A Study of Synchronous and Asynchronous Circuits in Monolithic 3D Integration

Xiameng Zhang

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Reading Committee:

Madhava Sarma Vemuri

Tadesse Ghirmai

Kaibao Nie

Seungkeun Choi

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

A Study of Synchronous and Asynchronous Circuits in Monolithic 3D Integration

Xiameng Zhang

Chair of the Supervisory Committee:
Madhava Sarma Vemuri
Department of Electrical and Computer Engineering

Monolithic three-dimensional integration (M3D) has emerged as a promising pathway for extending integrated-circuit scalability beyond conventional two-dimensional (2D) technology. By sequentially stacking active device layers and connecting them through fine-grained metal interlayer vias (MIVs), M3D can improve device density, reduce interconnect length, and enhance energy efficiency. This thesis investigates M3D from both application-driven and layout-methodology perspectives.

While M3D offers density and interconnect benefits, its sequential fabrication and vertical stacking introduce reliability concerns related to process variation, thermal effects, and timing uncertainty. To address these challenges, we first studied the quasi delay insensitive asynchronous circuits based on Null Conventional Logic (NCL). The asynchronous circuits address these challenges by eliminating the global clock and using local handshaking, making them robust to timing variations. To explore the complementary benefits of M3D and QDI design, this work proposes a transistor-level M3D methodology for static NCL threshold gates. Results show that M3D-NCL substantially reduces area while improving delay and power over 2D implementations.

The second part studies the MIV placement opportunities and design consideration which affect the area, delay, skew, and power of M3D standard-cell designs. A methodology is proposed to study and compare conventional 2D and M3D standard cells in terms of power, performance, and area (PPA). Using this methodology, standard cells are implemented in

both 2D and M3D, with different MIV placement strategies considered for the M3D case. Results show that the proposed designs achieve large area reduction with favorable delay, skew, and power trends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Madhava Sarma Vemuri, for his invaluable guidance, support, and mentorship throughout my graduate studies and the completion of this thesis. His expertise, insightful feedback, and encouragement greatly contributed to the development of this work.

I would also like to thank my committee members, Dr. Tadesse Ghirmai, Dr. Kaibao Nie and Dr. Seungkeun Choi for their valuable guidance, constructive suggestions, and thoughtful feedback, which helped improve the quality of this research.

I am especially grateful to Dr. Kushal K. Ponugoti and Dr. Ashiq A. Sakib for their continuous guidance, support, and insightful discussions regarding my work on asynchronous circuits. Their expertise and encouragement have been invaluable to my research progress in this area.

Special thanks are extended to my academic advisor, Beth Gee, for her continuous support, understanding, and assistance during my academic journey.

I am deeply grateful for all their support and encouragement.

Chapter 1

INTRODUCTION

The continued demand for energy-efficient and high-performance integrated circuits (ICs) has motivated both new circuit design paradigms and new physical integration technologies. For decades, conventional two-dimensional (2D) scaling has been the primary mechanism for increasing transistor density and reducing cost per function, following the trend commonly associated with Moore's law [1, 2]. However, as device dimensions approach physical and manufacturing limits, further scaling in planar technologies becomes increasingly difficult due to manufacturing complexity, interconnect limitations, routing congestion, power dissipation, and timing uncertainty [3]. To overcome these limitations, three-dimensional (3D) integration has emerged as a promising alternative by stacking active device layers vertically, thereby improving functional density, shortening interconnect lengths, and enhancing energy efficiency [4, 5]. Among different 3D approaches, conventional die stacking relies on Through-Silicon Vias (TSVs), which introduce significant area overhead and limit fine-grained integration [6, 7]. Monolithic three-dimensional integration (M3D) provides a more compact alternative by sequentially fabricating active device layers and connecting them through fine-grained metal interlayer vias (MIVs), enabling reduced footprint area, shorter wirelengths, and improved circuit performance [5,8,9]. At the same time, M3D introduces design challenges related to sequential fabrication, process variation, thermal effects, inter-tier parasitics, timing uncertainty, and MIV placement. These challenges motivate the two research directions explored in this thesis. First, because quasi-delay-insensitive (QDI) asynchronous circuits eliminate the global clock and rely on local handshaking, they provide a robust design paradigm for addressing timing uncertainty in M3D systems [10]. Null Convention Logic (NCL), as a mature QDI paradigm, offers robustness against process-voltage-temperature (PVT) variations and operating-condition changes, although its adoption has been limited by computer-aided design (CAD) constraints and larger

area footprint [11–13]. Second, because MIVs provide the primary vertical connections between device tiers, their placement strongly affects routing resources, parasitic resistance and capacitance, layout area, and circuit performance, making layout-level M3D methodology essential for translating vertical integration density into practical power, performance, and area (PPA) benefits [14]. Motivated by these observations, this thesis studies M3D from both an application-driven perspective, through M3D-based QDI/NCL asynchronous circuits, and a layout-methodology perspective, through MIV placement strategies for M3D standard-cell designs. The following sections provide the necessary background for these two research directions, first introducing QDI/NCL asynchronous circuits and their relationship with M3D, and then discussing M3D standard-cell design considerations with emphasis on MIV-based vertical interconnects and placement strategies.

M3D offers various abstract design styles, including block-level, gate-level, and transistor-level, as depicted in Fig. 1.1 [15–18]. In block-level design style, the conventional 2D IP blocks are partitioned into different layers, as depicted in Fig. 1.1a. Compared to other implementations, this style offers the least area savings, as it does not allow finer integration of Metal Inter-layer Vias (MIVs) [14]. In gate-level design style, as depicted in Fig. 1.1b, the gates are placed in multiple layers, and MIVs connect the input and output pins in different layers. This approach allows the utilization of existing libraries and has reported area savings of up to 50% [17]. However, it does not address the reliability concerns that result from sequential integration [19–21]. In a transistor-level design style, NMOS and PMOS are assigned to different layers, requiring only local interconnects to route the bottom layers, as depicted in Fig. 1.1c. This results in fewer manufacturing steps, resulting in controlling variations arising from M3D integration [22]. The role of MIV is to facilitate interconnections to active and gate regions of various devices in multiple layers. [14] indicates that transistor-level implementations can result in a 12% to 40% reduction in area. The theoretical 50% reduction of area is not observed due to a mismatch between transistor devices in different layers and additional area overhead caused by the presence of MIVs in the top layer. Additionally, a Keep-Out Zone (KOZ) around the MIV is a critical requirement to prevent it from biasing the adjacent devices in the top layer, which further reduces the area savings [9, 23].

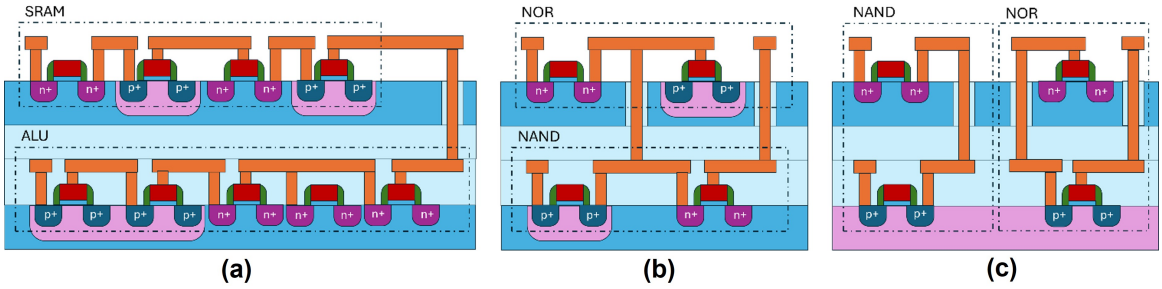


Figure 1.1: M3D implementation styles: (a) block-level, (b) gate-level, and (c) transistor-level.

The remainder of this thesis is organized as follows. Chapter 2 presents the study of M3D-based QDI asynchronous circuits using NCL. It introduces the proposed transistor-level M3D methodology for static NCL threshold gates, evaluates the resulting M3D-NCL cells against conventional 2D implementations, and demonstrates their use in an unsigned NCL array multiplier. Chapter 3 presents the study of M3D standard-cell design with different MIV placement strategies. It describes the methodology for comparing 2D and M3D standard cells and analyzes how internal and external MIV placements affect area, delay, skew, and power. Chapter 4 concludes the thesis by summarizing the main findings and discussing future research directions.

Chapter 2

MIDAS MONOLITHIC 3D INTEGRATION FOR DELAY INSENSITIVE ASYNCHRONOUS CIRCUITS

2.1 Background and Related Work

2.1.1 NCL Background

NCL has several unique features that separate it from its synchronous/Boolean counterparts. Unlike Boolean logic, NCL employs a 1-hot encoding scheme (1-of-n), most commonly the dual-rail (DR) logic (1-of-2), to encode data and eliminate timing references. In DR logic, two wires (or rails) are utilized to represent one bit of information, as opposed to a single wire in Boolean logic. A DR variable, D, can assume one of the three permissible values from the set {DATA0, DATA1, NULL}, where DATA0 ($D^1D^0 = 01$) and DATA1 ($D^1D^0 = 10$) correspond to Boolean logics 0 and 1, respectively. ($D^1D^0 = 00$) is a spacer state, commonly referred to as the NULL state, which indicates the unavailability of the data. ($D^1D^0 = 11$) constitutes an invalid state. The NCL framework resembles a traditional synchronous framework, with each stage consisting of an input and output NCL registration unit and one combinational logic (C/L) in between, as illustrated in Fig. 2.1a. The input wavefronts are, however, controlled by local handshaking signals, rather than a global clock signal. A 4-phase handshaking protocol is employed, in which the completion detection unit (C/D) and registers maintain an alternating sequence of NULL and DATA to differentiate between two unique data sets [11], as shown in Fig. 2.1(b).

