Chapter 3
Monolithic 3D Integration

Zvi Or-Bach

Abstract As the down-sizing of transistors has arrived at fundamental and practical limits, the technology direction with the largest potential for progress is the integration of transistors in the 3rd dimension on top of each other, maintaining and using the quality of monolithic, crystalline silicon in all successive transistor layers. After decades of exploratory research, monolithic 3D integration is now ready for cost-effective, large-scale implementation of nanoelectronic systems. It offers the largest gains in transistors-per-chip, it solves the on-chip interconnect and communication gridlock and thus the energy, speed and bandwidth problems. It opens a new era of effective industry networks for the sustained growth of the nanoelectronics economy. Monolithic 3D is already being adapted for mass production, in the non-volatile memory segment-3D NAND, and it can be expected that the other segments of the semiconductor industry will follow.

3.1 Why Monolithic 3D

The growth of Integrated Circuits has been driven primarily by the increase of device integration. This technological progress is based on Moore's Law, which is predicated on the notion that the optimum device integration would double the device count every two years. Moore's original prediction accounted for three mechanisms of improvement—decreasing the device (transistor) size, increasing the die size, and improvement of the circuit architecture. Yet, the primary mechanism used over the last 5 decades has been dimensional reduction. Every 2 years, a

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

B. Hoefflinger (ed.), *Chips 2020 Vol. 2*,
The Frontiers Collection, DOI 10.1007/978-3-319-22093-2_3

new technology node has been developed. Each new node is about 0.7× of the prior node for most critical device dimensions. This dimensional scaling is also known as Dennard scaling.

Dimensional scaling over the prior decades had the added benefits of reduction of cost per function, reduction in power, and increase in speed of device operation. Unfortunately dimensional scaling has reached the point of diminishing returns due to the escalating costs of implementation, as illustrated by Fig. 3.1.

These increasing challenges, which are directly related to dimensional scaling, are due mainly to:

1. Lithography
2. On-chip interconnect
3. Transistor variation.

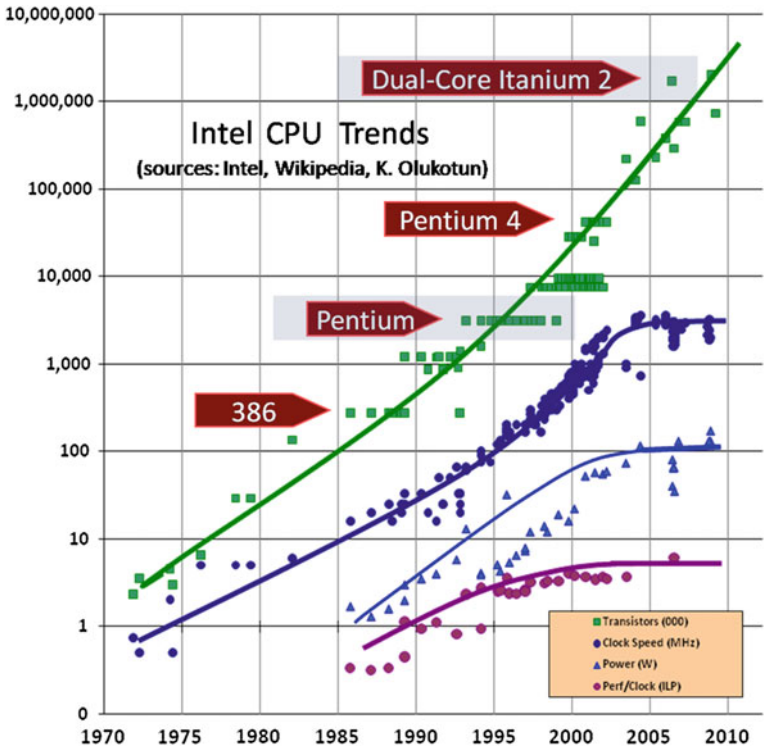


Fig. 3.1 Dimensional scaling is reaching the diminishing-return phase

3.1.1 Lithography

Dimensional scaling has been implemented by a 0.7× reduction of device critical dimensions. Accordingly, the lithography process needs to project smaller features every node, as illustrated in Fig. 3.2.

The industry kept moving to shorter wave lengths utilized in the lithography tool, but reached a technology limit at 193 nm using excimer lasers. The development of an Extreme Ultra Violet (EUV) lithography tool is an on-going major challenge and is not ready for production. Meanwhile, 193 nm immersion lithography is being used in a double and quad processing manner to mitigate the dimensional printing challenge, but at an escalating cost impact on the end device. This is illustrated in Fig. 3.3 as presented at the IEEE IITC 2014 workshop.

3.1.2 On-Chip Interconnect

Dimensional scaling improves transistor switching speed, but it increases the interconnect resistance, wire-to-wire capacitance, and overall interconnect RC. For over a decade, on-chip interconnects have been dominating device performance. First the industry changed interconnections from aluminum to copper, and then the inter-metal dielectric has been changed to low-K materials. Recently, the use of air-bridges was reported in reference to Intel's 14 nm logic process.

At IEDM 2013, Geoffrey Yeap, Qualcomm VP of Technology, stated in his invited talk: "As performance mismatch between transistors and interconnects continue to increase, designs have become interconnect-limited. Monolithic 3D (M3D) is an emerging integration technology poised to reduce the gap significantly between transistors and interconnect delays to extend the semiconductor roadmap way beyond the 2D scaling trajectory predicted by Moore's Law." Yeap provided the following chart—Fig. 3.4—to show the growing gap between transistor delay and interconnect delay.

3.1.3 Transistor Variation

Dimensional scaling has reached the point where some of the critical device dimensions are as small as only a few atomic layers. These and multiple other issues have caused a severe increase of across-the-die transistor variation, thus limiting the industry's ability to reduce the 6-transistor SRAM bit-cell size, as illustrated in the following table (source: imec): Fig. 3.5.

Clearly, below the 28 nm node, SRAM bit-cell scaling is slowing and falls far short of the 2X-per-node needed to maintain Moore's Law.

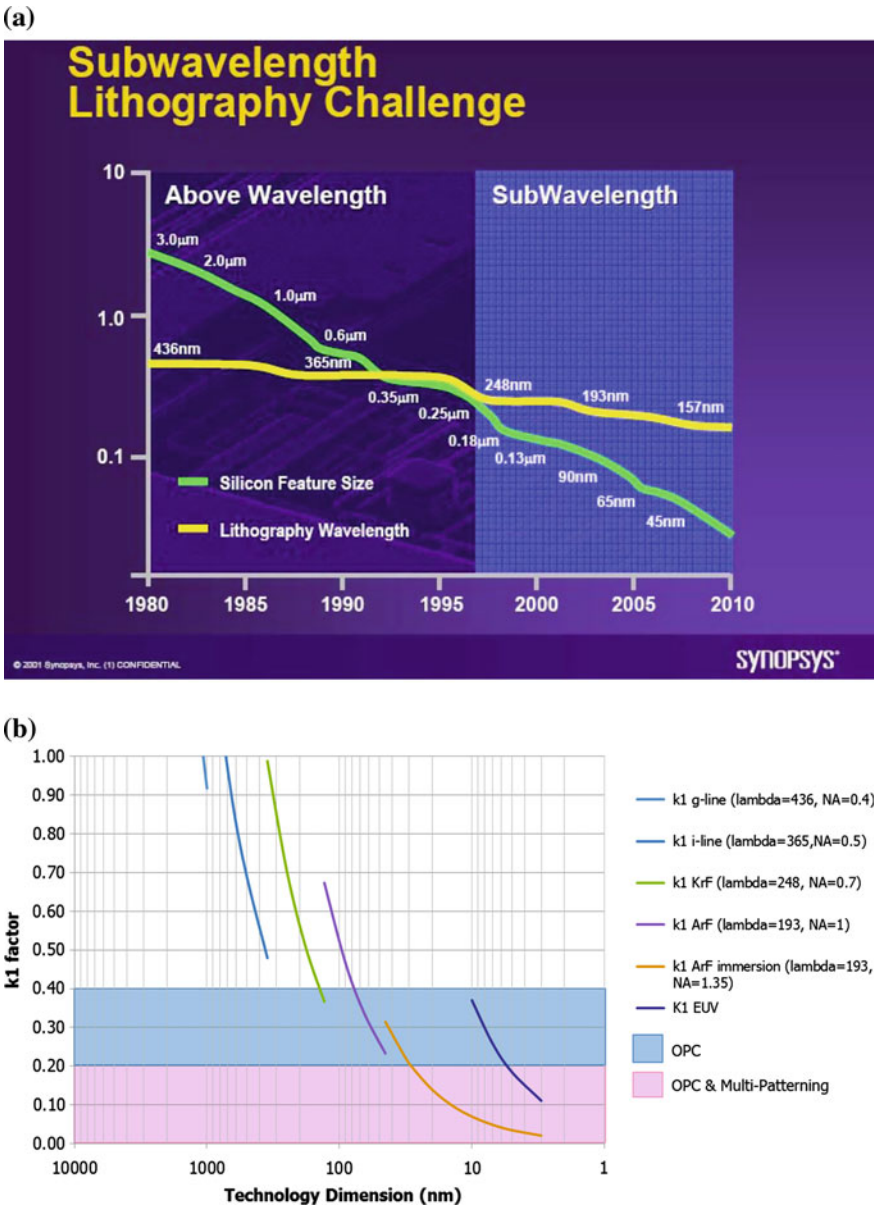


Fig. 3.2 **a** The lithography challenge. **b** Lithographic k1 correction factor by technology node

Monolithic 3D is well positioned to serve as an alternate path for industry's desire to increase device integration. It achieves increases in device integration by effectively having a smaller 2D die size, and, by "folding" the die, the on-chip interconnects are kept relatively short, thus enabling the increase of integration

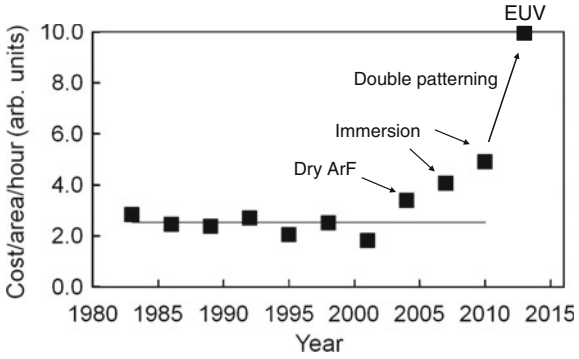


Fig. 3.3 Cost of lithography per wafer area per hour with dimension scaling

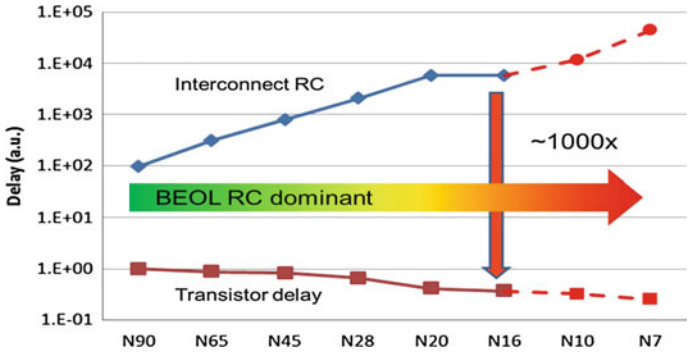


Fig. 3.4 On-chip interconnect delay scaling versus transistor delay

Early Production	2011-2012	2013-2014	2015-2016	2017-2018	2019-?
	22-20 nm	16-14 nm	10 nm	7 nm	5 nm
Memory (um2)	SRAM 0.09-0.08	SRAM 0.08-0.07	SRAM 0.06-0.05	SRAM < 0.05	SRAM < 0.05 (STT-MRAM)
Device	Planar, FinFET	FinFET, FDSOI	FinFET	FinFET (LOCSOI, GAA)	GAA FinFET (NW)
Gate EOT (nm)	HKMG 0.9	HKMG 0.8	HKMG 0.7	HKMG 0.7	HKMG 0.7

Fig. 3.5 SRAM bit-cell scaling

without the escalating costs and interconnect deficiencies associated with dimensional scaling.

3.2 Historical Review of Monolithic 3D Technologies

For many years, monolithic 3D was considered impractical due to the 400 °C temperature limit imposed by the aluminum or copper interconnect. This limitation led to the focus on Through-Silicon-Via (TSV) technology as the only viable path for 3D ICs. It now seems clear that the TSV process flow is intrinsically expensive and, accordingly, is being perpetually pushed to the future. TSVs allow stacking of fully processed devices using wafers that are thinned to about 50 μm.

In monolithic 3D, the upper transistor layers are orders of magnitude thinner, less than about 100 nm. Accordingly, the vertical connectivity is comparable to the horizontal connectivity and is many orders of magnitude better than for TSV technology. Some of the very early work was done in 1989 [1, 2] using selective epitaxial seeding from the bulk to build transistors over transistors without vertical interconnection in-between. In recent years, pioneering efforts were published providing practical paths for monolithic 3D logic devices [1, 3–8]. In the following, we present a brief overview of historical and current works on monolithic 3D technologies.

