains lumped hot spots, while chip 2 contains scattered hot spots. The graphs show maximum on-chip temperatures and significance of lateral heat flow modeling in chips 1 and 2.

evaluate the induced errors in on-chip temperatures when lateral modeling is turned off for the nanoscale layers. We simulate two scenarios for this study. In the first scenario, we simulate a 10 mm × 10 mm chip with a single hot spot at the center of each tier. Each hot spot is of size 200 μm × 200 μm with a power density of 1,600 W/cm^2. We set the background power density of the upper and bottom tiers to 50 and 40 W/cm^2, respectively. We report the maximum and mean temperatures across the two cases of modeling, i.e., with lateral modeling turned on and turned off. As shown in Figure 5, the maximum difference in on-chip temperature can reach ~ 4 °C for layers that are at ~ 90 μm distance from the heat sink. This difference is highest at the hot spot locations and the area around the hot spots. Note that the bottom and the upper tiers are at a distance of 76 and 77.48 μm from the heat spreader, respectively.

In the second scenario, we simulate two 10 mm × 10 mm chips (see Figure 6) with hot spots consuming ~ 10% of the total chip area. In both chips, we place hot spots in the upper tier to produce high on-chip temperatures. The background power density of the upper and bottom tiers is set to 25 and 20 W/cm^2, respectively, while the hot spot power density is set to 1,000 W/cm^2. The size of each hot spot in the chip with scattered hot spots is 200 μm × 200 μm. The first chip with lumped hot spots results in higher on-chip temperatures, while the second chip with scattered hot spots (i.e., with high power density spread across the chip) has lower temperature owing to higher lateral heat flow from hotter to cooler regions. When lateral modeling is turned off, the second chip is expected to show higher inaccuracy in on-chip temperature, as illustrated in Figure 6. According to this figure, for the layers away from the heat sink, the inaccuracy in on-chip temperatures increases to several degrees. Furthermore, the inaccuracy is observed to be the highest (~ 2 °C) in the areas next to the hot spots.

3.3 Thermal Impact of Interconnect Utilization in Mono3D ICs

In this section, we investigate the thermal impact of various interconnect utilization levels on lateral and vertical heat flow as well as on-chip temperatures. The utilization level of the interconnects is representative of the extent of parallelism within an application.

To perform this analysis, we first fix the total power budget of the Mono3D chip stack. We then vary the ratios of power distribution between the active tiers and the interconnects. We then run thermal simulations using HotSpot-6.0 to analyze how these utilization levels affect the thermal profile and lateral versus vertical heat flow in a Mono3D IC.

We set the power profile of the Mono3D IC to ensure that the maximum on-chip temperature does not exceed ~ 110 °C. We set the total power budget to 90 Watts and vary the ratio of power assigned to the interconnects versus the active tiers. The ratio represents three kinds of application behaviour: light (30:70), moderate (50:50) and extensive (70:30) parallelism. We then distribute the interconnect power linearly between metal layers 1-8, with metal 8 consuming the highest power. We test two hot spot scenarios: (i) high density hot spots at the center of both the tiers, and (ii) multiple high density hot spots in the upper tier alone (one at the center and four at the center of each quadrant of the upper tier). Each hot spot, in both the scenarios, is of size 400 μm × 400 μm.

In the scenario with a single hot spot on each tier, we set the hot spot power density to 1,300, 1,500, and 1,700 W/cm^2. For the multiple hot spot scenario, we set each hot spot power density to 1,600, 1,800 and 2,000 W/cm^2. We then run thermal simulations with HotSpot to obtain the steady-state thermal profile of the chip. Figures 7 and 8 show the on-chip temperature and the significance of modeling lateral heat flow with increasing distance from the heat sink, as the metal utilization increases from 30% to 70%. In both scenarios, we find that as the interconnect power increases, the difference in temperatures between w/ and w/o lateral heat flow modeling slightly increases. In addition, with increasing metal utilization, the maximum on-chip temperature also increases slightly (~ 1.2 °C). This behavior is because metal layers of the upper tier comprise global interconnect layers, which exhibit higher vertical thermal resistance. As a result, ignoring lateral heat flow induces a greater error in the reported temperatures of the Mono3D IC. Note that this error increases with increasing distance from the heat sink, showing that lateral heat flow modeling is relatively more important for layers that are farther away from the heat sink. In addition, due to strong vertical thermal coupling, the increase in maximum on-chip temperature is less. Furthermore, with lateral modeling enabled, it takes ~ 18 minutes to run a steady-state HotSpot simulation for a two-tier Mono3D IC with 100 × 100 grids (per layer), while it takes ~ 13 minutes to simulate the same Mono3D IC with lateral modeling disabled.

