# Monolithic 3D Integrated Circuits: Recent Trends and Future Prospects

Krithika Dhananjay, *Graduate Student Member, IEEE*, Prachi Shukla, *Graduate Student Member, IEEE*, Vasilis F. Pavlidis, *Senior Member, IEEE*, Ayse Coskun, *Senior Member, IEEE*, and Emre Salman, *Senior Member, IEEE*

***Abstract*—Monolithic 3D integration technology has emerged as an alternative candidate to conventional transistor scaling. Unlike conventional processes where multiple metal layers are fabricated above a single transistor layer, monolithic 3D technology enables multiple transistor layers above a single substrate. By providing vertical interconnects with physical dimensions similar to conventional metal vias, monolithic 3D technology enables unprecedented integration density and high bandwidth communication, which plays a critical role for various data-centric applications. Despite growing number of research efforts on various aspects of monolithic 3D integration, commercial monolithic 3D ICs do not yet exist. This tutorial brief provides a concise overview of monolithic 3D technology, highlighting important results and future prospects. Several applications that can potentially benefit from this technology are also discussed.**

***Index Terms*—Three-dimensional integrated circuits, monolithic integrated circuits, through-silicon vias, fabrication, physical design, thermal management, testing, neural networks, security, computer applications.**

## I. Introduction

As TWO-DIMENSIONAL geometry scaling of conventional transistors is coming to an end, International Roadmap for Devices and Systems (IRDS) predicts that monolithic 3D integration technology (here termed as Mono3D technology) will be one of the critical performance boosters [1]. Mono3D technology represents a radical departure from conventional fabrication technologies where transistors are first patterned via the front-end-of-line (FEOL) portion of the process, followed by the patterning of multiple interconnect layers via the back-end-of-line (BEOL) process. Unlike other vertical integration technologies, manufacturing multiple transistor layers on a single substrate in Mono3D technology exhibits unique opportunities for providing extremely dense ICs.

In the past two decades, the semiconductor industry has witnessed significant improvements in various forms of vertical integration technologies, such as the through silicon via (TSV) based 3D integration (also referred to as chip stacking) and lateral integration of multiple dies placed on the same interposer substrate for higher bandwidth communication (also referred to as 2.5D integration). The number of commercial applications that utilize these TSV-based and, relatively, more established forms of vertical integration, has steadily increased. For example, there are commercial FPGAs and GPUs that utilize interposer based 2.5D integration to increase density while significantly reducing the interconnect overhead [2]–[5]. Commercial products of TSV-based systems include primarily memory arrays, such as the Hybrid Memory Cube (HMC) [6] and High Bandwidth Memory (HBM) [7], which are multi-layer DRAM chips. Most recently, commercial integration of multiple logic chiplets has been demonstrated in a face-to-face configuration with TSVs [5].

Although TSV-based 3D technologies enable significant benefits in system-level performance, power consumption, and form factor, compared to typical 2D integration, these technologies suffer from a noticeable asymmetry between the transistor dimensions and the dimensions of the TSVs. The channel length of modern transistors has reached sub-10 nm dimensions, whereas the diameter of modern TSVs is in the range of several micrometers. This large gap is a significant limitation on the density/granularity of TSV-based die stacking. For example, a typical 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 [8], [9]. In the 7 nm technology node, a TSV is equivalent to approximately several thousand gates in terms of load capacitance, thereby consuming significant dynamic power and causing *RC* delay [10]. Monolithic 3D technology mitigates these problems by reducing the dimension of vertical interconnects, referred to as monolithic inter-tier vias (MIVs), down to nanometers, thereby enabling unprecedented levels of integration density and granularity [11]. Recognizing the high potential of Mono3D, this tutorial brief provides a concise overview on multiple facets of monolithic 3D integration technology while outlining existing research directions and highlighting future prospects.

Manuscript received November 2, 2020; revised December 13, 2020; accepted December 30, 2020. Date of publication January 13, 2021; date of current version February 26, 2021. This work was supported in part by the National Science Foundation under Grant CCF 1910075/1909027. This brief was recommended by Associate Editor Y. Ha. *(Corresponding author: Emre Salman.)*

Krithika Dhananjay and Emre Salman are with the Department of Electrical and Computer Engineering, Stony Brook University–SUNY, Stony Brook, NY 11794 USA (e-mail: emre.salman@stonybrook.edu).

Prachi Shukla and Ayse Coskun are with the Electrical and Computer Engineering Department, Boston University, Boston, MA 02215 USA.

Vasilis F. Pavlidis is with the Department of Computer Science, University of Manchester, Manchester M13 9PL, U.K.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2021.3051250.

Digital Object Identifier 10.1109/TCSII.2021.3051250

1549-7747 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.

TABLE I
A QUALITATIVE COMPARISON BETWEEN 2.5D, TSV-BASED 3D, AND MONO3D INTEGRATION TECHNOLOGIES

| Integration technology | Commercial Availability | CAD tools | Design-for-test |
|---|---|---|---|
| INTERPOSER-BASED 2.5D | NVIDIA and AMD GPUs [3], [4], Xilinx and Intel FPGAs [2], [5], Samsung I-Cube [21], TSMC CoWoS and InFO [22], [23]. | Existing 2D CAD tools used for die designs. Chip/package/board co-design for heterogeneous integration under development [24]. | Standardized test-port interface: IEEE 1838 [25], [26]. |
| DIE STACKED 3D | Commercial fabrication for memory stacking [6], [27], Intel's LakeField processor [28] and Samsung's X-Cube [29]. | Various CAD tools developed for 3D design. Chip/package/board co-design for heterogeneous integration under development [30], [31]. | Standardized test-port interface: IEEE 1838 [25]. |
| MONO3D | SRAM arrays and CNFET, NVM integration demonstrated [16], [20]. No commercial production yet. | Tools that are integrated with existing flows [32]–[34] have been proposed, but no standard CAD flow yet. | Low cost dual-BIST DfT architecture proposed for MIV-testing but not standardized [35]. |