3.2.1 Thin-Film Polysilicon-Based Monolithic 3D

The simplest approach for monolithic 3D technologies is to use Thin-Film-Transistors—TFT. Most common TFTs use polysilicon devices that could be directly deposited over an existing semiconductor wafer without exceeding the 400 °C temperature limit. TFT performance is inferior to mono-crystalline transistors but could be useful for some memory applications. An early attempt to build a 3D-FPGA using TFTs for the FPGA program memory was made by a start-up named Tier Logic in collaboration with Toshiba [9]. 3D-NAND made with poly TFTs, currently beginning volume production, is considered by most as the future path for non-volatile memories and will be detailed later in the chapter.

3.2.2 Crystalline Overlay

Forming a crystalline overlay can be done by crystallizing a prior deposited poly or amorphous layer. Most common crystallization techniques utilize laser-based melting and various re-crystallization techniques. Some use seeding from the underlying single crystal base to direct the crystal formation [10], others suggest the

use of a μ-Czochralski process which is based on pulsed-laser crystallization of a-Si [11]. And some let the layer crystallize as it may [6, 7]. While these techniques may form small regions of crystalized silicon, it seems that layer transfer techniques would be preferable for most 3D logic device applications due to the perfect uniformity of the transferred mono-crystal.

3.2.3 *Layer Transfer*

Layer transfer techniques became a commonly practiced semiconductor process mostly through the construction of Silicon-on-Insulator (SOI) wafers. The most common layer transfer technique is the ion-cut, also known as Smart-Cut®, invented by CEA Leti and used by Soitec over the last two decades as illustrated in Fig. 3.6.

An alternative layer transfer technique was invented at Canon named ELTRAN —Epitaxial Layer TRANsfer [12]. The ELTRAN layer transfer technique is done at below 400 °C. Variations on the ion-cut techniques were developed by Soitec and SiGen to allow the layer transfer at temperatures below 400 °C by the use of co-implant or mechanical force [13]. Another variation was invented by IBM and is used in the MIT Lincoln Lab [14] where an SOI wafer is processed with transistors, then flipped and bonded to a base wafer with transistors, and then the bulk of the SOI wafer is etched away using the oxide layer as an etch stop. This technique provides a solution that falls in-between TSV and monolithic 3D in terms of the density of vertical (inter-layer) interconnect.

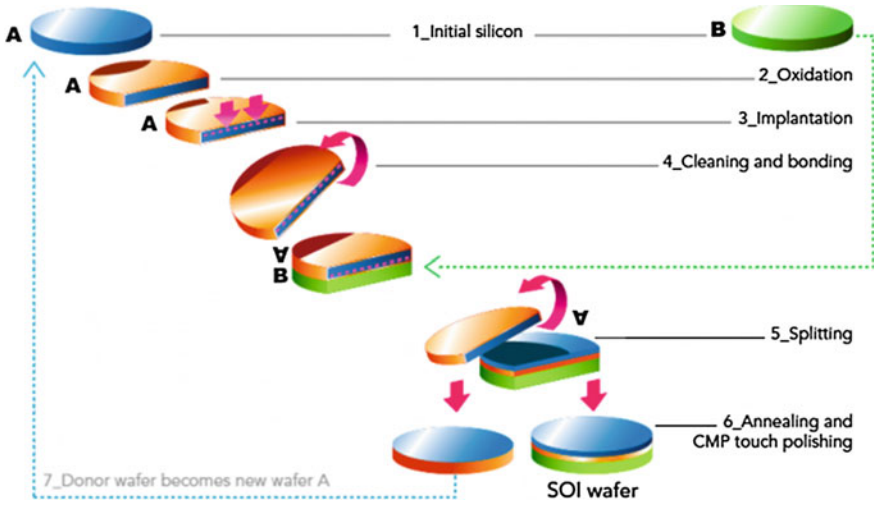


Fig. 3.6 Layer transfer using the smart-cut process

3.2.4 *Transistor Activation*

A key step in forming transistors in a mono-crystallized layer is doping activation. Doping activation requires more than 600 °C, typically 800 °C, which needs to be managed in order not to damage the underlying interconnection layers such as copper.

CEA Leti has been one of the more active organizations working on monolithic 3D technology, which they sometimes call Sequential 3D, for logic devices [15]. In its early work CEA Leti developed technologies, which did not use interconnection between the base layer and the upper transistor layer. Recently, CEA Leti has done work with the collaboration and support of IBM, ST Micro and Qualcomm, as illustrated in Fig. 3.7. In some of their recent work, refractory metals were used, such as tungsten, which can withstand high processing temperatures. Some of the upper-layer transistor processes were adapted to be performed at below 600 °C, and the lower transistor structures were modified to survive up to 600 °C. Lately, excimer laser annealing was integrated into the process flow to allow more flexibility for monolithic 3D integrations [16].

Another company, which has been active in the monolithic 3D space, is MonolithIC 3D Inc. The company published several process flows enabling upper-layer transistor activation without damaging the underlying interconnection layers.

3.2.4.1 The RCAT Process

Figure 3.8 describes the "RCAT process" [18], which constructs the RCAT transistors which have been commonly used in DRAM manufacturing since the 90 nm

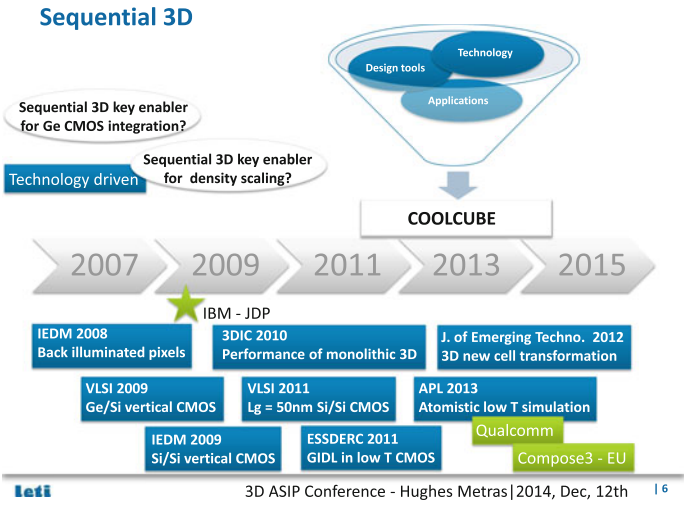


Fig. 3.7 CEA Leti collaborative road map for monolithic 3D technology [17]

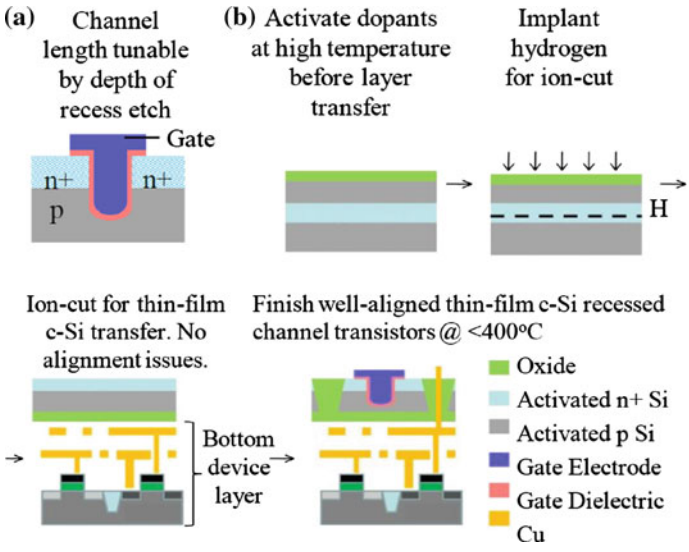


Fig. 3.8 **a** A recessed channel transistor. **b** Process flow for Monolithic 3D logic. Bottom device layer with Cu/low k does not see more than 400 °C

node. The RCAT transistor is competitive with standard planar transistors [19] and looks like the inverse of a FinFET. High-temperature dopant-activation steps are done before transferring bilayer n+/p silicon layers atop Cu/low-k using ion-cut. The transferred layers are un-patterned; therefore, no misalignment issues occur while bonding. Following bonding, sub-400 °C etch and deposition steps are used to define the recessed-channel transistor. This is enabled by the unique structure of the device. These transistor-definition steps can use the alignment marks of the bottom Cu/low-k stack since transferred silicon films are thin (usually sub-100 nm) and transparent. Sub-50 nm diameter through-layer connections can be produced due to the excellent alignment.

The key idea for the RCAT process is the activation of the semiconductor-layer doping prior to the layer-transfer step. This completely avoids thermally damaging the underlying-layer interconnect or transistors. Forming the RCAT transistor after the layer transfer uses etch and deposition processes, which do not require high temperatures. This type of flow could be used for other types of transistors such as the junction-less transistor (gated resistor) or for vertical transistors as demonstrated by Besang Inc.

3.2.4.2 The Gate Replacement Process

Recently, the industry has moved to Hi-K metal gates and later fully adopted the "gate last" (gate replacement) approach to avoid exposing the hafnium oxide to high temperatures. This could be used for forming monolithic 3D as illustrated in

Fig. 3.9. First, the dummy gate stack transistors are processed with no temperature restrictions on a donor wafer. Then, using ion-cut and a carrier wafer, a small slice of the donor wafer is transferred to the top of a base wafer. Gate replacement is then performed by removing the H+ damaged gate oxide and replacing it with a HKMG stack using low-temperature etch and deposition processes. This flow has one serious limitation—alignment. As the layer transfer process is now being done on a patterned layer, the transfer misalignment of ~ 1 µm would impact the second layer. This misalignment could be reduced in the case of a repeating pattern to the size of the repetition (100 s of nm).

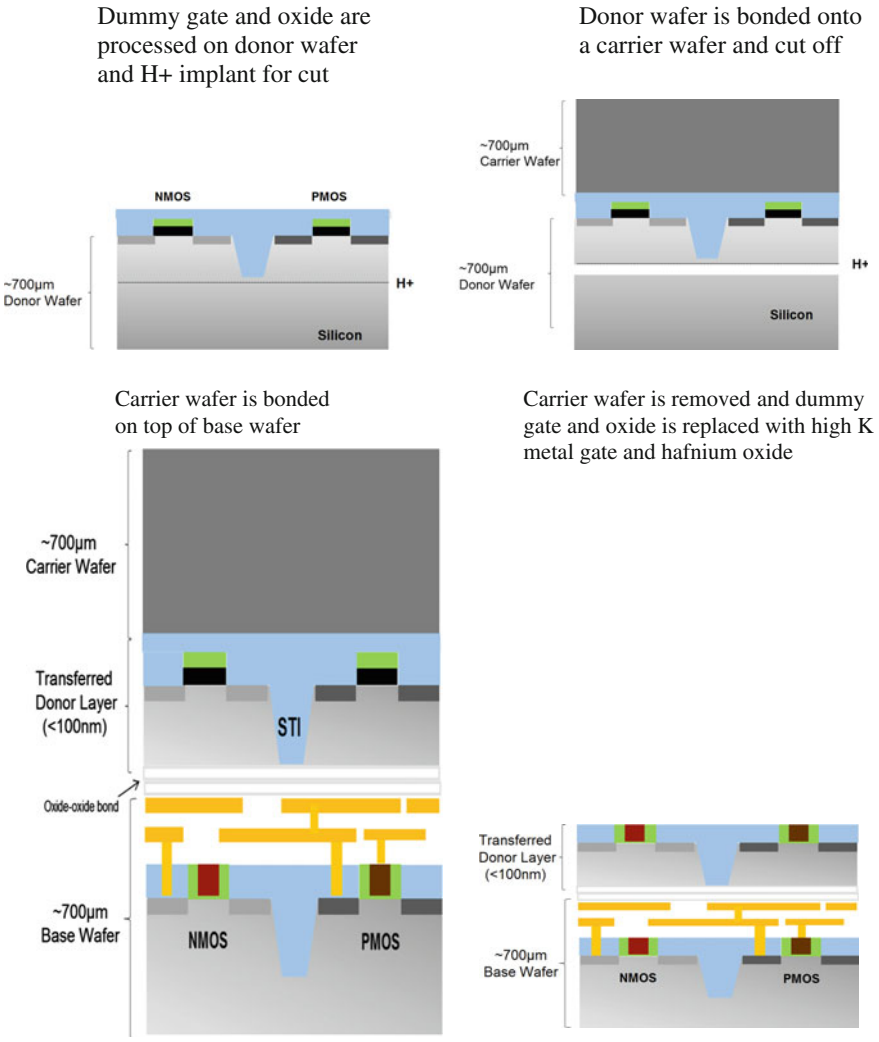


Fig. 3.9 Process flow for a gate replacement process

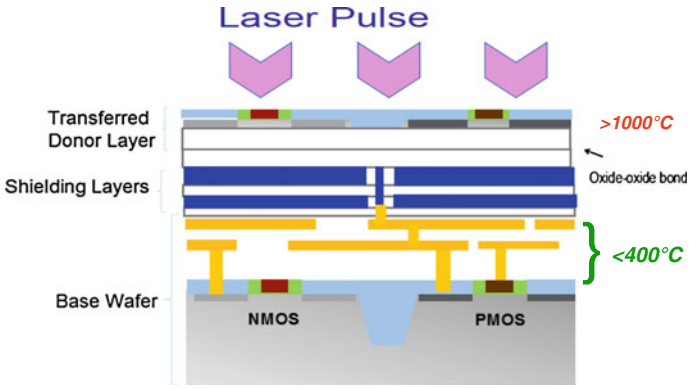


Fig. 3.10 Schematic of an example 3D-IC stack with shielding layers between stacked active layers, with laser annealing to activate the upper active layers

3.2.4.3 Laser-Annealing Process

The process utilizes laser annealing [5] for the 2nd-layer transistor activation while the base layers are protected by an inter-layer shield as illustrated in Fig. 3.10. This process is now becoming viable thanks to the fact that fully-depleted transistors can be integrated atop thin c-Si layers with relatively straightforward modifications of the gate-last CMOS process flow. The transferred donor layer is processed to form transistors with short pulsed-laser exposures providing, for example, annealing of process-induced damage and activation of dopants. It should be noted that the shield/heat sink layers are useful as Vss/Vdd planes and may also serve as a heat spreader and EMI shield.

