3D Nanofabric: Layout Challenges and Solutions for Ultra-scaled Logic Designs

Edouard Giacomin[1(✉)], Juergen Boemmels[2], Julien Ryckaert[2], Francky Catthoor[2,3], and Pierre-Emmanuel Gaillardon[1]

[1] University of Utah, Salt Lake City, UT, USA
{edouard.giacomin,pierre-emmanuel.gaillardon}@utah.edu
[2] IMEC, Leuven, Belgium
[3] KU Leuven, Leuven, Belgium

Abstract. In the past few years, novel fabrication schemes such as parallel and monolithic 3D integration have been proposed to keep sustaining the need for more cost-efficient integrated circuits. By stacking several devices, wafers, or dies, the footprint, delay, and power can be decreased compared to traditional 2D implementations. While parallel 3D does not enable very fine-grained vertical connections, monolithic 3D currently only offers a limited number of transistor tiers due to the high cost of the additional masks and processing steps, limiting the benefits of using the third dimension. This book chapter introduces an innovative planar circuit netlist and layout approach, enabling a new 3D integration flow called *3D Nanofabric*. The flow, consisting of N identical vertical tiers, is aimed at single instruction multiple data processor *Arithmetic Logic Units* (ALUs). By using a single metal routing layer for each vertical tier, the process flow is significantly simplified since multiple vertical layers can potentially be patterned at once, similar to the 3D NAND flash process. In our study, we thoroughly investigate the layout constraints arising from the *Nanofabric* flow and the non-crossing planar graph constraint and propose several techniques to overcome them. We then show that by stacking 32 layers to build a 32-bit ALU, the footprint is reduced by 8.7× compared to a conventional 7 nm FinFET implementation.

Keywords: 3D Logic Integration · Nanotechnologies · Emerging Technologies · Layout

1 Introduction

For many years, the semiconductor industry has continued to scale down the *Metal-Oxide-Semiconductor Field-Effect Transistor* (MOSFET) to increase the number of devices per area unit, thus enhancing the performances of *Integrated Circuits* (ICs). Novel transistor topologies have emerged in the past few years as an alternative to planar transistors, such as FinFETs [1]. They allow better electrostatic control, decreased leakage, and reduced short-channel effects, improving electrical performances. However, FinFETs still suffer from the short-channel

© IFIP International Federation for Information Processing 2021
Published by Springer Nature Switzerland AG 2021
A. Calimera et al. (Eds.): VLSI-SoC 2020, IFIP AICT 621, pp. 279–300, 2021.
https://doi.org/10.1007/978-3-030-81641-4_13

effect and other physical limitations, such as quantum effects [2], and can not be scaled indefinitely. Therefore, alternative routes are being investigated to: (i) first, keep pushing the cost scaling for a given performance and (ii) then pack more performance for the same cost to enable more functionality per area.

In particular, in recent years, three-dimensional integrated circuits (3D ICs) have been proposed [3–19]. A 3D IC is an integrated circuit manufactured by stacking silicon wafers, dies, or transistors. They are then interconnected vertically to achieve performance improvements at reduced power due to shorter interconnects than conventional 2D approaches. Furthermore, stacked device layers increase the number of transistors per unit footprint without requiring costly feature size reduction. In the past few years, two 3D integration schemes have emerged: parallel 3D [3–9], where wafers or dies are stacked and interconnected using *Through Silicon Vias* (TSVs) and bonding techniques, and monolithic 3D [10–19], where multiple layers of transistors and/or memory are deposited sequentially on top of one another on the same starting substrate.

While the large size of the TSVs limits the interconnection density of parallel 3D integration, monolithic 3D allows a finer interconnection granularity. However, state-of-the-art monolithic 3D works [10–19] are currently constrained by the number of active tiers (2–4), limiting the potential offered by 3D integration. In this book chapter, we extend our previous work [20] that introduces a new 3D integration scheme, called *3D Nanofabric*. The *Nanofabric* consists of N identical vertical tiers, each realizing the same logic function. As such, it can be used in *Single Instruction Multiple Data* (SIMD) processor *Arithmetic Logic Units* (ALUs), where each vertical tier is one ALU bit. We propose here to use a single metal routing layer at each vertical tier to greatly simplify the process flow, as multiple vertical layers can potentially be patterned at once. Note that the only practical way to process multiple vertical layers at once in a one-shot fashion, both for deposition and etching of materials, is to restrict the process flow to a single layer. This leads to a non-crossing planar graph requirement, which will be formulated a bit further. While we are aware of the challenges 3D technologies bring, such as thermal aspects including cooling, power distribution, yield, and reliability, those are out of the scope of this book chapter and are part of ongoing and future work. Instead, this book chapter focuses on the layout constraints and proves that conventional designs can be integrated into the *3D Nanofabric* flow, given the constraints of a planar graph without crossing wires within a vertical tier.

The contributions of this book chapter are:

- We introduce a novel 3D design style using a very simplified set of masks and describe a possible process flow that could enable a sufficiently high yield across all layers.
- We investigate the physical design constraints arising from our proposed *3D Nanofabric* flow.
- We propose several solutions at the gate and netlist levels to design complex logic gates under the different non-crossing planar graph layout constraints.
- We provide a footprint comparison of different conventional logic gates between our proposed *3D Nanofabric* and a 2D 7 nm FinFET implementation.

– At the circuit level, we show that by stacking up to 32 layers to build a larger 32-bit ALU, the footprint is reduced by 8.7× compared to a 2D planar 7 nm FinFET implementation.

The rest of this book chapter is organized as follows: Sect. 2, presents related work. Section 3 briefly presents the proposed *3D Nanofabric* concept and describes a possible technology process flow. Section 4 discusses the different physical design constraints of the *3D Nanofabric*. Section 5 proposes several solutions. Section 6 provides experimental footprint comparisons with a conventional 2D technology. Section 7 concludes this book chapter.

2 Background and Related Work

Our proposed *3D Nanofabric* aims at a similar objective as the 3D NAND, namely, to exploit repetitive vertical layers to decrease the footprint, but is targeted at logic applications. However, this can only be achieved by proposing a circuit netlist topology and layout that relies solely on a single layer where the device channel, poly, and metal wires are all embedded without any other crossing than the gate on top of the device channel. To the best of our knowledge, that is a crucial challenge that has not been enabled by any other proposed netlist approach in literature.

2.1 Parallel 3D

Parallel 3D integration [3–9], also called stacked 3D integration, refers to a 3D integration scheme in which devices on separate wafers are fabricated in parallel prior to a bonding or stacking step, as shown in Fig. 1(a). In this process, wafers, dies, or packages are vertically interconnected, allowing several partitioning schemes, such as subsystem, block or die. Parallel 3D can be realized by employing several techniques, like TSV [5–7] or bonding [8]. Bonding is used to join the surface of two wafers or chips using various chemical and physical processes [9], while TSVs are vertical connections that pass completely through a silicon wafer or a die. While TSVs allow a fine-grained integration of several dies into a single 3D stack, they also consume a significant area, which otherwise could be used for logic gates. As a result, the 3D interconnection density is considerably limited by the large size of the TSVs (μm range), and the maximum TSV density achievable today is around 10^5 vias/mm^2 [16].