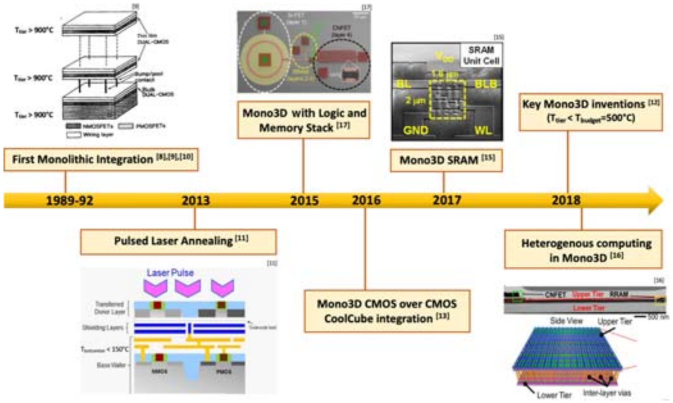

First Monolithic Integration [8],[9],[10]
Mono3D with Logic and Memory Stack [17]
Mono3D SRAM [15]
Key Mono3D inventions [12] ($T_{tier} < T_{budget}$=500°C)
1989-92
2013
2015
2016
2017
2018
Pulsed Laser Annealing [11]
Mono3D CMOS over CMOS CoolCube integration [13]
Heterogenous computing in Mono3D [16]
SRAM Unit Cell
BL
BLB
GND
WL
2 µm
Laser Pulse
CNFET
Upper Tier
RRAM
Lower Tier
Side View
Inter-layer vias

Fig. 1. Chronological timeline of primary developments in the fabrication process of MONO3D ICs.

## II. MONO3D FABRICATION PROCESS

A chronological timeline illustrating the primary developments in the fabrication process of MONO3D technology is shown in Fig. 1. The development of a sequential fabrication process to construct multiple transistor layers on a single substrate dates back to late 80s [12], [13]. These early studies relied on laser beam recrystallization to form multiple silicon-on-insulator (SOI) layers on top of the p-type substrate. A single device layer contained only one type of transistor, thereby reducing the number of process steps related to well formation and ion implantation. Since these layers were subject to 900°C temperature during device fabrication, doped polysilicon interconnects were utilized in these early implementations due to its ability to withstand high temperatures. Correct operation of the devices was demonstrated by measuring I-V characteristics in each layer. However, the SOI devices exhibited approximately 4× higher threshold voltage variation.

A primary challenge in building MONO3D structures is to ensure the reliability of devices within the bottom layer that can be degraded by the temperature and processing time required to fabricate high quality transistors at upper layers. Another related challenge is to manufacture multiple low resistivity interconnect (and low-k dielectric) layers for the first device layer since routability becomes an important concern due to very high density integration. These issues limit the processing temperature of upper layers below 500°C, referred to as the thermal budget constraint.

Several steps in traditional FEOL processes require temperatures much higher than 500°C, such as activation of the implanted impurities and annealing to fix the crystal defects encountered during ion implantation. One approach is to rely on pulsed laser annealing and activation, assuming that the thickness of the transistor silicon layer is sufficiently small (in the range of 20 nm) [14]. In this approach, lasers with pulse widths below 100 ns are used to momentarily increase the temperature of the higher layers above 1400°C. Thin shielding layers are used to ensure that the temperature that diffuses to the bottom layers remains below 150°C. This pulsed laser based recrystallization and activation enables low resistance polysilicon gates for the upper layers [15]. Another key fabrication-level enabler for MONO3D technology is high quality growth of silicon by using low temperature epitaxy [15]. In this approach, traditional surface preparation techniques were replaced with a low temperature process including a combination of dry and wet etch preparation. A silicon epitaxial layer was built with good selectivity and crystallinity at 500°C.

Another risk when fabricating upper layer devices is contamination since the wafer is reintroduced to FEOL processes after a contaminated backend process where interconnects and vias were built. A three step contamination control strategy was proposed [15], consisting of etching the wafer bevel edge (shown to be the most critical contamination source) to remove the deposited metals, decontamination via wet cleaning, and encapsulation of the wafer bevel.

These developments in manufacturing enabled the demonstration of MONO3D ICs with relatively low complexity. For example, transistor-on-transistor integration with alignment accuracy in the range of nanometers was shown [16]. Successful MONO3D integration of an SRAM array with sufficient stability has also been achieved [17], [18]. Finally, carbon nanotube field-effect transistors and non-volatile resistive memory were successfully integrated with conventional silicon based devices on a MONO3D platform [19], [20].

## III. MONO3D DESIGN METHODOLOGIES

In this section, the latest advances related to physical design, thermal integrity, and testing methodologies of MONO3D technology are described. A broad qualitative comparison of these challenges (including commercial status) for 2.5D, TSV-based 3D, and MONO3D integration is provided in Table I.

### A. Physical Design

Over the recent years, developing pseudo-3D tools for monolithic integration has gained significant attention [36]. These studies make some process or technology file

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.

modifications to “trick” the 2D engine to produce quality 3D designs.