The NCL logic, including registers and completion, comprises 27 fundamental gates with *hysteresis*, i.e., state holding capability. Each gate is characterized by unique input configurations and threshold levels. Together, these gates can implement any function of up to 4 variables. The threshold gates are classified into two categories: non-weighted and weighted. A non-weighted gate, also known as an *m-of-n* gate, is denoted as *THmn*, where n is the number of inputs and m ($1 \leq m \leq n$) is the threshold (the minimum number of inputs

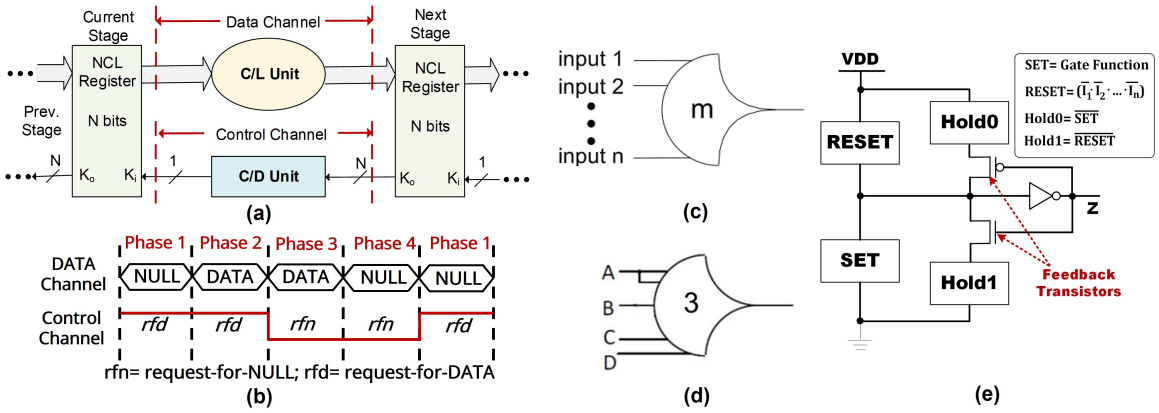


Figure 2.1: (a) NCL framework, (b) 4-phased handshaking scheme, (c) Generic non-weighted gate symbol, (d) TH34w2 weighted gate, and (e) CMOS static implementation template for NCL gates.

that must be asserted to assert the output) (Fig. 2.1c). A weighted gate is represented as $THmnWw_1w_2\ldots w_R$, where w_R $(1 < w_R \leq m)$ is the weight corresponding to input R (Fig. 2.1d). Fig. 2.1e illustrates the generic CMOS template for the static implementation of any NCL threshold gate. The template consists of four main blocks: *SET*, *RESET*, *Hold0*, and *Hold1*. The gate threshold function is implemented by the *SET* block, while the *RESET* block deasserts the output when all inputs are deasserted. The *Hold* blocks implement hysteresis functions that retain the output when neither *SET* nor *RESET* is active, utilizing two feedback transistors, as indicated in Fig. 2.1e. NCL gates can also be implemented in a semi-static manner, in which the *Hold* blocks are eliminated and replaced by a weak-feedback inverter arrangement [11]. While this decreases the transistor count as compared to the static version, it does not inherently result in a reduced silicon area. This is because pmos transistors require substantial sizing up to maintain the necessary driving strength for proper functionality. Furthermore, semi-static gates exhibit reduced speed and higher energy consumption per switching relative to their static counterparts [13]. Therefore, the research presented herein primarily focuses on the static implementations of NCL gates.

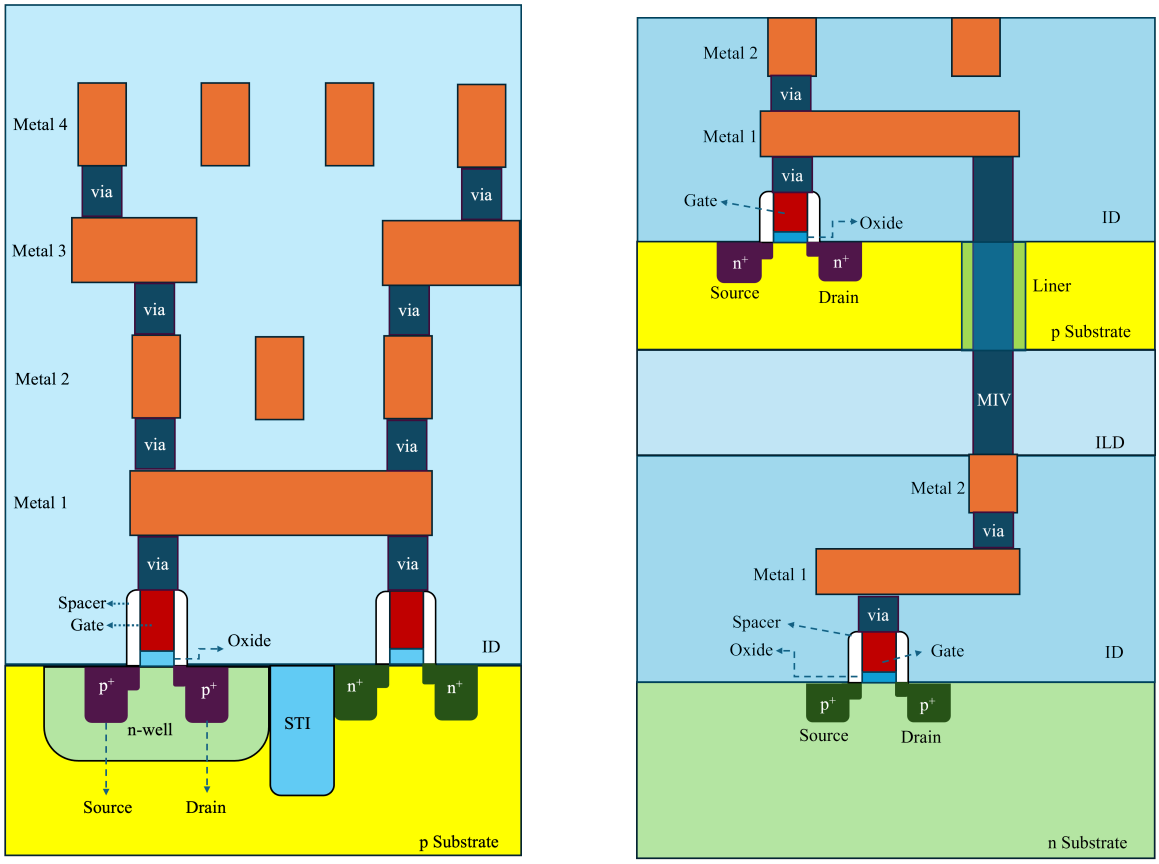
2.1.2 Related Works

NCL circuits must follow delay-insensitive (DI) constraints, such as *input-completeness* and *observability*[1], and maintain handshaking protocols for synchronization, leading to significant area overhead compared to their clocked counterparts. Over the years, many research efforts have focused primarily on NCL implementations with fewer constraints to enhance area efficiency. In [24], a modified NCL architecture, NCL_X, was proposed that allows the C/L unit to be relaxed, i.e., input-incomplete and unobservable, substantially reducing the area. However, the DI constraints are preserved by additional completion logic, resulting in a handshaking network that is both larger and slower. Multiple relaxation schemes were proposed in [25] and [26], aiming to achieve an optimal balance between strict and relaxed implementation. [27] is a more recent work that intelligently combines [24] and [25] and expands upon both to yield smaller and faster NCL circuits. At the circuit level, the minimization of the fundamental threshold gates' area was also the focus of numerous prior studies, both in CMOS [28], [29], [30] and beyond CMOS technologies [31], [32], [33]. However, those works relied on traditional 2D technology, limiting the potential for substantial area optimization. There exist very few works that have explored the potential of 3D integration in the context of clockless design. In [34], researchers have focused on the use of vertical integration techniques to design QDI asynchronous circuits using Through-silicon-via (TSV) based integration. However, TSVs are significantly larger than MIVs ($> 200\times$) and require large KOZs to limit the degradation of devices around TSVs [35]. Meanwhile, [36] introduces a gate-level M3D implementation methodology for desynchronized circuits and reports reductions in wirelength and area of up to 35% and 50%, respectively. However, as mentioned earlier, the transistor-level M3D integration provides better reliability than gate-level integration, and to the best of our knowledge, no prior research has been reported on the transistor-level M3D implementation for QDI NCL asynchronous circuits.

[1]Input completeness requires that no output of a C/L circuit can transition from Null-to-Data (or Data-to-Null) until all inputs have transitioned from Null-to-Data (or Data-to-Null). Observability requires that every switching gate must be observable at the output; i.e., switch at least one of the primary outputs.

2.2 Proposed Methodology

2.2.1 Process and Design Considerations



(a) 2D Process

(b) M3D Process

Figure 2.2: 2D and proposed M3D process comparison (not to scale)

Since there are no open-source process design kits (PDKs) for the M3D implementation of asynchronous logic, we have adapted the NCSU FreePDK45 model library and made realistic assumptions when designing M3D-IC for the Bulk-Si process [37]. In the conventional planar implementation, both the PMOS and NMOS transistors are fabricated on the same layer, as shown in Fig. 2.2(a). Silicon trench isolation (STI) limits the interactions of different PMOS and NMOS devices. For the proposed M3D implementation, we have

considered a two-tier transistor-level abstract style with NMOS and PMOS transistors in the top and bottom layers, respectively, as shown in Fig. 2.2(b) [5]. The interlayer dielectric (ILD) limits electrical coupling and improves thermal conductivity between layers [38]. The M3D follows a sequential integration process (i.e., the bottom, ILD, and top layers are created sequentially). Since the top layers are created on the fabricated bottom layer, a thermal budget ($< 500^{o}$C) is set to limit the degradation of transistor devices in the bottom layers [19–21]. Metal routing layers are created inside the interconnect dielectric (ID). MIVs connect the metal interconnects in both layers to facilitate interconnections between the top and bottom layers. An oxide liner is created to prevent the MIV from biasing the devices in the top layer. An additional keep-out-zone (KOZ) is also required for the Bulk-Si around MIV to limit the interactions of MIV and neighboring devices [39].

The process and design parameters considered for this work are presented in Table 2.1. The nominal values are extracted from the NCSU FreePDK45nm library [37]. We considered the length of the channel region to be 50 nm, and the length and width of the source and drain regions to be 90 nm. The width and pitch of the interconnect wires are 65 nm and 130 nm, respectively. We have considered the copper connecting region to have a sheet resistance of 0.38Ω/sq and a capacitance per unit length of 179.93 fF/mm, based on [37,40]. The metal vias connect 1) the interconnect to active regions (M1 - source/drain) and 2) the interconnect to the gate region (M1 - gate). We have considered the via resistance to be 6Ω, and the MIV resistance and capacitance to be 5.5Ω and 0.04 fF, respectively, which is consistent with [14]. The MIV thickness and the KOZ around the MIV are both considered to be 50nm, as per [39].

