Abstract

Monolithic 3D (MONO3D) integration technology has emerged as a promising alternative to conventional transistor scaling, addressing the limitations of traditional two-dimensional (2D) integration. Unlike conventional approaches where multiple metal interconnect layers are fabricated over a single active transistor layer, MONO3D enables the fabrication of multiple transistor layers on a single substrate. This approach utilizes vertical interconnects-monolithic inter-tier vias (MIVs)-that possess physical dimensions comparable to conventional metal vias. As a result, MONO3D offers unprecedented integration density and significantly higher bandwidth communication, which are essential for supporting a wide range of data-centric and high-performance applications. Despite increasing academic and industrial interest, the commercialization of MONO3D integrated circuits remains unrealized due to challenges in fabrication, thermal management, and design automation. This tutorial brief presents a comprehensive overview of MONO3D technology, covering recent advances in fabrication processes, design methodologies, and testing strategies. Additionally, the paper explores emerging application domains such as in-memory computing, deep neural network acceleration, and hardware security, which can greatly benefit from the unique features of MONO3D. The brief also outlines current limitations and identifies future research directions that are critical for enabling the widespread adoption of monolithic 3D integration in the semiconductor industry

Table of Content

List of Figures

List of Abbreviations

FEOL	Front-End-Of-Line
BEOL	Back-End-Of-Line
TSV	Through-Silicon Via
MONO3D	Monolithic 3D Integration
EDA	Electronic Design Automation
SICs	Stacked Integrated Circuits
DFT	Design-For-Test
ILD	Inter-Layer Dielectric
MAC	Multiply-Accumulate
DNNs	Deep Neural Networks
MIV	Monolithic Inter-Tier Via

1. Introduction

In traditional semiconductor manufacturing, the fabrication of integrated circuits (ICs) follows a well-established two-dimensional (2D) approach. This process begins with the front-end-of-line (FEOL) phase, where individual transistors are patterned on the surface of a silicon wafer. Following this, the back-end-of-line (BEOL) phase is used to deposit and pattern multiple layers of metal interconnects that connect the transistors into functional circuits. While this method has seen continuous scaling and optimization, it is now approaching physical and technological limits, particularly as device dimensions shrink into the sub-10 nanometer range.To address these scaling limitations, the semiconductor industry has increasingly turned toward three-dimensional (3D) integration technologies. In recent years, techniques such as through-silicon via (TSV)-based 3D stacking and 2.5D interposer-based integration have enabled significant advancements in bandwidth, form factor, and performance. These methods involve stacking multiple dies or placing them side-by-side on a shared substrate, connected through large vertical or lateral interconnects. Despite their benefits, these approaches are constrained by the relatively large size of TSVs and the complexity of aligning and bonding multiple dies, which limits their granularity and integration density.

Monolithic 3D integration (MONO3D) represents a breakthrough in this context. Unlike conventional 3D integration methods, MONO3D technology enables the sequential fabrication of multiple layers of transistors on a single silicon substrate. Each layer is processed using standard CMOS techniques, with planarization and thin-film deposition steps in between, allowing for precise layer-to-layer alignment and tight vertical integration. This process eliminates the need for die stacking or TSVs, offering a new level of granularity in vertical design.

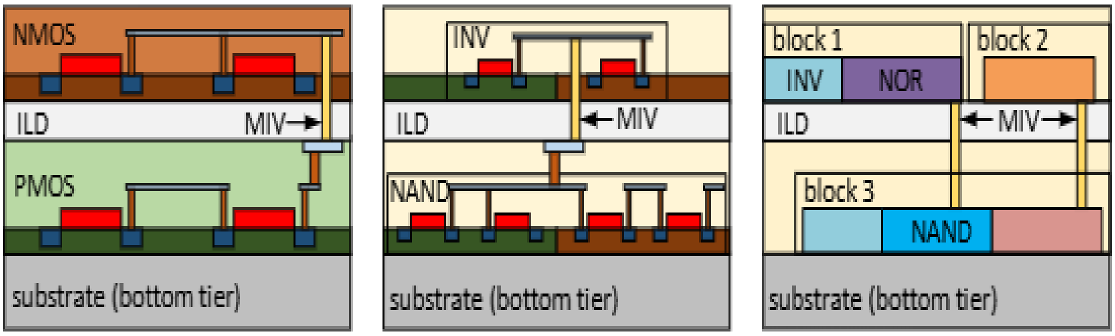


Figure 1.1: Three different design styles for monolithic 3D (M3D) technology

Monolithic 3D integration can be implemented at the transistor, gate, and block levels, each offering different levels of granularity and design flexibility. In transistor-level monolithic 3D integration, which is the focus of this study, nMOS and pMOS transistors of a circuit are placed in separate tiers-typically, the pull-down network (nMOS transistors) resides in one tier, while the pull-up network (pMOS transistors) is located in another. This fine-grained approach allows for intra-cell connections using monolithic inter-tier vias (MIVs) and provides opportunities for independent optimization of device characteristics in the top and bottom tiers. Despite its complexity, this level of integration can still be supported by existing electronic design automation (EDA) tools with some modifications.

In gate-level monolithic 3D integration, entire standard cells or gates are split across tiers within a functional block. MIVs are used for inter-cell communication, enabling a moderate level of 3D integration that balances design complexity and performance benefits. Finally, block-level monolithic 3D integration adopts a more coarse-grained strategy where different functional blocks of the IC, such as memory or logic modules, are placed in different tiers. This approach simplifies design and manufacturing but offers less opportunity for the performance and area optimizations achievable with finer-grained techniques.

An essential feature of MONO3D is the use of monolithic inter-tier vias (MIVs), which are significantly smaller than TSVs and are formed with nanometre-scale dimensions. These vias enable dense and energy-efficient vertical connections between transistor layers, supporting finer partitioning of logic and enabling more compact, power-efficient circuit designs. The compatibility of MONO3D with standard CMOS processes also makes it an attractive option for future large-scale manufacturing. As the demand for higher performance, lower power, and increased functionality continues to grow, monolithic 3D integration is poised to play a central role in advancing integrated circuit design beyond the limits of traditional 2D scaling and heterogeneous 3D integration techniques.