2.2 Monolithic 3D

Monolithic 3D [10–19] refers to multiple transistor tiers and/or memory cells vertically stacked sequentially on the same starting substrate, as shown in Fig. 1(b). More particularly, the bottom transistor tier is first processed with or without interconnects, called *Intermediate Back-End-Of-Line* (iBEOL). The top tier is then processed, followed by a contact processing step. This 3D integration

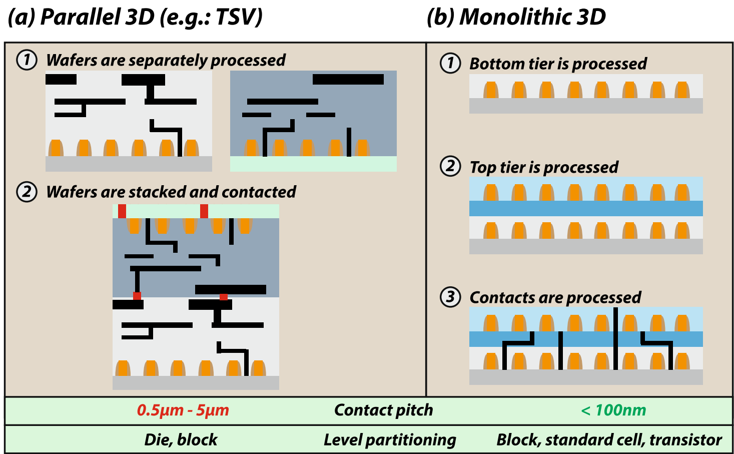

Fig. 1. 3D integration schemes: (a) Parallel integration (e.g.: TSV); (b) Monolithic integration.

scheme can achieve a very low 3D contact pitch, similar to a standard contact (<100 nm), since the devices can be stacked using the lithography precision aliment. Compared to the parallel integration scheme, monolithic 3D achieves a larger interconnection density (up to 2×10^7 vias/mm^2), using conservative 65 nm design rules [11]. However, monolithic 3D is challenging as processing the top tier can damage the bottom tier [13], so low thermal budget devices, such as junction-less devices [14], are required for the top tier. Moreover, it is difficult to obtain a stable iBEOL between the two tiers as copper metallization can contaminate the bottom tiers [13]. Monolithic 3D opens several opportunities, such as stacking 2 nodes $N-1$ instead of a node N [17], in a Logic-on-Logic or Memory-on-Logic way [10], or more disruptive approaches where emerging technologies can be stacked on top of CMOS [18,19]. However, only four active tiers have been demonstrated up to this date [19], limiting the benefits of using the third dimension. Besides, for pure homogeneous logic stacking, the potential cost benefits of these approaches to extend to more than 2 layers appears to be even more limited [10,12].

2.3 Other Logic 3D Technologies

Recent works proposed to use gate-all-around devices in an array fashion [21,22] to further decrease the footprint compared to parallel or monolithic 3D. In Skybridge 3D [22], a junctionless vertical nanowire template structure is employed to design static logic gates. As the template is pre-doped with p and n-type horizontal stripes, any static CMOS gate can be designed by forming the pull-up and pull-down networks through series and parallel devices. For instance, series networks are built with series devices implemented on a single nanowire, while parallel connections are achieved using devices on different nanowires. Similarly

to [22] a *Stacked Horizontal Nanowire based 3-D IC* (SN3D) was introduced [21], where junctionless horizontal nanowires are employed. Each static CMOS gate can be designed by stacking several nanowires on top of each other. Common contact and horizontal insulation features are used to connect or isolate the different source and drain regions to realize series and parallel connections. While these works showed a significant footprint reduction (5.5–40× s) compared to conventional 2D and monolithic 3D implementations, the number of masks and processing costs remain high as they have to be accumulated for every sequential layer that is added. This implies that no real cost scaling is feasible in this way. Hence, that is not compatible with our objectives, as introduced in the introduction.

2.4 3D NAND Memory

3D NAND memory has been proposed [23–25] to sustain the continuous demand for data storage. This 3D technology consists of many same vertical layers, stacked on top of each other and processed in a single shot. Since it has a highly repetitive mask set, 3D NAND technology is very cost-effective. Recently, up to 128 vertical layers have been demonstrated for the 3D NAND [24], resulting in a minimal footprint per stored bit. As a result, 3D NAND is currently replacing 2D NAND in the SSD market. While our proposed *3D Nanofabric* is aimed at logic applications and not memory, it uses a similar concept to 3D NAND as it consists of repetitive vertical layers stacked on top of each other, where multiple layers can be patterned at once. However, the complexity and challenges for this logic extension are highly non-trivial, as we will show in the rest of this book chapter. Hence, several disruptive novel aspects have to be employed to enable this.

3 Proposed *3D Nanofabric* Concept

In this section, we briefly summarize the proposed *3D Nanofabric* concept and then present a possible fabrication flow.

3.1 General Overview

The goal of our proposed 3D Nanofabric is to substantially decrease the manufacturing costs so that scaling many layers becomes truly attractive. This is achieved by: (i) considerably reducing the area by stacking many layers vertically; (ii) using a simplified process flow where all vertical layers can be patterned at once, similarly to what 3D NAND has achieved. While inspired by the 3D NAND process flow, our proposed 3D Nanofabric is aimed at logic applications. The proposed *3D Nanofabric* consists of N identical stacked vertical tiers, depicted in Fig. 2(a). In other words, the *3D Nanofabric* is a 3D ALU where each tier is an ALU bit. Hence, it is aimed at realizing SIMD processor datapaths,

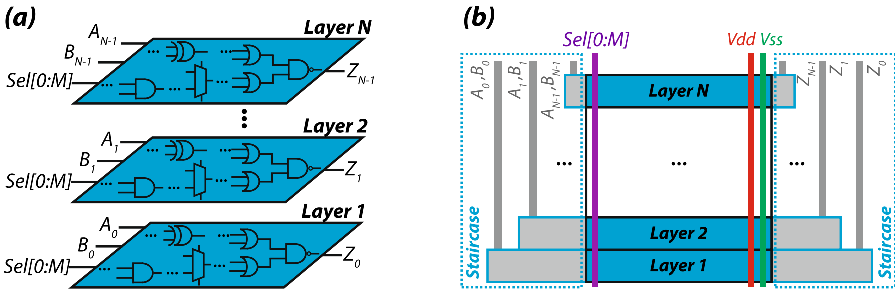

Fig. 2. *3D Nanofabric* concept: (a) Identical transistor tiers; (b) Cross-section general organization.

where the datapath is composed of an array of 3D ALUs. The way the *Nanofabric* communicates with the other parts of the processor (control, memory, *etc.*) is out of the scope of this book chapter and will be investigated in future work. To stack many vertical layers, we propose here to use a very restricted set of masks (i.e., only a single metal routing track), which allows multiple layers to be patterned at once during fabrication, as will be explained in Sect. 3.2. As shown in Fig. 2(b), the global signals which are shared among all the vertical layers, such as the select signals $Sel[0:M]$ (M depending on the number of operations the ALU can realize) or V_{dd} and V_{ss}, are provided through vertical pillars. The other signals (inputs and outputs of each ALU slice) are fed independently to each vertical layer from the side, using staircase-like structures similar to 3D NAND [26] chips. To stack many layers, we propose here to use a very restricted set of masks (i.e., only a single metal routing track) on top of using physically identical vertical tiers. This small set of masks and layout regularity enables a low-cost manufacturing process flow, in which multiple layers can be patterned at once, as will be explained in the next subsection.