Shrunk-2D (S2D) [32] was the first commercial-quality tool in which the entire design area is shrunk by 50% by scaling the dimensions of the original 2D chip and the standard cells by a factor of $1/\sqrt{N}$, where $N$ is the number of tiers. The metal width and pitch are also scaled by the same factor, keeping the $RC$ per unit length fixed. The 2D P&R engine is applied to this shrunk design before the cells are blown up and partitioned to their respective tiers. This tool, however, does not estimate $RC$ parasitic impedance accurately and also expects the commercial 2D engines to handle dimensions smaller than their capability. Another tool called Compact-2D [33] was proposed with more accurate timing characterization and lesser legalization issues than S2D. However, it does not support simultaneous timing closure for all the tiers, and hence, leading to performance degradation. Alternatively, Cascade2D tool [34] performs a design-aware partitioning at the RTL stage. The placement, routing and timing closure are simultaneous for both the tiers, which significantly improves the timing characteristics of the design, unlike the other two flows.

A detailed comparative study was performed between the three algorithms on RISC-V Rocketcore processor using identical 3D footprints [37]. The comparison showed that though, S2D and Compact-2D have large total negative slacks compared to Cascade2D, the total wirelength and overall power consumption for Cascade 2D was 21.4% and 7.5% greater than conventional 2D, respectively. However, the total wirelength of S2D and Compact-2D was, respectively, 12.2% and 10.5% lower than the conventional 2D. The overall power consumption of S2D and Compact-2D was also shown to be 7.1% and 8% lower than conventional 2D, respectively. This increase in the wirelength and power is attributed to the tier-partitioning strategy and MIV planning strategies in Cascade2D being unsuitable for flattened gate-level designs.

Another important physical design issue in MONO3D ICs is routing congestion since the chip footprint is significantly reduced (20% to 40%), but the overall number of available metal layers does not significantly increase [38]–[40]. This issue was investigated by Yan *et al.* [40], [41] by developing a MONO3D cell library and two-layer PDK in 45 nm technology node [42]. The pull-down nMOS transistors of the cells and the I/O pins are placed on the upper layer whereas the pull-up pMOS transistors are placed on the bottom layer. The connections among the layers are achieved via intra-cell MIVs, thereby permitting the use of existing 2D placement tools. An example is provided in Fig. 2, where a 2D and MONO3D implementations of a 128-point FFT operating at 500 MHz are illustrated. It was demonstrated that increasing the number of routing tracks per cell from 8 to 10 slightly increases the die-level footprint of the 128-point MONO3D FFT core, but reduces the overall wirelength (due to less routing congestion). Furthermore, the timing characteristics are significantly enhanced since the coupling capacitances are reduced. As compared to 2D, a MONO3D FFT core (with 10 routing tracks) reduces the footprint by approximately 37% while also achieving approximately 13% reduction in overall power consumption. Using this library, the effect of successive in-place optimization (IPOs) on partitioned tiers produced either by a greedy bin-based Fidducia-Mattheyses or a displacement-based legaliser is presented [43], where a single BEOL connecting the tiers and few IPOs yield higher PPA for the explored circuits.

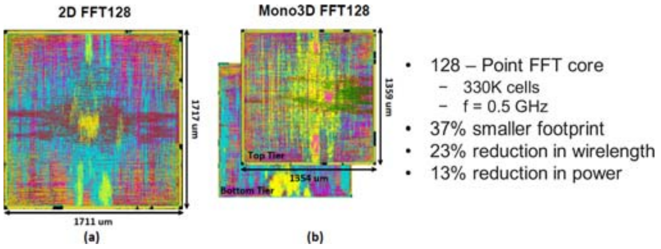


Fig. 2. A fully placed and routed 128-point FFT in 45 nm technology using: (a) conventional 2D, (b) transistor-level MONO3D technology with two layers.

Although there have been several academic efforts to develop complete end-to-end design flows, developing commercially available tools that leverage the integration, power, and performance benefits provided by MONO3D still remains an open problem. Thermal aspects are an integral part of the design process and should be considered while optimizing for other parameters, as explored in the following section.

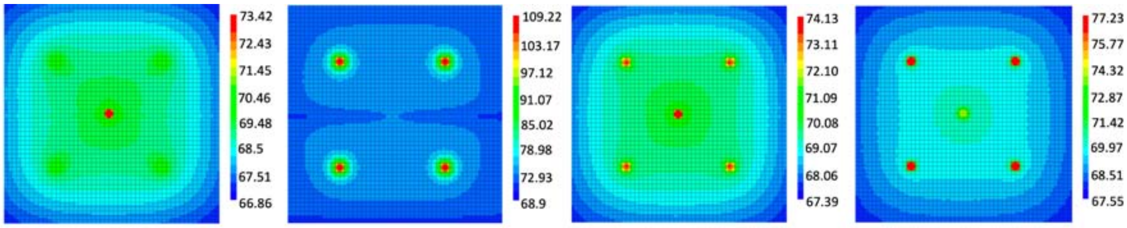
### B. Thermal Integrity

A major limitation in 3D ICs is effective heat dissipation from tiers away from the heat sink/spreader. MONO3D systems face additional thermal issues due to the higher device integration density, routing congestion, and strong inter-tier thermal coupling. These characteristics further differentiate MONO3D technology from TSV-based 3D ICs. Several works have presented techniques to alleviate thermal challenges in MONO3D systems. Careful design of power delivery networks (PDN) can help reduce chip temperatures by up to 5% [44] while ignoring the effect of PDNs leads to overestimated chip temperatures [45]. Nano-pillars were placed at design time in transistor-level MONO3D systems for heat dissipation from selected hot spot regions, and 53% reduction in temperature was shown through finite element method simulations [46].

A two-tier MONO3D system is modeled by Samal *et al.* [47]. Strong inter-tier thermal coupling (i.e., similar temperature across tiers) is demonstrated due to a thin dielectric in-between and a non-linear regression model is constructed to estimate chip temperatures. However, the lateral heat flow is ignored due to the thin tiers. In comparison, they show that the tier away from the heat spreader for a similar TSV-based system, is hotter than that in the MONO3D system due to higher vertical resistance (since vertical heat flow path is longer) and presence of a thick bonding layer between tiers.