2. Literature survey

[1] **"An overview of the development of a GPU with inte grated HBM on silicon interposer," in Proc. IEEE Electron. 2017, Compon. Technol. Conf.**

The paper presents the technical implementation of a GPU integrated with High Bandwidth Memory (HBM) using a 2.5D silicon interposer. This architecture enables ultra-wide I/O interfaces between the GPU and HBM stacks, significantly improving memory bandwidth while reducing power consumption. The silicon interposer serves as a passive substrate with fine-pitch wiring, facilitating high-speed signaling and efficient thermal dissipation. The design includes through-silicon vias (TSVs) in the HBM stacks and micro-bumps connecting the GPU and memory to the interposer, allowing tight integration in a compact footprint. This configuration addresses the bottlenecks of traditional memory systems by delivering enhanced performance and power efficiency, critical for modern high-performance computing applications.

[2] **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.**

The paper presents IEEE Std P1838, a Design-for-Test (DfT) standard for 2.5D, 3D, and 5.5D stacked integrated circuits (SICs). It introduces a modular test architecture to support efficient and interoperable testing of dies from multiple vendors. Key components include a serial control mechanism (IEEE 1149.1), a Die Wrapper Register (IEEE 1500), and a Flexible Parallel Port (FPP) for scalable test access. These features enable high-speed testing across different stages like pre-bond, mid-bond, and post-bond. The standard addresses complex test challenges in advanced multi-die systems.

[3] **"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.**

The paper explores metal stack and partitioning techniques for optimizing monolithic 3D IC designs. It examines how different metal layer configurations affect performance, power efficiency, and thermal management. The study highlights the influence of partitioning on interconnectivity and signal integrity. Simulation results demonstrate that optimal metal stack configurations improve power dissipation and heat management. The paper provides insights for achieving efficient, manufacturable 3D IC designs with balanced performance and complexity.

3. Milestones in Mono3D Technology

The image illustrates the historical timeline and technological evolution of monolithic 3D (Mono3D) integration from its early developments in the late 1980s to major advancements up to 2018. The timeline starts with the **first monolithic integration (1989–1992)**, where high-temperature processes (>900°C) were required, limiting stacking flexibility.

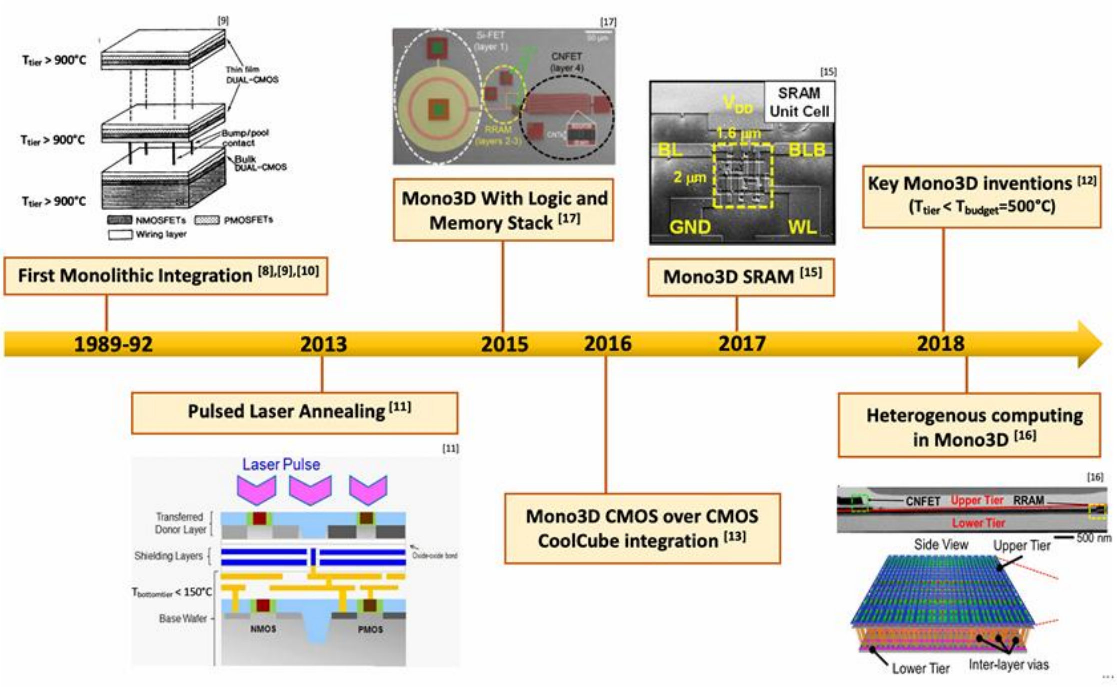


Figure 3.1: Chronological timeline of primary developments in the fabrication process of MONO3D ICs.

A major breakthrough in 2013 was the introduction of Pulsed Laser Annealing, enabling low thermal budget processing (Tmax < 500°C), essential for stacking transistors without damaging the underlying layers.

In 2015, Mono3D with logic and memory stack was demonstrated, followed by Mono3D CMOS over CMOS integration (CoolCube) in 2016, which further refined 3D stacking with better performance and efficiency.

By 2017, Mono3D SRAM integration showed practical implementations in memory, and in 2018, the field reached a stage of heterogeneous computing in Mono3D, allowing different functionalities (logic, memory, etc.) to be stacked vertically for high-density and high-performance systems.

4. Open-Source Cell Library for Monolithic 3D ICs

The primary characteristics of the proposed cell library are described in Section 4.1. The design flow to integrate the proposed library into the design process is discussed in Section 4.2. Cell-level simulation results and comparison of 3D cells with 2D cells are provided in Section 4.3.

4.1 Library Development

An open-source standard cell library was developed specifically for transistor-level monolithic 3D (M3D) integrated circuits using 45nm technology. The library is built upon the FreePDK45 process design kit (PDK), a widely-used open-source 45nm PDK developed by North Carolina State University (NCSU) [21, 27].

Architecture and Design Methodology

Mono3D comprises two vertically integrated tiers:

- The top tier incorporates the nMOS transistors, forming the pull-down networks of CMOS gates.
- The bottom tier includes the pMOS transistors, forming the pull-up networks.