2.2.2 Overview of Selected NCL Threshold Gates

As mentioned in Section 2.1.1, the 27 NCL gates can implement any function of up to 4 variables. In this work, we focus on 6 specific gates: TH22 ($PMOS$: 6, $NMOS$: 6), TH24 ($PMOS$: 13, $NMOS$: 13), TH34 ($PMOS$: 13, $NMOS$: 11), TH24comp ($PMOS$: 9, $NMOS$: 9), THand0 ($PMOS$: 10, $NMOS$: 10), and TH54w322 ($PMOS$: 11, $NMOS$: 10). These gates are chosen as they cover both weighted and non-weighted gates of varying

Table 2.1: Process and design parameters used in the study for 2D and M3D implementation

Notation	**Description**	**Value**
L_G	Channel length	50 nm
l_{src}	Length of source/drain region	50 nm
w_{src}	Width of source/drain region	90 nm
t_{ILD}	Interlayer dielectric thickness	120 nm
t_{miv}	MIV thickness	50 nm
w_{M1}	Width of M1 interconnect	65 nm
p_{M1}	Pitch of M1 interconnect	130 nm
w_{gate}	Width of gate region	50 nm
$R_{int,sq}$	Sheet resistance of interconnect	0.38 Ω/sq
R_{via}	Via resistance	6 Ω
C_{int}	Interconnect capacitance per unit length	179.93 fF/mm
R_{MIV}	MIV resistance	5.5 Ω
C_{MIV}	MIV capacitance	0.04 fF

complexities and transistor counts. By analyzing this set, we can better understand how gate complexity affects performance and integration in our proposed M3D design. Each gate has two to four inputs (a, b, c, and d) and one output (Z). The TH22 gate is the simplest gate in the set that behaves like a 2-input Muller C-element [41], asserting/deasserting the output only when both inputs are asserted/deasserted; otherwise, holding the previous value. Based on the threshold, the SET function for this gate can be expressed as $Z = ab$. The TH24 gate, with four inputs and a threshold of two, is the largest NCL gate, which gets asserted when at least two of its inputs are asserted, making the SET expression to be $Z = ab + ac + ad + bc + bd + cd$. Similarly, the TH34 gate with a threshold of three

has the following *SET* function: $Z = abc + abd + acd + bcd$. The *SET* functions of the TH24comp and THand0 gates are $Z = ac + ad + bc + bd$ and $Z = ab + bc + ad$, respectively. The weighted gate in our study is the TH54w322 gate. The gate has four inputs, with the first input a being assigned a weight of three, inputs b and c each being assigned a weight of two, and the remaining input d having a weight of one. The output is asserted when either a and b are asserted, a and c are asserted, or when all of b, c and d are asserted, satisfying the threshold of five (i.e., $Z = ab + ac + bcd$).

2.2.3 Parasitic Model Used in the Study

In this section, we have used the standard TH22 gate to explain the parasitic model of conventional 2D and the proposed M3D implementation. For both cases, there are 4 common metal routing scenarios, such as: 1. VDD-to-Node, 2. Node-to-GND, 3. Node-to-Node, 4. Input-to-Node as shown in Fig. 2.3. The VDD-to-Node and Node-to-GND connection scenario interconnects power rails such as VDD and GND to the transistor devices' active regions (i.e., the source and drain). The Node-to-Node connects the active regions of PMOS and NMOS devices. The Input-to-Node connects the inputs, such as a, b, c, and d, to the gate regions of the devices. RC parasitics, such as R_{wire} and C_{wire}, for the interconnect regions are calculated using the standard cell height. The wires are considered to be extended for half the height of the cell in scenarios that connect to the power and ground rails [42]. In the M3D implementation, since the PMOS and NMOS devices are placed in different layers, MIV is used to connect the interconnect regions in different layers. Subsequently, the associated RC parasitics, such as R_{MIV} and C_{MIV} are added into the parasitic model of the M3D implementation, as shown in Fig. 2.3(b). The scaling factor α is used to characterize the reduction of RC parasitics from metal interconnect routing, which is a consequence of the improved routing congestion caused by the reduced cell height in both layers. *It is important to note that we have associated the scaling factor to interconnect scenarios such as Node-to-Node and Input-to-Node, as these two are more substantially impacted by M3D implementation than other interconnect scenarios.* Interconnect capacitances, such as fringing capacitance and metal coupling capacitance, are ignored to limit the complexity of the

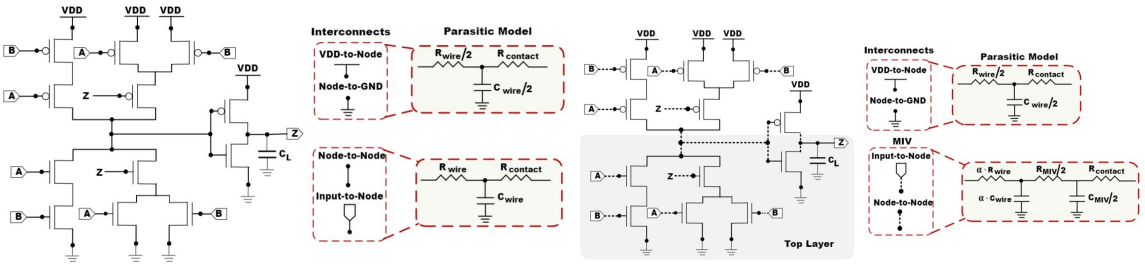
design.

2.3 Simulation Results and Performance Assessment

2.3.1 Systematic Study of PPA Metrics of NCL Standard Cells

Based on the parasitic model described in the previous section, we designed both 2D and M3D NCL standard cells, as outlined in Section 2.2.2, utilizing a commercial SPICE circuit simulator. Custom cells with a height of 14 tracks were designed using the model libraries provided by the NCSU FreePDK45nm. A 1 fF load capacitance has been considered for each gate. A systematic variation in wirelength reduction, ranging from 20% to 40%, was conducted to examine the effects of M3D integration. This variation is represented by the parameter α, which quantifies the extent of physical reduction due to vertical stacking. Specifically, the α values of 0.8, 0.7, and 0.6 correspond to the wirelength reductions of 20%, 30%, and 40%, respectively. These values were chosen to capture both conservative and aggressive levels of 3D integration and to assess how different degrees of vertical compaction affect key performance metrics. Table 2.2 reports the power, performance, and area (PPA) metrics when the wire length reduction is 30%, i.e., ($\alpha = 0.7$). These metrics were chosen because they collectively represent the temporal and dynamic behavior of the test circuits. Delay and skew influence timing robustness and throughput, while power is critical for energy efficiency and thermal reliability, especially in vertically stacked architectures.

The simulation results indicate that the propagation delay (T_D) is reduced by approximately 11% on average in the M3D implementations relative to the 2D counterparts, with the TH22 gate exhibiting the most notable improvement of around 12%. The average output skew (T_S) was reduced by 10.2% on average for the M3D case. The average power consumption of the gates was lowered by around 10% on average. The TH22 and TH34 gates have shown the highest power savings, reaching up to 11%, in comparison to others. The total area footprint of the 2D implementations is compared with that of the M3D implementations, accounting for the additional MIV area overhead for the Bulk-Si process. The results indicate a 44% decrease in average area utilization, with TH24 exhibiting the maximum reduction of 46%.



(a) 2D Implementation (b) M3D Implementation

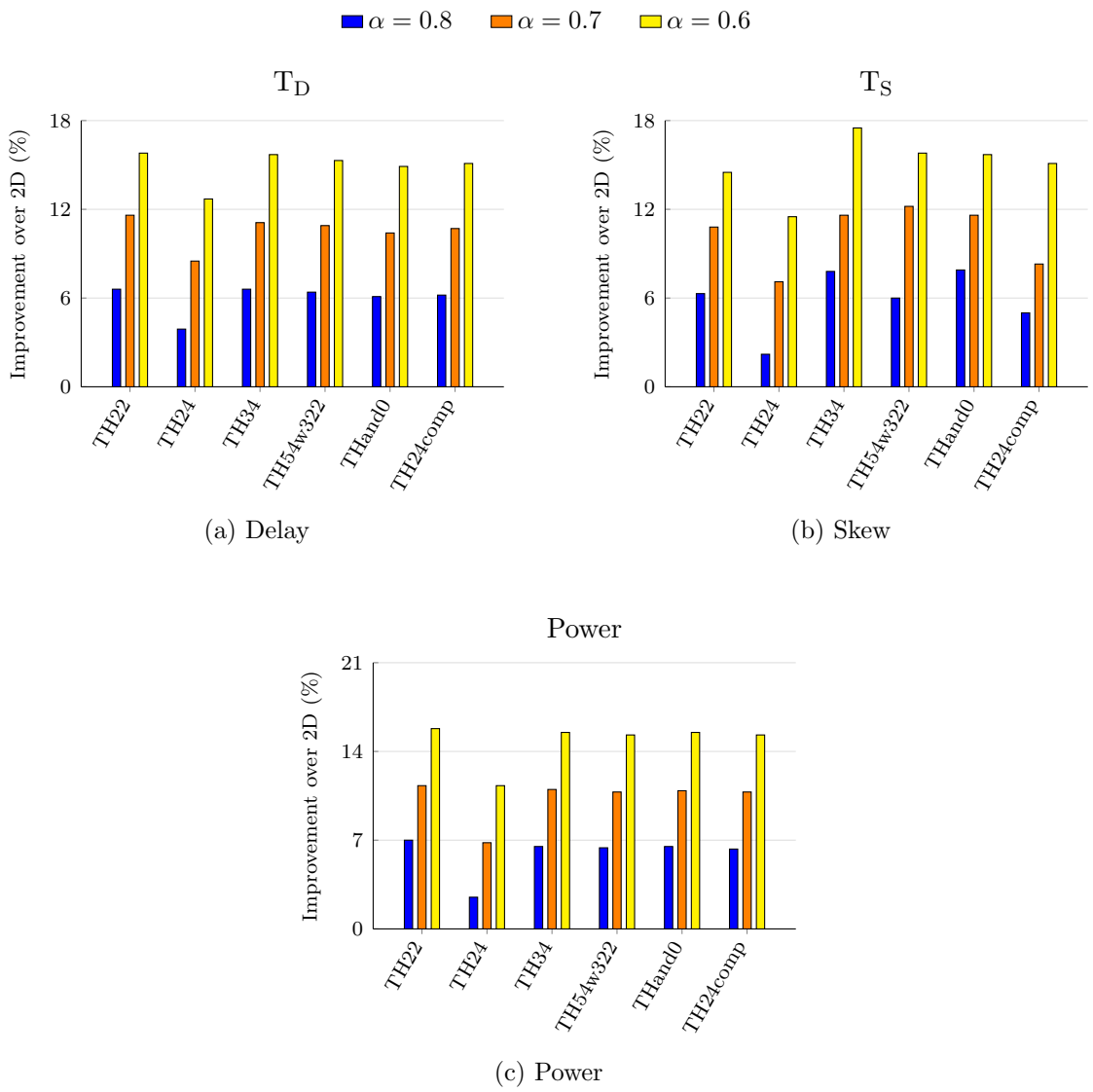
Figure 2.3: 2D and M3D implementation of TH22 circuit schematic with the RC prasitic model

Figure 2.4 compares the PPA metrics of the different NCL gates across varying scaling factors (α). We observe that the highest performance gains are achieved when the wirelength is reduced by 40% ($\alpha = 0.6$). In this configuration, the benefits of shorter interconnects and reduced parasitics are most pronounced. The delay and skew improve by up to 16% and 18%, respectively, while the power consumption shows a reduction of up to 16% relative to the baseline 2D implementation. These improvements come from placing logic elements closer together in the M3D layout, which shortens the distance signals need to travel and reduces the energy during switching.

