2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS) | 979-8-3503-8717-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/MWSCAS60917.2024.10658875

Co-design for Heterogeneous Integration: High Level Decisions to the Rescue

Daniel Xing, Abir Akib, Yuntao Liu, and Ankur Srivastava
Institute for Systems Research
University of Maryland, College Park
College Park, Maryland, USA
{dxing97,aakib,ytliu,ankurs}@umd.edu

***Abstract*—The impact of informed high level decisions on the performance-power efficiency of semiconductors is well known. From the context of heterogeneous integration for 3D IC, new innovations are needed which close the loop from architecture to device; from multiphysics to high level abstraction; from performance to security. Such considerations will be presented.**

***Index Terms*—heterogeneous integration, 3D IC, chiplets, co-design, high level synthesis, thermal management**

I. Introduction

3D integration technology expands circuit design into the third dimension by vertically stacking multiple functional device layers and interconnecting them using Through-Silicon-Vias (TSVs), as shown in Fig. 1. The vertical stacking structure is an attractive option for increasing transistor density. It also reduces interconnect wirelength hence scaling down power and delay. The reduction in interconnect wirelength can be leveraged by implementing a more highly connected architecture without increasing power or delay. Moreover, 3D integration allows separate layers to be fabricated using disparate materials and technologies. Heterogeneous integration optimizes existing System-on-Chip (SoC) designs by integrating components of different novel technologies into a single chip.

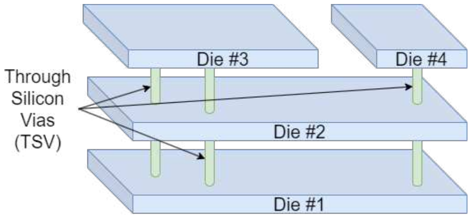

Fig. 1. An illustration of 3D heterogeneous integration where each die can be manufactured by a different fab, connected to each other through TSVs, and packaged in the same chip.

Tight integration of heterogenous chips or chiplets using advanced packaging schemes would require detailed analysis and co-design for performance, power efficiency, manufacturability, yield and quality, lifetime reliability, security and privacy. Heterogeneous integration of chiplets would subject them to substantial environmental extremes during fabrication and use. Their reliability, lifetime, and other physical metrics need detailed analyses, qualifications, and certification. The intricate interdependencies of these properties is illustrated in Fig. 2. In this paper, we discuss possible co-design approaches to 3D heterogeneous integration that account for the architecture, physical design, thermal management, and security issues of 3D ICs.

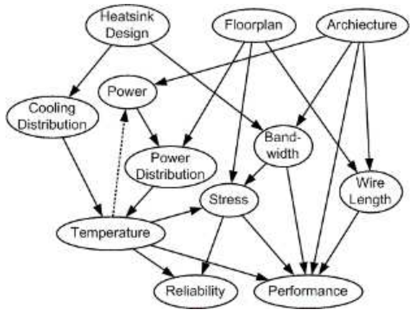

Fig. 2. Interdependencies between various metrics of 3D heterogeneous that need to be addressed by co-design approaches.

II. Architectural/Physical Co-design for 3D ICs

The typical IC design flow is partitioned into multiple steps (e.g. architectural design, RTL synthesis, circuit design, physical design, etc.) so that each step can be tackled and solved individually. However, that means high level architectural design decisions and lower level physical design decisions for 3D ICs are typically abstracted away from each other. While low level design steps such as global placement have had many recent innovations that improve physical design quality without explicit high level design information [1]–[5], there has been much less work on integrating high level architectural decisions with 3D IC's unique considerations.

One work on integrating high level design into the 3D IC design flow proposed integrating high level synthesis (HLS) as a subroutine during 3D floorplanning [6]. Since HLS can change the timing, area, and power characteristics of each floorplanned datapath by adjusting resource allocations and scheduling in order to adjust the PPA of floorplan modules, PPA adjustment can be integrated as a probabilistic step during simulated annealing.