One critical fabrication constraint for monolithic 3D integration is the thermal budget. The top tier must be processed at temperatures below 500–600 °C to prevent damage to the transistors in the previously fabricated bottom tier [15]. Consequently, due to lower thermal processing, the device quality in the top tier may degrade. To mitigate performance impacts, pMOS transistors—which inherently have lower mobility—are placed in the bottom tier, where they can be fabricated under optimal thermal conditions.

Each standard cell includes Monolithic Inter-Tier Vias (MIVs) to connect transistors between the two tiers. These MIVs have a width of 50 nm and a height of 215 nm. The device characteristics (e.g., transistor models, metal layers) remain identical to those in the original 2D FreePDK45 library, meaning the impact of thermal degradation is not directly modeled. However, the library structure is designed to allow the integration of novel device models and manufacturing steps to better reflect advanced M3D integration processes. This flexibility enables researchers to explore system-level impacts of various device behaviors and process variations.

The bottom tier is allocated two metal layers—metal1_btm and metal2_btm—primarily used for intra-cell routing. An inter-layer dielectric (ILD) of 100 nm thickness separates the two tiers, minimizing inter-tier capacitive coupling, as validated. The top tier retains the full 10 metal layers as provided by the 2D FreePDK45 library. Intra-cell connections that span both tiers are established using MIVs, which are strategically placed to minimize interconnect length and reduce the overall cell height.

The current version of the Mono3D library includes 20 standard cells, which are AND2X1, INVX2, AOI21X1, INVX4, BUFX2, LATCHNEG, BUFX4, MUX2X1, CLKBU1, NAND2X1, CLKBUF2, NOR2X1, CLKBUF3, OAI21X1, DFFPOSX1, OR2X1, FILL, XNOR2X1, INVX1 and XOR2X1 all cells are designed using a full-custom design methodology with a cell stacking technique, allowing for efficient vertical integration of devices.

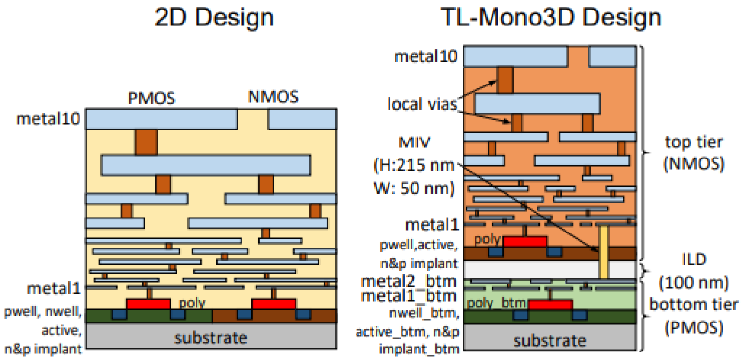


Figure 4.1: Conventional 2D and transistor-level monolithic 3d technology with two tiers.

Unlike previous monolithic 3D (M3D) designs where the power and ground rails from different tiers overlap [22, 23], the Mono3D library adopts a non-overlapping rail configuration. Specifically, the power rail is positioned at the top of the bottom tier, while the ground rail is located at the bottom of the top tier. These tier-specific rails are connected to the overall system-level power distribution network using power and ground rings that are implemented during the placement and routing stages of physical design. To support vertical integration, a dedicated routing track is allocated for intra-cell monolithic inter-tier vias (MIVs), which are strategically distributed within each standard cell. This design choice helps minimize interconnect length, reduce routing congestion, and optimize the overall cell height. While the standard 2D NandGate library employs 14 routing tracks per cell [28], the Mono3D library introduces three distinct variants-Mono3D v1, v2, and v3-with 8, 9, and 10 routing tracks per cell.

4.2 Design Flow

The design flow adopted in this work for monolithic 3D integration involves several key modifications to standard toolchains and processes to accommodate the unique structural and electrical characteristics of 3D ICs. A new technology file (.tf) was created for the Mono3D cell library, which defines all the necessary new layers such as additional interconnect metals, vias, inter-layer dielectric (ILD), and monolithic inter-tier vias (MIVs). Alongside this, a modified display resource file (.drf) was generated to facilitate the full-custom layout development of 3D standard cells. The extracted netlist includes MIVs and accurately reflects the interconnection between nMOS transistors in the top tier and pMOS transistors in the bottom tier.

To ensure accurate parasitic modeling, the RC extraction rule file was updated to recognize the newly introduced metal layers, device tiers, and MIVs. In this process, various parasitic components such as intrinsic plate capacitance, fringe capacitance, and near-body coupling capacitance were considered between silicon and metal, and between adjacent metal layers. Each MIV was characterized with a resistance of 5.5 ohms and a capacitance of 0.04 femtofarads, based on previously reported experimental data. It is important to note that tier-to-tier coupling capacitance was excluded from the extraction model, as it has been experimentally shown to be negligible with a 100 nm thick ILD.

Following RC extraction, each 3D cell was characterized using the Encounter Library Characterizer (ELC) tool to generate accurate timing and power lookup tables. To verify the integrity of this characterization, the extracted netlists were also simulated using HSPICE. The resulting .lib file was then converted into the .db format, making it compatible with commercial electronic design automation (EDA) tools for circuit synthesis, placement, clock tree synthesis, and routing. Since all the I/O pins of the Mono3D cells are placed in the top tier, conventional physical design tools can be utilized without additional customization.