The design of the shielding layer would be highly dependent on the type of laser used for the annealing process taking into account the thickness of the silicon and oxide layers. It has been shown [2] that, for excimer laser annealing with $\sim$100 ns pulses, there is actually no requirement for a shielding layer as the heat absorbed by the silicon layer dissipates to less than 400 °C before it reaches the underlying interconnect layers.

The laser annealing technique could be used as complementary processing to any of the other monolithic 3D process flows. In general, annealing techniques are useful to repair structural damage associated with other processes, and, accordingly, they could be used afterward.

3.3 Precision Bonders—A Game Changer for Monolithic 3D

The monolithic 3D flows presented in the previous section are all practical and would allow monolithic 3D ICs with good performance and competitive cost. Yet they do imply a transistor processing flow that is different from the one already

developed and matured in 2D. In fact, all monolithic 3D IC process flows presented in the past required new front-end-of-line process development. Such new development efforts represent an additional barrier for adopting monolithic 3D technology. As it often happens in the semiconductor industry, the improvement in existing processing equipment or the development of a new type of equipment opens the door to new devices or improvements in our ability to process new devices. Past wafer bonders have been limited to about 1 μm wafer-to-wafer misalignment. At the 2014 IEEE S3S conference, two companies presented wafer bonders with about 200 nm wafer-to-wafer misalignment [20, 21]. These types of wafer bonders enable a game change in monolithic 3D manufacturing [22]. For the first time, monolithic 3D technology could be integrated within almost any semiconductor manufacturing facility using their existing transistor processes. In the following section, we provide the description of one such process flow.

3.3.1 Monolithic 3D IC Using Precision Bonders

The flow in Fig. 3.11 is based upon what we call 'gate-replacement' [23] processes, and it leverages the precision-bonder alignment accuracy. In step 1, a 'donor' wafer will be used to process a transistor layer labeled Stratum 3. The existing front-end process could be used. Alternatively for a gate-last flow, the donor wafer would be held before the gate-replacement phase. Then H+ would be implanted at the desired depth ($\sim$100 nm) in preparation for the layer-transfer step. In step 2, the donor wafer is bonded (oxide to oxide) to a 'carrier wafer' and ion-cut off. This bonding step does not require precise alignment. In step 3, the carrier wafer can now be annealed to repair the potential H+ implant damage. In step 4, the donor wafer is now processed to form Stratum 2. The existing front-line process can be used including FinFET or any other available front-line process. The choice of the transistor and the architecture for strata 2 and 3 should consider the need for vertical isolation in-between them. Note that between the transferred layer and the carrier wafer there is an oxide layer, which would be an excellent etch-stop allowing the transfer onto the target layer without the need for ion-cut. A preferred strategy is to use Stratum 2 for the high-performance circuits while Stratum 3 would be used for support of less sensitive circuits. All high-temperature steps should be completed at this point, since in the following step, interconnects are added. In step 5, add contacts and at least one metal layer. In step 6, bond (oxide to oxide or metal to metal) to the target wafer using the precise bonder alignment with less than 200 nm misalignment. Now grind and etch off the carrier wafer. (Not presented here are options to remove the carrier wafer for reuse.) In step 7, the dummy gate and the gate oxide of Stratum 3 can now be replaced, and connections can be made between Stratum 2, Stratum 3 and the underlying target wafer. Alignment and via processing are done just as between conventional BEOL metal layers, as the transferred layer is very thin ($\sim$100 nm).

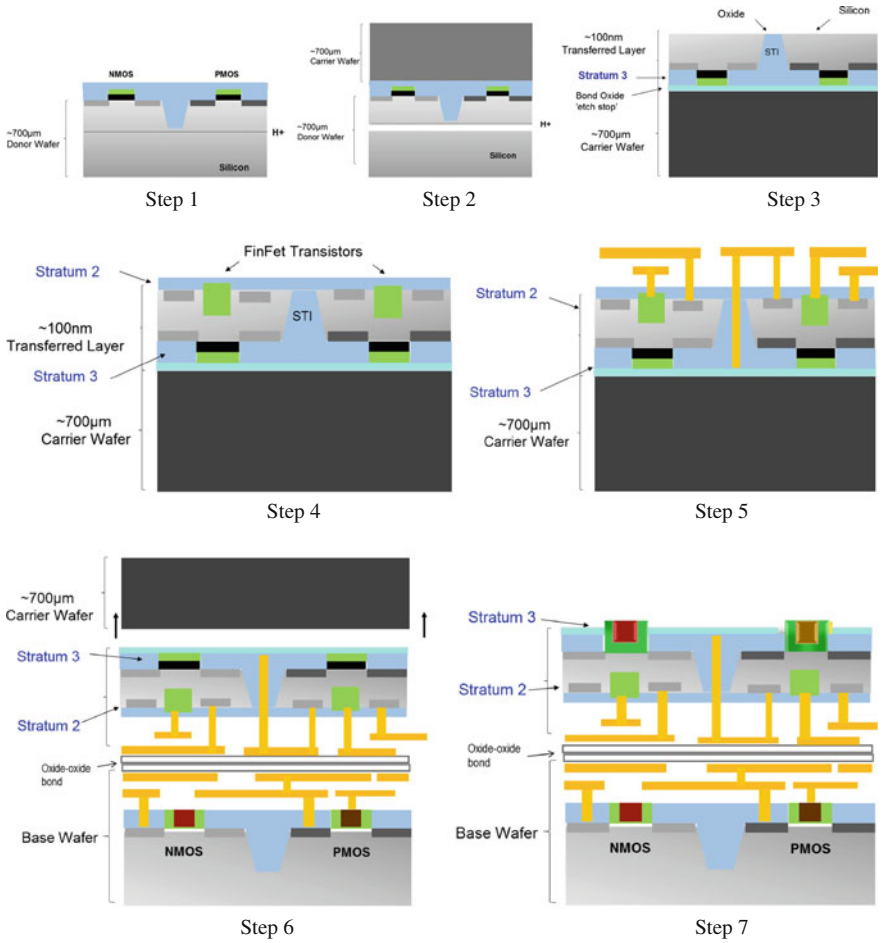


Fig. 3.11 Process flow for gate-replacement process and precise bonding

3.3.2 Smart Alignment

Having a thin transferred layer allows the through-layer via to be as small as a conventional BEOL interconnection via (∼50 nm). Yet the 200 nm bonding alignment window would appear to require a landing pad of 200 nm by 200 nm for each vertical connection. In addition to wafer-to-wafer misalignment, we also need to account for reticle-to-reticle misalignment. This would be highly dependent upon whether or not both wafers come from the same process line using the same lithography tool/stepper. The total misalignment across the wafer between the Stratum 1 and Strata 2/3 would include the reticle-to-reticle misalignment, the wafer bonding misalignment and some other error factors. Assuming that the total across wafer misalignment is less than 300 nm, then it would seem that a bonding pad of

300 nm × 300 nm at the top of Stratum 1 would be needed to allow a safe Strata 2/3 to Stratum 1 via connection. With Smart Alignment the connection is made by two perpendicular 300 nm long strips as seen in Fig. 3.12. The vertical strip is part of the top layer of the target (bottom) wafer.

After bonding, the through-layer via would be aligned to the target wafer in the Y direction and to the transferred layer in the X direction as seen in Fig. 3.13. The top connection strip could then be processed, aligned to the transferred (top) layer. This alignment scheme reduces the vertical connection overhead to a minimum, and it allows for multiple vertical connections per unit area of 300 nm × 300 nm.

3.3.3 Strata 2, 3—Examples

Figure 3.14 illustrates one example for circuit allocation for Stratum 2 and Stratum 3 with an intrinsic vertical isolation. For Stratum 2, the most advanced devices for forming high-speed logic could be used such as FinFET transistors. The SRAM for the high-speed logic circuit could be placed onto the close-by Stratum 3. A compelling option for the SRAM would be the use of Zeno technology [24] where a two-stable-state, single-transistor SRAM is enabled by a deep-implant back-bias. The vertical isolation is achieved by the back-bias. The FinFET transistor by design is also isolated from the substrate. This use of Strata 2 and 3 is compelling as the memory creates no obstructions and offers a very short fetch-path for memory access. Such a dual functional layer (Strata 2 + Strata 3) could be a product offered by itself as an add-on to many designs and single-chip systems.

Such process-flows with a dual functional layer could enable many new innovative devices such as:

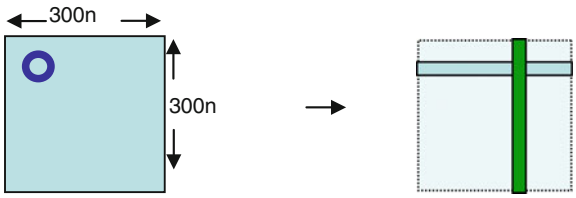


Fig. 3.12 Smart alignment

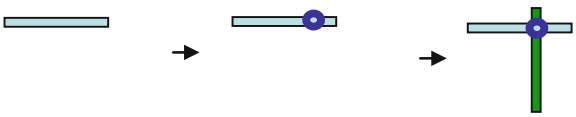


Fig. 3.13 Smart alignment

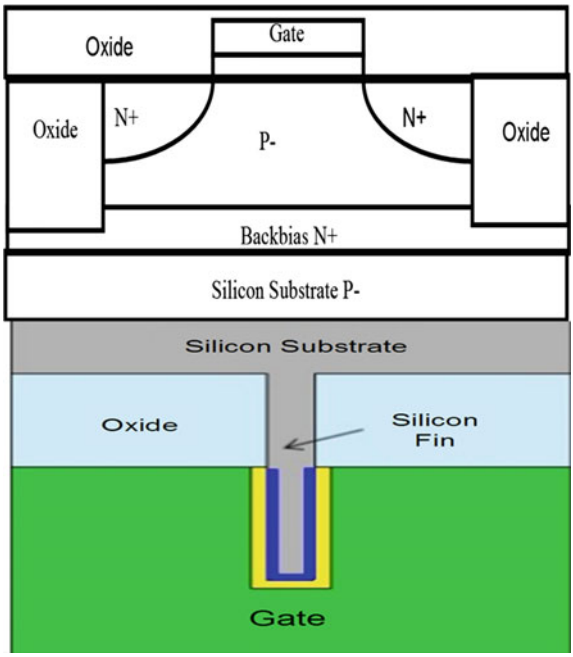


Fig. 3.14 1T SRAM over FinFET

- An image sensor on Stratum 3 with pixel electronics on Stratum 2 could provide an unparalleled dynamic range for cameras.
- A full redundancy layer [25] on Stratum 3 provides redundancy to Stratum 2, allowing almost unlimited logic integration on huge dies, essentially a server farm on a chip.
- A configurable logic fabric as an add-on…

3.3.4 Monolithic 3D Cost Estimates

It is well known that high cost is the number-one issue which slows down the adoption of 3D ICs based on TSV. The proposed monolithic 3D flow has the potential to overcome this barrier as it avoids the use of a thick layer with lengthy etch and deposition processes. In fact, it can provide circuit fabrics for two strata for a cost that is less than one wafer substrate. The donor wafer is reusable, and the cost of the first ion-cut is estimated to be less than $60 [26]. The carrier wafer could be reusable or utilize an inexpensive test wafer costing about $30. The estimated per-wafer cost of precision bonding is less than $20. Other steps involved in layer transfer, cleaning, etch, etc., are estimated at about $30 total. The costs for transistor formation for Strata 2 and 3 and their associated interconnects are no different from any other circuit fabrication costs. Accordingly, we estimate that the cost structure

is comparable with the fabrication cost of 2D devices. Yet having the overall design built in a 3-strata fabric provides huge power, performance, and cost benefits.