4 OPPORTUNITIES TO ENHANCE THERMAL INTEGRITY IN MONO3D ICS

Some of the unique Mono3D technology characteristics can be leveraged to mitigate thermal issues. For example, the bottom tier in existing Mono3D processes has limited number of metal layers (typically in the range of 2 to 3) since additional metal layers introduce significant fabrication challenges for top tier devices [4, 5]. Thus, a majority of the hot spots are expected to be within the top tier. Furthermore, logic gates within the bottom tier are likely to rely on the metal layers located within the top tier, particularly for global interconnections. Joule heating will therefore be more dominant within the top tier. These characteristics can be leveraged

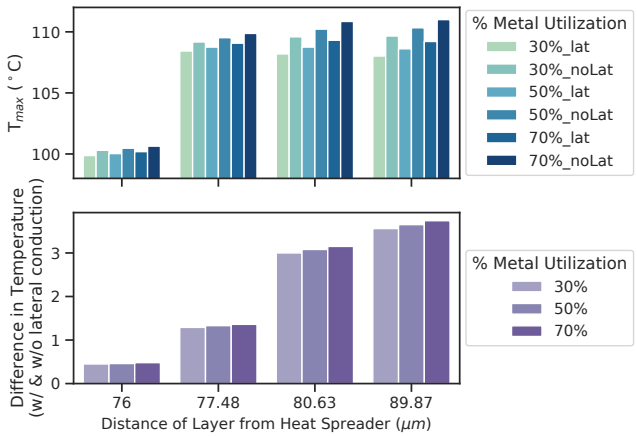

Figure 7: Single hot spot scenario (hot spot power density = 1,700 W/cm^2): Significance of lateral heat flow modeling with increasing metal utilization.

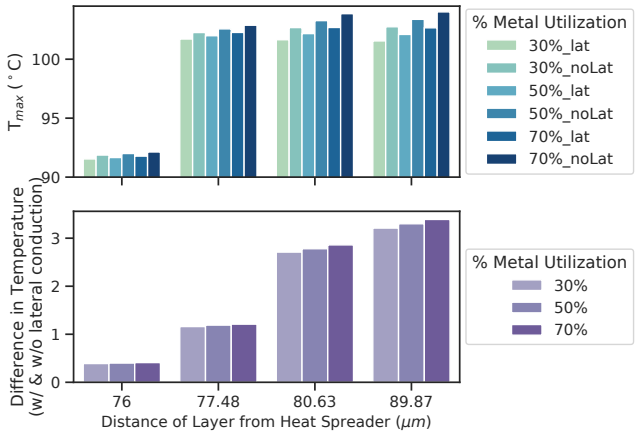

Figure 8: Multiple hot spot scenario (hot spot power density = 2,000 W/cm^2): Significance of lateral heat flow modeling with increasing metal utilization.

to reduce computational complexity of temperature analysis by nonuniform grid sizes depending upon the tier.

