Chapter 8
Monolithic 3D Integration—An Update

Zvi Or-Bach

8.1 Precise Bonder Enables Monolithic 3D Integration

The 3D IC space is considered to have two main branches—Through Silicon Via—"TSV" and Monolithic 3D. Some call the first branch '3D Parallel' and the second branch '3D Sequential.' The key differentiating aspect is the vertical connectivity density or pitch as is illustrated in Fig. 8.1, taken from a recent article [1] entitled "CoolCube™: More than a True 3D VLSI Alternative to Scaling."

Now that advanced precision bonders such as EVG-GEMINI® FB XT [2] and TEL-Synapse™ Si [3] are at the 50 nm (3σ) alignment precision range, a 3D Parallel integration flow could enable 50 nm like vertical pitch, which represents the

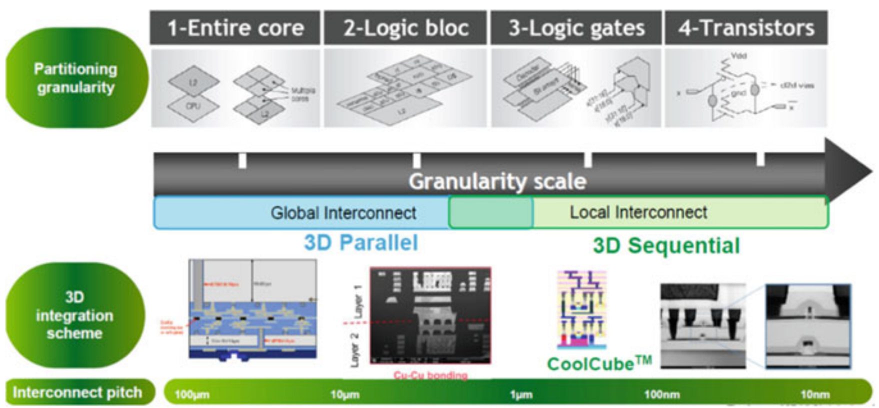


Fig. 8.1 Two 3D VLSE complementary approaches by CEA-Leti

Z. Or-Bach (✉)
MonolithIC 3D Inc., 3555 Woodford Drive, San Jose, CA 95124, USA
e-mail: Zvi@MonolithIC3D.com

© Springer Nature Switzerland AG 2020

B. Murmann and B. Hoefflinger (eds.), *NANO-CHIPS 2030*,
The Frontiers Collection, https://doi.org/10.1007/978-3-030-18338-7_8

Monolithic (or Sequential) 3D level of vertical connectivity. Such bonding precision could be assisted and enhanced by other technologies such as "Smart Alignment" [4, Chap. 3.3.2], Staggering [5, 6, Fig. 21A–C] and "Electronic Alignment" [7, Figs. 1–3].

8.2 Thinning the Transferred Layer—"Cut-Layer"

A second enabling technology for monolithic 3D using precision bonders is wafer thinning technology, especially for applications of more than two levels. To achieve high-density vertical connectivity, one needs to have a through silicon via with a diameter far less than 1 μm, compared to the >5 μm diameter of the common TSV technologies. For small diameter through silicon vias, the silicon layer needs to be very thin, as the aspect ratio for etching and filling such a via needs to be less than 1–10. In common TSV technologies, the transfer wafer is first thinned by backgrinding to a thickness of about 50 μm. It was found that thinning below 50 μm makes handling of the wafer unpractical—hence the >5 μm via diameter of common TSV technologies.

However, for the monolithic 3D application the thinning would take place after the transferred wafer has been bonded to the target wafer, thus achieving mechanical stability from the target wafer. In many applications, the desired thinning could be to 50 nm or even less. Such aggressive thinning would require a built-in control to avoid over thinning. Currently, without a built-in control, manufacturers avoid thinning below 10 μm. We can call such a built-in control a 'Cut-Layer.' One such built-in control is the BOX (Buried Oxide) of SOI wafer as was invented by IBM [8] and been used for many years by MIT Lincoln Lab [9] (Fig. 8.2).

SOI wafers are widely available these days at multiple technology nodes and wafer fabs, which could encourage a smooth adoption of precision wafer bonders for monolithic 3D applications.

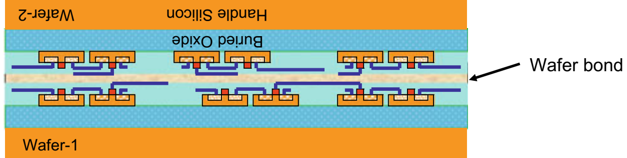
One disadvantage of SOI wafers is the relative high price of SOI substrates. A few innovative alternatives for the BOX as a 'cut-layer' are presented in the following.