Two-tier MONO3D and TSV-based 3D IC are modeled along with the metal layers [48] using HotSpot-6.0 [49] (an architecture level thermal simulator). Synthetic power profiles are simulated by placing one hot spot at the center of the bottom tier (closer to heat sink) and four hot spots in the upper tier (away from heat sink), each with a power density of 750 W/cm$^2$. The results are shown in Fig. 3. It can be observed that thermal coupling helps in effective heat dissipation of the

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.



(a) Bottom tier (TSV-based 3D) (b) Upper tier (TSV-based 3D) (c) Bottom tier (MONO3D) (d) Upper tier (MONO3D)

Fig. 3. Thermal maps of the TSV-based 3D and MONO3D ICs.

upper tier to the heat sink in MONO3D (note the lower temperature in Fig. 3(d) than in Fig. 3(b)). Alternatively, thermal coupling may cause hot spots in the upper tier that form similar hot spots in the bottom tier (note the additional hot spots in Fig. 3(c) in comparison to Fig. 3(a)). Higher localization of the hot spots in the MONO3D IC can also be seen, thus demonstrating lower lateral thermal coupling. Several power profiles are also simulated by varying hot spot distribution, size, and area. Despite the limited lateral heat flow, this flow of heat should not be ignored as it may under-estimate hot spots by ≈4°C in tiers away from heat sink. These results demonstrate the specific thermal challenges that should be considered while designing MONO3D ICs to ensure thermal integrity.

### C. Design-for-Test

MONO3D ICs exhibit unique test challenges due to multiple transistor layers and dense MIVs. Aggressive scaling of the thickness of the inter-layer dielectric and the increased densities of MIVs (30 million per mm$^2$) [50] are the primary cause of functional and timing faults in MONO3D systems [51]. Typical fault models for testing the MIVs are similar to the interconnect fault models in 2D ICs and can be broadly categorized as shorts, opens and stuck-at-faults. For the testing of such faults, the conventional ATPG-based methodologies tend to fall short primarily because all the test vectors should be propagated through the multiple tiers and the MIVs, adding significant amount of constraints to the tool, demanding built-in self-test (BIST) solutions for MONO3D technology [52].

A solution was proposed to enhance the observability and controllability of MIVs by using a die-wrapper register cell on both ends of MIV [53]. The BIST technique proposed by Koneru *et al.* employs interface scan cells connected to twisted-ring counter on a dedicated test layer targeting the testing of opens and shorts in MIVs [54]. The cost per die savings were inferred to be as high as 40% when compared to the similar technique employed for TSV-based 3D IC that mandates the insertion of die wrapper register. However, both of these methods incur a significant area overhead. Recently, a low-cost dual-BIST architecture was proposed using XOR gates to detect opens, shorts and stuck-at faults [52]. In that work, only two test patterns are used for exhaustive testing of all MIV fault scenarios combined and the test responses are compacted using 2-bit signatures at the output. The maximum area and power overheads of the BISTed designs reported are 2.6% and 9% when compared to a standard design [35].

The TSV-based pre-bond strategies cannot be adapted to MONO3D testing predominantly because bare MIVs cannot be exposed and the existing wafer-probing techniques cannot support the MIV pitch that is on the order of 100 to 200 nm [52]. The post-bond strategies also cannot be extended to MONO3D because of the significant increase in the number of MIVs used. Consequently, the above mentioned works have proposed several novel BIST solutions for MONO3D technology, which continues to be an active field of research.

## IV. EMERGING APPLICATIONS FOR MONO3D

### A. Near- and In-Memory Computing

Computing near- or in-memory is a promising and popular approach to surmount the challenges relating to the well-known “memory wall” issue. Consequently, several methods that enable computing in the DRAM memory chips [55], [56] or near non-volatile memories [57] have been developed. The high MIV density of MONO3D circuits makes them well suited for these computing paradigms.

Due to the physical proximity of the device layers in MONO3D circuits, the two aforementioned approaches are hardly distinguishable. Rather a better way to discern the use of these two paradigms in MONO3D circuits is to observe the related partition granularity [58]. Thus, near-memory computing is better served by MONO3D technology if block-level partition is utilized. Alternatively, in-memory computing is underpinned by transistor-level integration. The former exploits the massive interconnect density enabled by the MIVs and the low delay of these interconnects. On the other hand, the latter can utilize one of the device layers to deploy the additional bit-wise logic operations without disrupting the density of the memory array and maintaining most of the original performance.

A recent example of near-memory computing with MONO3D combines a processing tier with a layer of resistive RAM (ReRAM) on top, extending the RADAR accelerator architecture in three physical dimensions. The use of MONO3D improves the execution of the Smith-Waterman algorithm by three orders of magnitude compared to 2D implementation [59], while an effort to speed up the same algorithm on a FPGA [60] yields a considerably lower improvement of 330×.

Alternatively, a cell enabling in-memory computing through transistor-level partitioning has been presented [61]. The cell consists of nine transistors (9T) and can implement

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.

NAND/AND, NOR/OR, and XOR/XNOR operations, in addition to storing a bit, within a memory cycle. The key idea of this 9T cell is to maintain the storage (SRAM) cell in one of the device layers and implement the remaining three transistors used for the logic operations in the upper device layer. Although this structure does not result in the minimum footprint, a savings of 51% is achieved as compared to a 2D implementation of the same 9T cell. Furthermore, since the memory array is not disrupted, the memory density compared to a 2D SRAM array remains unaltered. These recent examples demonstrate the presently unexplored potential of using MONO3D for upcoming non von Neumann architectures.