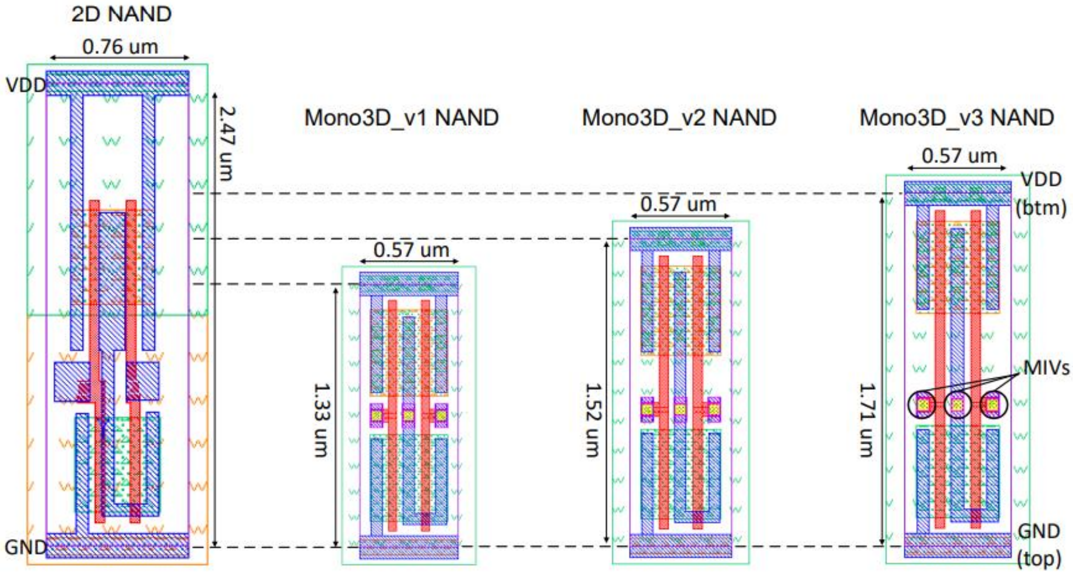


Figure 4.2: Comparison of the layout views of a NAND gate

For physical verification, design rule check (DRC), layout versus schematic (LVS), and parasitic extraction (PEX) were all performed using Calibre [29]. The DRC rule file was modified to incorporate the new elements introduced by the monolithic 3D structure, including additional metal layers, MIVs, and ILD. For instance, the minimum spacing between two MIVs was defined as 120 nm, resulting in an MIV pitch of 170 nm. The LVS rule file was also updated to enable the tool to distinguish transistors placed in different tiers, ensuring that the device hierarchy and connectivity between tiers were correctly verified.

4.3 Cell-Level Evaluation

The cell-level area improvements achieved through monolithic 3D (M3D) integration are illustrated in Figure 6. As shown, the reduction in cell area varies significantly based on the specific standard cell and Mono3D variant. For Mono3D v1, area reductions range from 6.5% to 64.1%, while Mono3D v2 sees reductions between -6.9% and 59.0%, and Mono3D v3 ranges from -13.5% to 53.8%. On average, Mono3D v1, v2, and v3 achieve area reductions of 32%, 22%, and 14%, respectively. Negative percentages indicate cases where the cell area increased compared to the 2D counterpart, typically occurring when the decrease in cell height leads to a disproportionate increase in cell width. Although cell height is significantly reduced in the 3D versions, the overall area reduction is less dramatic due to slight increases in cell width required to accommodate MIVs and additional intra-cell routing within the compressed footprint.

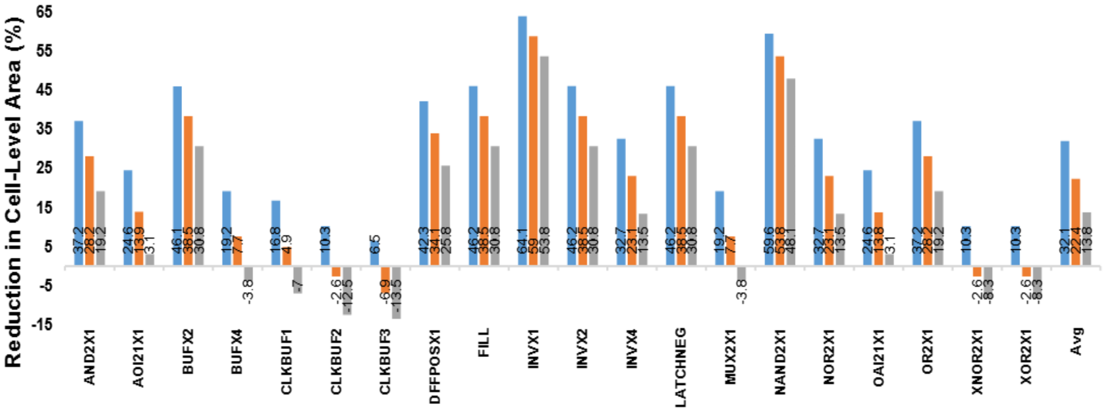


Figure 4.3: The percentage area reduction for each monolithic 3D cell compared to 2D cells is reported for Mono3D v1, v2, and v3.

To evaluate delay and power performance, HSPICE simulations were conducted on extracted 3D netlists and compared to conventional 2D technology. Simulations were performed at a supply voltage of 1.1 V, transition time of 50 ps, and temperature of 27°C. The results, summarized in Table II, show that Mono3D v1 cells experience an average increase in propagation delay of 2.15%, with Mono3D v2 and v3 showing increases of 3.22% and 3.78%, respectively. In contrast, average power consumption is reduced by 0.93% in Mono3D v1, 0.46% in Mono3D v2, and 0.08% in Mono3D v3 compared to 2D cells. The slight increase in delay is attributed to the denser layout of 3D cells, which introduces additional coupling capacitance and MIV impedance. However, some cells-such as the D-Flip-Flop-benefit from reduced delay and power consumption due to their longer average interconnect lengths, which are effectively optimized in the monolithic 3D architecture. It is important to note that changes in delay and power at the cell level are highly dependent on specific cell layouts, interconnect configurations, and the number and placement of MIVs.

5. Mono-3D Fabrication Process

A chronological timeline illustrating the primary developments in the fabrication process of MONO3D technology is shown in Fig. 1. The idea of sequentially constructing multiple transistor layers on a single substrate date back to the late 1980s, when initial studies demonstrated the use of laser beam recrystallization to form multiple silicon-on-insulator (SOI) layers over a p-type substrate. These early implementations typically included a single type of transistor per device layer, which helped reduce the complexity of well formation and ion implantation processes. The devices were fabricated under high temperatures, reaching up to 900°C, and utilized doped polysilicon interconnects due to their high thermal stability. Proper operation of the transistors was confirmed through I-V measurements, although a major drawback was a roughly 4× increase in threshold voltage variation in the SOI devices.