Furthermore, unlike TSV-based 3D ICs, Mono3D technology permits circuit partition at different granularities (transistor-level, gate-level, and block-level) with important implications to on-chip temperature. For example, a transistor-level partitioning can achieve the highest integration density, but it can also cause hot spots that span multiple tiers since an aggressor block that generates significant heat needs to be partitioned into two tiers. This flexibility in partitioning can be utilized to develop a thermally-optimum distribution of transistor-, gate-, and block-level partitioning for a Mono3D chip, given a set of system-level parameters. This approach can play a key role to mitigate the effects of strong inter-tier thermal coupling while also ensuring interconnect latency constraints. A typical example is to partition a processing unit with memory where the two elements should have spatial proximity to

ensure fast read/write cycles, but also sufficient thermal isolation to prevent higher leakage current due to elevated temperatures of the memory.

As demonstrated in the previous section, in Mono3D ICs, the vertical heat flow is relatively stronger as compared to lateral flow. Thus, there is room for improvement by optimizing the on-chip interconnects to ensure sufficient lateral heat flow. This characteristic is highly distinct from TSV-based 3D ICs where the number of TSVs has considerable impact on on-chip temperature since vertical heat flow is highly restricted. Alternatively, in Mono3D ICs, it is expected that the on-chip global interconnects (particularly the power and ground networks that do not introduce switching activities) play an important role to enhance lateral heat dissipation. This characteristic is supported by simulations in the previous section where the difference in temperature between w/ and w/o lateral heat flow modeling increases with increasing interconnect power.

The strong vertical thermal coupling and limited lateral coupling also motivate thermally-aware design-time and runtime optimizations specific to Mono3D. For example, a floorplanning technique can co-optimize both tiers together to avoid placing high-density blocks in the same vertical line. Similarly, runtime techniques (such as task allocation or operating mode of the cores) can also be designed to avoid very high temperatures by monitoring the cores on both tiers together. In addition, we demonstrate that the lateral heat flow between the hot spots and the cooler region in Mono3D is confined. This behavior can be leveraged in floorplanning because two neighbouring hot spots may not have a significant impact on one another.

5 CONCLUSIONS

We model a flip-chip two-tier Mono3D IC and a TSV-based 3D IC to compare the thermal characteristics of the two technologies. These models include the metal layers, bulk silicon, active tiers, bonding layer and ILD. It is observed that Mono3D technology exhibits strong vertical thermal coupling with relatively homogeneous temperature distribution across both the bottom and the upper tiers. This characteristic is in contrast with the TSV-based 3D IC, where the upper tier reaches very high temperatures due to the presence of the bonding layer and a longer resistive path to the heat sink. We also demonstrate that despite the limited lateral flow of heat in Mono3D (due to thin tiers), ignoring lateral heat propagation for the layers farther from the heat sink can lead to inaccuracies in estimating the on-chip temperatures (up to $\sim 4\ ^{\circ}C$). In addition, as the interconnect utilization increases (i.e., more Joule heating due to metal layers), both the maximum on-chip temperature and the significance of modeling lateral heat flow slightly increase. Finally, we discuss opportunities to enhance thermal integrity in Mono3D ICs by leveraging some of its unique characteristics.

ACKNOWLEDGMENTS

This project has been partially funded by the NSF CRI (CI-NEW) grant #1730316.