### B. Deep Neural Networks

Deep Neural Networks (DNNs) have become popular for a wide range of applications, such as image recognition or speech recognition [62]. DNN accelerators are also designed to satisfy the high throughput and memory bandwidth demands of DNNs [63]–[66]. However, high energy consumption in DNNs (due to heavy computation and data movement) is a major design concern. Since MONO3D offers high interconnect density and power/performance benefits over 2D, there is a growing interest towards designing energy-efficient MONO3D DNN accelerators.

Yu *et al.* reduced the memory footprint of CNNs by encoding the activation and weight for sparsity, and then interfacing the CNN accelerator with a MONO3D non-volatile RAM to significantly improve bandwidth using MIVs [67]. The reported average improvement in execution time, power dissipation, and energy efficiency are up to 16×, 4.5×, and 69×, respectively, for CNN training/inference w.r.t. Nvidia GeForce GTX 1080 Ti. While some works (e.g., [67], [68]) have utilized the high-density MIVs in a block-level integration to highlight the benefits of MONO3D, other works (e.g., [69], [70]) have used various partitioning techniques to achieve an energy-efficient ASIC design. Chang *et al.* have performed gate-level partitioning for the MAC units and block-level for the SRAMs for two DNN architectures [69]. 22.3% and 6.2% improvement in iso-performance power consumption and performance, respectively, were shown against a 2D baseline. Do *et al.* designed a two-layer GPU scratchpad memory using MONO3D to enable fast SRAM access by enabling concurrent row and column accesses, thus improving the system performance by 46.3% [70]. These works benefit from the reduced wirelength, but do not consider thermal issues.

A temperature-aware optimization flow to design a near-optimal energy-efficient systolic DNN accelerator in MONO3D mobile systems, under user-specified performance/thermal constraints is presented [71]. The optimization flow highlights the performance and temperature tradeoffs by showing that DNNs with a large number of MAC operations and higher dynamic power (such as VGG19 [72]) can result in high chip temperatures, which further increases temperature-dependent leakage due to tight inter-tier thermal coupling. The net increase in total power can result in thermal violations. While all of these works have demonstrated the benefits offered by MONO3D, there is still scope for developing standard performance and power models to help in designing DNN accelerators using different MONO3D partitioning techniques, utilizing the available bandwidth, while also accounting for thermal awareness.

### C. Hardware Security

The past decade has experienced a proliferation of research literature related to the hardware security aspects of 3D ICs [73], [74]. A majority of the work specific to MONO3D is based on circuit-level obfuscation techniques that broadly encompass logic locking and layout camouflaging. Camouflaging technique is used to thwart reverse engineering attacks by making modifications to the layout. A layout-level camouflaging technique is proposed for transistor-level MONO3D circuits by using dummy contacts [75]. The camouflaged MONO3D ISCAS '89 benchmark circuits are shown to exhibit, on average, power and area savings of, respectively, 6.3% and 47.5% compared to a conventional 2D IC at the expense of 27.4% timing degradation. Logic locking employs key-controlled logic gates for protection against IP piracy and reverse engineering attacks. In a recent work, the pull-down and pull-up networks of MONO3D cells in separate tiers are locked using parallel or serial locking transistors and camouflaged contacts, independently [76]. Therefore, one leaking tier does not compromise the entire design.

Due to exacerbated thermal issues, 3D systems (and particularly MONO3D ICs) can be more vulnerable to thermal side channel analysis attacks and thermal covert channels [77]. These attack models have not received much attention for MONO3D ICs. Recently, a thermal-aware side channel shielding technique was proposed for 3D ICs by generating custom activity patterns [78]. Another work developed thermally-aware floorplanning strategies to mitigate the thermal side-channel leakage [79]. However, both of these works are based on TSV-based 3D ICs. Similarly, covert communication through lateral thermal coupling between two cores has been demonstrated on a multicore 2D system with throughputs on the order of 50 bits per second [80]. Even though thermal side channel leakage or thermal covert channels are relatively low bandwidth processes, the strong vertical thermal coupling between the layers of a MONO3D IC (see Section III-B and Fig. 3) can potentially increase this capacity. For example, consider a MONO3D multicore system where one of the cores has access to sensitive information. This information can be covertly transmitted to another core located at the bottom layer. Such attack is possible provided that a temperature-based communication channel is established between those two cores by leveraging strong thermal crosstalk. This phenomenon can be a critical security breach for MONO3D ICs, particularly for sandboxed systems where the data that belongs to an application or core should be protected from other applications and cores.

## V. CONCLUSION AND FUTURE PROSPECTS

In this tutorial brief, the current status of MONO3D technology is presented with emphasis on both fabrication- and design-level developments. The primary process challenges

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.

related to thermal budget constraint, high quality silicon growth, and contamination are discussed, including existing approaches to these issues. At the design-level, important results on design automation, thermal integrity, and design-for-test are provided with emphasis on unique MONO3D characteristics. Three specific applications that can potentially benefit from MONO3D technology are also presented.

Lack of experimental results is a primary limitation related to most of the existing work on MONO3D ICs. Similarly, the fabrication-level constraints and design-level tools/methodologies are not sufficiently coupled, thereby making these methods less applicable. A stronger interaction between fabrication and design is anticipated and required in the future, as the MONO3D process matures and more opportunities arise for fabrication. These experimental results can also facilitate the detailed characterization and modeling of inter-tier process variations, which remain a primary concern for MONO3D technology. Existing cost models for MONO3D circuits can also benefit from these results. In the mean time, cross-layer design methods that go beyond physical design automation will enable system-level design space exploration while simultaneously considering important design objectives such as efficiency, performance, and thermal integrity. These methodologies and models should properly consider the fundamental partition approaches that MONO3D supports as each of these approaches will most likely be better suited for specific applications.