One of the major challenges in MONO3D fabrication is ensuring the reliability of the devices in the bottom layers, which can degrade due to the thermal exposure required for building transistors in the upper layers. Additionally, routability becomes a significant issue due to the high density of integration, necessitating the incorporation of multiple low-resistivity interconnect and low-k dielectric layers for the first device layer. These concerns enforce a strict thermal budget constraint, limiting the processing temperature for upper layers to below 500°C.

Many conventional front-end-of-line (FEOL) steps, such as impurity activation and annealing to repair crystal defects, typically require temperatures much higher than 500°C. To overcome this, pulsed laser annealing has been explored as a viable alternative. In this method, laser pulses with durations below 100 nanoseconds are used to momentarily elevate the temperature of the upper layers above 1400°C. Thin shielding layers are employed to ensure that the underlying layers remain below 150°C. This enables effective recrystallization and activation, allowing for the formation of low-resistance polysilicon gates in the upper transistor layers. Another key enabler is the development of high-quality silicon epitaxy at low temperatures. This process involves replacing traditional surface preparation techniques with a combination of dry and wet etching, enabling the growth of high-crystallinity and selective silicon epitaxial layers at around 500°C.

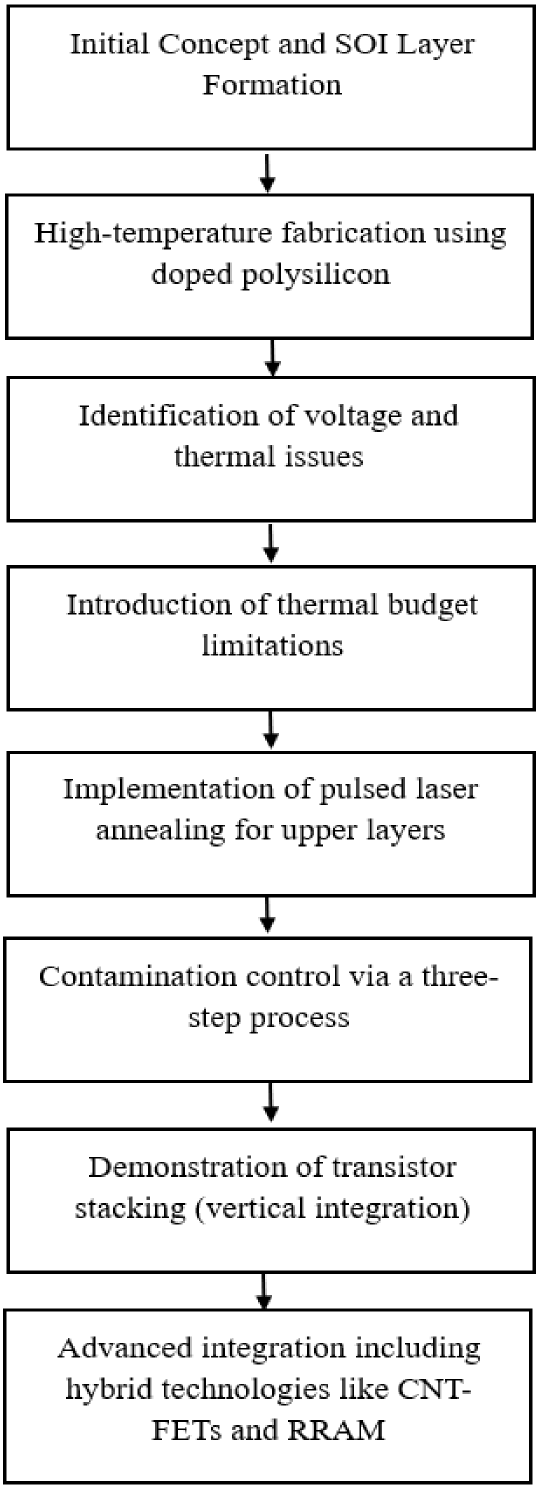

Initial Concept and SOI Layer Formation
High-temperature fabrication using doped polysilicon
Identification of voltage and thermal issues
Introduction of thermal budget limitations
Implementation of pulsed laser annealing for upper layers
Contamination control via a three-step process
Demonstration of transistor stacking (vertical integration)
Advanced integration including hybrid technologies like CNT-FETs and RRAM

Figure 5.1: Flow Chart

Fabrication of upper-layer devices also presents contamination risks, especially when wafers are reintroduced to FEOL processes after undergoing backend processing involving interconnects and vias. To address this, a three-step contamination control strategy has been proposed. First, the wafer's bevel edge, identified as the most critical contamination source, is etched to remove residual metals. Second, a wet cleaning process is applied to decontaminate the wafer. Finally, the bevel is encapsulated to prevent future contamination during subsequent steps.

These advancements have collectively enabled the demonstration of MONO3D integrated circuits (ICs) with relatively low design complexity. Notably, transistor-on-transistor integration has been achieved with nanometer-scale alignment accuracy. Further milestones include the successful MONO3D integration of SRAM arrays exhibiting stable operation, as well as the incorporation of heterogeneous technologies-such as carbon nanotube field-effect transistors and non-volatile resistive memory-alongside conventional silicon-based devices on a unified MONO3D platform.

6. Mono3D Design Methodologies

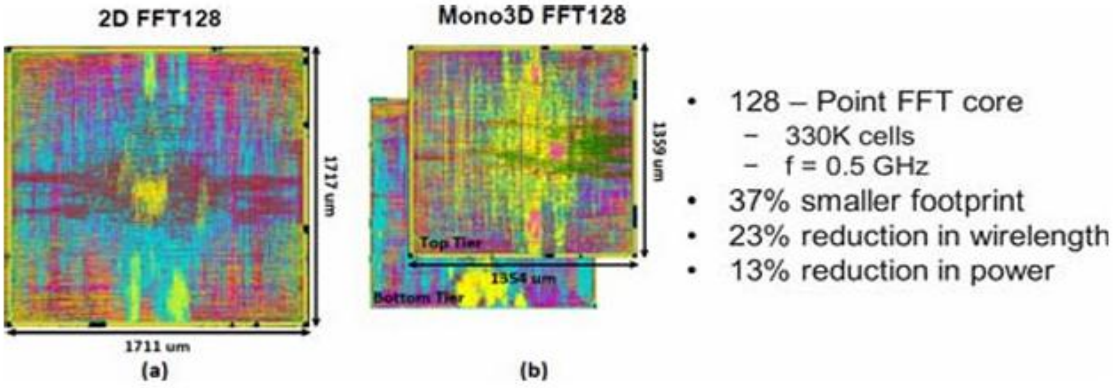
In this section, the latest advances related to physical design, thermal integrity, and testing methodologies of MONO3D tech nology are described. A broad qualitative comparison of these challenges (including commercial status) for 2.5D, TSV-based 3D, and MONO3D integration is provided in Table

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].