In contrast, when the wirelength reduction is limited to 20% ($\alpha = 0.8$), the overall improvements are modest. All the performance parameters, i.e., delay, skew, and power, are improved by 6% on average. These results suggest that without compromising the area benefits, limited reductions in wirelengths may result in reduced power and performance benefits arising from M3D.

2.3.2 Performance Evaluation for Large-Scale NCL Circuits

The goal of this section is to present the impact of M3D integration on large-scale NCL circuits and provide a comparative analysis of performance against their 2D counterparts. A QDI 4×4 NCL unsigned multiplier was synthesized at the gate level for this purpose. A Boolean RTL netlist was first generated using Yosys [43], an open-source RTL synthe-



(a) Delay

(b) Skew

(c) Power

Figure 2.4: Delay, skew, and power improvements of the M3D implementations over the 2D implementations under varying scaling factors (α).

sizer. The generated Boolean netlist was then converted to an NCL gate-level netlist by the UNCLE synthesis tool, using the tool's intermediate output before the addition of registration and completion logic [44]. The multiplier created utilized multiple threshold gates, including the TH22, TH24comp, and THand0 NCL gates, from Table 2.2. The circuit was

Table 2.2: Conventional 2D vs M3D implementations of static NCL threshold gates ($\alpha = 0.7$)

Gate	Case	T_D (ps)	T_S (ps)	Power (µW)	Area (µm^2)
TH22	2D	96.4	69.6	1.74	0.2052
	M3D	85.2	62.1	1.55	0.1226
	% Improvement	**11.6%**	**10.8%**	**11.3%**	**43.9%**
TH24	2D	154	95.7	2.07	0.4446
	M3D	141	88.9	1.93	0.2598
	% Improvement	**8.5%**	**7.1%**	**6.8%**	**46.1%**
TH34	2D	208	124	2.19	0.4104
	M3D	185	110	1.95	0.2598
	% Improvement	**11.1%**	**11.6%**	**11%**	**41.6%**
TH54w322	2D	164	110	2.00	0.3591
	M3D	146	96.7	1.79	0.2206
	% Improvement	**10.9%**	**12.2%**	**10.8%**	**42.7%**
THand0	2D	158	105	1.98	0.3420
	M3D	141	92.8	1.77	0.2010
	% Improvement	**10.4%**	**11.6%**	**10.9%**	**44.9%**
TH24comp	2D	153	102	2.01	0.3078
	M3D	137	93.7	1.79	0.1814
	% Improvement	**10.7%**	**8.3%**	**10.8%**	**44.3%**
	% Average	**10.5%**	**10.2%**	**10.3%**	**43.9%**

then converted to SPICE models using the threshold gate implementations based on Section 2.2.3. The C/L of the 4x4 NCL multiplier requires a total of 2124 transistors.

We evaluate the performance of the M3D-NCL multiplier in terms of area, power, and delay, and compare it with that of its 2D counterpart. The M3D-NCL multiplier demonstrates a 44.5% reduction in area utilization compared to the 2D version. We report the average power and delay of the designed multipliers using multiple sets of random input

combinations. For the nominal case ($\alpha = 0.7$), the 2D and M3D versions consumed 56µW and 46.7µW, respectively, indicating an improvement of ∼17%. The 2D-NCL multiplier demonstrates a worst-case delay of 1.98ns, while the M3D-NCL multiplier has a delay of 1.37ns, demonstrating an improvement of 30.8%. Fig. 2.5 also presents the % improvement in power and delay for varying scaling factors. Finally, this analysis highlights the potential of M3D IC technology to address the area overheads associated with asynchronous circuits, while also improving other performance parameters. By enabling compact layouts and efficient interconnect structures, M3D integration can pave the way towards making asynchronous logic more viable for future energy-efficient computing systems.

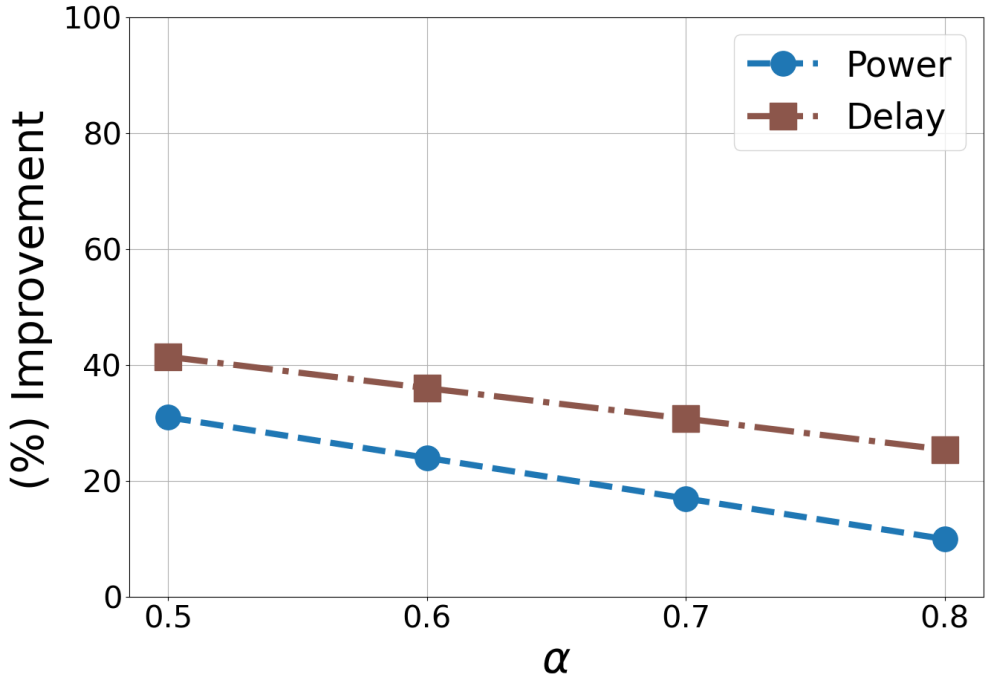


Figure 2.5: (%) Improvement of M3D implementation over 2D for Multiplier circuit with varying α

Chapter 3

MONOLITHIC 3D INTEGRATION FOR SYNCHRONOUS CIRCUITS

3.1 Background and Related Work

In M3D integration, devices are created sequentially on different layers and connected using MIVs and metal interconnects in the subsequent top and bottom layers [45]. Compared with conventional TSV-based 3D integration, M3D technology uses much smaller vertical interconnects facilitated by the thin inter-layer stack [17,45]. Consequently, MIV dimensions are up to 200× smaller than the TSV, resulting in significant area benefit [35]. Since the top device layers are fabricated sequentially, M3D follows a strict thermal budget ($< 500^o$), to avoid performance degradation in the bottom layer devices. Prior M3D studies based on CoolCube and sequential 3D integration typically rely on low-temperature processing for the top layer [20, 46, 47].

There are different abstraction styles, including block-level, gate-level, and transistor-level in M3D integration [15–17]. In block-level M3D design, large functional blocks or IP modules are assigned to different layers, and MIVs are mainly used for inter-block connection [15]. This style is compatible with existing 2D design methodologies but offers relatively limited fine-grained layout optimization, since intra-block routing remains largely constrained by the 2D IP block. In gate-level M3D design, existing 2D standard cells are distributed across multiple layers, and MIVs connect cell pins, or local interconnects between layers [16]. This approach can reduce wirelength and routing congestion while still preserving much of the conventional standard-cell design flow. However, it does not fully extend the granularity of M3D and often requires extra processing steps to fabricate the different active devices in the top and bottom layers. In contrast, transistor-level M3D design, PMOS and NMOS devices are partitioned across different layers [17,45,48]. This style provides the finest granularity and can achieve higher layout density, but it also introduces more frequent MIV usage and stronger sensitivity to MIV placement, routing overhead, and

device-level interactions [17, 45].

Despite the benefits of increased granularity, we do not see 50% area savings in transistor-level M3D due to uneven width sizing of PMOS and NMOS devices in subsequent layers. Placement of MIV at the transistor-level can significantly affect area savings and metal interconnect utilization. Most recent works in transistor-level M3D have utilized MIV in the external placement scenario [17, 22, 45, 49]. To address this gap, we systematically study internal MIV placement and compare it with external placement to evaluate its area, power, and performance benefits in transistor-level M3D.

3.2 M3D Process

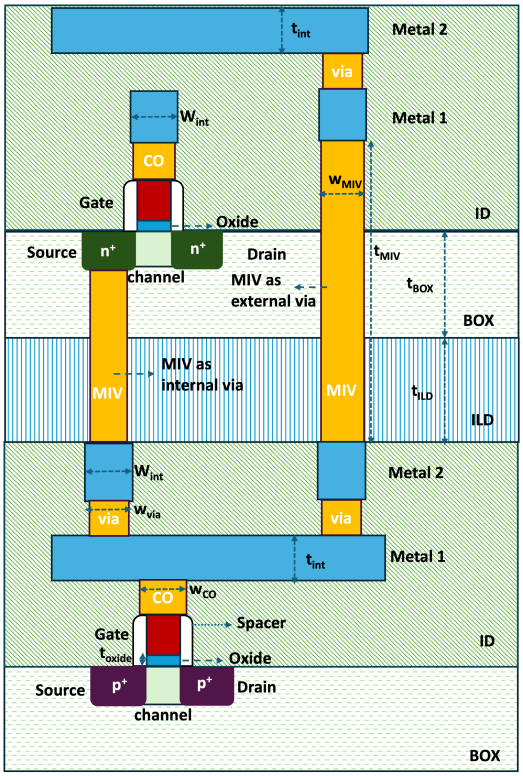


Figure 3.1: 2-layer M3D process used in the study.