## REFERENCES

[1] *International Roadmap for Devices and Systems*. Accessed: Dec. 3, 2020. [Online]. Available: https://irds.ieee.org

[2] *Xilinx-Virtex-7-2000T*. Accessed: Dec. 3, 2020. [Online]. Available: https://www.xilinx.com/video/fpga/virtex-7-2000t-asic-prototyping-emulation.html

[3] C.-C. Lee *et al.*, "An overview of the development of a GPU with integrated HBM on silicon interposer," in *Proc. IEEE Electron. Compon. Technol. Conf.*, 2016, pp. 1439–1444.

[4] *Nvidia Tesla P100*. Accessed: Dec. 3, 2020. [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-p100/

[5] *Embedded Multi-Die Interconnect Bridge*. Accessed: Nov. 2, 2020. [Online]. Available: https://www.intel.com/content/www/us/en/foundry/emib.html

[6] J. T. Pawlowski, "Hybrid memory cube (HMC)," in *Proc. IEEE Hot Chips Symp.*, Aug. 2011, pp. 1–24.

[7] D. U. Lee *et al.*, "25.2 A 1.2V 8GB 8-channel 128GB/s high-bandwidth memory (HBM) stacked dram with effective microbump I/O test methods using 29nm process and TSV," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2014, pp. 432–433.

[8] H. Wang, M. H. Asgari, and E. Salman, "Compact model to efficiently characterize TSV-to-transistor noise coupling in 3D ICs," *Integr. VLSI J.*, vol. 47, no. 3, pp. 296–306, Jun. 2014.

[9] S. M. Satheesh and E. Salman, "Power distribution in TSV-based 3D processor-memory stacks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 4, pp. 692–703, Dec. 2012.

[10] Q. Xie, X. Lin, Y. Wang, S. Chen, M. J. Dousti, and M. Pedram, "Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 8, pp. 761–765, Aug. 2015.

[11] S. Wong, A. El-Gamal, P. Griffin, Y. Nishi, F. Pease, and J. Plummer, "Monolithic 3D integrated circuits," in *Proc. Int. Symp. VLSI Technol. Syst. Appl.*, Apr. 2007, pp. 1–4.

[12] T. Kunio, K. Oyama, Y. Hayashi, and M. Morimoto, "Three dimensional ICs, having four stacked active device layers," in *Int. Tech. Dig. Electron Devices Meeting*, 1989, pp. 837–840.

[13] Y. Hayashi *et al.*, "A new three dimensional IC fabrication technology, stacking thin film DUAL-CMOS layers," in *Int. Electron Devices Meeting Tech. Dig.*, 1991, pp. 657–660.

[14] B. Rajendran, A. K. Henning, B. Cronquist, and Z. Or-Bach, "Pulsed laser annealing: A scalable and practical technology for monolithic 3D IC," in *Proc. IEEE Int. 3D Syst. Integr. Conf.*, 2013, pp. 1–5.

[15] L. Brunet *et al.*, "Breakthroughs in 3D sequential technology," in *Proc. IEEE Int. Electron Devices Meeting*, 2018, pp. 7.2.1–7.2.4.

[16] L. Brunet *et al.*, "First demonstration of a CMOS over CMOS 3D VLSI CoolCube$^{\mathrm{TM}}$ integration on 300mm wafers," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 1–2.

[17] W. Chen *et al.*, "A dual-split-controlled 4P2N 6T SRAM in monolithic 3D-ICs with enhanced read speed and cell stability for IoT applications," *IEEE Electron Device Lett.*, vol. 39, no. 8, pp. 1167–1170, Aug. 2018.

[18] T.-T. Wu *et al.*, "High performance and low power monolithic three-dimensional sub-50 nm poly Si thin film transistor (TFTs) circuits," *Sci. Rep.*, vol. 7, p. 1368, May 2017.

[19] T. F. Wu *et al.*, "Hyperdimensional computing exploiting carbon nanotube FETs, resistive RAM, and their monolithic 3D integration," *IEEE J. Solid-State Circuits*, vol. 53, no. 11, pp. 3183–3196, Nov. 2018.

[20] M. M. Shulaker, T. F. Wu, M. M. Sabry, H. Wei, H.-S. P. Wong, and S. Mitra, "Monolithic 3D integration: A path from concept to reality," in *Proc. IEEE Design Autom. Test Europe Conf. Exhibition*, Mar. 2015, pp. 1197–1202.

[21] *2.5D Interposer(I-Cube) Development*. Accessed: Dec. 3, 2020. [Online]. Available: www.samsungfoundry.com

[22] Y.-L. Chuang *et al.*, "Unified methodology for heterogeneous integration with CoWoS technology," in *Proc. IEEE Electron. Compon. Technol. Conf.*, 2013, pp. 852–859.

[23] C. C. Liu *et al.*, "High-performance integrated fan-out wafer level packaging (InFO-WLP): Technology and system integration," in *Proc. Int. Electron Devices Meeting*, 2012, pp. 14.1.1–14.1.4.

[24] M. A. Kabir, D. Petranovic, and Y. Peng, "Coupling extraction and optimization for heterogeneous 2.5D chiplet-package co-design," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design*, 2020, pp. 1–8.

[25] M. Hutner, R. Sethuram, B. Vinnakota, D. Armstrong, and A. Copperhall, "Special session: Test challenges in a chiplet marketplace," in *Proc. IEEE VLSI Test Symp.*, 2020, pp. 1–12.