Our recent work [7] proposes a more formal approach to datapath architecture synthesis integrated with 3D global placement. Instead of separating HLS from physical design, we propose a novel physically-aware binding method that takes placement-derived timing constraints into consideration when performing HLS operation binding, and a new dynamic net weighting approach applicable to global placement tools that adjust placement cost of each global net based on the net's likelihood of violating timing constraints.

To make HLS *physically*-aware, we first extract wirelengths from an existing 3D placement of hardware resources and calculate estimates of module-module wirelength and TSV RC delays. We then incorporate these physically-realizable delays along with a given scheduled dataflow graph (SDFG) as constraints into a integer-linear program (ILP) that finds a operation binding that minimizes a global timing metric (either total negative slack or worst negative slack). The resulting architecture is guaranteed to optimally minimize the chosen timing metric for a given 3D module placement and SDFG.

To make global placement *architecture*-aware, we take the timing constraints generated from our binding method and convert them into an equivalent wirelength that has the same RC delay. We then modify global placement by adding a weight term to each net's contribution to total wire cost that dynamically increases or decreases based on the global placer's current solution. To account for TSVs, we weight each TSV by its estimated RC delay contribution. Additionally, since many analytical global placers rely on gradient descent, our proposed weighting method is fully differentiable and is therefore applicable to a wide range of global placement methods.

III. Thermal Considerations for 2.5D and 3D IC

One of the key benefits of 3D ICs is high integration density. But since the power density rises proportionally to the number of layers in the stack, power delivery and thermal management become more complicated [8]. An effective thermal management method is to integrate thermal-aware floorplanning and placement at early design stages. Ma et. al. pointed out that traditional physical design methods where chiplets are closely packed to minimize wirelength end up creating thermally inefficient designs [9]. A thermally-aware chiplets placement method for heterogeneous 2.5D systems is thus proposed which inserts spacings between chiplets to jointly minimize temperature as well as total wirelength. Knechtel et. al. propose a block alignment methodology during floorplanning which shorten and optimize wire length and help planning massive interconnects [10]. A die partitioning method has been proposed in [11] which aims to minimize the number of TSVs and place hot and cool blocks alternately to reduce peak on-chip temperature. Zhang et. al. have proposed a 3D global routing algorithm with insertion of thermal wires and vias to minimize peak chip temperature [12]. Another effective thermal management technique is designing multi-story power delivery network where each die contain a separate power supply to distribute the loads more effectively, but at a higher area overhead [13].

For thermal management of 3D ICs, some previous works focus on designing efficient cooling systems. Micro-channel cooling has been identified to have a great potential in removing heat from 3D ICs but comes at a cost of high liquid pumping power. Thermal TSVs are efficient in conducting heat across different layers creating a more uniform temperature distribution but their efficacy is limited by the nature of heat sink. In [14], a hybrid cooling system has been proposed which uses micro-channel based liquid cooling as heat removing agent and thermal TSVs as heat conduction paths to the micro channel structures. A non-uniformly distributed micro-channel cooling system has been proposed in [15] to reduce pumping power. A dynamic thermal management scheme has also been proposed which uses thermal sensors to track chip power profile and dynamically controls temperature by tuning fluid flow rate. Such dynamic power and temperature management schemes are capable of tuning frequency, voltage, reconfigurable micro-architectural parameters etc at runtime to ensure thermal safety. The dynamic thermal management scheme proposed in [16] introduces a thermally aware job scheduling technique as a low overhead solution to reduce thermal problems. In [17], run-time thermal effects are considered for making power budgeting decisions of the different cores to ensure each core can maximize their performance without exceeding total power budget of the chip. At the same time, memory access behavior and heterogeneous cooling efficiency of 3D ICs are considered while making thread migration decisions.

IV. Security Considerations for 3D Heterogeneous Integration

Security and trust issues with the semiconductor supply chain have always been a concern of US chip design companies that are mostly fabless and need to outsource the fabrication to off-shore foundries [18]. Piracy, counterfeiting, and malicious modifications of the design are some of examples of possible attacks.