Fig. 3.1 illustrates the two-layer M3D process used in this work. We have used a $28nm$ process for the study and made realistic design assumptions based on existing 2DIC imple-

mentations and recent demonstrations in M3D, since there are no mature processes for M3D design and implementation [50, 51]. In the transistor-level abstraction style, the NMOS devices are placed in the top layer, while the PMOS devices are placed in the bottom layer [17, 52]. The design and process parameters are presented in Table 3.1. The interlayer dielectric (ILD) separates the top and bottom device layers within the buried oxide (BOX). The ILD region also acts as an electrical insulator, limiting electrical coupling between the regions. The BOX and ILD regions are assumed to be $100nm$ [53].

Table 3.1: Process parameters used in the design of two-layer M3D

Notation	**Description**	**Value**
t_{BOX}	Height of buried oxide	100 nm
t_{ILD}	Height of ILD region	100 nm
t_{ox}	Oxide thickness	1.5 nm
t_{MIV}	Thickness of MIV	400 nm
W_{MIV}	Width of MIV	42 nm
w_{CO}	Width of CO	42 nm
W_{via}	Width of via	42 nm
W_{int}	Width of M1 & M2 metal	50 nm
t_{int}	Thickness of M1 & M2 metal	95 nm
$d_{int,pitch}$	Minimum M1 & M2 pitch	100 nm
$d_{via,pitch}$	Minimum via pitch	112 nm
$d_{MIV,pitch}$	Minimum MIV pitch	112 nm
R_{MIV}	Resistance of MIV	3.8 Ω
C_{MIV}	Capacitance of MIV	40 aF

In the proposed M3D process, each layer has local interconnect-stack regions comprising metal routing layers and vias created inside an interconnect dielectric (ID) region. We have assumed similar design and process dimensions for both layers and used polysilicon material (PO) for the gate. The contact (CO) connects the gate region to Metal1 (M1). The interconnect (via) provides the vertical connection between Metal1 (M1) and Metal2 (M2) within the same device layer. The MIV is introduced as the inter-layer vertical via

that connects the top and bottom device layers through the ILD and BOX regions. For the design, Silicon dioxide (SiO_2) is used as the dielectric for the BOX, ILD, and ID regions, and Copper (Cu) is used for the metal interconnect layers and the MIV.

The width and thickness of the MIV are denoted by W_{MIV} and t_{MIV}. The width of the interconnect via regions and MIV is assumed to be $42nm$. The separation between the M1 regions and the BOX region is assumed to be $200nm$. The minimum width and thickness of the M1/M2 region are denoted by W_{int} and t_{int}, and are assumed to be $50nm$ and $95nm$ respectively. The minimum pitch of the regions signifies the minimum separation between two identical layers. The pitch of the M1 and M2 regions is denoted by $d_{int,pitch}$ and is assumed to be $100nm$. The pitch of the via and MIV are denoted by $d_{via,pitch}$, $d_{MIV,pitch}$, and are both assumed to be $112nm$.

3.3 MIV Placement Scenarios

In the standard-cell design, we have identified 2 possible placement scenarios, specifically 1) External placement and 2) Internal placement, as shown in Fig. 3.2. Using these scenarios, the interconnection between the 1) gate and 2) active (source or drain) regions of the top and bottom transistors can be facilitated for different layers. In traditional standard-cell design, input and output pins connect to these gates and active regions. Often, these regions share the same pin terminal. Since the transistors are partitioned across different layers, we require an MIV to enable shared pin connections. Possible interconnection combinations among these regions include gate-gate, gate-active, and active-active across both layers. To this end, we have assumed 2 metal routing layers (M1 and M2) on the top and bottom layers; this can be increased to alleviate metal routing congestion in both regions.

1. *External Placement*: In this scenario, MIV connects the highest metal routing layers (M2) in the bottom layer with the lowest metal routing layer (M1) in the top-layer of M3D, as shown in Fig. 3.2(a). To facilitate these connections, MIV passes through the top layer, resulting in area overhead due to its presence in the top region. For all of the interconnection combinations mentioned earlier, an area overhead is observed. As the total number of pins in the standard cell design increases, the number of MIVs

increases, thereby increasing area overhead and reducing area savings. Additionally, routing these external MIVs requires additional metal routing resources to connect the intended regions of the transistor, increasing routing congestion. Consequently, it adds parasitic resistance and capacitance to the signal path, degrading the performance metrics such as propagation delay, slew, and power consumption.

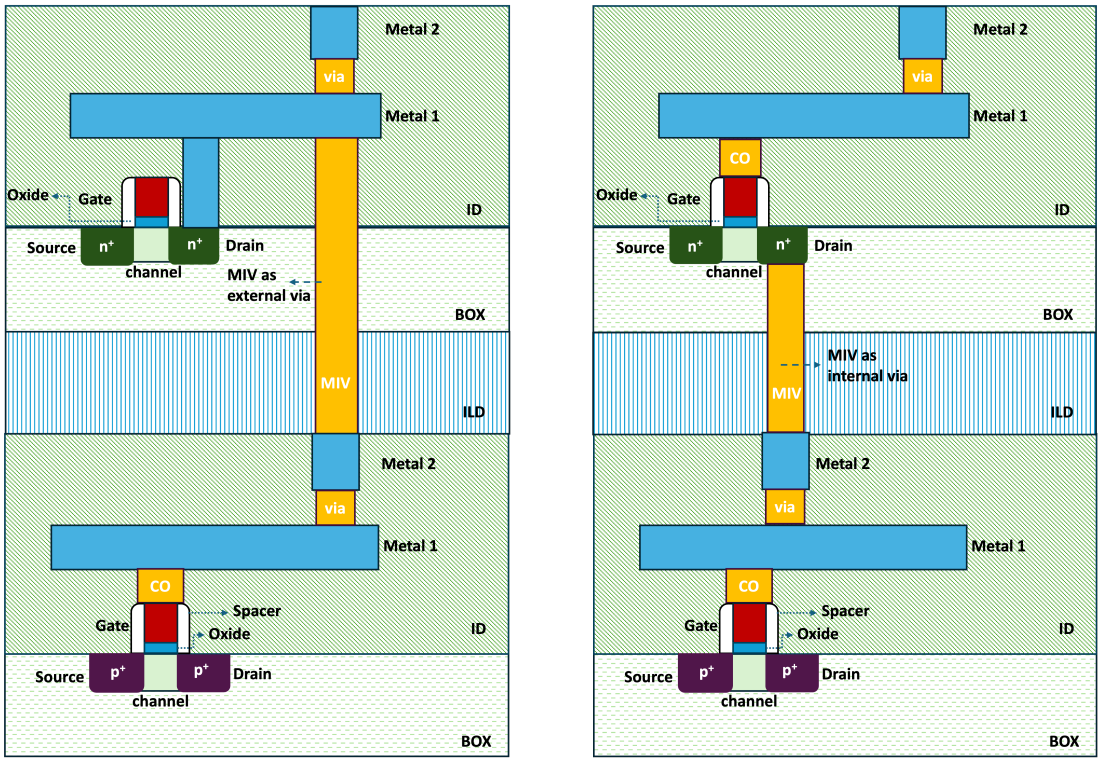
2. *Internal Placement*: In this scenario, it allows MIV to be opportunistically connected directly to the active and gate regions in the top layer as an internal contact. This configuration enables a compact vertical connection between device layers, thereby reducing routing complexity and eliminating the additional area overhead associated with external placement, as shown in Fig. 3.2(b). Consequently, the reduced routing lowers the overall parasitic resistance and capacitance in the signal path, improving propagation delay, signal slew, and power consumption.

3.4 Proposed Methodology

This section presents the proposed methodology for creating the 2-layer transistor-level M3D. We have used the process and design metrics mentioned in Tab. 3.1 to create the standard cell design in M3D using a $28nm$ process. The standard cell characterization flow used in the study is presented in Fig. 3.4. We have used Synopsys's commercial CAD tools in the design flow. The design flow is explained below:

3.4.1 Technology File Extension

Since there are no existing process design kits (PDKs) and standard cell implementations for the $28nm$ M3D process, we used existing PDKs and extended them to incorporate M3D design elements. Towards this end, we first modified the technology files (.tf) and display resource files (.drf) to include layout and design information for the top and bottom-layer transistors, metal routing layers, and the MIV layers. These modifications were made using the Synopsys custom compiler CAD tool. The modification mainly included three aspects. First, a new set of bottom-layer definitions was introduced to complement the original top-layer device and local interconnect definitions, thereby explicitly distinguishing the top and



(a) External placement

(b) Internal placement

Figure 3.2: Comparison of MIV placement scenarios in 2-layer M3D: (a) external placement, where the MIV connects M2 in the bottom layer to M1 in the top layer, and (b) internal placement, where the MIV connects directly to the active region.

bottom layers in the layout editor. Second, a dedicated MIV structure and its associated via definitions were added to establish electrical connectivity between the bottom metal layer M2 and the top metal layer, M1. Third, the corresponding display properties, physical attributes, routing specifications, and geometric rule constraints of the newly introduced layers were incorporated into the technology file to ensure consistent recognition during layout editing, routing, and subsequent extraction. With these extensions, the original 2D technology description was transformed into a two-layer layout framework that supports bottom-layer device formation, intra-layer routing, and vertical connections between the

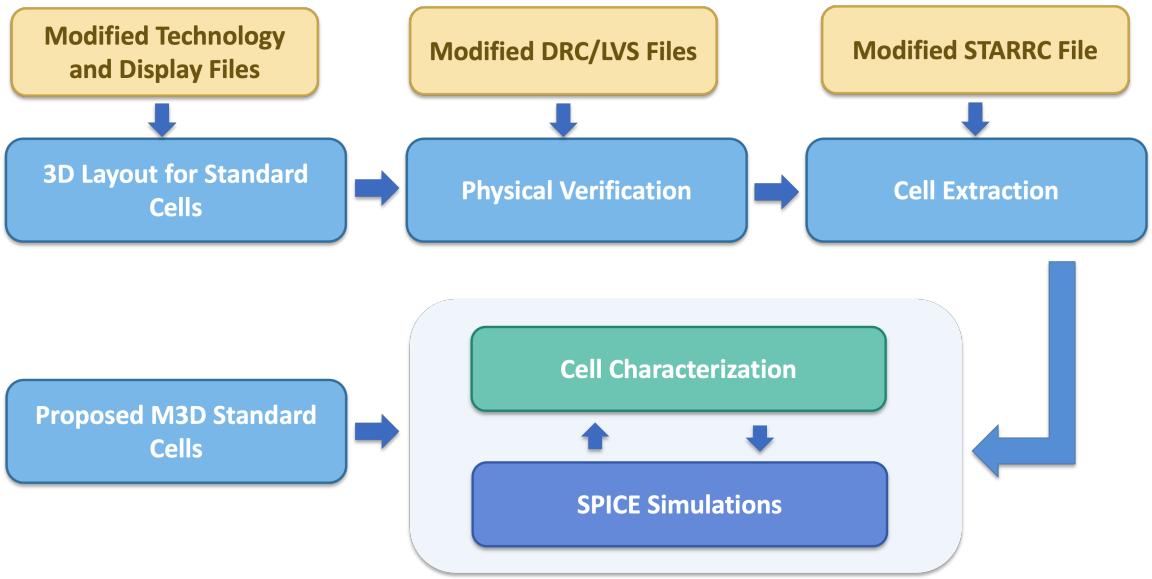


Figure 3.4: Standard Cell Characterization

bottom and top metal layers, thereby providing the foundation for subsequent DRC/LVS modifications and M3D standard-cell implementation.