[26] R. Wang and K. Chakrabarty, "Tackling test challenges for interposer-based 2.5-D integrated circuits," *IEEE Design Test*, vol. 34, no. 5, pp. 72–79, Oct. 2017.

[27] J.-S. Kim *et al.*, "A 1.2 V 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4 × 128 I/Os using TSV based stacking," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 107–116, Jan. 2012.

[28] W. Gomes *et al.*, "8.1 lakefield and mobility compute: A 3D stacked 10nm and 22FFL hybrid processor system in $12\times12\mathrm{mm}^2$, 1mm package-on-package," in *Proc. IEEE Int. Solid State Circuits Conf.*, 2020, pp. 144–146.

[29] *Samsung X-Cube*. Accessed: Dec. 3, 2020. [Online]. Available: https://news.samsung.com/global/samsung-announces-availability-of-its-silicon-proven-3d-ic-technology-for-high-performance-applications

[30] T. Brandtner, K. Pressel, N. Floman, M. Schultz, and M. Vogl, "Chip/package/board co-design methodology applied to full-custom heterogeneous integration," in *Proc. IEEE Electron. Compon. Technol. Conf.*, 2020, pp. 1718–1727.

[31] I. H.-R. Jiang, Y.-W. Chang, J.-L. Huang, and C.-P. Chen, "Intelligent design automation for 2.5/3D heterogeneous SoC integration," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design*, 2020, pp. 1–7.

[32] 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 Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017.

[33] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 6, pp. 1151–1164, Jun. 2020.

[34] K. Chang *et al.*, "Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, 2016, pp. 1–8.

[35] A. Chaudhuri *et al.*, "Advances in design and test of monolithic 3-D ICs," *IEEE Design Test*, vol. 37, no. 4, pp. 92–100, Aug. 2020.

[36] A. Todri-Sanial and C. S. Tan, *Physical Design for 3D Integrated Circuits*. Boca Raton, FL, USA: CRC Press, 2016.

[37] H. Park, B. W. Ku, K. Chang, D. E. Shim, and S. K. Lim, "Pseudo-3D approaches for commercial-grade RTL-to-GDS tool flow targeting monolithic 3D ICs," in *Proc. Int. Symp. Phys. Design*, 2020, pp. 47–54.

[38] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power benefit study for ultra-high density transistor-level monolithic 3D ICs," in *Proc. ACM Design Autom. Conf.*, May 2013, pp. 1–10.

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.

[39] J. Shi, D. Nayak, M. Ichihashi, S. Banna, and C. A. Moritz, "On the design of ultra-high density 14nm Finfet based transistor-level monolithic 3D ICs," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Jul. 2016, pp. 449–454.

[40] C. Yan and E. Salman, "Mono3D: Open source cell library for monolithic 3-D integrated circuits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 3, pp. 1075–1085, Mar. 2018.

[41] C. Yan and E. Salman, "Routing congestion aware cell library development for monolithic 3D ICs," in *Proc. Int. Conf. Rebooting Comput.*, Nov. 2017, pp. 1–4.

[42] C. Yan, S. Kontak, H. Wang, and E. Salman, "Open source cell library Mono3D to develop large-scale monolithic 3D integrated circuits," in *Proc. IEEE Proc. Int. Symp. Circuits Syst.*, May 2017, pp. 1–4.

[43] N. K. Sketopoulos, C. P. Sotiriou, and V. F. Pavlidis, "Metal stack and partitioning exploration for monolithic 3D ICs," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Jul. 2020, pp. 398–403.

[44] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, "Full chip impact study of power delivery network designs in monolithic 3D ICs," in *Proc. IEEE/ACM Proc. Int. Conf. Comput.-Aided Design*, 2014, pp. 565–572.

[45] H. Wei, T. F. Wu, D. Sekar, B. Cronquist, R. F. Pease, and S. Mitra, "Cooling three-dimensional integrated circuits using power delivery networks," in *Proc. IEEE Int. Electron Devices Meeting*, 2012, pp. 14.2.1–14.2.4.

[46] M. A. Iqbal and M. Rahman, "New thermal management approach for transistor-level 3D integtration," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf.*, 2017, pp. 1–3.

[47] S. K. Samal, S. Panth, K. Samadi, M. Saeidi, Y. Du, and S. K. Lim, "Adaptive regression-based thermal modeling and optimization for monolithic 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 35, no. 10, pp. 1707–1720, Oct. 2016.

[48] P. Shukla, A. K. Coskun, V. F. Pavlidis, and E. Salman, "An overview of thermal challenges and opportunities for monolithic 3D ICs," in *Proc. Great Lakes Symp. VLSI*, 2019, pp. 439–444.

[49] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 5, pp. 501–513, May 2006.

[50] C. Liu and S. K. Lim, "A design tradeoff study with monolithic 3D integration," in *Proc. Int. Symp. Qual. Electron. Design*, 2012, pp. 529–536.

[51] K. Chang, A. Koneru, K. Chakrabarty, and S. K. Lim, "Design automation and testing of monolithic 3D ICs: Opportunities, challenges, and solutions: (Invited paper)," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, 2017, pp. 805–810.

[52] A. Chaudhuri, S. Banerjee, H. Park, B. W. Ku, K. Chakrabarty, and S. Lim, "Built-in self-test for inter-layer vias in monolithic 3D ICs," in *Proc. IEEE Eur. Test Symp.*, 2019, pp. 1–6.

[53] E. J. Marinissen, T. McLaurin, and H. Jiao, "IEEE Std P1838: DfT standard-under-development for 2.5D-, 3D-, and 5.5D-SICs," in *Proc. IEEE Eur. Test Symp.*, 2016, pp. 1–10.