3.2 Potential *3D Nanofabric* Process Flow

In this section, we briefly describe a possible technological solution for manufacturing the proposed *3D Nanofabric*. Based on the Coventor® modeling software, the process flow has been used to derive the design and layout rules presented in this section and employed to obtain the results of Sect. 6. Note that a more complete and thorough process flow study is out of the scope of this book chapter. While a simple solution would be to create the structure sequentially layer-by-layer, this would not be cost-effective at all as most steps would have to be repeated for each layer. Instead, we propose a solution that only uses a single metal routing layer and patterns multiple vertical layers at once.

The gate-forming processing flow steps are illustrated in Fig. 3(a)–(h). As shown in Fig. 3(a), the flow starts by depositing the layer-stack: for each vertical tier, we deposit an active layer (blue), a sacrificial layer (green), which will become the gate (dummy-gate, referred to as *GATE_INTEND*), and an

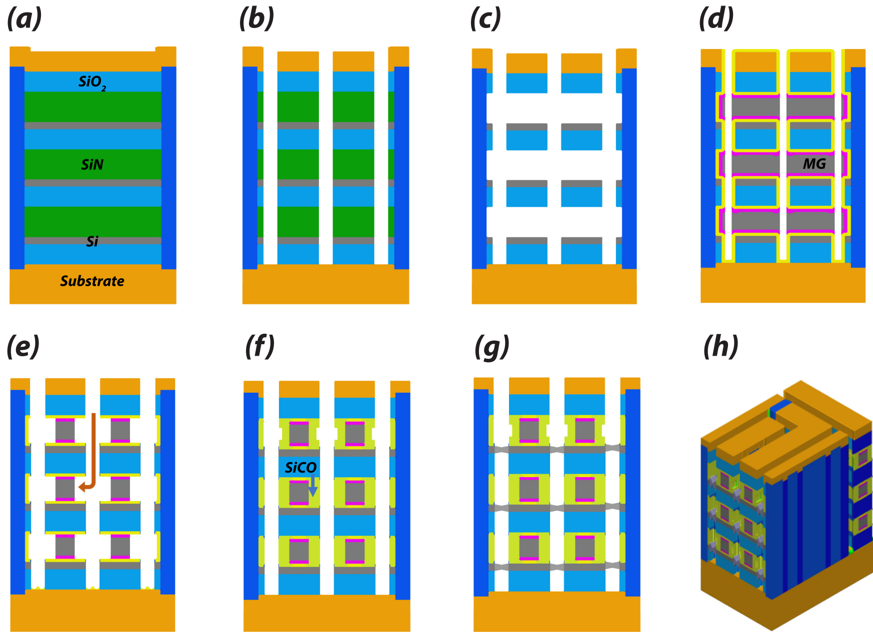

Fig. 3. *3D Nanofabric* gate forming process flow steps: (a) Layer-stack deposition; (b) Trenches creation, where source and drain regions will be formed; (c) Dummy-gate removal; (d) Gate-oxide and metal-gate filling; (e) Metal recess; (f) Spacer fill and etch-back; (g) Metal lines formation; (h) Resulting 3D structure. (Color figure online)

inter-dielectric layer (grey). Note that for the sake of the readability of the figure, only 3 active layers are depicted, but the described process is extendable to N tiers. While there are multiple possible options for creating active layers, we propose here to use a layer transfer of crystalline silicon, as it is done for *Silicon On Insulator* (SOI) processes. Those SOI-like silicon devices are well understood and have good electrical characteristics. The sacrificial layer may be a nitride layer, such as S_iN or some other material that can be etched with a sufficient selectivity with respect to the active and the inter-dielectric layers. The process relies on an indirect fabrication of the gates, which are formed in a collateral fashion. As depicted in Fig. 3(b), the layer-stack is patterned through an etching process by forming trenches through where source and drain regions will be formed. As such, a high-aspect-ratio etch is employed, such as reactive-ion etch or any suitable dry etching process. As a result, the layer-stack is then partitioned by a number of sub-stacks separated by trenches, referred to as channel-islands. The dummy-gate material is then removed by using a selective isotropic etch process, as shown in Fig. 3(c). As a considerable amount of material is removed from the layer-stack, a mechanical support is required for the active layer and inter-dielectric layer. This is achieved by the design rule that

every gate-island is abutting a vertical support wall of an insulating material (referred to as $OXWALL$), such as S_iO_2. As illustrated in Fig. 3(d), the gate dielectric and gate electrode materials are then formed through conformal deposition in the cavities obtained from removing the dummy-gate. Note that the trenches can then subsequently be re-etched (by reusing the previous hard mask of Fig. 3(b)) to remove the gate electrode material filling them. Then, insulating sidewall spacers are formed as follows: first, the metal gate lines are recessed from the side (Fig. 3(e)) by an isotropic metal etch, and then, the formed cavities are filled with the spacer material (Fig. 3(f)). As earlier, the excess material in the trenches is removed by an anisotropic high-aspect-ratio etch using the same hard mask of Fig. 3(b). Then, source and drain regions are formed at the end of the channel portion facing the trenches. These regions are doped using in-situ epitaxy doping. The next step is to form the wiring lines and vertical pillars (i.e., vias, referred to as $CONT_VERT$). For the vias, vertical holes are etched through the whole layer-stack. For the wiring lines (referred to as $METAL_LINE$), holes are formed, which are used as filling ports for the metal lines. The metal lines are filled over the whole length of the line through these filling ports. Therefore, a very conformal deposition is needed to avoid pinch-off. The metal is then removed from the plugs (referred to as $METAL_CUT$) and refilled with a dielectric to cut the wiring metal lines at specific locations. As shown in Fig. 3(g), the wiring lines extend across and over the source and drain regions of the active semiconductor.

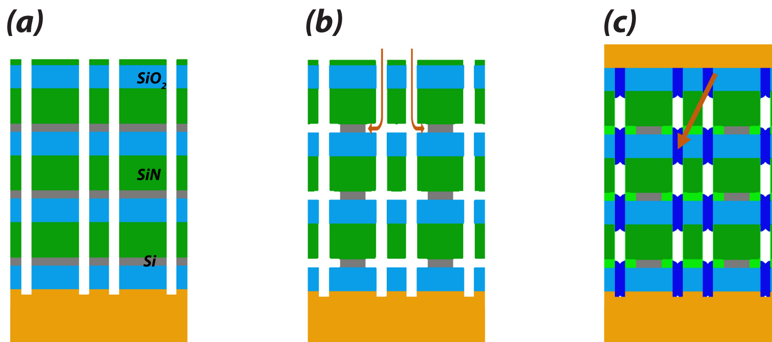

Fig. 4. *3D Nanofabric* active patterning process flow steps: (a) Cut a narrow gap; (b) Silicon isotropic etch; (c) Oxide selective deposit to recreate the vertical gating.