3.4.2 DRC/LVS Rule Modification

For the physical verification flow, both the DRC and LVS rule files were revised to support the newly introduced M3D design layers. In the DRC file, the main modification was the addition of rules to support the bottom-layer objects used in the layout implementation, including poly, active, contact, metal, and via layers, as well as the MIV. Using these DRC checks, the bottom-layer design could be verified against spacing, enclosure, and area constraints. In the LVS file, the modification primarily focused on enabling accurate device recognition and connectivity extraction for the newly added bottom-layer structures. This included incorporating the bottom-layer device-related layers into the LVS recognition flow and extending the connectivity definition to ensure that the vertical connection through the MIV was correctly captured in the extracted netlist. Through these changes, the original 2D layout verification environment was extended into an M3D-aware flow that reliably checks

and validates the proposed two-layer layouts.

3.4.3 M3D Cell Layout Construction

Using the standard cells from [54], we characterized seven standard cells: INV1X1, NAND2X1, OR2X1, AND2X1, NOR2X1, MUX21X1, and AOI21X1, in conventional 2D implementations using the existing PDK, which served as the reference baseline for area, delay, slew, and power comparisons in Section 3.5. We used a cell-folding technique to implement the M3D standard cells as shown in Fig. 3.5. In this approach, the PMOS and NMOS transistors of each 2D cell as shown in Fig. 3.5(a) are partitioned into separate vertical layers: PMOS in the bottom layer and NMOS in the top layer, effectively folding the planar cell into the vertical dimension. Then, each 2D object in the bottom layer was replaced with a corresponding object from the modified PDK.

As mentioned in Section 3.3, for the external placement version, we created a via stack that connects the bottom-layer M2 to the top-layer M1 using MIV to connect the gate and active regions on the top layer. Then a CO contact was used to connect to different gates or active regions using M1 in the top-layer as shown in Fig. 3.5(b). A design rule was created to specify a minimum separation between MIV-PO region in the top layer and the MIV-active regions for external placement. However, for internal placement, to facilitate the direct connection to the PO and active regions, we have relaxed this constraint. The internal MIV can directly connect the bottom side of the active and PO regions, as shown in Fig. 3.5(c).

3.4.4 Post-Layout Parasitic Extraction

For post-layout evaluation, parasitic extraction was performed for three layout styles: conventional 2D, M3D with external placement, and M3D with internal placement. Using the existing interconnect technology files (.itf), parasitic resistance and capacitance were extracted from each layout using the StarRC extraction tool [55] to generate the corresponding SPICE netlists. For the M3D cases, the top and bottom layouts were extracted separately, and their SPICE subcircuits were connected using MIVs to form the complete

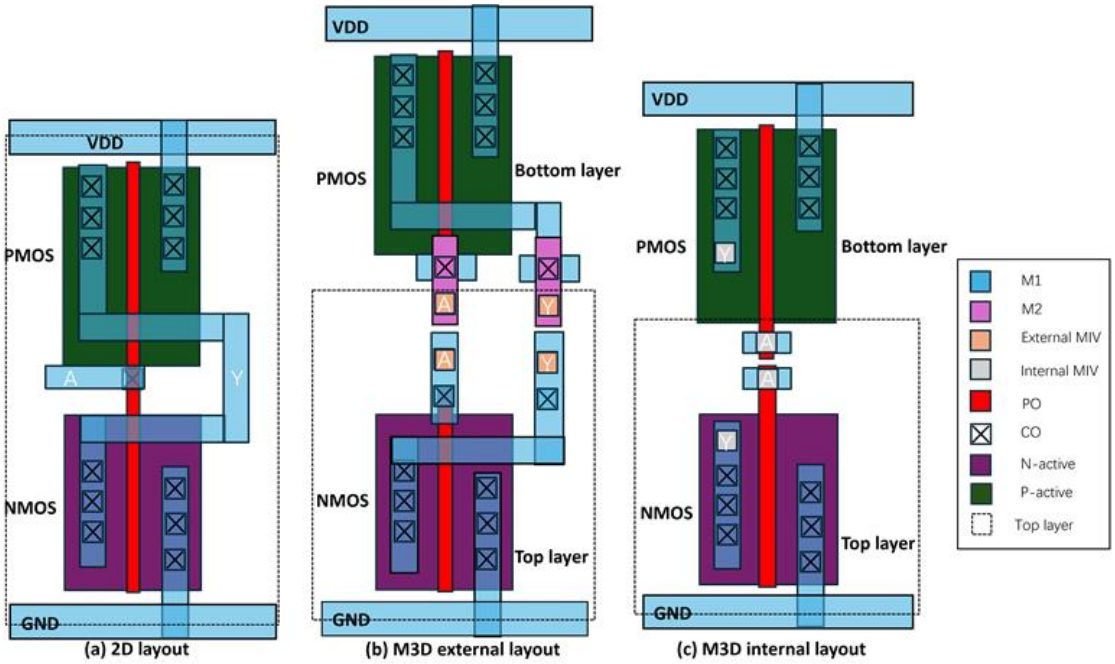


Figure 3.5: Comparison of standard cell layouts for the INV1X1 gate: (a) conventional 2D implementation, (b) M3D implementation with external MIV placement, and (c) M3D implementation with internal MIV placement. The dashed line indicates the area footprint in the top layer.

post-layout M3D netlist. The MIV parasitic resistance and capacitance were explicitly included in the characterization flow; based on its physical dimensions, the MIV resistance was assumed to be 3.8Ω and capacitance as $40aF$ [14]. To simplify the extraction, the parasitic capacitance of the ILD regions was assumed to be negligible [56]. Based on the extracted netlists, HSPICE simulations were performed for all input transition combinations to evaluate propagation delay, signal slew, and power for all three implementations.

3.5 Simulation Results and Performance Assessment

This section presents the performance, power, and area comparison of the standard cells characterized using the extraction flow presented in Section 3.4. Based on the extracted

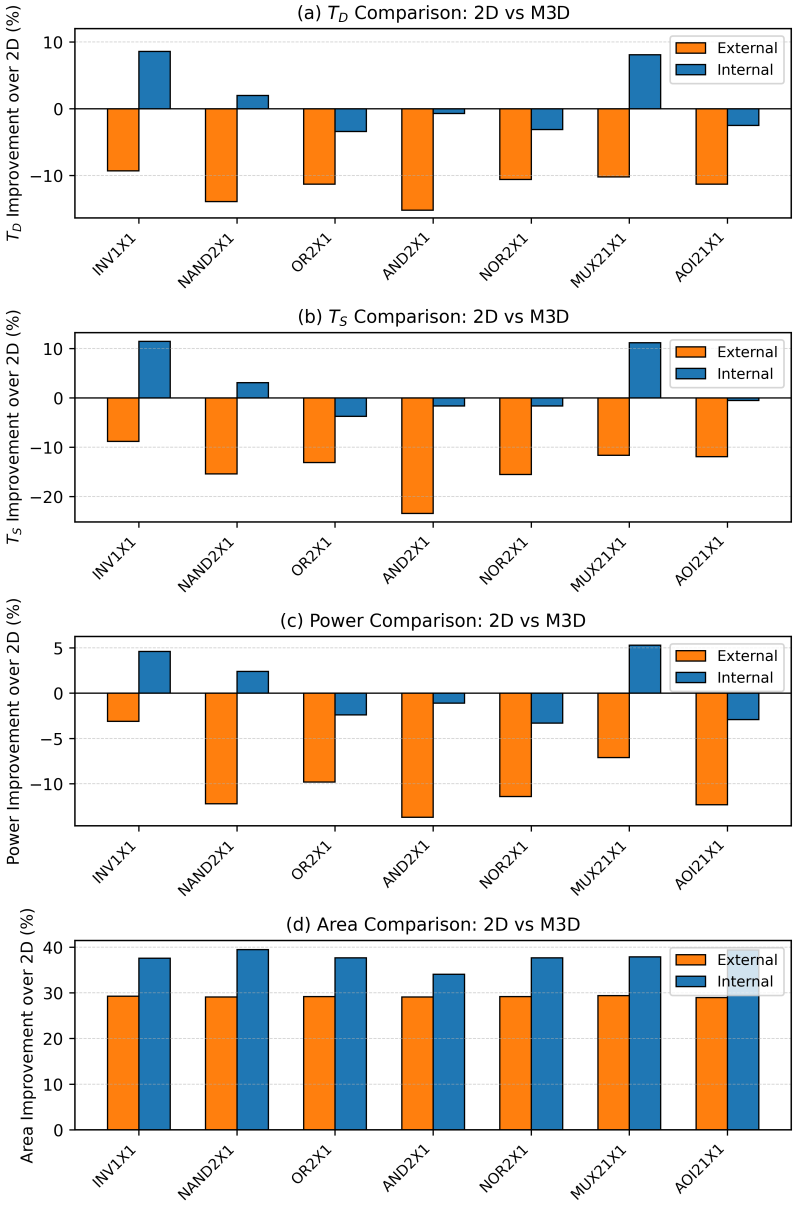


Figure 3.6: PPA improvement of M3D standard-cell implementations over the 2D baseline for (a) propagation delay T_D, (b) signal slew T_S, (c) power, and (d) area. Positive values indicate improvement and negative values indicate degradation relative to the 2D baseline.

netlists, circuit simulations were performed for all input transitions, and the resulting delay, slew, and power were evaluated. Table 3.2 compares the conventional 2D implementations

with the M3D designs using external-MIV and internal-MIV placements for seven representative logic cells. The worst propagation delay (T_D), signal slew (T_S), and average power (Power) were extracted from HSPICE-based transient simulations and summarized in the table. The area metrics for M3D were extracted using the top-layer layout dimensions, which determine the overall cell footprint, and compared with the corresponding 2D layout dimensions.