[54] A. Koneru and K. Chakrabarty, "An inter-layer interconnect BIST solution for monolithic 3D ICs," in *Proc. IEEE VLSI Test Symp.*, 2018, pp. 1–6.

[55] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable in-situ accelerator," in *Proc. IEEE/ACM Int. Symp. Microarchit.*, Oct. 2017, pp. 288–301.

[56] S. Angizi and D. Fan, "ReDRAM: A reconfigurable processing-in-DRAM platform for accelerating bulk bit-wise operations," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2019, pp. 1–8.

[57] W. Huangfu, S. Li, X. Hu, and Y. Xie, "RADAR: A 3D-ReRAM based DNA alignment accelerator architecture," in *Proc. ACM/ESDA/IEEE Design Autom. Conf.*, Jun. 2018.

[58] V. Pavlidis, I. Savidis, and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*. Cambridge, MA, USA: Morgan Kaufmann, 2017.

[59] D. C. Stow, I. Akgun, W. Huangfu, Y. Xie, X. Li, and G. H. Loh, "Efficient system architecture in the era of monolithic 3D: Dynamic inter-tier interconnect and processing-in-memory," in *Proc. ACM/IEEE Design Autom. Conf.*, Jun. 2019, p. 100.

[60] X. Jiang, X. Liu, L. Xu, P. Zhang, and N. Sun, "A reconfigurable accelerator for Smith–Waterman algorithm," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 54, no. 12, pp. 1077–1081, Dec. 2007.

[61] F.-K. Hsueh *et al.*, "TSV-free finfet-based monolithic 3D+-IC with computing-in-memory SRAM cell for intelligent IoT devices," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2017, pp. 306–309.

[62] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2553–2561.

[63] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. Annu. Int. Symp. Comput. Archit.*, 2017, pp. 1–12.

[64] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[65] T. Chen *et al.*, "DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," *ACM SIGARCH Comput. Archit. News*, vol. 42, no. 1, pp. 269–284, 2014.

[66] A. Parashar *et al.*, "SCNN: An accelerator for compressed-sparse convolutional neural networks," *ACM SIGARCH Comput. Archit. News*, vol. 45, no. 2, pp. 27–40, 2017.

[67] Y. Yu and N. K. Jha, "SPRING: A sparsity-aware reduced-precision monolithic 3D CNN accelerator architecture for training and inference," *IEEE Trans. Emerg. Topics Comput.*, early access, Jun. 18, 2020, doi: 10.1109/TETC.2020.3003328.

[68] Y. Yu and N. K. Jha, "A monolithic 3D hybrid architecture for energy-efficient computation," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 4, no. 4, pp. 533–547, Oct.–Dec. 2018.

[69] K. Chang, D. Kadetotad, Y. Cao, J.-S. Seo, and S. K. Lim, "Monolithic 3D IC designs for low-power deep neural networks targeting speech recognition," in *Proc. IEEE/ACM ISLPED*, 2017, pp. 1–6.

[70] C. T. Do, J. H. Choi, Y. S. Lee, C. H. Kim, and S. W. Chung, "Enhancing matrix multiplication with a monolithic 3D based scratchpad memory," *IEEE Embedded Syst. Lett.*, early access Jun. 12, 2020, doi: 10.1109/LES.2020.3001954.

[71] P. Shukla, S. S. Nemtzow, V. F. Pavlidis, E. Salman, and A. K. Coskun, "Temperature-aware optimization of monolithic 3D deep neural network accelerators," in *Proc. ACM Asia South Pac. Design Autom. Conf.*, 2021, pp. 709–714.

[72] V. J. Reddi *et al.*, "MLPerf inference benchmark," 2019. [Online]. Available: arXiv:1911.02549.

[73] J. Dofe, Q. Yu, H. Wang, and E. Salman, "Hardware security threats and potential countermeasures in emerging 3D ICs," in *Proc. Great Lakes Symp. VLSI*, 2016, pp. 69–74.

[74] Y. Xie, C. Bao, C. Serafy, T. Lu, A. Srivastava, and M. Tehranipoor, "Security and vulnerability implications of 3D ICs," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 2, no. 2, pp. 108–122, Apr.–Jun. 2016.

[75] C. Yan, J. Dofe, S. Kontak, Q. Yu, and E. Salman, "Hardware-efficient logic camouflaging for monolithic 3-D ICs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 6, pp. 799–803, Jun. 2018.

[76] J. Dofe, C. Yan, S. Kontak, E. Salman, and Q. Yu, "Transistor-level camouflaged logic locking method for monolithic 3D IC security," in *Proc. IEEE Asian Hardw.-Oriented Security Trust*, 2016, pp. 1–6.

[77] R. J. Masti, D. Rai, A. Ranganathan, C. Müller, L. Thiele, and S. Capkun, "Thermal covert channels on multi-core platforms," in *Proc. USENIX Security Symp.*, 2015, pp. 865–880.

[78] P. Gu, D. Stow, R. Barnes, E. Kursun, and Y. Xie, "Thermal-aware 3D design for side-channel information leakage," in *Proc. IEEE Int. Conf. Comput. Design*, 2016, pp. 520–527.

[79] J. Knechtel and O. Sinanoglu, "On mitigation of side-channel attacks in 3D ICs: Decorrelating thermal patterns from power and activity," in *Proc. ACM/EDAC/IEEE Design Autom. Conf.*, 2017, pp. 1–6.

[80] D. B. Bartolini, P. Miedl, and L. Thiele, "On the capacity of thermal covert channels in multicores," in *Proc. Eur. Conf. Comput. Syst.*, 2016, pp. 24:1–24:16.

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:26:24 UTC from IEEE Xplore. Restrictions apply.