6.1 Physical Design

In recent years, the development of pseudo-3D tools for monolithic 3D (MONO3D) integration has attracted significant attention. These approaches typically involve modifying technology or process files to adapt existing 2D engines for generating efficient 3D layouts. One such tool, Shrunk-2D (S2D), represents the first commercial-quality solution in this area. It operates by shrinking the entire design and standard cell dimensions by a factor of 1/√N, where N is the number of tiers. The metal width and pitch are similarly scaled to maintain a consistent resistance-capacitance (RC) per unit length. The 2D place-and-route (P&R) engine is applied to the scaled design, which is then expanded and partitioned into different tiers. However, S2D is limited in accurately estimating RC parasitics and often requires commercial 2D engines to handle geometries beyond their capability. To overcome these limitations, Compact-2D was proposed, which provides improved timing characterization and fewer cell legalization issues. Nonetheless, it lacks support for simultaneous timing closure across tiers, potentially degrading overall performance. Another approach, Cascade2D, performs design-aware RTL-level partitioning and allows concurrent placement, routing, and timing closure across tiers, significantly enhancing timing performance.

A comparative analysis of S2D, Compact-2D, and Cascade2D using the RISC-V Rocketcore processor under identical 3D footprints revealed key trade-offs. Although S2D and Compact-2D suffer from higher total negative slack compared to Cascade2D, their total wirelengths were 12.2% and 10.5% lower than conventional 2D, respectively. Moreover, their power consumption was reduced by 7.1% and 8%, respectively. On the other hand, Cascade2D demonstrated better timing characteristics but at the cost of 21.4% higher wirelength and 7.5% higher power consumption than 2D, mainly due to its tier-partitioning and MIV (Monolithic Inter-tier Via) strategies being less effective for flat gate-level designs.



- 128 – Point FFT core
  - 330K cells
  - f = 0.5 GHz
- 37% smaller footprint
- 23% reduction in wirelength
- 13% reduction in power

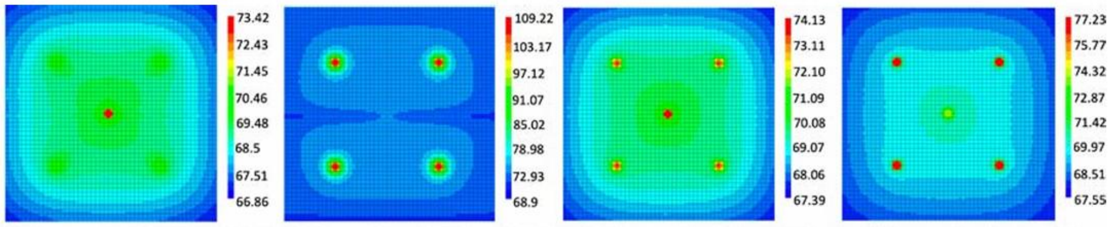
Figure 6.1: A fully placed and routed 45 nm technology using: (a) conventional 2D, (b) transistor-level MONO3D technology with two layers.

Routing congestion is another major concern in MONO3D ICs, as chip footprints are significantly reduced (by 20% to 40%) without a proportional increase in available metal layers. This issue was addressed by Yan et al. through the development of a MONO3D cell library and a two-layer process design kit (PDK) in 45 nm technology. In this architecture, nMOS transistors and I/O pins are placed on the upper layer, while pMOS transistors are placed on the bottom layer. These are connected using intra-cell MIVs, enabling continued use of existing 2D placement tools. A demonstration involving a 128-point FFT design running at 500 MHz illustrated the effectiveness of this approach. Increasing routing tracks per cell from 8 to 10 slightly increased the overall footprint but reduced wirelength due to decreased congestion and improved timing owing to lower coupling capacitance. Compared to the 2D implementation, the MONO3D version achieved a 37% reduction in footprint and 13% lower power consumption.

Using this MONO3D cell library, further optimization through successive in-place optimizations (IPOs) was explored. Partitioning methods such as greedy bin-based Fidducia-Mattheyses and displacement-based legalizers were employed, and a single back-end-of-line (BEOL) connection between tiers was shown to yield better power, performance, and area (PPA) outcomes. Despite promising academic advancements in end-to-end MONO3D design flows, the development of robust, commercially available EDA tools that fully harness the integration, performance, and power benefits of MONO3D remains an open challenge. Additionally, thermal considerations must be integrated into the physical design process to ensure optimal results across all critical design metrics.

6.2 Thermal Integrity

One of the major limitations of 3D integrated circuits (ICs) is the difficulty in dissipating heat effectively from tiers that are farther away from the heat sink or heat spreader. In the case of monolithic 3D (MONO3D) ICs, thermal issues are further exacerbated due to higher device integration densities, increased routing congestion, and strong inter-tier thermal coupling. These characteristics distinguish MONO3D technology significantly from through-silicon via (TSV)-based 3D ICs. Several techniques have been proposed in the literature to address thermal challenges in MONO3D systems. For instance, the careful design of power delivery networks (PDNs) has been shown to reduce overall chip temperature by up to 5%. Conversely, neglecting the impact of PDNs can lead to an overestimation of chip temperatures. In another approach, nano-pillars were strategically placed during the design phase in transistor-level MONO3D systems to dissipate heat from specific hot spot regions, resulting in up to 53% temperature reduction as demonstrated through finite element method simulations.