Figure 3(h) shows the resulting 3D structure. Gate lines extend across and over the channel region portions of the horizontal channel transistors. The gate lines and wiring lines are arranged side-by-side and their separation is ensured by the spacers. The single layer of the gate lines and wiring lines of each logic cell

of each device tier is readily visible in the figure, indicating a common geometric horizontal plane intersected by all gate lines and wiring lines of each logic cell.

Similar to the gate patterning, the *ACTIVE* layer employs sideways processing. After the active patterning, we need to "repair" the inter-dielectric layer, as the gate is crossing over the edge of the active. "Repairing" can only be done over small distances, so the initial patterning does not use final dimensions. The active design will be upsized until only a small gap is left. As illustrated in Fig. 4(a), this gap is etched into the layer-stack with another high-aspect-ratio etch, which exposes the active on the sidewall. As shown in Fig. 4(b), a high selective silicon etch is then used to trim the active silicon to the target size. This way of patterning will impact some design restrictions on the *ACTIVE* layer: the distance between two *ACTIVE*-polygons should either be the nominal value or be big enough to fit a double-gap into it. Once the active is patterned, the inter-dielectric layer gap is closed by selective deposition of oxide on oxide, as depicted in Fig. 4(c).

3.3 *3D Nanofabric* Layout Examples

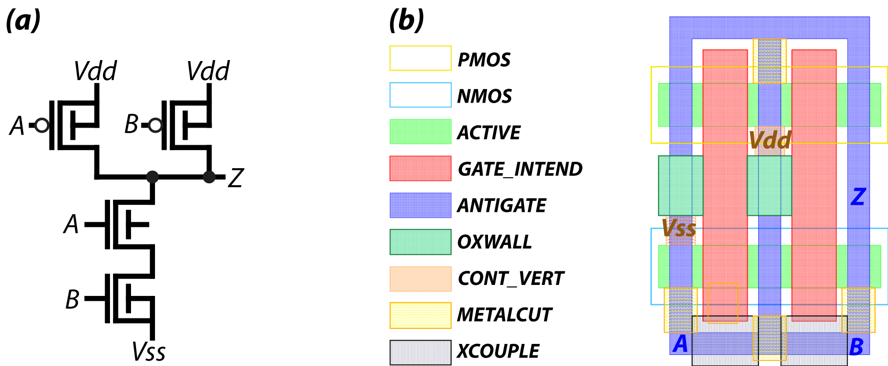

Fig. 5. NAND2: (a) Schematic; (b) Layout with layer legend.

The layout of a conventional NAND2 gate is depicted in Fig. 5(b) to illustrate our proposed flow. As discussed, each gate (*GATE_INTEND*) is surrounded by a *METAL_LINE* layer. As such, some metal breakers are required (*METAL_CUT*) to achieve all the different connections. The *XCOUPLE* layer is used to connect the gate and the routing layer (*METAL_LINE*). Two *OXWALL* squares in direct contact with the gate can be observed and are used to mechanically support the vertical structure. They also act as metal routing breakers. Besides, the V_{dd} and V_{ss} supply lines are fed through vertical pillars (brown *CONT_VERT* squares) to the logic gate. Note that, as explained earlier, the *GATE_INTEND* layer is not a physical mask as the gates are formed indirectly throughout the flow. This layer is only shown here for layout purposes to ease the design step. Also, the

XCOUPLE layer is used to form a connection between the *GATE_INTEND* and *METAL_LINE* layers.

Figure 6(a)–(c) depicts the layouts of an INV, NAND3, and SRAM6T cells, respectively. As can be observed, such simple gates can be efficiently designed with the proposed *3D Nanofabric* as their internal organization is straightforward, resulting in compact gates. This is because gates such as AND, OR, NAND, or NOR simply require a stack of series transistors and a stack of parallel transistors, so most of the source and drain regions can be shared. However, as will be described in the next sections, more complex logic gates require specific techniques to be designed and will result in an area overhead compared to traditional 2D layouts. Note that the SRAM 6T uses both vertical and horizontal gate patterns to result in a more compact gate.

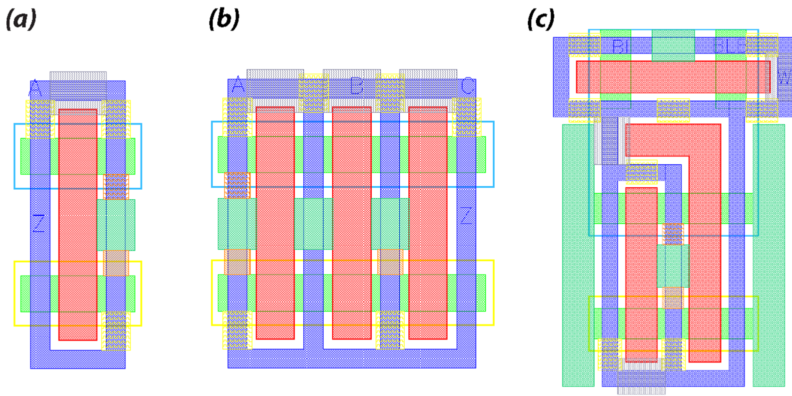

Fig. 6. Layout of various cells using the proposed *Nanofabric* rules: (a) INV; (b) NAND3; (c) SRAM6T.

4 Layout Challenges

In this section, we describe the different layout challenges arising from the technology assumptions and the *Nanofabric* manufacturing flow. It will be shown that these challenges are very different from the ones that have to be dealt with in the 3D NAND case.

4.1 Gate Layer Forming

In conventional 2D technologies (planar or FinFET), the metal routing layers often span across unrelated gate and active layers, as they are distinct from a processing point of view, as shown in Fig. 7, depicting a conventional planar layout

of an AOI211 gate. In the proposed *Nanofabric* flow, this is not possible due to the fabrication process: as explained in Sect. 3.2, the *GATE_INTEND* layer is derived from a boolean operation on the *ANTIGATE* layer. As such, it is strictly impossible to have the *ANTIGATE* layer spanning on the *GATE_INTEND* layer, as it is the case in traditional 2D designs, which limits the freedom in terms of physical layout. Also, as depicted in Fig. 5(b), the *GATE_INTEND* layer has to be surrounded on all sides by the *ANTIGATE*. As a result, some breakers have to be employed to achieve distinct connections on the different source and drain sides. While it is not an issue for a simple gate like the NAND2, it brings some challenges for more complex gates like the XOR2 or AO22.