REFERENCES

[1] Why Monolithic 3D? http://www.monolithic3d.com/why-monolithic-3d.html
[2] P. Batude et al. 2008. Enabling 3D Monolithic Integration. *ECS Transactions* 16, 8 (2008), 47–54.
[3] P. Batude et al. 2009. GeOI and SOI 3D monolithic cell integrations for high density applications. In *IEEE Symposium on VLSI Technology*. 166–167.
[4] P. Batude et al. 2014. 3D sequential integration opportunities and technology optimization. In *IEEE Proc. of International Interconnect Technology Conference*. 373–376.
[5] P. Batude et al. 2015. 3DVLSI with CoolCube process: An alternative path to scaling. In *IEEE Symposium on VLSI Technology*. T48–T49.
[6] G. Berhault et al. 2016. 3DIP: An iterative partitioning tool for monolithic 3D IC. In *IEEE Proc. of International 3D Systems Integration Conference (3DIC)*. 1–5.
[7] A. Guler and N. K. Jha. 2018. Hybrid monolithic 3D IC floorplanner. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 26, 10 (Oct. 2018), 1868–1880.
[8] R. Ho, K. W. Mai, and M. A. Horowitz. 2001. The future of wires. *Proc. of the IEEE* 89, 4 (2001), 490–504.
[9] M. A. Iqbal and M. Rahman. 2017. New thermal management approach for transistor-level 3-D integtration. In *IEEE Proc. of SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. 1–3.
[10] K. M. Kim et al. 2016. Four-tier monolithic 3D ICs: Tier partitioning methodology and power benefit study. In *ACM Proc. of International Symposium on Low Power Electronics and Design (ISLPED)*. 70–75.
[11] J. Meng, K. Kawakami, and A. K. Coskun. 2012. Optimizing energy efficiency of 3D multicore systems with stacked DRAM under power and thermal constraints. In *ACM Proc. of Design Automation Conference (DAC)*. 648–655.
[12] Z. Or-Bach. 2013. The monolithic 3D advantage: Monolithic 3D is far more than just an alternative to 0.7 x scaling. In *IEEE Proc. of the International 3D Systems Integration Conference (3DIC)*. 1–7.
[13] S. Panth, K. Samadi, Y. Du, and S. K. Lim. 2015. Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)* 34, 4 (Apr. 2015), 540–553.
[14] G. Pares et al. 2011. Through silicon via technology using tungsten metallization. In *IEEE Proc. of International Conference on IC Design & Technology*. 1–4.
[15] V. F. Pavlidis, I. Savidis, and E. G. Friedman. 2017. *Three-dimensional integrated circuit design*. Newnes.
[16] I. Pletea, Z. Wurman, Z. Or-Bach, and V. Sontea. 2015. Monolithic 3D layout using 2D EDA for embedded memory-rich designs. In *IEEE Proc. of SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. 1–2.
[17] S. K. Samal et al. 2014. Full chip impact study of power delivery network designs in monolithic 3D ICs. In *IEEE/ACM Proc. of International Conference on Computer-Aided Design (ICCAD)*. 565–572.
[18] S. K. Samal et al. 2016. Adaptive regression-based thermal modeling and optimization for monolithic 3D ICs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35, 10 (Oct. 2016), 1707–1720.
[19] S. K. Samal et al. 2016. Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs. In *ACM Proc. of International Conference on Computer-Aided Design (ICCAD)*. 129.
[20] S. M. Satheesh and E. Salman. 2012. Power distribution in TSV-based 3D processor-memory stacks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2, 4 (Dec. 2012), 692–703.
[21] M. Schultz et al. 2016. Embedded two-phase cooling of large three-dimensional compatible chips with radial channels. *Journal of Electronic Packaging* 138, 2 (Apr. 2016), 021005.
[22] K. Skadron et al. 2003. Temperature-aware microarchitecture. In *IEEE Proc. of International Symposium on Computer Architecture (ISCA)*. 2–13.
[23] G. J. Snyder et al. 2006. Hot spot cooling using embedded thermoelectric coolers. In *IEEE Semiconductor Thermal Measurement And Management Symposium*. 135–143.
[24] H. Wei et al. 2012. Cooling three-dimensional integrated circuits using power delivery networks. In *IEEE International Electron Devices Meeting*. 14–2.
[25] Q. Xie et al. 2015. Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries. *IEEE Transactions on Circuits and Systems II: Express Briefs* 62, 8 (Aug. 2015), 761–765.
[26] C. Yan, S. Kontak, H. Wang, and E. Salman. 2017. Open source cell library Mono3D to develop large-scale monolithic 3D integrated circuits. In *IEEE Proc. of International Symposium on Circuits and Systems*. 1–4.
[27] C. Yan and E. Salman. 2017. Routing congestion aware cell library development for monolithic 3D ICs. In *International Conference on Rebooting Computing*. 1–4.
[28] C. Yan and E. Salman. 2018. Mono3D: Open source cell library for monolithic 3D integrated circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers* 65, 3 (Mar. 2018), 1075–1085.