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

Figure 6.2: Thermal maps of the TSV-based 3D and MONO3D ICs

Samal et al. modelled a two-tier MONO3D system and demonstrated strong inter-tier thermal coupling caused by the presence of a thin dielectric layer between tiers. A non-linear regression model was employed to estimate chip temperatures under such conditions. However, this model neglected lateral heat flow due to the extremely thin tier structure. In contrast, TSV-based 3D systems exhibited greater vertical thermal resistance and higher temperatures in tiers farther from the heat spreader, primarily due to a thicker bonding layer and a longer vertical heat flow path. To further evaluate thermal behavior, both two-tier MONO3D and TSV-based systems were modeled using HotSpot-6.0, an architectural-level thermal simulator. Synthetic power profiles were created by placing one hot spot in the bottom tier (closer to the heat sink) and four hot spots in the top tier (farthest from the heat sink), each with a power density of 750 W/cm^2. Results showed that thermal coupling in MONO3D aided heat dissipation from the upper tier to the heat sink, as evidenced by the lower temperatures recorded. However, the same thermal coupling can also lead to propagation of hot spots from the upper tier to the lower tier, highlighting the bidirectional nature of thermal influence.

Additionally, the MONO3D ICs demonstrated higher hot spot localization and reduced lateral thermal coupling. Various power profiles were simulated by altering the distribution, size, and area of hot spots. While lateral heat flow was found to be limited in MONO3D, it remains significant enough that ignoring it could lead to an underestimation of temperatures by approximately 4°C in tiers located away from the heat sink. These observations underscore the importance of considering specific thermal behaviors and constraints when designing MONO3D ICs to ensure proper thermal integrity and system reliability.

6.3 Design-for-Test

MONO3D integrated circuits (ICs) present distinct testing challenges due to the presence of multiple stacked transistor layers and a dense network of monolithic inter-tier vias (MIVs). The aggressive scaling of the inter-layer dielectric thickness and the extremely high MIV density—reaching up to 30 million MIVs per mm^2—are primary contributors to functional and timing faults in MONO3D systems. Fault models for testing MIVs are largely based on interconnect fault models in traditional 2D ICs and are typically classified into shorts, opens, and stuck-at faults. However, conventional automatic test pattern generation (ATPG)-based methods are often inadequate, as test vectors must traverse multiple tiers and MIVs, imposing significant constraints on the testing tools. This limitation has necessitated the development of built-in self-test (BIST) solutions specifically tailored for MONO3D technology.

To improve MIV observability and controllability, one approach involves using die-wrapper register cells at both ends of an MIV. A notable BIST method proposed by Koneru et al. includes the use of interface scan cells connected to a twisted-ring counter on a dedicated test layer, specifically targeting opens and shorts in MIVs. This technique has demonstrated potential cost savings of up to 40% per die when compared to similar methods used for TSV-based 3D ICs, which require additional die-wrapper registers. Despite the benefits, these methods introduce notable area overhead. To address this, a more efficient dual-BIST architecture has recently been proposed that utilizes XOR gates for the detection of opens, shorts, and stuck-at faults. This architecture achieves complete fault coverage using just two test patterns and compresses the test output into compact 2-bit signatures. Reported designs using this dual-BIST approach show maximum area and power overheads of only 2.6% and 9%, respectively, compared to non-BIST implementations.

Importantly, traditional pre-bond test strategies used for TSV-based 3D ICs are unsuitable for MONO3D ICs because bare MIVs cannot be exposed, and wafer-probing technologies cannot support MIV pitches as small as 100 to 200 nm. Post-bond test strategies also fail to scale effectively due to the vastly higher number of MIVs in MONO3D systems. As a result, several novel BIST methodologies have been proposed to address these limitations. Despite significant progress, testing MONO3D ICs remains an active area of research, with continued efforts required to develop scalable, low-cost, and effective test solutions.

7. Emerging Applications for Mono3D

Emerging applications for MONO3D include areas where high integration density, low power consumption, and high bandwidth are critical. These applications span near- and in-memory computing, deep neural network (DNN) accelerators, and hardware security solutions. MONO3D's fine-grained vertical integration enables new computing architectures and improved performance in data-intensive and security-sensitive systems.

7.1 Near- and In-Memory Computing

Near- and in-memory computing have emerged as promising approaches to mitigate the limitations imposed by the traditional "memory wall"—a bottleneck in computing systems caused by the latency and energy required for data movement between the processor and memory. MONO3D integrated circuits are particularly well-suited for these paradigms due to their extremely high MIV (Monolithic Inter-tier Via) density and the physical proximity of their stacked device layers. In MONO3D architectures, the distinction between near- and in-memory computing becomes less clear because of the minimal vertical separation between logic and memory layers. Instead, these paradigms are better distinguished by their partitioning granularity. Near-memory computing typically employs block-level partitioning, leveraging the dense, low-delay MIVs for fast inter-layer communication. In contrast, in-memory computing adopts transistor-level partitioning, allowing bit-wise logic operations to be executed directly within the memory layer by utilizing additional transistors on adjacent tiers.

A noteworthy example of near-memory computing using MONO3D involves the integration of a processing tier with a resistive RAM (ReRAM) layer, extending the RADAR accelerator into three dimensions. This architecture significantly boosts the execution speed of the Smith-Waterman algorithm—an essential computation in bioinformatics—by three orders of magnitude compared to a 2D implementation. In comparison, an FPGA-based acceleration of the same algorithm achieved only a 330× improvement. On the other hand, an innovative in-memory computing cell architecture featuring nine transistors (9T) has been proposed, capable of performing logical operations (e.g., NAND, NOR, XOR) along with bit storage within a single memory cycle. The design maintains the memory array's density by placing the logic transistors on a separate tier, achieving a 51% area saving over its 2D counterpart while preserving performance. These advancements underscore the untapped potential of MONO3D architectures for enabling non-Von Neumann computing paradigms.

7.2 Deep Neural Networks (DNNs)

Deep Neural Networks (DNNs) are central to modern AI applications such as image classification, speech recognition, and natural language processing. These workloads demand high computational throughput and memory bandwidth, which has motivated the development of specialized DNN accelerators. However, DNNs are also energy-intensive due to extensive data movement and arithmetic operations. MONO3D ICs present a compelling solution to this problem by offering superior interconnect density and enhanced power/performance trade-offs compared to traditional 2D architectures. Consequently, recent efforts have focused on designing MONO3D-based DNN accelerators to improve energy efficiency.