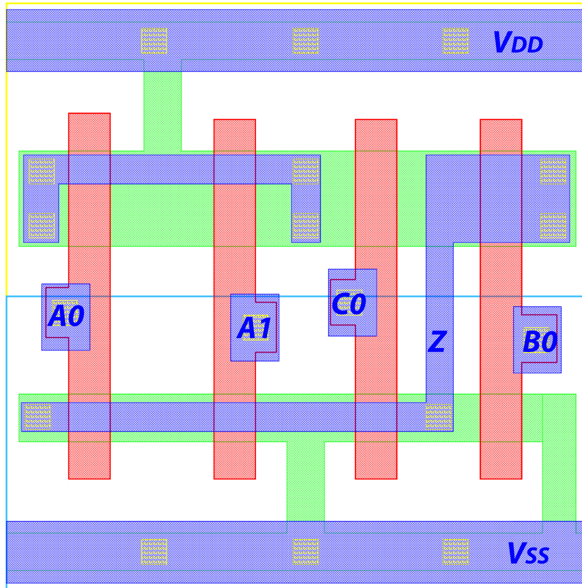

Fig. 7. Conventional planar layout of a AOI211 gate.

4.2 Single Metal Level Routing

As discussed earlier, the main layout limitation is that only a single metal routing track can be used within the *Nanofabric*, which considerably restricts the physical design. This means that when designing, no upper metal level layers can be used in case of metal crossing in high congestion areas. Without any crossing possibility, it means that complex gates, such as XOR2 or the FA are challenging to design. However, it is still possible to design such kinds of gates, and some solutions are proposed in the next section. Another requirement arising from the single metal rule is that the standard cell input and output pins have to be located on the border to be accessed externally, as illustrated in Fig. 5(b). Since

the flow only uses a single routing metal layer, there is no way to access the pins located in the center of the cell through higher levels of metals, as is the case in conventional 2D designs. As an example, the inputs *A1* and *C0* in Fig. 7 would make the cell non-routable using the proposed *Nanofabric* flow.

5 Layout Solutions

In this section, we present the algorithm, consisting of several steps, used in the *Nanofabric* to overcome the planar non-crossing layout restrictions. We first describe each step with examples and then provide the complete algorithm.

5.1 Step 1: Resolving Loops at the Cell Level

The first step to resolve metal crossing is to make sure that no metal loop is present within a single logic cell. Several techniques are employed:

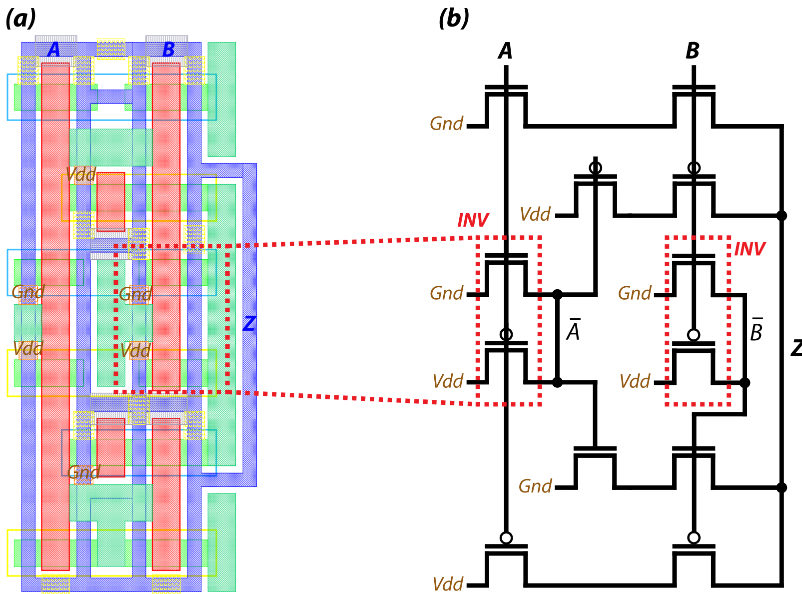

Fig. 8. XOR2 logic gate: (a) Layout using the proposed *Nanofabric* rules; (b) Transistor-level schematic. Note that the schematic is identical to a traditional static CMOS XOR2 gate.

Transistor Placement and Stacking: Due to the non-conventional way of designing logic cells, there is more freedom to move the transistors vertically and horizontally, instead of having fixed top *p-well* and bottom *n-well* zones

as in traditional 2D designs. While this is not the case for simple gates like the NAND2, more complex gates will require such arrangement, as depicted in Fig. 8 for a XOR2. Due to the complexity of the XOR2 cell and the non-crossing planar graph constraints, the transistor sharing the same gate signals (mainly A and B) are all stacked on each other to relieve congestion within the cell. In particular, the internal inverters are also stacked as they share the same inputs as the XOR2 gate. Note that unlike conventional design styles, there is no fixed height for the different logic cells, as complex gates such as the XOR2 will require a larger height due to the transistor stacking. Therefore, more different design styles are possible for a given cell, depending on its desired shape and internal structure.

Vertical Signals: Global signals, including V_{dd}, V_{ss}, or the ALU control signals shared among all the vertical layers to perform the same logic function, are provided to the *Nanofabric* through vertical pillars. In particular, unlike conventional 2D designs, the standard cell power supply grid lines are removed. This relieves metal routability since those signals will not block the metal routing layer. Note that for an ALU, the primary inputs and outputs are independent for each vertical tiers. Hence, they cannot be provided through vertical pillars spanning among all tiers. Instead, similarly to the 3D NAND process [26], staircase-like structures are employed to convey all the signals to the appropriate tier independently.

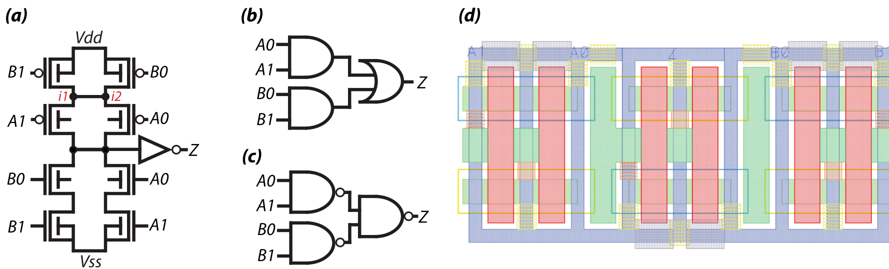

Fig. 9. AO22: (a) Transistor-level based schematic; (c) Gate-level based schematic using AND/OR gates; (d) Gate-level based schematic using NAND gates; (d) Layout using NAND gates with the proposed *Nanofabric* rules.

Gate-Based Logic Cells: A solution to design complex gates is to use gate-level based designs instead of transistor-level based designs. For the AO22 gate, which transistor level-based design is depicted in Fig. 9(a), the different connections, notably *i1* and *i2*, make it impossible to be designed using the proposed *Nanofabric* flow. Since each gate has to be surrounded by the metal layer, and there is only a single metal layer, these kinds of connections where 4 transistors share the same drain or source are particularly challenging. However, using the gate-level based design shown in Fig. 9(b) greatly simplifies the routing and

makes it possible by merely cascading basic gates (NAND2, NOR2, *etc.*). While the gate-level based design uses more transistors (18 instead of 10), it can be rearranged using De Morgan's equation, as shown in Fig. 9(c), and only uses 2 more transistors than the transistor-level based implementation. The layout of AOI22 gate based on NAND2 gates is depicted in Fig. 9(d).