3.4 EDA for Monolithic 3D

Design tools and algorithms for 3D ICs have been a subject of research work for many years. Some of that work had been summarized in related books [27], and relevant conferences [28]. Academic work, which resulted in many papers and tools, is being done at GeorgiaTech at the GTCAD Laboratory. The challenge is with respect to a commercial tool that would be ready for use by the semiconductor engineering and design community. Here we would have the classical chicken-egg challenge. The commercial EDA industry would wait for the design market to be large enough to justify the attention and the required investment, which is hard to have without the design tools to support such commercial design efforts.

This has been recognized by CEA Leti, which has done research work [29–31] allowing the use of commercial 2D EDA tools for a specific class of 3D designs. In one such case, one stratum has been allocated just for the drive and repeater while the other stratum carries the logic cell with minimum drive as illustrated in Fig. 3.15. The 2D tool would be used to place the 'modified logic cells' for the logic stratum, and the proper drive for each of the logic cells would be placed accordingly in the drive stratum [29].

A more flexible approach, named CELONCEL (Cell-on-cell) stacking [30, 31], allows cells to be placed on top of each other considering the pin-access issues. A physical design tool (CELONCEL$_{PD}$) was proposed that transforms the monolithic 3D placement problem into a virtual 2D problem solved using existing 2D placers. A highly parallelizable zero-one linear program formulation is used for layer assignment followed by a linear-program-based minimum perturbation for a high-quality 3D layout. Figure 3.16 illustrates the general EDA flow. It includes first deflating the cell library by 50 %, then using a commercial EDA placer and then re-inflating the cells after splitting them to two layers.

In the last section of this chapter, multiple monolithic 3D advantages are presented. Many of these leverage the aspect that each stratum would be processed independently—this is often referred to as heterogeneous integration. Accordingly, in many of those, the monolithic 3D design would be a 2D design using 2D EDA with some limited constraints related to the other strata. Such could be:

1. Memory over/under logic. A Memory process is preferably different from a logic process. In a design, which uses some very large memory blocks, those blocks could be either manually placed or assisted by a floor-planner tool. Then a 2D tool could be used to place and route the logic fabric with the memory pins serving as virtual I//O constraints.
2. Image sensor with pixel electronics. These could become mostly a full-custom design where each pixel has the same deep pixel electronics.

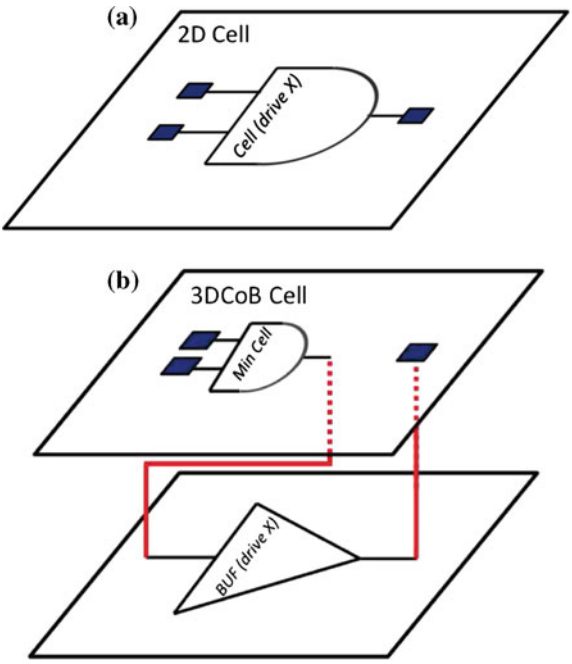


Fig. 3.15 Cell-on-Buffer (CoB) concept: **a** conventional 2D cell, **b** its equivalent 3D CoB cell

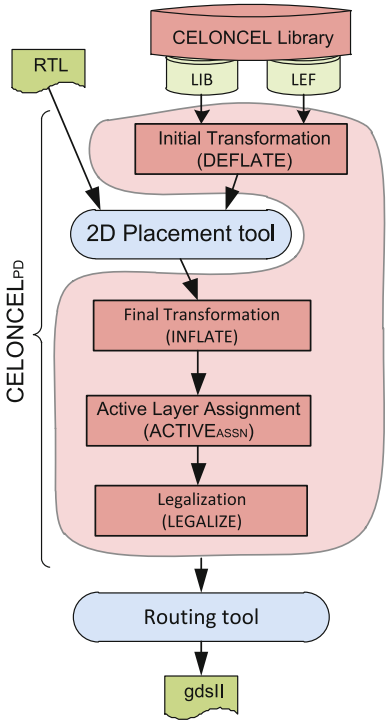


Fig. 3.16 Logic-to-layout using 2D commercial EDA for cell-on-cell

3. 3D FPGA. A full-custom-design technique could be used for the regular terrain of the programmable logic array.

3.5 Managing the Heat

Several questions concern the heat-removal aspect for 3D IC's. The first question relates to having more transistors in a smaller space. While more complex circuits present an ever-increasing power challenge, having them built in monolithic 3D is an important part of the solution as it is well documented that 80 % of the power consumption is due to on-chip connectivity [32]. The more interesting question relates to the fact that Stratum 2 transistors are thermally isolated (surrounded by oxide) and without direct access to the silicon bulk for heat removal as is illustrated in Fig. 3.17.

Figure 3.18 illustrates the solution of using the power-delivery network (PDN) for heat removal. This work was reported in IEDM 2012 [33]. Having Stratum 2 only about 1 micron away from the bulk allows a very effective heat removal path through the power-delivery network (Fig. 3.19).

The PDN would enable removing the overall heat, but there could still be specific hot spots and active areas that might not have a thermal connection path to the PDN. The first step would be to add a heat-spreader layer to even out hot spots. In the previously presented Fig. 3.9, a heat spreader is illustrated underneath the transistor layer of the upper strata. Such a heat spreader could be used to shield the interconnect layers from top heat, act as power delivery and provide EMI/RFI isolation. The heat-spreader layer could be constructed from copper with thin isolation from the transistor layer above, or from less common, but even better heat-conducting material such as graphene or CVD diamond. In some cases, the power-delivery path could be too far from an active heat generating source, and some active cells might not have a power connection at all.

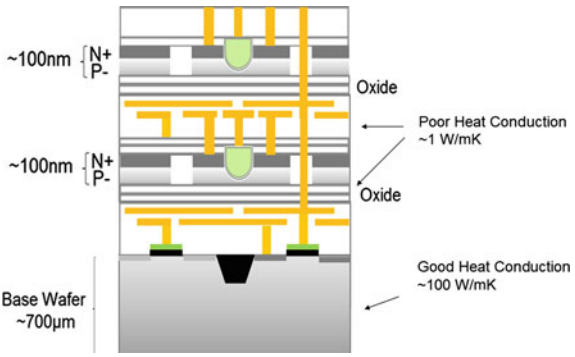


Fig. 3.17 The inner-strata transistor has no natural path for heat removal

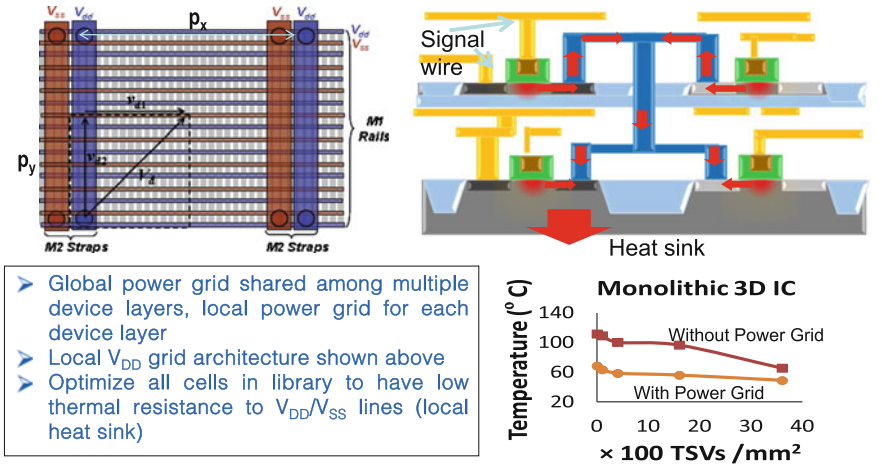


Fig. 3.18 Heat removal by the power-delivery network (PDN) [33]. © IEEE 2012

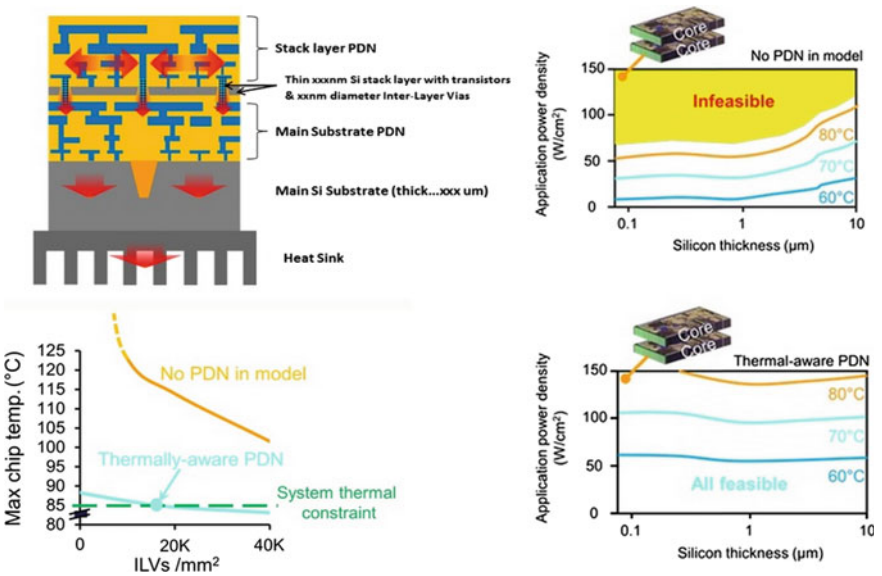


Fig. 3.19 Heat removal by the power-delivery network (PDN) [33]. © IEEE 2012

Figure 3.20 illustrates an important part of the heat removal for monolithic 3D—thermally conducting, electrically isolated contacts [34]. A simple solution could be a reverse-biased contact (illustrated in Fig. 3.20) for a common cell such as a transmission gate, which has no electrical connections to the PDN.

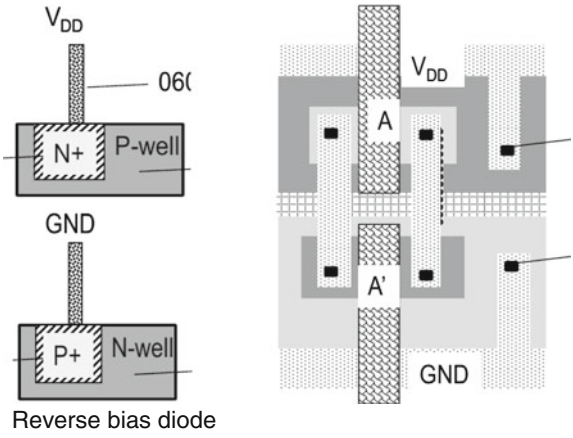


Fig. 3.20 Thermally conducting, electrically isolated contacts

3.6 3D Memories: 3D NAND,…

3.6.1 Introduction to BiCS

In landmark papers [35, 36], Toshiba introduced in 2007 a new approach to memory processing, which they call—BiCS—Bit-Cost Scalable Flash Memory. The key idea of BiCS is to share lithography and processing for multiple layers of processing. By properly architecting a memory-design and -processing flow, many layers could be processed together allowing scaling of the device into the vertical direction while keeping the incremental cost very low as illustrated in Fig. 3.21 [37].

It is now clear that 3D NAND, also called V NAND, is the scaling path for the Non-Volatile (NV) Memory industry. In mid-2013, Samsung announced the mass production of 24-layer 3D NAND, and, in 2014, a second generation with 32 layers has been released to the market. Figure 3.22 is a cartoon by Samsung illustrating the change to monolithic 3D already taking place for memory products.