One such work by Yu et al. employs sparsity encoding of CNN activations and weights, interfaced with a MONO3D-based non-volatile RAM to maximize bandwidth through MIVs. This resulted in execution time, power dissipation, and energy efficiency improvements of up to 16×, 4.5×, and 69×, respectively, when compared to Nvidia's GTX 1080 Ti GPU. Other studies have explored different MONO3D partitioning techniques to optimize accelerator performance. For example, Chang et al. used gate-level partitioning for Multiply-Accumulate (MAC) units and block-level partitioning for SRAM arrays across two DNN architectures. Their approach demonstrated 22.3% power savings and 6.2% performance improvement over 2D baselines. Do et al. proposed a two-layer GPU scratchpad memory using MONO3D to allow simultaneous row and column access, boosting system performance by 46.3%. However, these studies largely overlook thermal issues, which are a critical concern in 3D ICs due to strong inter-tier thermal coupling.

To address this, a temperature-aware optimization flow was introduced for designing energy-efficient systolic DNN accelerators in MONO3D environments under specified performance and thermal constraints. This study highlighted the trade-off between thermal impact and performance, particularly for high-power DNNs like VGG19. Increased dynamic power leads to thermal hotspots, raising leakage currents and total power, thereby risking thermal violations. These studies demonstrate MONO3D's effectiveness in DNN acceleration but also underscore the need for standard models that incorporate power, performance, and thermal considerations.

7.3 Hardware Security

The advancement of MONO3D ICs has also sparked interest in the domain of hardware security, which is critical for protecting intellectual property, sensitive data, and system integrity. Most security research for MONO3D focuses on circuit-level obfuscation, primarily through logic locking and layout camouflaging techniques. Layout camouflaging makes reverse engineering difficult by embedding dummy contacts into the circuit layout. In transistor-level MONO3D circuits, camouflaged ISCAS '89 benchmark circuits showed average area savings of 47.5% and power savings of 6.3% compared to 2D designs, albeit with a 27.4% performance degradation.Logic locking involves embedding key-controlled logic gates in a design to prevent unauthorized usage. In a recent MONO3D implementation, pull-up and pull-down networks in different tiers were locked independently using serial and parallel locking transistors along with camouflaged contacts. This architecture ensures that leaking information from one tier does not compromise the entire system. Despite these advancements, MONO3D ICs remain vulnerable to thermal side-channel attacks and thermal covert channels, which exploit the strong vertical thermal coupling inherent in 3D integration.

While such attacks have been studied in TSV-based 3D ICs and 2D multicore systems, limited work has focused on MONO3D-specific vulnerabilities. For example, a temperature-aware shielding method has been proposed for TSV-based designs using custom activity patterns to mask power consumption. Another approach uses thermal-aware floorplanning to reduce leakage. These techniques, however, need to be adapted to MONO3D architectures, where thermal crosstalk is more pronounced. Consider a MONO3D multicore system: if one core has access to sensitive data, it could covertly transmit information to another core on a lower tier via a thermally established communication path, posing a significant threat in sandboxed systems. As MONO3D ICs become more prevalent, developing robust thermal isolation and hardware-level protection mechanisms will be essential to ensure system security.

8. Future Scope

Looking ahead, the future of MONO3D IC technology lies in the convergence of fabrication innovation and advanced design techniques. As the technology matures, a stronger synergy between process development and design methodology is essential. Future work must focus on bridging this gap through co-development frameworks that allow simultaneous consideration of fabrication constraints and design goals.

One of the most critical areas for future research is the comprehensive characterization of inter-tier process variations. Detailed experimental studies are needed to accurately model the performance, reliability, and variability of MONO3D layers. These models will also enhance the precision of cost estimations, thermal analysis, and performance predictions for MONO3D systems.

Additionally, the development of robust, cross-layer design automation frameworks is expected to play a pivotal role in mainstream MONO3D adoption. These frameworks should enable full system-level exploration, incorporating considerations such as energy efficiency, thermal integrity, and real-time performance constraints. Special emphasis must be placed on the various partitioning techniques supported by MONO3D (e.g., block-level, gate-level, transistor-level), as each is best suited to particular application domains.

Moreover, with the rise of AI, IoT, and edge computing applications, MONO3D technology presents a promising path toward achieving ultra-compact, low-power, and high-throughput processing platforms. By leveraging its architectural advantages, MONO3D can contribute significantly to the development of non-von Neumann computing paradigms, secure hardware architectures, and high-efficiency neural accelerators.

In summary, MONO3D technology holds immense promise, and its future advancement depends on fostering a collaborative ecosystem that unites material science, process engineering, electronic design automation, and system-level architecture.

Conclusion

Monolithic 3D (MONO3D) integrated circuit technology represents a significant advancement in the field of VLSI design and fabrication. This tutorial has presented a comprehensive overview of the current status of MONO3D, highlighting developments in both fabrication and design methodologies. Key process-level challenges—such as thermal budget limitations, high-quality monocrystalline silicon growth, and contamination during layer fabrication—have been explored, along with existing strategies to mitigate these obstacles.

On the design side, advancements in automation tools, thermal management techniques, and testing methodologies specific to MONO3D architectures have been detailed. The unique vertical integration and high density of inter-layer vias (MIVs) in MONO3D ICs enable superior system performance, compactness, and energy efficiency when compared to traditional 2D and TSV-based 3D ICs. Furthermore, this work has illustrated how MONO3D can significantly benefit three key application areas: near- and in-memory computing, deep neural network (DNN) acceleration, and hardware security.

Despite these promising developments, a major limitation remains the lack of extensive experimental validation across many studies. This gap restricts the practical applicability and wider adoption of MONO3D circuits. Additionally, the disconnect between fabrication constraints and design-level methodologies impedes fully optimized MONO3D solutions. Therefore, a tighter integration between process engineering and circuit design is critical to unlocking the full potential of MONO3D ICs.

References

[1] 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.

[2] W. Gomes et al., “8.1 lakefield and mobility compute: A 3D stacked 10nm and 22FFL hybrid processor system in 12×12mm2, 1mm package-on-package,” in Proc. IEEE Int. Solid State Circuits Conf., 2020.

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

[4] 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.

[5] 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.