Propagate an Internal Signal Using Inverters: Another way of resolving specific metal crossing is to propagate the signal within the logic gate. As depicted in Fig. 10, transistor $N1$ and $P1$ are driven by input A, while transistor $N2$ is driven by $\bar{A}$. One way to achieve this without crossing is to use a first inverter to generate $\bar{A}$ to drive transistor $N2$. Then, another inverter is used to invert signal $\bar{A}$ back to A to drive transistor $P1$. That way, signal A is propagated internally within the logic gate.

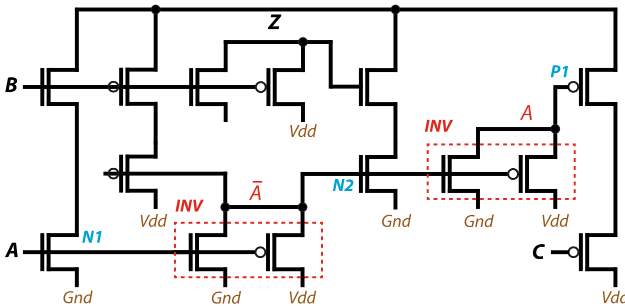

Fig. 10. Signal propagation using internal inverters. Here, signal A is propagated within the logic cell.

5.2 Step 2: Resolving Loops at the Netlist Level

Once all logic gates do not contain any internal metal loop, they can be used to build more complex blocks, such as a complete ALU. Duplicated gates can be used to resolve any additional metal loop in the netlist when connecting the different gates. As illustrated in Fig. 11(a), the input arrangement of the AO22 gate is causing a metal crossing, and there is no way to move the gates to overcome this issue. This metal crossing can be resolved by duplicating the OR2 gate (in blue) on the side. As depicted in Fig. 11(b), its output can now be connected to the AO22 gate without being confined, as it was the case before. Note that while it brings an area overhead, duplicating logic gates will always resolve any crossing issue as the gates can be duplicated up to the netlist primary inputs.

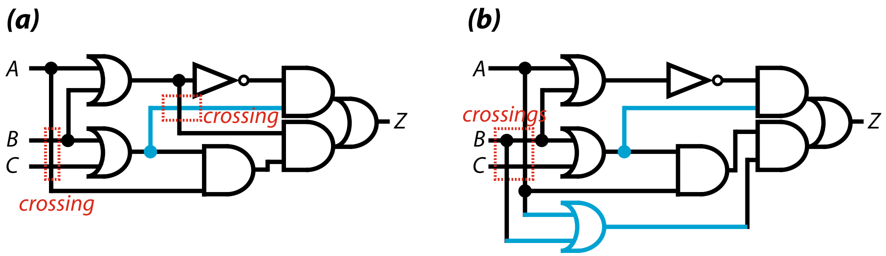

Fig. 11. Logic circuit schematic: (a) Containing 2 metal crossings; (b) Alleviating 1 metal crossing through duplicated inputs from the staircase. (Color figure online)

5.3 Step 3: Duplicating Signals Through Staircases and Vertical Signals

As explained in Sect. 3.1, each 2D layer will receive its primary inputs from its sides. However, the first logic level of the ALU may require some inputs to be fed to several parallel gates, implying possible metal crossing, as shown in Fig. 12(a). In this example, input B is driving three parallel gates. However, since there is no way to place them next to each other, the B metal wire has to cross inputs A and C. Since the primary inputs of each 2D layer are provided through a vertical staircase, they can be duplicated to be fed to more gates in the ALU. As depicted in Fig. 12(b), both metal crossings can be resolved by duplicating the primary inputs A and B. Besides, as using step 2 might also result in several duplicated primary inputs, the staircase will be able to feed them to the ALU while avoiding metal crossing. As the control signals are provided through vertical pillars, those can also be easily duplicated if they need to control several logic gates.

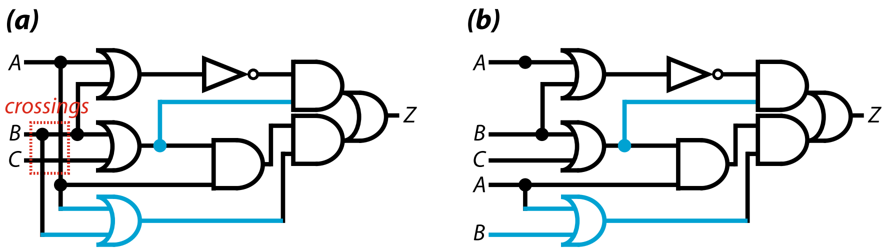

Fig. 12. Logic circuit schematic: (a) Containing 1 metal crossings; (b) Alleviating the metal crossing by using duplicated inputs from the staircase.

5.4 Non-crossing Planar Graph Algorithm

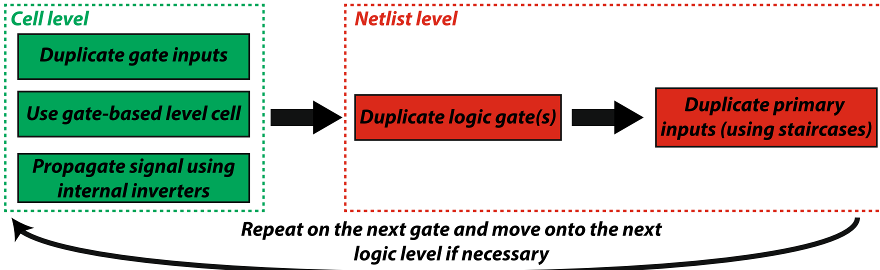

Fig. 13. Non-crossing planar graph algorithm illustration.

In this sub-section, we present the complete algorithm, illustrated on Fig. 13, to produce the layout for an ALU netlist while only using a single metal layer. The algorithm, described with more details in Algorithm 1, consists of all the previous layout solutions presented in Sect. 5 combined. The algorithm starts from one of the last logic gate (producing an output) and propagates backward through the netlist. It first solves the internal gate crossings for each gate, before solving the metal loops at the netlist level (between several gates). Once all the gates of a given logic level have been treated, it moves to the previous logic level until it reaches the primary inputs. If necessary, those primary inputs are duplicated through the staircases or the vertical signals. Here, we assume that the netlist does not contain feedback loops. While feedback loops are generally present in sequential circuits, the goal here is to design combinational ALUs for SIMD processors, so it is unlikely to happen. Besides, a proper synthesis of the ALU function would also eliminate the feedback loops within the netlist.

6 Experimental Results

In this section, we first describe our experimental methodology and then demonstrate the footprint benefits of our proposed *3D Nanofabric*.

6.1 Experimental Methodology

We developed an in-house PDK for the *3D Nanofabric* flow for the footprint evaluations, following the technological assumptions presented in Sect. 3.2. For the 2D baseline, we considered 2 cases: (a) the ASAP 7 nm FinFET design kit from ASU [27] and (b) an in-house FinFET IN7 node. For a fair area comparison, transistors are minimum sized in all cases. For *both 2D* cases, the ALU area values were obtained after synthesis by using the complete available logic libraries. For the 3D case, an extra step is performed to draw the layout by hand, following the novel approach described above.