3.6.2 3D-NAND

Since the first BiCS disclosure, multiple 3D Memories with shared lithography have been made public. It seems that each and every memory vendor has its own architecture claiming it is the best. Most of these architectures use polysilicon transistors, which are good enough for most memory products and easier to process for a multi-layer 3D architecture. The following description presents one [38] of these architectures. The process-flow in Fig. 3.23 shows how shared lithography and processing could be used. Multiple oxide/nitride steps are used with multi-states charge trap dielectric and poly-silicon channel. The polysilicon-channel

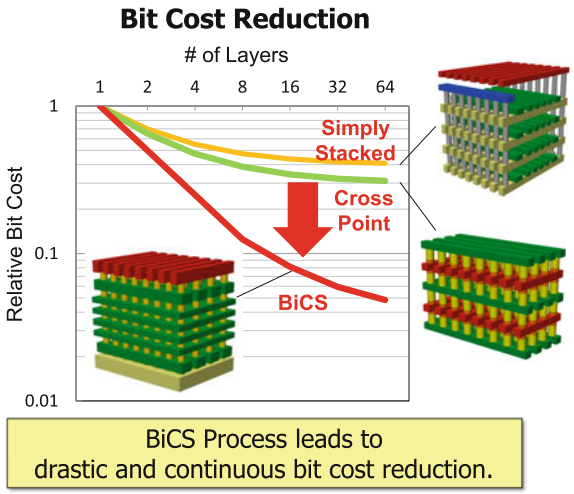


Fig. 3.21 Bit-cost reduction [37]. © IEEE 2014

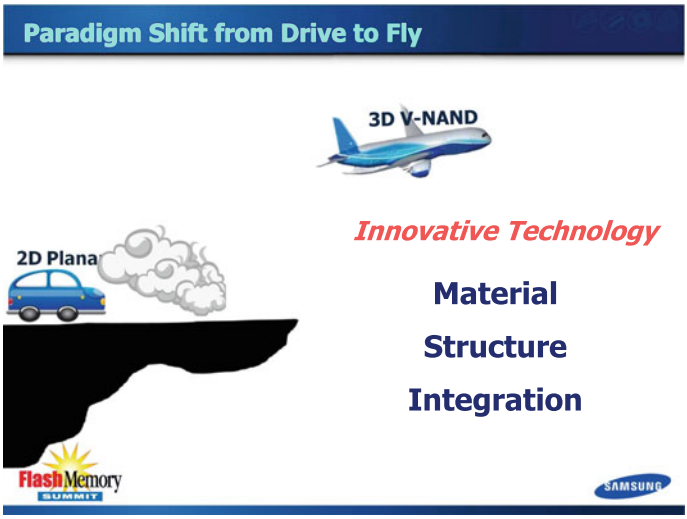


Fig. 3.22 Paradigm shift from 2D scaling to 3D scaling (*Source* SAMSUNG)

performance is acceptable since the grain size is large compared to cell dimensions. The use of an excimer laser to further increase the grain size has been shown to improve the performance. The device uses gate-all-around transistors, which provides excellent electrostatic channel control (Fig. 3.24).

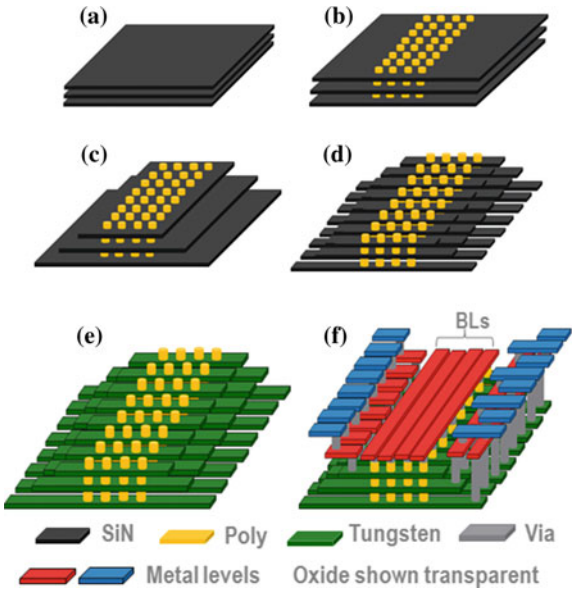


Fig. 3.23 Process-flow for 3D NAND Flash [39]. © IEEE 2009. **a** Deposit multiple Sio_2/SiN layers, **b** Etch hole and deposit channel poly (shared litho step), **c** Make staircase pattern for contacts (shared litho step), **d** WL cut etch (shared litho step), **e** Using flow in Fig. 3.24, make charge trap flash dielectrics and gate/WL, **f** BEOL

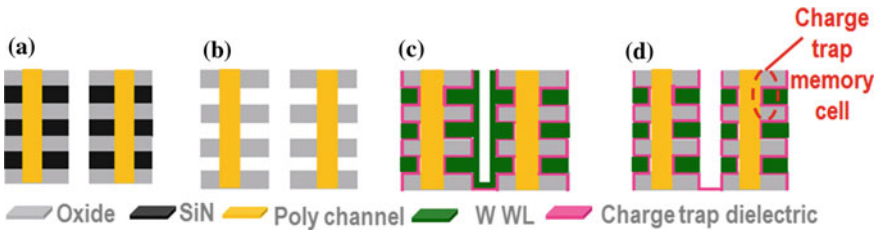


Fig. 3.24 Charge-trap dielectric and electrode definition [39]. © IEEE 2009. **a** After step (d) of Fig. 3.23, **b** Wet etch SiN, **c** Deposit charge trap dielectric and WL, **d** Gate node separation

3.6.3 Making Contact Without Adding Lithography Steps

An important complementary technology allows for making contacts to each of the stacked layers without the need for an individual lithography step. Figure 3.25 illustrates the contact-formation flow using an approach similar to 'spacer technology'. It utilizes isotropic etch/slim followed by anisotropic etch (RIE) to form a staircase structure, which, with one lithography step, can then contact each of the memory stacked planes.

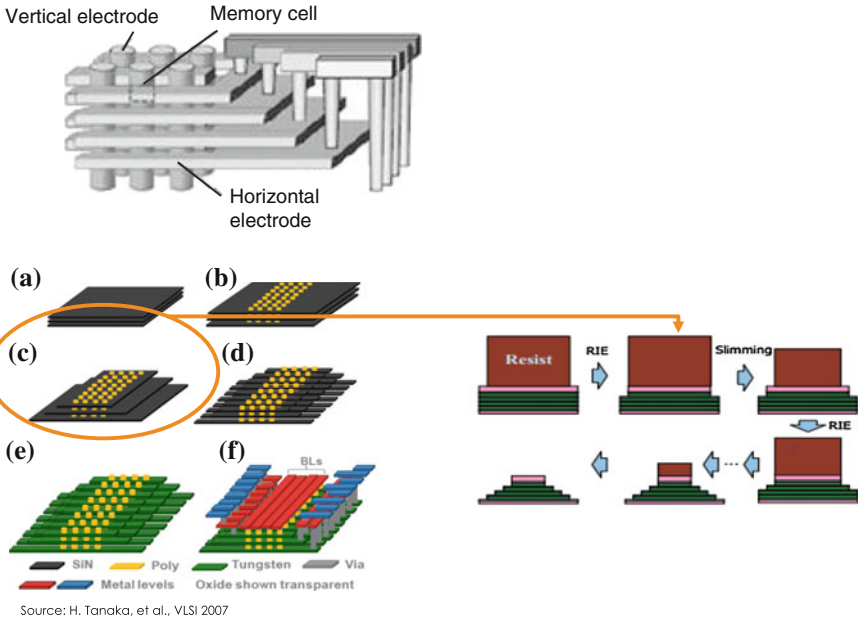


Fig. 3.25 Staircase patterns for contacts of the individual stacked layers. © IEEE 2007. **a** Deposit multiple Sio_2/SiN layers, **b** Etch hole and deposit channel poly (shared litho step), **c** Make staircase pattern for contacts (shared litho step), **d** WL cut etch (shared litho step), **e** Using flow in Fig. 3.24, make charge trap flash dielectrics and gate/WL, **f** BEOL

3.6.4 3D-NOR Flash

The NAND architecture became the most popular NV Memory architecture by providing the highest memory density and accordingly the lowest cost per bit. In some memory technologies, there is a need for a NOR architecture such as in 3D-DRAM [40] and RRAM [38]. These architectures could use polysilicon or mono-crystallized silicon. Figure 3.26 illustrates such a NOR architecture for RRAM application.

3.7 Advanced Work—Non-silicon Monolithic 3D

3.7.1 III–V Semiconductor 3D Integration

At SMART LEES, Singapore, a manufacturing facility is being put in place [41] to integrate III–V crystal layers with foundry-processed silicon offering truly heterogeneous integration as illustrated in Fig. 3.27.

Here, the heterogeneous integration is in two aspects. First is the ability to mass-produce some base generic function on a CMOS wafer by conventional foundries. Then a far smaller facility processes the monolithic 3D integration of

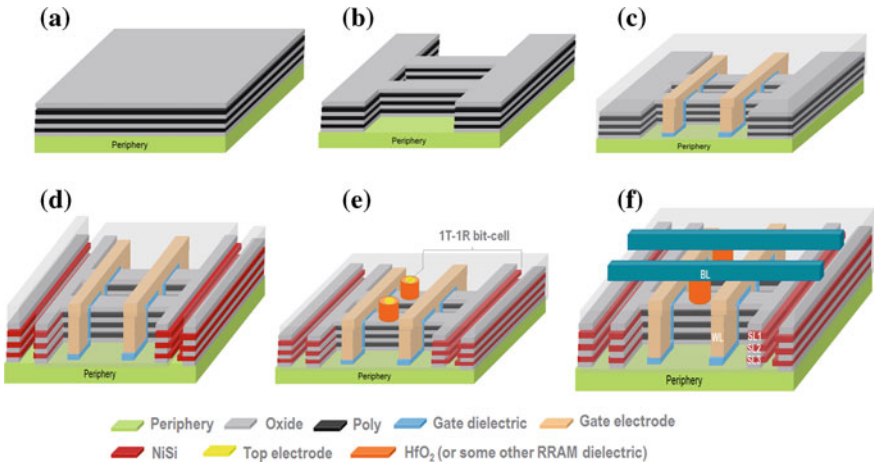


Fig. 3.26 3D RRAM—NOR Architecture [38]. © IEEE 2014. **a** Deposit multiple Sio_2/poly Si layers. Or use ion-cut to make Sio_2/c-Si layers, **b** Pattern (shared litho step), **c** Form gate of select transistors (shared litho step), **d** Pattern SL, then silicide (shared litho step), **e** Form RRAM dielectric and electrode for multi level 1T-1R cells (shared litho step), **f** Form BLs

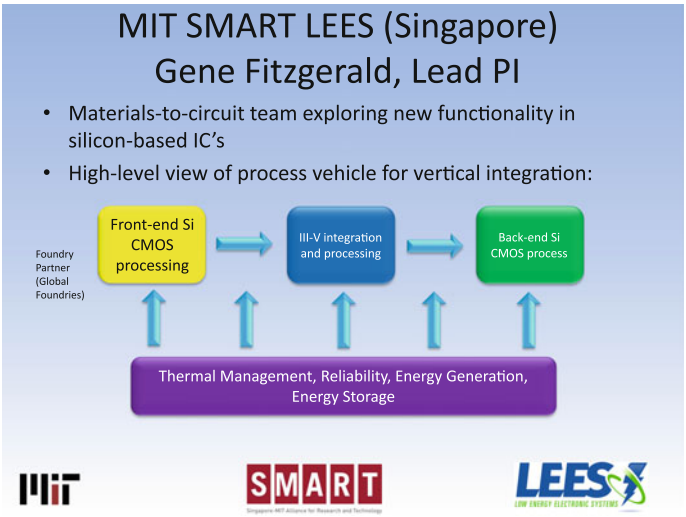


Fig. 3.27 Heterogeneous integration of III–V over silicon from foundries [41]. © IEEE 2014

upper strata of a lower-volume structure, customizing the device. The upper strata could be another type of crystal such as III–V materials. The III–V materials could be processed by epitaxial growth over base silicon materials and utilizing ion-cut to transfer over oxide and eventually on top of a foundry-base silicon wafer, like in Fig. 3.28.

- Layer transfer is key to removing thickness required for dislocation engineering

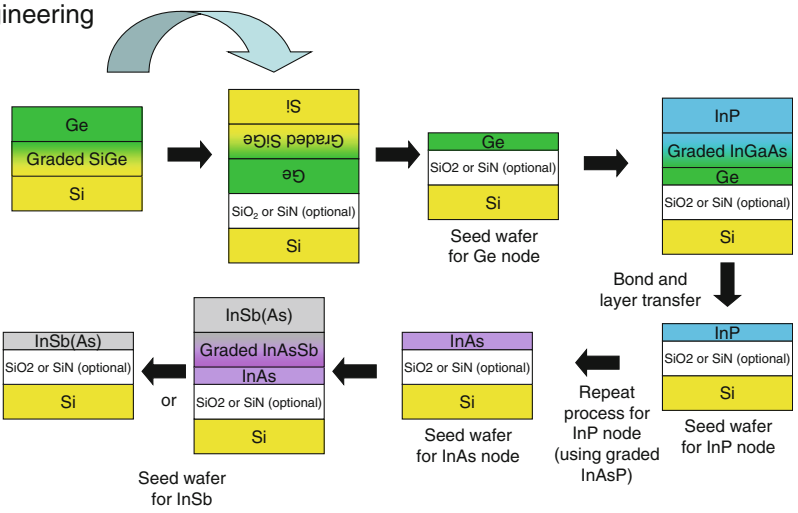


Fig. 3.28 Using layer-transfer (Ion-cut) for heterogeneous integration [41]. © IEEE 2014

3.7.2 Monolithic 3D Integration of Semiconductor, Carbon Nano Tube, STT MRAM and RRAM

A research effort going on at Stanford [42] and other leading US universities is looking to provide 1000× improvements over conventional semiconductor technology by leveraging monolithic 3D integration of semiconductor strata with a CNT STT-MRAM and R-RAM stratum—(Figs. 3.29, 3.30).