In terms of T_D, T_S, power, and area, the results show that both M3D styles provide substantial area savings over the 2D baseline. In particular, the M3D implementations with internal MIVs achieve the most favorable overall tradeoff, providing large area reduction together with modest average improvements in delay, slew, and power. In contrast, the external-MIV implementations preserve the area benefit of vertical integration but introduce noticeable degradation in timing and power. The detailed comparisons are given below.

3.5.1 Performance Comparison:

Table 3.2 shows the T_D and T_S results for all seven standard cells under both M3D placement scenarios. For external placement, all standard cells used in the study exhibited degradation in both T_D and T_S. Based on the results, the average degradation was 11.7% in T_D and 14.2% in T_S. The worst case T_D degradation of 15.2% occurred in AND2X1, while the worst case T_S degradation of 23.4% also occurred in AND2X1. These penalties are primarily attributed to the additional parasitic resistance and capacitance introduced by the via stack and the associated metal routing layers required to reach the target gate and active regions in the top layer.

For internal MIV placement, an average improvement of 1.3% T_D and 2.6% T_S was achieved over the 2D baseline as shown in Figs. 3.6(a) and 3.6(b). The best-case T_D improvement of 8.6% was reported for INV1X1, followed by MUX21X1 at 8.1% for the internal MIV placement. The best-case T_S improvements of 11.5% and 11.2% were observed for INV1X1 and MUX21X1, respectively. However, cells with more complex topologies, including NOR2X1, AND2X1, AOI21X1, and OR2X1, exhibited marginal T_D and T_S degradations ranging from 0.5% to 3.7%, due to increased MIV connectivity overhead relative to

the routing reduction achieved.

3.5.2 Power Comparison:

For external MIV placement, all seven cells exhibited higher power consumption than the 2D baseline, with an average degradation of 9.9%. The worst-case penalty of 13.7% occurred for AND2X1, followed by AOI21X1 at 12.3% and NAND2X1 at 12.2%. This increase is consistent with the timing degradation observed under external placement, due to higher parasitic effects at the cells' input and output pins, increased overall switching delays, and increased dynamic power consumption.

Under internal MIV placement, an average power improvement of 0.4% was achieved over the 2D baseline as shown in Fig. 3.6(c). The best-case improvement of 5.3% was observed for MUX21X1, followed by INV1X1 at 4.6% and NAND2X1 at 2.4%. However, cells including NOR2X1, AND2X1, AOI21X1, and OR2X1 exhibited marginal power degradations ranging from 1.1% to 3.3%, consistent with the minor T_D and T_S penalties observed for the same cells in the performance comparison.

3.5.3 Area Comparison:

Both internal and external MIV placements achieve substantial area reductions relative to the 2D implementation, as shown under the area metric in Table 3.2. For external MIV placement, an average area reduction of 29.2% was achieved across all cells. This area reduction is due to the cell-folding technique and is consistent throughout, driven by the mismatch between PMOS and NMOS transistor sizing in standard-cell implementations [22] and the additional area required for external MIV placement. However, the proposed internal placement achieved a significant reduction in area of 37.7% compared to 29.2% for the external placement, in the average case. For the internal placement, the best-case reduction of 39.5% and 39.4% was observed for NAND2X1 and AOI21X1, respectively, while the lowest reduction of 34.1% was observed for AND2X1.

The comparison between the two scenarios is presented in Fig. 3.6(d). Overall, the internal MIV placement scenario achieved a superior area efficiency across all layouts. It

achieved an additional average area savings of 8.5% compared to the external MIV placement scenario. As discussed in Section 3.3, the internal MIV placement eliminated the use of any external via stack, directly reducing routing overhead and increasing layout compaction.

Table 3.2: Performance, power and area comparison of conventional 2D with different M3D placement scenarios

Gate	**Case**	T_D **(ps)**			T_S **(ps)**			**Power** (μ**W**)			**Area** (μm^2)		
		2D	**External**	**Internal**	**2D**	**External**	**Internal**	**2D**	**External**	**Internal**	**2D**	**External**	**Internal**
INV1X1	Value	7.43	8.12	6.79	26.46	28.80	23.42	0.89	0.91	0.85	1.44	0.75	0.58
	% Improvement	–	**-9.3%**	**8.6%**	–	**-8.8%**	**11.5%**	–	**-3.1%**	**4.6%**	–	**29.3%**	**37.6%**
NAND2X1	Value	40.32	45.91	39.53	14.68	16.94	14.23	2.36	2.65	2.30	2.71	1.35	1.12
	% Improvement	–	**-13.9%**	**2.0%**	–	**-15.4%**	**3.1%**	–	**-12.2%**	**2.4%**	–	**29.1%**	**39.5%**
NOR2X1	Value	40.86	45.19	42.11	14.51	16.75	14.75	2.46	2.74	2.54	2.71	1.31	1.08
	% Improvement	–	**-10.6%**	**-3.1%**	–	**-15.5%**	**-1.6%**	–	**-11.4%**	**-3.3%**	–	**29.2%**	**37.7%**
AND2X1	Value	48.67	56.05	49.00	11.02	13.60	11.19	3.32	3.77	3.35	2.94	1.47	1.22
	% Improvement	–	**-15.2%**	**-0.7%**	–	**-23.4%**	**-1.6%**	–	**-13.7%**	**-1.1%**	–	**29.1%**	**34.1%**
AOI21X1	Value	50.28	55.93	51.51	16.06	17.97	16.15	2.66	2.99	2.74	3.22	1.86	1.59
	% Improvement	–	**-11.3%**	**-2.5%**	–	**-11.9%**	**-0.5%**	–	**-12.3%**	**-2.9%**	–	**29.0%**	**39.4%**
OR2X1	Value	49.43	55.03	51.11	11.50	13.01	11.93	3.45	3.79	3.54	3.26	1.58	1.24
	% Improvement	–	**-11.3%**	**-3.4%**	–	**-13.1%**	**-3.7%**	–	**-9.8%**	**-2.4%**	–	**29.2%**	**37.7%**
MUX21X1	Value	54.89	60.47	50.44	39.51	44.09	35.09	2.53	2.71	2.40	3.47	1.78	1.45
	% Improvement	–	**-10.2%**	**8.1%**	–	**-11.6%**	**11.2%**	–	**-7.1%**	**5.3%**	–	**29.4%**	**37.9%**
	% Average	–	**-11.7%**	**1.3%**	–	**-14.2%**	**2.6%**	–	**-9.9%**	**0.4%**	–	**29.2%**	**37.7%**

Chapter 4

CONCLUSION

This thesis studies both synchronous and asynchronous circuit design in M3D. As ICs continue to face limitations from conventional 2D scaling, M3D provides a promising path for improving integration density and interconnect efficiency through vertical device stacking and fine-grained MIVs. However, the benefits of M3D depend strongly on circuit style, device-tier partitioning, inter-tier connectivity, and layout-level design decisions. Therefore, this thesis examines how M3D can be applied to different classes of digital circuits and how its physical design choices affect circuit-level behavior.

For asynchronous circuits, this thesis focuses on QDI designs based on NCL. M3D introduces reliability concerns related to sequential fabrication, process variation, thermal effects, and timing uncertainty. QDI asynchronous circuits can help address these concerns by eliminating the global clock and relying on local handshaking, making them robust to timing variations. At the same time, M3D can benefit NCL-based asynchronous circuits by reducing their area overhead, shortening interconnects, and improving delay and power. This part of the thesis studies the complementary relationship between M3D and QDI asynchronous design through M3D-based NCL threshold gates and an unsigned NCL array multiplier.

For synchronous circuits, this thesis focuses on M3D standard-cell design and layout methodology. MIV placement is treated as a key design factor because it directly affects routing resources, layout footprint, parasitic loading, delay, skew, and power. A methodology is developed to compare conventional 2D and M3D standard cells and to evaluate how different MIV placement strategies influence PPA. This study shows how layout-level decisions determine whether the density advantages of M3D can be translated into practical circuit benefits. Together, these two studies evaluate M3D as both a circuit-level opportunity and a layout-level design methodology for future digital systems.

BIBLIOGRAPHY

[1] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[2] Chris A Mack. Fifty years of Moore's law. *IEEE Transactions on semiconductor manufacturing*, 24(2):202–207, 2011.

[3] Bei Yu, Xiaoqing Xu, Subhendu Roy, Yibo Lin, Jiaojiao Ou, and David Z Pan. Design for manufacturability and reliability in extreme-scaling VLSI. *Science China Information Sciences*, 59(6):061406, 2016.

[4] Jinwoo Kim, Lingjun Zhu, Hakki Mert Torun, Madhavan Swaminathan, and Sung Kyu Lim. A PPA study for heterogeneous 3-D IC options: Monolithic, hybrid bonding, and microbumping. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 32(3):401–412, 2023.

[5] Madhava Sarma Vemuri and Umamaheswara Rao Tida. FDSOI Process Based MIV-transistor Utilization for Standard Cell Designs in Monolithic 3D Integration. In *IEEE 36th International System-on-Chip Conference (SOCC)*, pages 1–6, 2023.

[6] Umamaheswara Rao Tida and Madhava Sarma Vemuri. Efficient metal inter-layer via utilization strategies for three-dimensional integrated circuits. In *IEEE 33rd International System-on-Chip Conference (SOCC)*, pages 195–200. IEEE, 2020.

[7] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Efficient and Scalable MIV-transistor with Extended Gate in Monolithic 3D Integration. In *IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4, 2023.

[8] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Dual-Purpose Metal Interlayer Via Utilization in Monolithic Three-Dimensional (M3D) Integration. In *IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 424–427. IEEE, 2020.

[9] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Modeling and Design of Dual-purpose MIV in Monolithic 3D IC. *IEEE Access*, 2024.