Algorithm 1: *3D Nanofabric* non-crossing planar graph algorithm.

```
Starts at the output node (last level of logic depth);
Logic_level = Get_Total_Nb_Logic_Levels();
while (Logic_level != 1) do
    Number_gates = Get_Current_Logic_Level_Nb_Gates();
    while (Number_gates != 1) do
        if Current_gate has internal crossings then
            Duplicate_Gate_Inputs();
            Use_Gate_Based_Logic_Cell();
            Propagate_Signal_Using_Inverters();
        else
            Use_Transistor_Based_Logic_Cell();
        end
        Number_gates = Number_gates - 1;
    end
    if Crossing between gates then
        Duplicate_Gate();
    end
    Logic_level = Logic_level - 1;
end
if Crossing between primary inputs then
    Duplicate_Signals();
end
```

6.2 Logic Gate Area Comparison

Table 1 shows the area of a few conventional logic gates, using the proposed *3D Nanofabric* flow compared to other technologies.

Table 1. Logic gates area (in μm^2) using ASAP7, IN7 and the proposed *3D Nanofabric* process.

Gate	ASAP7	IN7	*3D Nanofabric*
INVD1	0.044	0.016	0.029
NOR2D1	0.058	0.024	0.041
AO22D1	0.092	0.040	0.127
XOR2D1	0.117	0.072	0.083
NOR3D1	0.073	0.032	0.052
Average	0.077 (−17%)*	0.037 (+1.8×)*	0.066

* *3D Nanofabric* area overhead/reduction, when compared to ASAP7 and IN7 respectively.

As expected, compared to a highly and aggressively optimized IN7 library, using the *3D Nanofabric* process brings an area overhead (1.8× on average) due to the non-crossing rule, which requires extra transistors or spacing for complex gates. In particular, the area overhead is even more significant for gate-level based cell such as the AO22 gate due to the additional transistors. Note that the logic gate area is reduced (17% on average) compared to ASAP7 since the proposed *Nanofabric* allows us to design compact gates, as the *nmos* and *pmos* transistors can be placed closer to each other. Besides, the significant difference between ASAP7 and IN7 is from the fact that IN7 is equivalent to a commercial foundry 5 nm technology node due to its aggressive dimensions and multiple design boosters enabling a 6-track library, while ASAP7 can only achieve 7.5 tracks.

6.3 ALU Footprint Comparison

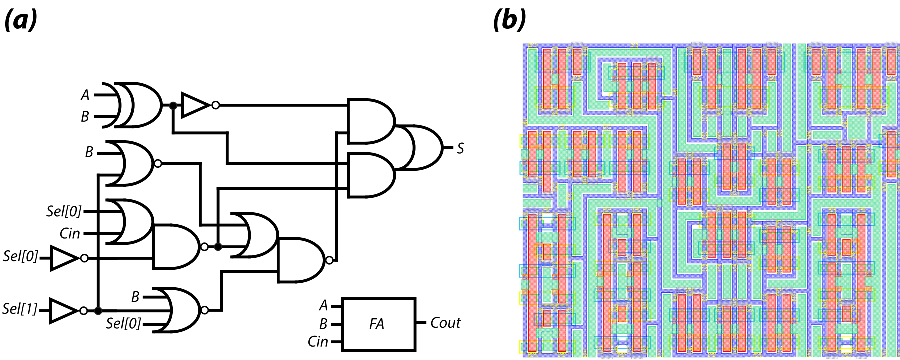

Fig. 14. 1-bit basic ALU: (a) Schematic; (b) Layout view using the proposed *3D Nanofabric* rules and process flow.

In this section, we consider a basic 1-bit ALU, whose schematic is depicted in Fig. 14(a). This 1-bit ALU is capable of performing the following operations:

- $A + B + C_{in}$
- $A \& B$
- $A | B$
- $A\hat{}B$

We study the area of the 1-bit ALU for all 3 cases using ASAP7, IN7, and the proposed *3D Nanofabric*. For both ASAP7 and IN7, we only considered gates using minimum sized transistors, as we made the same assumption for the proposed *3D Nanofabric*. The layout of the 1-bit ALU using the *3D Nanofabric* is shown in Fig. 14(b). Note the presence of several *OXWALL* regions, which fill the extra empty spaces required to route the single metal level layer. Here, there is no need for dummy-poly as in a FinFET technology where the gate is needed to define the Source-Drain. Instead, the empty spaces are filled with

the *OXWALL* dielectric layer. Also, the gate to gate distance is always enforced (36 nm) to ensure that all the gate are aligned, so the layout is fully regular. As shown in Table 2, ASAP7 and IN7 have a 1.6× and 3.7× smaller area than the proposed *3D Nanofabric*, respectively, for the 1-bit ALU as some gates have to be duplicated to avoid crossing. Besides, some extra space is required for routing where 2D processes use higher metal layers.

Table 2. *3D Nanofabric* ALU footprint compared to ASAP7 and IN7 for an N-bit ALU.

Number of bits N	Footprint (in μm^2)*		
	ASAP7	IN7	3D
1	0.787 (+1.6×)	0.338 (+3.7×)	1.257
2	1.822 (−1.4×)	0.758 (+1.7×)	1.257
3	2.186 (−1.7×)	1.193 (+1.05×)	1.257
4	2.668 (−2.1×)	1.516 (−1.2×)	1.257
8	4.765 (−3.8×)	2.991 (−2.4×)	1.257
16	11.033 (−8.8×)	5.539 (−4.4×)	1.257
24	16.169 (−12.9×)	8.265 (−6.6×)	1.257
32	21.257 (−16.9×)	10.999 (−8.7×)	1.257

* Also shows the *3D Nanofabric* footprint overhead/reduction, when compared to ASAP7 and IN7 respectively.

Note that while a single layer brings some area overhead due to the layout constraints, the main goal of the proposed flow is to stack many vertical layers, to achieve a footprint reduction. By going to 3D to build larger ALUs, we can observe considerable footprint gains. This is because stacking 4 vertical layers in 3D has the same footprint as a single layer, while the area of the 2D implementation increases for each additional bit. In particular, when going to 2 and 4 layers, we can already remark some footprint reduction when using the proposed *Nanofabric* flow when compared to ASAP7 (45%) and IN7 (20%), respectively. More importantly, using 32 vertical layers to build a 32-bit ALU reduces the footprint even further by a factor of 16.9× and 8.7× when compared to ASAP7 and IN7, respectively. We believe that stacking 32 vertical layers is a fair assumption, as current 3D NAND processes have demonstrated up to 128 stacked layers [24]. Hence, we can expect that a higher number of vertical layers could be considered once the technology is more mature in the long term. Note that while the results presented in this section are for the specific ALU depicted in Fig. 14(a), similar results are expected when considering different ALU designs.

7 Conclusion

In this book chapter, we introduced a novel 3D design flow called *3D Nanofabric*. The flow consists of several identical stacked logic layers, making it well suited

for SIMD processor applications where many basic regular ALUs are repeated. We first proposed a possible fabrication flow and described how multiple vertical layers could be patterned at once to define the transistor structures. We then thoroughly investigated the layout constraints of the *Nanofabric* flow and proposed several solutions to overcome them so that basic ALUs can be designed. We showed the 32-bit ALU footprint is reduced by a factor of 8.7× compared to a traditional 2D approach using a 7 nm FinFET technology, when using 32 vertical layers. We believe that this novel 3D approach enables cost-effective 3D scaling as it enables more performant circuits at a smaller footprint with reduced production cost.