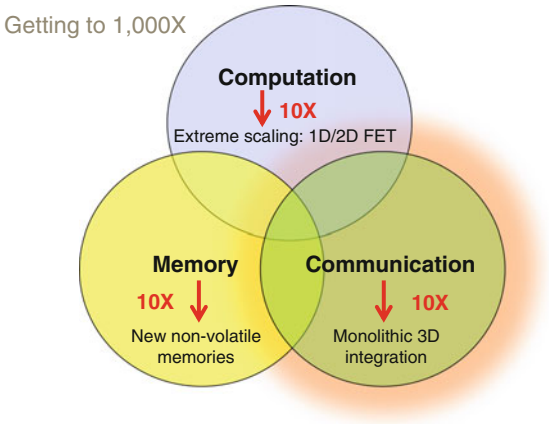


Fig. 3.29 1000× improvements [42]

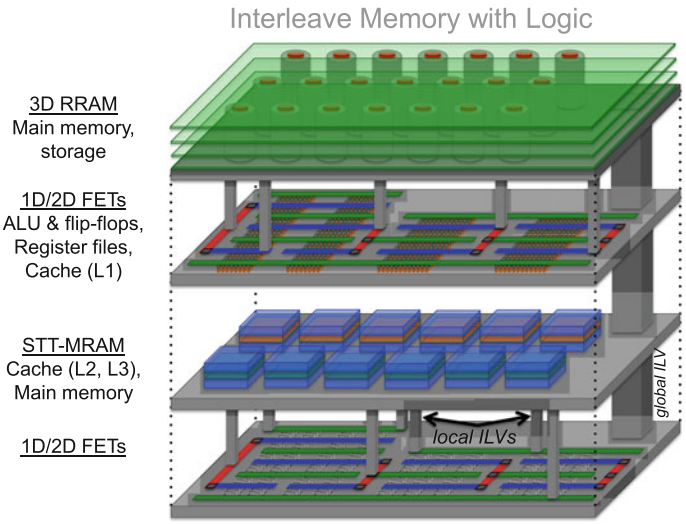


Fig. 3.30 Monolithic 3D integration of Semiconductor, Carbon Nano Tube, STT RAM and RRAM [42]. © IEEE 2014

3.8 The Monolithic 3D Advantages

3.8.1 Introduction

The most important aspect of scaling is the exponential increase of device integration. In this section, we cover some of the other advantages of monolithic 3D integration.

3.8.2 Reduction in Die Size and Power

3.8.2.1 Reduction in Die Size

Dimensional scaling has always been associated with increased wire resistivity and capacitance—see Fig. 3.4. Every node of dimensional scaling is associated with larger output drivers and more buffers and repeaters. Figure 3.31 illustrates the rapid increase of the number of transistors associated with the increased interconnect challenge.

Reduced interconnect length due to 3D stacking provides a reduction of buffers and the average transistor size. MonolithIC 3D Inc. released an open-source high-level simulator IntSim v2.0 to simulate a given design's expected size and

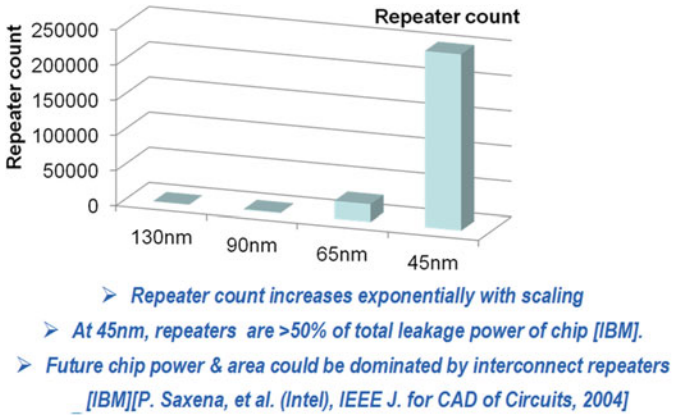


Fig. 3.31 Repeater and buffers consume escalating parts of the end-device power and area. *Source* IBM Power processors R. Puri et al. SRC Interconnect Forum, 2006

power based on process parameters and the number of strata. Using the simulator, we can see in Fig. 3.32 that a 2D design of 50 mm^2 area with an average gate-size of W/L = 6, will only need an average gate-size of W/L = 3 and accordingly only 24 mm^2 of total circuit area, if folded into two strata (the footprint will be therefore just 12 mm^2).

These results are in-line with many other monolithic-3D research results.

22nm node 600MHz logic core	2D-IC	3D-IC 2 Device Layers	Comments
Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire cap. to drive
Die Size (active silicon area)	50mm^2	24mm^2	3D-IC → Shorter wires → smaller gates → lower die area → wires even shorter 3D-IC footprint = 12mm^2
Power	Logic = 0.21W	Logic = 0.1W	Due to smaller Gate Size
	Reps. = 0.17W	Reps. = 0.04W	Due to shorter wires
	Wires = 0.87W	Wires = 0.44W	Due to shorter wires
	Clock = 0.33W	Clock = 0.19W	Due to less wire cap. to drive
	Total = 1.6W	Total = 0.8W	

Fig. 3.32 Monolithic 3D folding can reduce the required silicon by 50 %

3.8.2.2 Reduction in Power

Figure 3.33 illustrates that interconnect is now dominating about 80 % of device power.

Monolithic 3D enables the folding of a circuit, with each stratum only about 1 μ above or below its neighbor, combined with a very rich vertical connectivity between the strata—with the potential to strongly impact the 10 % of wires that consume more than 90 % of the device active power—Fig. 3.34.

3.8.3 Significant Advantages for Using the Same Fab and Design Tools

3.8.3.1 Depreciation

With dimensional scaling, every technology/process node requires a significant capital investment for new processing equipment—Fig. 3.35, significant R&D spending for new transistor process and device development—Fig. 3.36, and the building of an ever more complex and costly library and EDA flow.

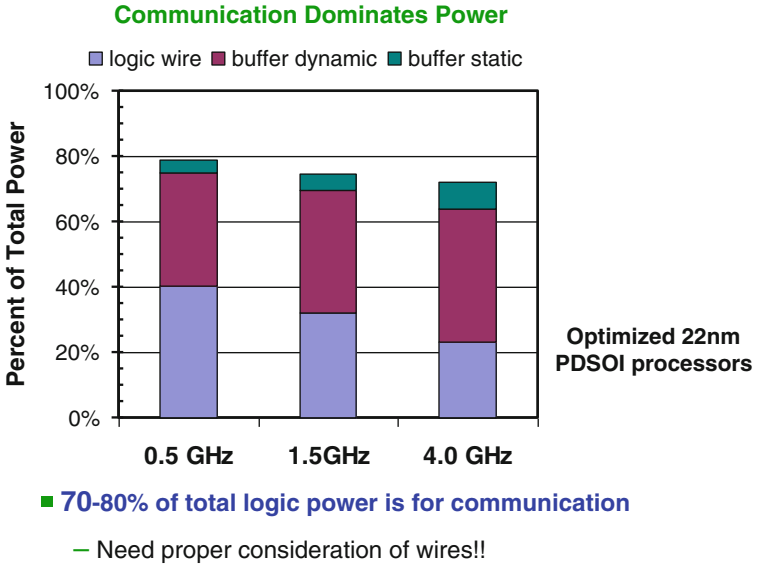
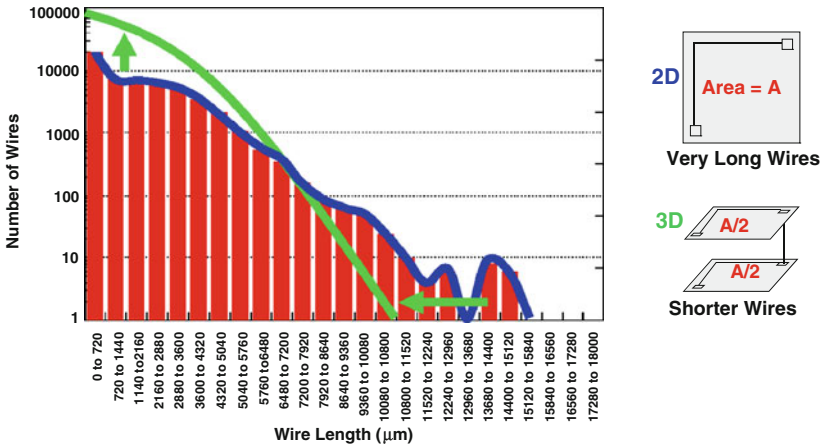


Fig. 3.33 Interconnect responsible for about 80 % of device power (IBM, Short-Course IEDM 2012)



- **>50% of active power (switching) dissipation is in microprocessor interconnects**
- **>90% of interconnect power is consumed by only 10% of the wires**

Fig. 3.34 Monolithic 3D can reduce the power and cost attributes by the long wires, MIT Lincoln Lab (After K. Guarini, IBM Semi R&D Center HPEC 2006)

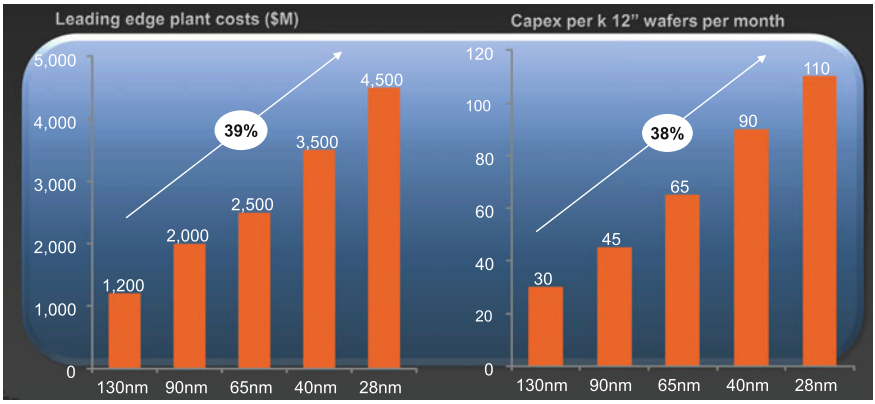


Fig. 3.35 Escalating Capex costs with dimensional scaling (*Source* World Fab Watch)

With monolithic 3D, these costs are not required as dimensions are maintained for multiple generations, and only the number of strata or layers is increased.

If the industry could use the same equipment and the same transistors and libraries for 4 years instead of 2, then all these costs could be depreciated over a longer time, with the resultant significant cost benefit.

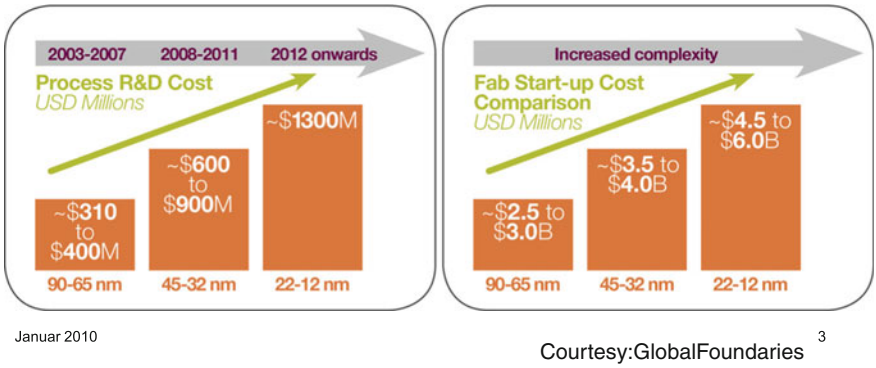


Fig. 3.36 Escalating process R&D costs with dimensional scaling

3.8.3.2 Learning Curve—Yield

Using the same transistor tools and EDA for longer periods has an additional important benefit. Learning curve equals yield improvement. With dimensional scaling, we face the predicament that, by the time we know how to manufacture a process node well, that learning quickly becomes obsolete as we quickly move on to the next node.

With monolithic 3D, the learning of the previous node stacking is directly utilized on the integration development of more strata, rather than on new materials, design-tool issues, etc. Figure 3.37 illustrates the dimensional scaling trend as each node of scaling is taking longer and costing more to get to a mature yield ('ramped-up').