While 3D integration is initially designed to improve chip performance, it has presented various potentials in countering security threats with its built-in advantages, such as improving security-sensitive operations performance, split manufacturing for hardware IP protection, and reduced side-channel exposure [19]. For example, the reduction of memory latency achieved using heterogeneous integration also provides a major defense against side-channel attacks like cache-timing side channel attacks against cryptography modules. These attacks make use of differences in access latency between memory tiers inside the memory hierarchy [20]. These latency variations can be significantly decreased by using stacked memory structures, which will improve encryption and decryption speed and strengthen the system's defense against cache-based timing side channel attacks. If the design house has access to a trusted foundry, then 3D integration allows designs to be split into a "control" layer manufactured by the trusted foundry and a "computation" layer manufactured by an untrusted external foundry [21]. In such a scheme, the parts of the circuit requiring the highest performance are produced with cutting-edge technology. The control layer produced by the trusted foundry includes security features such as oversight hardware to see internal signals on the computation layer which would be inaccessible to traditional external monitors. At the minimum, the secure die can contain routing information. Since the untrusted foundry can guess standard interconnects with high accuracy [22], the designer must control the physical placement of through-silicon vias for obfuscation [23].

However, challenges in testing complexities and supply chain vulnerabilities persist, demanding innovative solutions. Mechanisms to establish trust in the packaging facility need to be established to ensure that only trusted chips are integrated into the 3D IC. To address the detection challenges, novel forensic infrastructure in 3D IC needs to be developed, and we remain optimistic that some of the area and performance improvement can be traded for security, and the security gain of 3D HI should outweigh the new challenges.

V. Conclusion

A co-design framework for 3D heterogeneous integration should close the loop between the computational, physical, thermal aspects of the system. The computational view can be characterized by the SDFG which is capable of capturing the architectures and running applications. Physical view includes floor planning, wire delay modeling, power modeling, and 3D TSV reliability modeling. The thermodynamic view incorporates static and dynamic thermal modeling for a myriad of cooling solutions. The performance, efficiency, and reliability of any heterogeneous system depend on its architecture, physical configuration (layout, packaging), and the cooling (thermal) solution. The security against any attack surface needs to be considered at each of each views. Incorporating HLS in co-design can help substantially improve performance by navigating the intricate interdependence between cyber, physical, thermodynamic, and reliability aspects of the system.

Acknowledgment

This work is supported by the NSF under Award 1953285.

References

[1] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017, conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[2] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs," in *Proceedings of the 2018 International Symposium on Physical Design*, ser. ISPD '18. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 90–97. [Online]. Available: https://dl.acm.org/doi/10.1145/3177540.3178244

[3] D. H. Kim, K. Athikulwongse, and S. K. Lim, "Study of Through-Silicon-Via Impact on the 3-D Stacked IC Layout," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 5, pp. 862–874, May 2013, conference Name: IEEE Transactions on Very Large Scale Integration (VLSI) Systems. [Online]. Available: https://ieeexplore.ieee.org/document/6268361

[4] J. Lu, H. Zhuang, I. Kang, P. Chen, and C.-K. Cheng, "ePlace-3D: Electrostatics based Placement for 3D-ICs," in *Proceedings of the 2016 on International Symposium on Physical Design*. Santa Rosa California USA: ACM, Apr. 2016, pp. 11–18. [Online]. Available: https://dl.acm.org/doi/10.1145/2872334.2872361

[5] M.-K. Hsu, V. Balabanov, and Y.-W. Chang, "TSV-Aware Analytical Placement for 3-D IC Designs Based on a Novel Weighted-Average Wirelength Model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 4, pp. 497–509, Apr. 2013, conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[6] Y. Chen, G. Sun, Q. Zou, and Y. Xie, "3DHLS: Incorporating high-level synthesis in physical planning of three-dimensional (3D) ICs," in *2012 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Mar. 2012, pp. 1185–1190, iSSN: 1558-1101.