References

1. Natarajan, S., et al.: A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size. In: 2014 IEEE International Electron Devices Meeting, San Francisco, CA, pp. 3.7.1–3.7.3 (2014). https://doi.org/10.1109/IEDM.2014.7046976
2. Colinge, J.P.: FinFETs and Other Multi-Gate Transistors, 1st edn. Springer, Boston (2007). https://doi.org/10.1007/978-0-387-71752-4
3. Yoon, S.W., Yang, D.W., Koo, J.H., Padmanathan, M., Carson, F.: 3D TSV processes and its assembly/packaging technology. In: 2009 IEEE International Conference on 3D System Integration, San Francisco, CA, pp. 1–5 (2009). https://doi.org/10.1109/3DIC.2009.5306535
4. Chua, T.T., et al.: 3D interconnection process development and integration with low stress TSV. In: 2010 Proceedings 60th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, pp. 798–802 (2010). https://doi.org/10.1109/ECTC.2010.5490728
5. Van Olmen, J., et al.: 3D stacked IC demonstration using a through silicon via first approach. In: 2008 IEEE International Electron Devices Meeting, San Francisco, CA, pp. 1–4 (2008). https://doi.org/10.1109/IEDM.2008.4796763
6. Beyne, E., et al.: Through-silicon via and die stacking technologies for microsystems-integration. In: 2008 IEEE International Electron Devices Meeting, San Francisco, CA, pp. 1–4 (2008). https://doi.org/10.1109/IEDM.2008.4796734
7. Chaabouni, H., et al.: Investigation on TSV impact on 65nm CMOS devices and circuits. In: 2010 International Electron Devices Meeting, San Francisco, CA, pp. 35.1.1–35.1.4 (2010). https://doi.org/10.1109/IEDM.2010.5703479
8. Ruythooren, W., Beltran, A., Labie, R.: Cu-Cu bonding alternative to solder based micro-bumping. In: 2007 9th Electronics Packaging Technology Conference, Singapore, pp. 315–318 (2007). https://doi.org/10.1109/EPTC.2007.4469706
9. Zheng, Z., et al.: Demonstration of ultra-thin buried oxide germanium-on-insulator MOSFETs by direct wafer bonding and polishing techniques. Appl. Phys. Lett. **109**(2), 023503 (2016). https://doi.org/10.1063/1.4955486
10. Batude, P., et al.: Advances, challenges and opportunities in 3D CMOS sequential integration. In: 2011 International Electron Devices Meeting, Washington, DC, pp. 7.3.1–7.3.4 (2011). https://doi.org/10.1109/IEDM.2011.6131506
11. Brunet, L., et al.: First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers. In: 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, pp. 1–2 (2016). https://doi.org/10.1109/VLSIT.2016.7573428

12. Mallik, A., et al.: The impact of sequential-3D integration on semiconductor scaling roadmap. In: 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, pp. 32.1.1–31.1.4 (2017). https://doi.org/10.1109/IEDM.2017.8268483
13. Brunet, L., et al.: Breakthroughs in 3D Sequential technology. In: 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, pp. 7.2.1–7.2.4 (2018). https://doi.org/10.1109/IEDM.2018.8614653
14. Vandooren, A., et al.: 3D sequential stacked planar devices on 300 mm wafers featuring replacement metal gate junction-less top devices processed at 525°C with improved reliability. In: 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, pp. 69–70 (2018). https://doi.org/10.1109/VLSIT.2018.8510705
15. Liu, C., Lim, S.K.: A design tradeoff study with monolithic 3D integration. In: Thirteenth International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, pp. 529–536 (2012). https://doi.org/10.1109/ISQED.2012.6187545
16. Andrieu, F., et al.: A review on opportunities brought by 3D-monolithic integration for CMOS device and digital circuit. In: 2018 International Conference on IC Design & Technology (ICICDT), Otranto, pp. 141–144 (2018). https://doi.org/10.1109/ICICDT.2018.8399776
17. Gitlin, D., Vinet, M., Clermidy, F.: Cost model for monolithic 3D integrated circuits. In: 2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Burlingame, CA, pp. 1–2 (2016). https://doi.org/10.1109/S3S.2016.7804408
18. Sabry Aly, M., et al.: Energy-efficient abundant-data computing: the N3XT 1,000 x. Computer **48**(12), 24–33 (2015). https://doi.org/10.1109/MC.2015.376
19. Shulaker, M., et al.: Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. Nature **547**(7661), 74–78 (2017). https://doi.org/10.1038/nature22994
20. Giacomin, E., Boemmels, J., Ryckaert, J., Catthoor, F., Gaillardon, P.: Layout considerations of logic designs using an N-layer 3D Nanofabric process flow. In: 28th IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), Salt Lake City, UT, USA, 5–7 October 2020
21. Macha, N.K., Iqbal, M.A., Rahman, M.: Fine-grained 3-D CMOS concept using stacked horizontal nanowire. In: 2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), Beijing, pp. 151–152 (2016). https://doi.org/10.1145/2950067.2950079
22. Li, M., Shi, J., Rahman, M., Khasanvis, S., Bhat, S., Moritz, C.A.: Skybridge-3D-CMOS: a vertically-composed fine-grained 3D CMOS integrated circuit technology. In: 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, pp. 403–408 (2016). https://doi.org/10.1109/ISVLSI.2016.56
23. Kang, D., et al.: 13.4 a 512Gb 3-bit/cell 3D 6th-generation V-NAND flash memory with 82MB/s write throughput and 1.2Gb/s interface. In: 2019 IEEE International Solid- State Circuits Conference - ISSCC, San Francisco, CA, USA, pp. 216–218 (2019). https://doi.org/10.1109/ISSCC.2019.8662493
24. Siau, C., et al.: 13.5 a 512Gb 3-bit/cell 3D flash memory on 128-wordline-layer with 132MB/s write performance featuring circuit-under-array technology. In: 2019 IEEE International Solid-State Circuits Conference - ISSCC, San Francisco, CA, USA, pp. 218–220 (2019). https://doi.org/10.1109/ISSCC.2019.8662445
25. Shibata, N., et al.: 13.1 a 1.33Tb 4-bit/cell 3D-flash memory on a 96-word-line-layer technology. In: 2019 IEEE International Solid-State Circuits Conference - ISSCC, San Francisco, CA, USA, pp. 210–212 (2019). https://doi.org/10.1109/ISSCC.2019.8662443

26. Jang, J., et al.: Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra high density NAND flash memory. In: 2009 Symposium on VLSI Technology, Honolulu, HI, pp. 192–193 (2009)
27. Clark, L.T., et al.: ASAP7: a 7-nm finFET predictive process design kit. Microelectron. J. **53**, 105–115 (2016). https://doi.org/10.1016/j.mejo.2016.04.006. ISSN: 0026-2692