8.2.1 SiGe

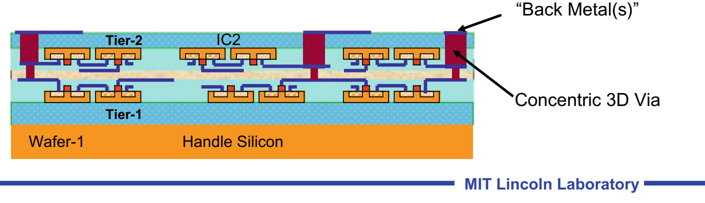
Silicon Germanium—"SiGe" is well-known material in silicon-based semiconductor devices. It has been used over the years for multiple applications including as an alternative channel material or as a way to form stress. It is a well characterized material which can be epitaxially grown. Additionally, there are well-known etch process both wet and dry to allow a selective etch of SiGe versus Silicon. The use of SiGe as an etch stop layer for 3D using layer transfer has been proposed many years

3-D Circuit Integration Flow-2

- **Invert, align, and bond Wafer-2 to Wafer-1**



- **Remove handle silicon from Wafer-2, etch 3D vias, deposit and CMP damascene tungsten interconnect metal**



Fermilab -31
CLK 2/28/2007
MIT Lincoln Laboratory

Fig. 8.2 Slide illustrating the use of SOI wafers, having the BOX as a ‘Cut-Layer’

ego [10]. Recently, SiGe has been used for next generation device Nanowire/Nanosheet for which SiGe could be selectively dry etched in a multilayer structure to allow a gate-all-around structure to be formed (Fig. 8.3).

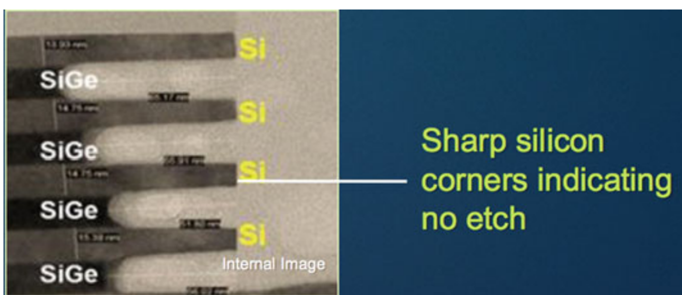
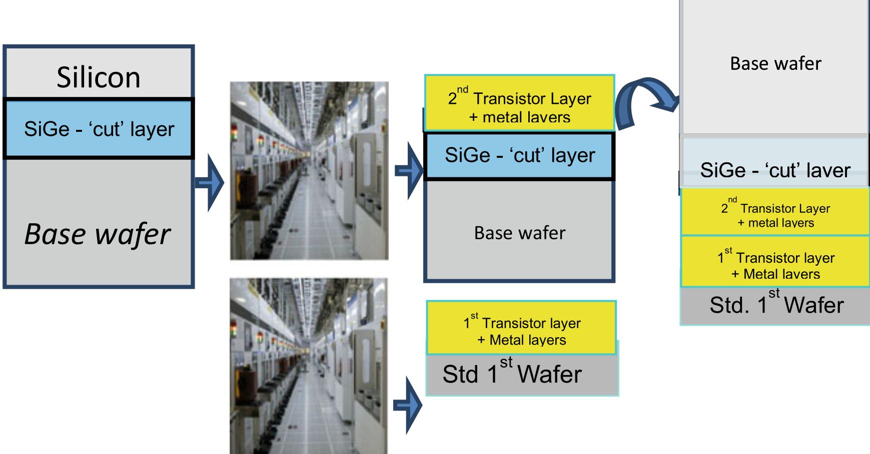


Fig. 8.3 SEM cross-section showing excellent etch of SiGe within alternating Si/SiGe layers, as will be needed for gate-all-around (GAA) horizontal nanowire (NW) transistor formation. *Source* Applied Materials



'Cuttable' Substrate ⇒ Std. Fab ⇒ 2nd Wafer ⇒ Flip Over 1st Wafer precisely bond (<50nm), grind &etch to "Cut Layer", connect/form connections

Fig. 8.4 Monolithic 3D integration flow using precise wafer bonder and wafers with built-in SiGe "cut-layer"

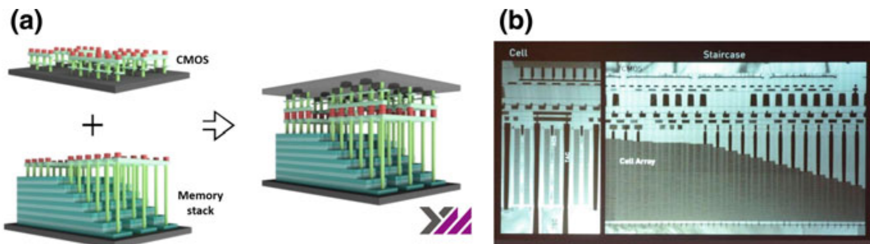


Fig. 8.5 **a** The Xtacking flow. **b** SEM of Xtacking device

"Under Xtacking addressing and I/O circuits are made on a separate wafer *(180 nm)* to the vertically stacked NAND cells and then bonded to them face-to-face through millions of **vertical** vias at the wafer-scale to complete the memory."