[7] D. Xing and A. Srivastava, "A High Level Approach to Co-Designing 3D ICs," in *2024 61st ACM/IEEE Design Automation Conference (DAC)*, Jun. 2024, pp. 1–6.

[8] J. Knechtel and J. Lienig, "Physical design automation for 3d chip stacks: Challenges and solutions," in *Proceedings of the 2016 on International Symposium on Physical Design*, ser. ISPD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 3–10. [Online]. Available: https://doi.org/10.1145/2872334.2872335

[9] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, and A. Joshi, "TAP-2.5D: A Thermally-Aware Chiplet Placement Methodology for 2.5D Systems," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Feb. 2021, pp. 1246–1251, iSSN: 1558-1101.

[10] J. Knechtel, E. F. Y. Young, and J. Lienig, "Planning massive interconnects in 3-d chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 11, pp. 1808–1821, 2015.

[11] C. Jang and J.-w. Chong, "Thermal-aware floorplanning with min-cut die partition for 3d ics," *ETRI Journal*, vol. 36, no. 4, pp. 635–642, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.14.0113.1204

[12] T. Zhang, Y. Zhan, and S. S. Sapatnekar, "Temperature-aware routing in 3d ics," in *Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, ser. ASP-DAC '06. IEEE Press, 2006, p. 309–314. [Online]. Available: https://doi.org/10.1145/1118299.1118377

[13] P. Jain, T.-H. Kim, J. Keane, and C. H. Kim, "A multi-story power delivery technique for 3d integrated circuits," in *Proceeding of the 13th international symposium on Low power electronics and design (ISLPED '08)*, 2008, pp. 57–62.

[14] B. Shi, A. Srivastava, and A. Bar-Cohen, "Hybrid 3d-ic cooling system using micro-fluidic cooling and thermal tsvs," in *2012 IEEE Computer Society Annual Symposium on VLSI*, 2012, pp. 33–38.

[15] B. Shi and A. Srivastava, "Cooling of 3d-ic using non-uniform micro-channels and sensor based dynamic thermal management," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2011, pp. 1400–1407.

[16] A. K. Coskun, J. L. Ayala, D. Atienza, T. S. Rosing, and Y. Leblebici, "Dynamic thermal management in 3d multicore architectures," in *2009 Design, Automation Test in Europe Conference Exhibition*, 2009, pp. 1410–1415.

[17] K. Kang, J. Kim, S. Yoo, and C.-M. Kyung, "Runtime power management of 3-d multi-core architectures under peak power and temperature constraints," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 6, pp. 905–918, 2011.

[18] A. Haramboure, G. Lalanne, C. Schwellnus, and J. Guilhoto, "Vulnerabilities in the semiconductor supply chain," 2023. [Online]. Available: https://www.oecd-ilibrary.org/content/paper/6bed616f-en

[19] Y. Xie, C. Bao, C. Serafy, T. Lu, A. Srivastava, and M. Tehranipoor, "Security and vulnerability implications of 3d ics," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 2, pp. 108–122, 2016.

[20] C. Bao and A. Srivastava, "Reducing timing side-channel information leakage using 3d integration," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 4, pp. 665–678, 2019.

[21] J. Valamehr, T. Sherwood, R. Kastner, D. Marangoni-Simonsen, T. Huffmire, C. Irvine, and T. Levin, "A 3-d split manufacturing approach to trustworthy system development," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 4, pp. 611–615, 2013.

[22] J. Rajendran, O. Sinanoglu, and R. Karri, "Is split manufacturing secure?" in *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2013, pp. 1259–1264.

[23] Y. Xie, C. Bao, and A. Srivastava, "Security-aware design flow for 2.5 d ic technology," in *Proceedings of the 5th International Workshop on Trustworthy Embedded Devices*, 2015, pp. 31–38.