3.8.4 Heterogeneous Integration

3D IC enables far more than an alternative for increased integration. It provides another dimension of design flexibility as shown in Fig. 3.38.

A well-known aspect of this flexibility is the ability to split the design into layers, which can be processed and operated independently, and still be tightly interconnected—especially for monolithic 3D as was presented in Sect. 3.7.

3.8.4.1 Logic, Memory, I/O

Let's start with quoting Mark Bohr, in charge of Intel's process development:

"One important perspective is that chip technology is becoming more heterogeneous. If you go back 10 or 20 years ago, it was homogeneous. There was a

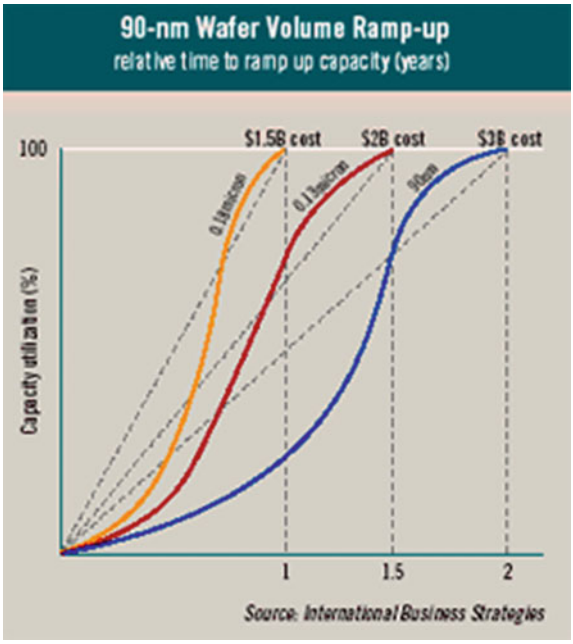


Fig. 3.37 The learning curve with dimension scaling

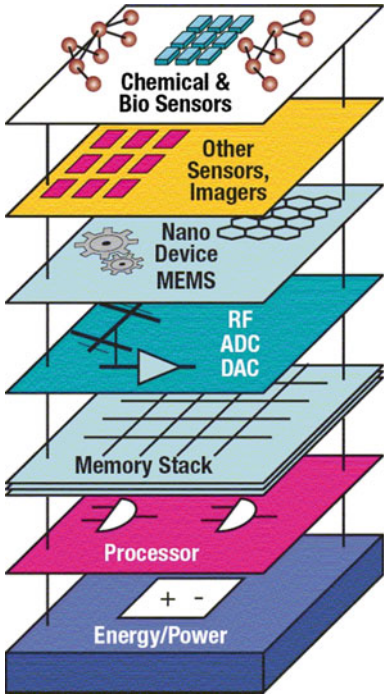


Fig. 3.38 Heterogeneous integration

CMOS transistor, it was the same materials for NMOS and PMOS, maybe different dopant atoms, and that basic CMOS transistor fit the needs of both memory and logic. Going forward, we'll see chips and 3D packages that combine more heterogeneous elements, different materials, and maybe transistors with very different structures whether they're for logic or memory or analog. **Combining these very different devices onto one chip or into a 3D stack—that's what we'll see**. It will be heterogeneous integration" (added emphasis).

The most important market for semiconductor products is smart mobility. For this market, the SoC device needs to integrate many functions, such as logic, memory, and analog. In most cases, the pure high-performance logic would be about 25 % of the die area, 50 % of the area would be memory (Fig. 3.39), and the rest would be analog functions such as I/O, RF, and sensors.

In 2D, all the functions need to be processed together and bear the same manufacturing costs. In a monolithic 3D-IC stack using heterogeneous integration, each stratum is processed in an optimized flow, allowing for a significant cost reduction and no loss in optimized performance for each function type.

3.8.4.2 Strata of Logic

The logic itself can be constructed better using heterogeneous integration. In many cases, only a portion of the logic needs to be high performance while other portions could be more cheaply done using an older process node. Other scenarios could include designing different strata with different supply voltages for power savings, different numbers of metal-interconnect layers, or other variations of the design space.

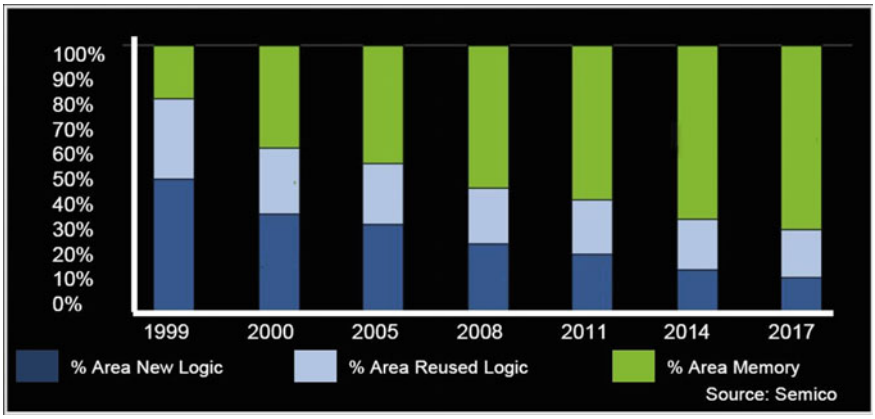


Fig. 3.39 Embedded memory to occupy about 70 % of SoC die area

3.8.4.3 Strata of Different Substrate Crystals and Fabrication Processes

3D enabled heterogeneous integration can be used as was presented in Sect. 3.7. Some layers can utilize silicon while others may use compound semiconductors. Some layers could be image sensors or other types of electro-optical structures and so forth.

3.8.5 Multiple Layers Processed Simultaneously—BiCS

An extremely powerful, unique advantage of monolithic 3D is the option to process multiple layers in parallel with one lithography step, as was detailed in Sect. 3.6. This option is most natural for regular circuits such as memory, but it is also available for other types of circuits with the right architectures.

The first merchants to recognize this option, and who are moving to monolithic 3D, are the NAND Flash vendors as seen in Fig. 3.40.

One of the clear future trends is the increase of content, often described by terms such as 'big data' and 'abundant data'. This trend certainly illuminates the future

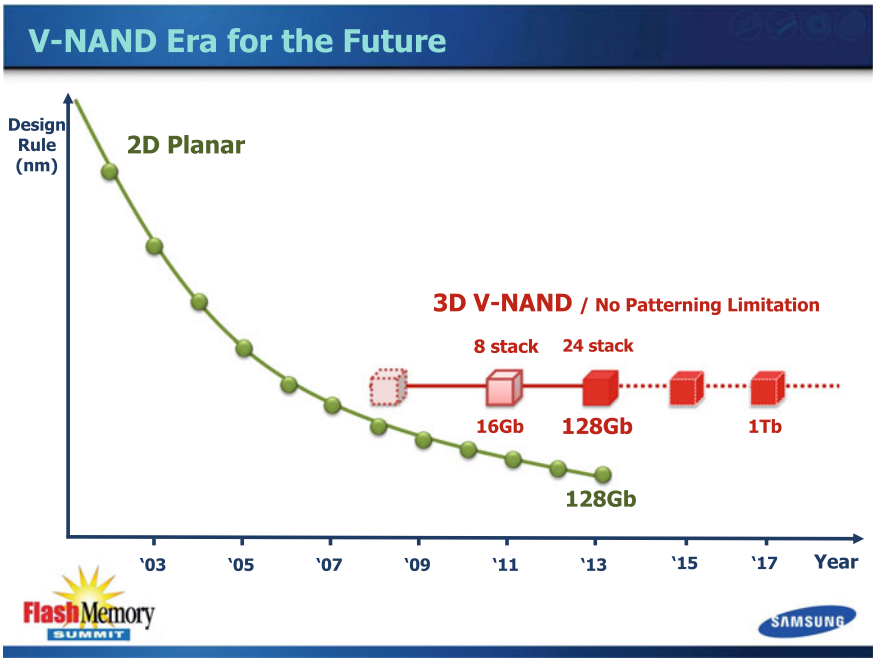


Fig. 3.40 For memory the future scaling is monolithic 3D (*Source* Flash Memory Summit and Samsung)

path for electronic systems. Monolithic 3D is well positioned to provide the best scaling path by integrating large amounts of 3D memory with logic providing a 1000× improvement in cost and power as previously discussed.

3.8.6 *Logic Redundancy Allowing 100× Integration with Good Yield*

The strongest value of an IC is the integration of many functions in one device. This is and will be the most important driver of Moore's Law, because, by integrating functions into one IC, we achieve orders-of-magnitude benefits in power, speed, and cost. At any given technology node, the limiting factor to integration is yield. As yield relates strongly to device area, most vendors are trying to limit the die size to about 50–100 mm^2. Some product applications require an extremely large die of over 600 mm^2, but those are rare and high value-added cases because the yield goes down exponentially as die size grows.

While memory redundancy is common in the IC industry, logic redundancy is only (and sparingly) used in a few FPGAs—no solution has been found after the failure of Trilogy, where "Triple Modular Redundancy" was employed systematically. Every logic gate and every flip-flop was triplicated with binary two-out-of-three voting at each flip-flop. Quoting Gene Amdahl: "*Wafer scale integration will only work with 99.99 % yield, which won't happen for* ***100 years.***" (*Source: Wikipedia*).

A unique advantage of monolithic 3D is the ability to construct redundancy for circuits including logic, with minimal impact on the design process and while maintaining circuit performance, such as shown in Fig. 3.41.

There are three primary ideas visible here:

- Swap at logic cone granularity.
- Redundant logic cone/block directly above, so no performance penalty.
- Negligible design effort, since the redundant layer is an exact copy.

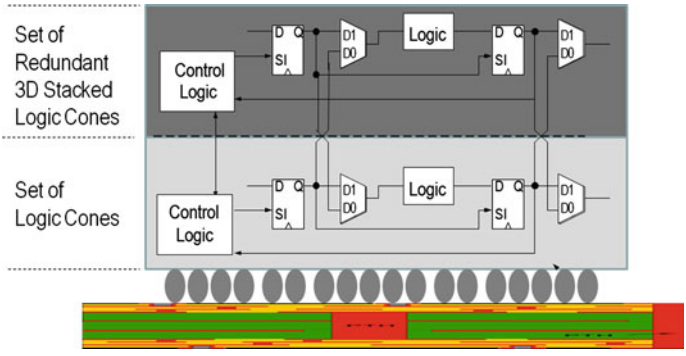


Fig. 3.41 Monolithic 3D enabling logic redundancy and repair

The new concept leverages two important technology breakthroughs:

The first is the scan-chain technology that enables a circuit test where faults are identified at the logic cone level. The second is the monolithic 3D IC, which enables a fine-grained redundancy: replacement of a defective logic cone by an identical logic cone that is only ~ 1 μm above.

Accordingly, by just building the same circuit twice, one on top of the other, with minimal overhead, every fault could be repaired by the replacement logic cone above. Such repair should have a negligible power penalty and a minimal cost penalty, whenever the base-circuit yield is as low as 50 %. There should be almost no extra design cost, and many additional benefits can be obtained.

This redundancy technique could be used also to repair faults throughout the device life-time, including in the field, which is a powerful advantage for some applications.

In today's designs, we expect more than one million flip-flops (and their logic cones). Consequently, if we expect one defect, then a device with redundancy layer would work unless the same cone is faulty on both layers, which, probability-wise, would be one in a million!

The ultra-integration value could be as much as:

- $\sim$10X Advantage of 3D WSI versus 2D @ Board Level
- $\sim$10X Advantage of 3D WSI versus 2D @ Rack Level
- $\sim$10X Advantage of 3D WSI versus 2D @ Server-Farm Level

Overall, a $\sim 1000\times$ advantage is possible, all due to shorter wires. Instead of placing chips on different packages, boards and racks, we integrate on the same stacked huge chip.

3.8.7 3D-FPGA

Dimensional scaling is associated with escalating mask-set and design costs. Yet designers choose in most case to use old process nodes rather than an FPGA [43]. As a result, the most popular node for design currently is 130 nm, a node that is trailing the leading edge by about 6 process nodes. 3D FPGAs could significantly reduce this huge gap by drastically reducing area- and cost-penalties. Accordingly, a 3D FPGA would have the opportunity to increase innovation and growth in the semiconductor industry at large.

3D-FPGAs could simplify the use of anti-fuse technology for high-density programmable interconnects with the additional benefits of programming the interconnection layers from the top, providing an order of magnitude density improvement.

An additional advantage of a 3D FPGA is in having two compatible products. The prototype device would have extra strata for the interconnect programming, which could be removed for further cost reduction in volume production, illustrated in Fig. 3.42.