- "YMTC pushed its "pitches at several microns" down to about 100 nm for use in 3D NAND." [13]
- "YMTC has started delivering samples of its 64-layer 3D NAND chip with volume production likely to kick off in the third quarter of 2019, … Xtacking architecture is already adopted in the company's 64-layer 3D NAND engineering samples. Xtacking enables YMTC's 64-layer 3D NAND to be competitive with the available 96-layer 3D NAND solutions, … company expects its monthly production capacity

to hit 100,000 wafers after moving 64-layer 3D NAND technology to volume production." [14]

8.3 The Precision Bonder Based Monolithic 3D Advantages

The 3D Parallel using a precision bonder and 'Cuttable' wafer provides attractive advantages compared to Sequential 3D while keeping the equivalent vertical connectivity.

- **Standard Fab process for all levels**—The nature of the parallel flow is that each level is being processed by itself and accordingly its thermal budget is not impacting the other levels in the stack. This is extremely important advantage as the present IC fabrication complexity forces a vendor to resist any process change.
- **Heterogeneous Integration**—These days fabrication facilities are being designed and constructed to support a specific type/class of products such as a specific technology node, a specific type of circuit—logic, memory, analog, power, RF, …, a specific type of substrate—Bulk Silicon, SOI, …. In parallel 3D mix and match of different types of wafers in the stack provides an unparalleled advantage. Some specific applications will be covered in Chaps. 11 and 15.
- **Time to Market**—In parallel 3D, all levels could be fabricated in parallel and then stacked to form the 3D IC. With today's complex processing advanced node parallel processing could take 3 months. For a 3D IC with four levels the sequential processing could take more than a year which might introduce an unacceptable time to market challenge. While in 3D Parallel the fabrication of even a ten-level 3D IC stack could be done in less than five months.
- **Per Level Testing**—For parallel 3D, each level could be tested before being added to the stack to reduce the risk of losing the 3D IC because of a defective level. While random defects are still likely and should be managed by redundancy or other techniques, a total level failure could be managed.

In short: Precision wafer bonders with a 'Cuttable' wafer provide a very attractive technology for Monolithic 3D integration and could enable a broad industry adoption of monolithic 3D IC technology.

8.4 Update on Sequential Monolithic 3D

The research activity for Sequential Monolithic 3D is ongoing by the world leading semiconductor labs with good device demonstration as reported in IEDM 2018.

8.5 Update on 3D Heterogeneous Integration

The old International Technology Roadmap for Semiconductors (ITRS) has ceased, acknowledging the sunsetting of Moore's law and ITRS issued in 2016 its final roadmap. A new initiative for a more generalized semiconductor road-mapping was started through the IEEE's Rebooting Computing initiative, called the International Roadmap for Devices and Systems (IRDS). One part of this new IRDS roadmap under IEEE has been the Heterogeneous Integration Roadmap that recently released its 2019 Edition [18]. This new report references the opportunities with 3D integration associated with heterogeneous integration similar to those covered in Chap. 15.

Additionally, many foundries have embarked on an effort to add wafer stacking technologies, and specifically hybrid bonding, to their offering. GlobalFoundries recently issued a press release about their collaboration with ARM to demonstrate High-Density 3D Stack Test Chips for High Performance Compute Applications, stating "the companies validated a 3D Design-for-Test (DFT) methodology, using GF's hybrid wafer-to-wafer bonding that can enable up to 1 million 3D connections per mm^2, extending the ability to scale 12 nm designs long into the future" [19]. This followed TSMC announcing a similar type of collaboration [20]. The TSMC program, called SoIC for 3D Integration, is a part of TSMC's advanced packaging options as presented in Fig. 8.6.

Technology	2.5D	3D-IC	SoIC
Structure cross-section	SoC1 SoC2 µbump BEOL Interposer	SoC1 µbump SoC2	SoC1 SoIC Bond SoC2
Interconnect	µbump + BEOL	µbump	SoIC bond
Chip Distance	~100 µm	~30 µm	0
Bong-pad Pitch	36µm (1.0X)	36µm (1.0X)	9µm (0.25X)
Speed	0.01X	1.0X	11.9X
Bandwidth Density	0.01X	1.0X	191.0X
Power Efficiency (Energy/bit)	22.9X	1.0X	0.05X

Fig. 8.6 Comparison of multi-die integration technologies. 2.5D and 3D-IC use backend equipment, SoICs frontend (wafer fab) technology. In SoIC, there is virtually no distance between integrated chips. It achieves a very small bond-pad pitch of 9 µm for good scalability. *Source* TSMC