[10] K.M. Fant and S.A. Brandt. NULL Convention Logic/sup TM/: a complete and consistent logic for asynchronous digital circuit synthesis. In *Proceedings of International Conference on Application Specific Systems, Architectures and Processors: ASAP '96*, pages 261–273, 1996.

[11] Scott Smith and Jia Di. *Designing asynchronous circuits using NULL convention logic (NCL)*. Springer Nature, 2022.

[12] Jia Di and Scott C Smith. *Asynchronous Circuit Applications*. Institution of Engineering and Technology, 2019.

[13] Danylo Khodosevych and Ashiq A. Sakib. Evolution of NULL Convention Logic Based Asynchronous Paradigm: An Overview and Outlook. *IEEE Access*, 10:78650–78666, 2022.

[14] Chen Yan and Emre Salman. Mono3D: Open Source Cell Library for Monolithic 3-D Integrated Circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(3):1075–1085, 2018.

[15] Shreepad Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim. High-density integration of functional modules using monolithic 3D-IC technology. In *2013 18th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 681–686. IEEE, 2013.

[16] Shreepad A Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In *Proceedings of the 2014 international symposium on Low power electronics and design*, pages 171–176, 2014.

[17] Young-Joon Lee and Sung Kyu Lim. Ultrahigh density logic designs using monolithic 3-D integration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(12):1892–1905, 2013.

[18] Xiameng Zhang, Kushal K. Ponugoti, Ashiq A. Sakib, and Madhava Sarma Vemuri. Monolithic 3D Integration for Null Convention Logic (NCL)-Based Asynchronous Circuits. In *2026 27th International Symposium on Quality Electronic Design (ISQED)*, pages 1–8, 2026.

[19] Maud Vinet, Perrine Batude, Claire Fenouillet-Beranger, Laurent Brunet, Vincent Mazzochi, Cao-Minh Vincent Lu, Fabien Deprat, Jessy Micout, Bernard Previtali, Paul Besombes, et al. Opportunities brought by sequential 3D CoolCube™ integration. In *2016 46th European Solid-State Device Research Conference (ESSDERC)*, pages 226–229. IEEE, 2016.

[20] Hoonhee Han, Rino Choi, Seong-Ook Jung, Sung Woo Chung, Byung Jin Cho, Sheng-Chi Song, and Changhwan Choi. Low temperature and ion-cut based monolithic 3D process integration platform incorporated with CMOS, RRAM and photo-sensor circuits. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 15–6. IEEE, 2020.

[21] Bipin Rajendran, Rohit S Shenoy, Daniel J Witte, Nehal S Chokshi, Robert L DeLeon, Gary S Tompa, and R Fabian W Pease. Low thermal budget processing for sequential 3-D IC fabrication. *IEEE Transactions on Electron Devices*, 54(4):707–714, 2007.

[22] Bon Woong Ku, Taigon Song, Arthur Nieuwoudt, and Sung Kyu Lim. Transistor-level monolithic 3D standard cell layout optimization for full-chip static power integrity. In *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE, 2017.

[23] Madhava Sarma Vemuri, Tanvir Ahmed, and Umamaheswara Rao Tida. Compact 6T-SRAM using bottom-gate transistor in FD-SOI process for Monolithic-3D Integration. In *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 725–729. IEEE, 2024.

[24] Alex Kondratyev and Kelvin Lwin. Design of asynchronous circuits by synchronous CAD tools. In *Proceedings of the 39th annual Design Automation Conference*, pages 411–414, 2002.

[25] Cheoljoo Jeong and Steven M Nowick. Optimization of robust asynchronous circuits by local input completeness relaxation. In *2007 Asia and South Pacific Design Automation Conference*, pages 622–627. IEEE, 2007.

[26] Cheoljoo Jeong and Steven M Nowick. Block-level relaxation for timing-robust asynchronous circuits based on eager evaluation. In *2008 14th IEEE International Symposium on Asynchronous Circuits and Systems*, pages 95–104. IEEE, 2008.

[27] Alexander Bodoh and Ashiq A. Sakib. ASCEND: Advanced Synthesis, Circuit Exploration, and Design Optimization for NCL Circuits. In *2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 1–4, 2024.

[28] Farhad A Parsan and Scott C Smith. CMOS implementation of static threshold gates with hysteresis: A new approach. In *2012 IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC)*, pages 41–45. IEEE, 2012.

[29] Matheus Trevisan Moreira, Michel Arendt, Fernando Gehm Moraes, and Ney Laert Vilar Calazans. Static Differential NCL Gates: Toward Low Power. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 62(6):563–567, 2015.

[30] Kelby Haulmark, Wassim Khalil, William Bouillon, and Jia Di. Comprehensive Comparison of NULL Convention Logic Threshold Gate Implementations. In *2018 New Generation of CAS (NGCAS)*, pages 37–40, 2018.

[31] Ashiq A. Sakib and Scott C. Smith. Implementation of Static NCL Threshold Gates Using Emerging CNTFET Technology. In *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 1–4, 2020.

[32] Ashiq A. Sakib, Abir A. Akib, and Scott C. Smith. Implementation of FinFET Based Static NCL Threshold Gates: An Analysis of Design Choice. In *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 37–40, 2020.

[33] Yu Bai, Ronald F DeMara, Jia Di, and Mingjie Lin. Clockless spintronic logic: A robust and ultra-low power computing paradigm. *IEEE Transactions on Computers*, 67(5):631–645, 2017.

[34] Francis Sabado II. *Asynchronous 3D (Async3D): Design Methodology and Analysis of 3D Asynchronous Circuits*. University of Arkansas, 2017.

[35] Sandeep Kumar Samal, Deepak Nayak, Motoi Ichihashi, Srinivasa Banna, and Sung Kyu Lim. Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology. In *2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pages 1–2. IEEE, 2016.

[36] Neela Lohith Penmetsa, Christos Sotiriou, and Sung Kyu Lim. Low Power Monolithic 3D IC Design of Asynchronous AES Core. In *2015 21st IEEE International Symposium on Asynchronous Circuits and Systems*, pages 93–99, 2015.

[37] NC State University. FreePDK45 - 45nm Open-Source PDK, 2025. Accessed: 2025-05-21.

[38] A. W. Topol, D. C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong. Three-dimensional integrated circuits. *IBM Journal of Research and Development*, 50(4.5):491–506, 2006.

[39] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Metal Inter-layer Via Keep-out-zone in M3D IC: A Critical Process-aware Design Consideration. In *24th International Symposium on Quality Electronic Design (ISQED)*, pages 1–8. IEEE, 2023.

[40] Viet Nguyen, P. Christie, A. Heringa, Aatish Kumar, and R. Ng. An analysis of the effect of wire resistance on circuit level performance at the 45-nm technology node. pages 191 – 193, 07 2005.

[41] David E Muller. Asynchronous logics and application to information processing. In *Symposium on the Application of Switching Theory to Space Technology*, pages 289–297. Stanford University Press, 1962.

[42] Young-Joon Lee, Patrick Morrow, and Sung Kyu Lim. Ultra high density logic designs using transistor-level monolithic 3D integration. In *Proceedings of the International Conference on Computer-Aided Design*, pages 539–546, 2012.

[43] Claire Wolf. Yosys Open SYnthesis Suite. `https://yosyshq.net/yosys/`.

[44] Ryan A Taylor and Robert B Reese. Uncle—Unified NCL Environment—an NCL design tool. IET, 2019.

[45] Kyungwook Chang, Kartik Acharya, Saurabh Sinha, Brian Cline, Greg Yeric, and Sung Kyu Lim. Impact and design guideline of monolithic 3-d ic at the 7-nm technology node. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 25(7):2118–2129, 2017.

[46] Claire Fenouillet-Beranger, Bernard Previtali, Perrine Batude, Fabrice Nemouchi, Mikaël Cassé, Xavier Garros, Lucie Tosti, Nils Rambal, Dominique Lafond, Hugo Dansas, et al. FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration. In *2014 44th European Solid State Device Research Conference (ESSDERC)*, pages 110–113. IEEE, 2014.

[47] Perrine Batude, C Fenouillet-Beranger, L Pasini, V Lu, F Deprat, L Brunet, B Sklenard, F Piegas-Luce, M Cassé, B Mathieu, et al. 3dvlsi with coolcube process: An alternative path to scaling. In *2015 Symposium on VLSI Technology (VLSI Technology)*, pages T48–T49. IEEE, 2015.

[48] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Small footprint 6t-sram design with miv-transistor utilization in m3d-ic technology. In *2023 IEEE 41st International Conference on Computer Design (ICCD)*, pages 118–125, 2023.

[49] Madhava Sarma Vemuri and Umamaheswara Rao Tida. Design approaches and consideration for a reliable and efficient Monolithic 3D Integration. In *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 533–538. IEEE, 2024.

[50] International roadmap for devices and systems. More Moore, 2024.

[51] P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, J. Jopling, C. Kenyon, S-H. Lee, M. Liu, S. Lodha, B. Mattis, A. Murthy, L. Neiberg, J. Neirynck, S. Pae, C. Parker, L. Pipes, J. Sebastian, J. Seiple, B. Sell, A. Sharma, S. Sivakumar, B. Song, A. St. Amour, K. Tone, T. Troeger, C. Weber, K. Zhang, Y. Luo, and S. Natarajan. High performance 32nm logic technology featuring 2nd generation high-k + metal gate transistors. In *2009 IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2009.

[52] Chang Liu and Sung Kyu Lim. A design tradeoff study with monolithic 3D integration. In *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, pages 529–536. IEEE, 2012.

[53] Camila Toledo de Carvalho Cavalcante. *Study, fabrication and characterization in view of 3D Sequential Integration*. Theses, Université Grenoble Alpes [2020-....], May 2021.

[54] R. Goldman, K. Bartleson, T. Wood, K. Kranen, V. Melikyan, and E. Babayan. 32/28nm Educational Design Kit: Capabilities, deployment and future. In *2013 IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia)*, pages 284–288, 2013.

[55] Synopsys. StarRC: Golden signoff parasitic extraction, 2026. Accessed: Jun. 3, 2026.

[56] P Batude, M Vinet, A Pouydebasque, C Le Royer, B Previtali, C Tabone, L Clavelier, S Michaud, A Valentian, O Thomas, et al. GeOI and SOI 3D monolithic cell integrations for high density applications. In *2009 Symposium on VLSI Technology*, pages 166–167. IEEE, 2009.