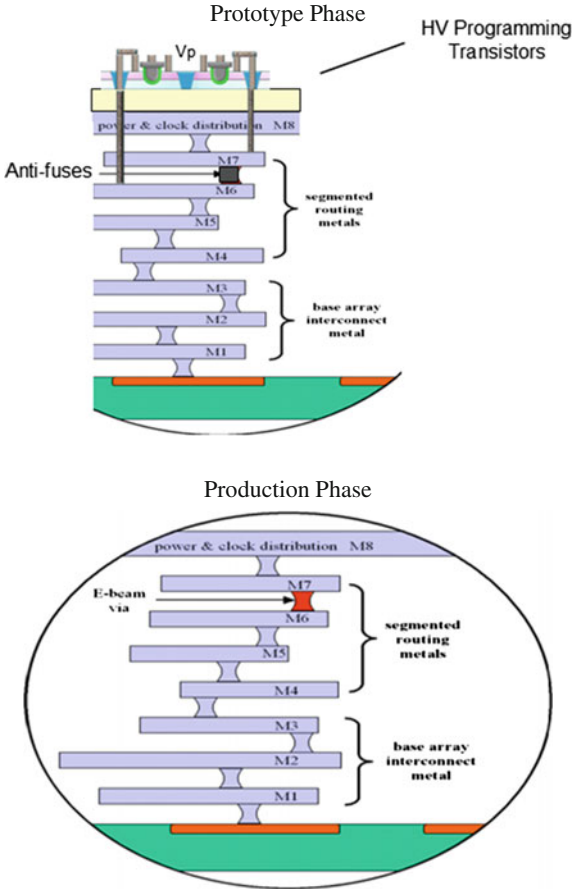


Fig. 3.42 The extra layers associated with interconnect programming could be removed for the volume part

3.8.8 Modular Platform

The 3D monolithic device would be a good fit for platform-based designs, wherein some parts of the device are used by all customers while other parts are tailored to a specific market/customer segment as illustrated in Fig. 3.43.

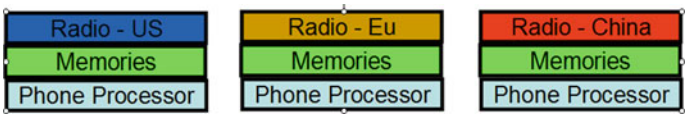


Fig. 3.43 3D-enabled modularity

Such a system architecture could be inexpensively used in many market segments and with multiple variations. One interesting application could be in the FPGA sector where the same platform could come with many flavors of memories and I/O.

3.8.9 Stacked Layers Are Naturally SOI

The upper layers of monolithic 3D devices are naturally Silicon-On-Insulator (SOI). The advantages of SOI are well-established, they increase with scaling, and they include:

- 90 % lower junction capacitance
- Near-ideal sub-threshold swing
- Reduced device cross-talk
- Lower junction-leakage
- Effective back-bias and multi-Vt options
- Multiple-gate operation for superb electrostatic channel control.

3.8.10 Local Interconnect Above and Below Transistor Layer

Improving on-chip interconnects is critical for enabling the increase in gate count. Simply adding interconnect layers provides limited improvement as each additional layer also adds to blockages/congestion in the intermediate layers created by the need to traverse them up and down the stack. In a monolithic 3D approach, interconnect can be formed and effectively used both above and below the transistor layer, thus doubling interconnect accessibility.

3.8.11 Re-buffering Global Interconnect by Upper Strata

Via blockage resulting from global interconnect buffering is growing exponentially as shown in Fig. 3.44. In addition to the reduction in buffers due to the significant reduction in the average wire-length in a 3D stack, moving those buffers to the upper stratum can effectively address the problem. Using such repeaters on a separate upper stratum does not add to routing congestion on the lower—and congested—metal layers, and it allows the utilization of a greater fraction of the active area.

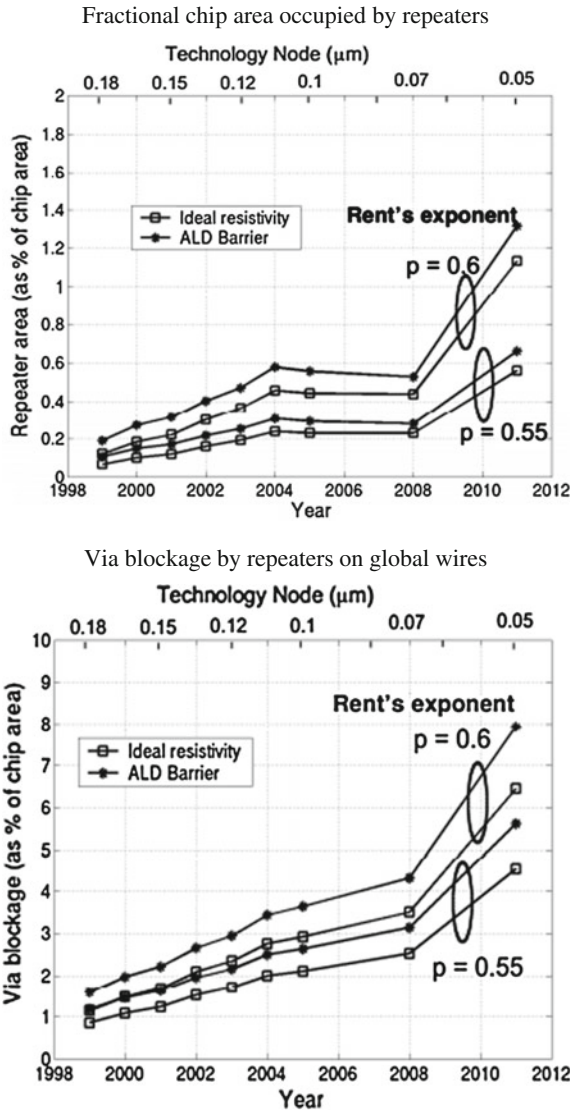


Fig. 3.44 Escalating area penalty for repeaters (*Source* Prof. Saraswat, Stanford Univ., EE311: Interconnect Scaling, slide 46)

3.8.12 Other Ideas

3.8.12.1 Image Sensor with Pixel Electronics

The image-sensor industry has moved to back-side illumination to increase the image-sensor area utilization. By adding the option for multiple strata, many additional benefits could be gained such as multi-spectrum day/night with extremely high dynamic range and other options as illustrated in Fig. 3.45.

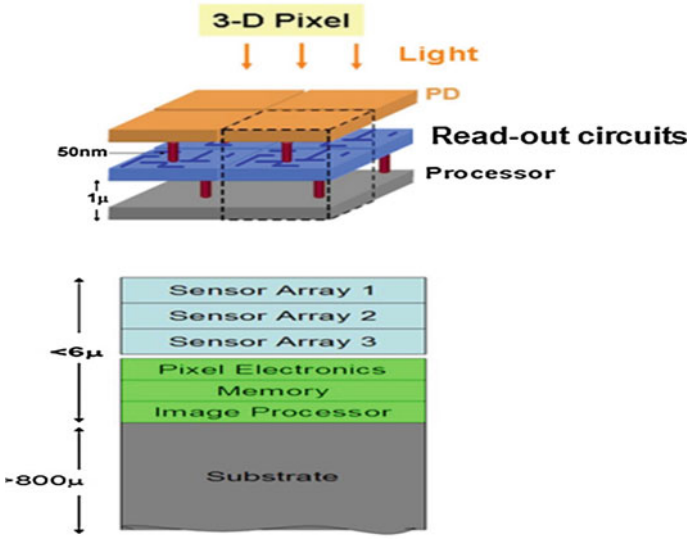


Fig. 3.45 Monolithic 3D image sensor

Pixel electronics behind every pixel could enable a very high dynamic range by counting and resetting individual sensors.

3.8.12.2 Micro-display

The display-market is always looking to reduce power and size while increasing the resolution and brightness. Monolithic 3D could provide ultra-high resolution with

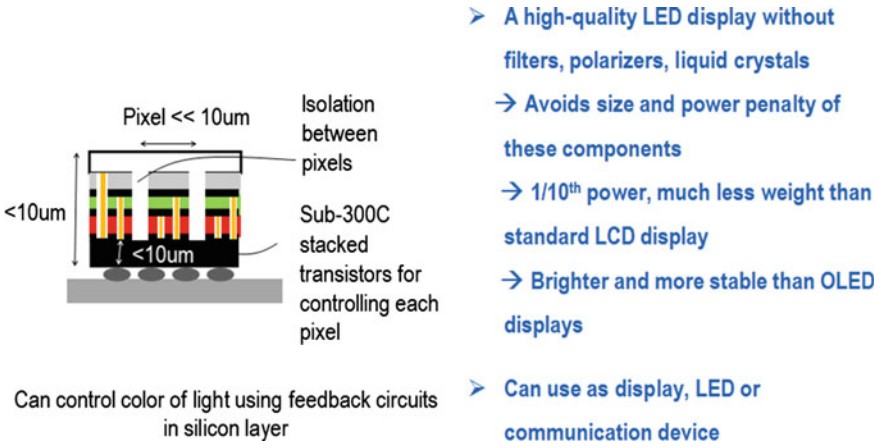


Fig. 3.46 Monolithic 3D micro-display

extreme power-efficiency and minimal size, by combining drive electronics with strata of different-color light-emitting diodes as is illustrated in Fig. 3.46.

3.9 Conclusion

Monolithic 3D is a disruptive semiconductor technology. It builds on existing infrastructure and know-how, and it can bring to the hightech industry many more years of continuous progress. While it provides all advantages, once provided by dimensional scaling, monolithic 3D offers many additional options and benefits. Best of all, monolithic 3D can be used in conjunction with dimensional scaling.

This chapter presented various techniques for monolithic 3D processing as well as multiple applications for monolithic 3D devices. It should be noted that the processes and applications could be mixed and matched in various ways to support future market needs.

References

1. Zingg, R.: Stacked CMOS inverter with symmetric device performance. IEDM (1989)
2. Roos, G.: Complex 3D CMOS circuits based on a triple-decker cell. J. Solid-State Circuits **27**, 1067 (1992)
3. Or-Bach, Z.: The monolithic 3D advantage: monolithic 3D is far more than just an alternative to 0.7× scaling. In: IEEE 3DIC Conference (2013)
4. Batude, P. et al.: Direct bonding: a key enabler for 3D monolithic integration. In: Proceedings of the Electro-Chemical Society (ECS) spring meeting, vol. 16, pp. 47 (2008)
5. Rajendran, B.: Pulsed laser annealing: a scalable and practical technology for monolithic 3D IC. In: IEEE 3DIC Conference (2013)
6. Yang, C.-C.:. Record-high 121/62 μA/μm on-currents 3D stacked epi-like Si FETs with and without metal back gate, paper 29.6 IEDM (2013)
7. Shen, C.-H.: Monolithic 3D chip integrated with 500 ns NVM, 3 ps logic circuits and SRAM. IEDM (2013)
8. Lee, S.-Y.: Wafer bonding method, US Patent 7,470,142
9. Natio, T.: World's first monolithic 3D-FPGA with TFT SRAM over 90 nm 9 layer Cu CMOS, VLSI Technology (2010)
10. Wong, S.: Monolithic 3D integrated circuits VLSI-TSA (2007)
11. Ishihara, R.: Monolithic 3D-ICs with single grain Si thin film transistors ULSI 2011
12. Yonehara, T.: Monolithic 3D-ICs with single grain Si thin film transistors. JSAP International No. 4, July 2001
13. Sadaka, M.: Smart stacking™ and smart Cut™ technologies for wafer level 3D integration. ICICDT (2013)
14. Topol, A.W.: Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs). IEDM (2005)
15. Vinet, M.: Monolithic 3D integration: a powerful alternative to classical 2D scaling. IEEE S3S (2014)
16. Deleonibus, S.: Future challenges and opportunities for heterogeneous process technology. Towards the thin films, zero intrinsic variability devices, zero power Era, IEDM (2014)

17. ASIP Conf.—Hughes Metras, Dec. 12, 2014
18. Sekar, D.C.: Monolithic 3D-ICs with single crystal silicon layers. IEEE 3DIC Conference (2011)
19. Kim, J.Y.: The breakthrough in data retention time of DRAM using recess-channel-array transistor (RCAT) for 88 nm feature size and beyond. Symposium on VLSI Technology (2003)
20. Uhrmann, T.: Monolithic IC integration key alignment aspects for high process yield. IEEE S3S (2014)
21. Sugaya, I.: New precision alignment methodology for CMOS wafer bonding. IEEE S3S (2014)
22. Or-Bach, Z.: Precision bonders—a game changer for monolithic 3D. IEEE S3S (2014)
23. Or-Bach, Z.: Practical process flows for monolithic 3D. IEEE S3S (2013)
24. Widjaja, Y.: Method of maintaining the state of semiconductor memory having electrically floating body transistor. US Patent 8,514,623
25. http://www.monolithic3d.com/ultra-large-integration—redundancy-and-repair.html
26. http://www.monolithic3d.com/blog/how-much-does-ion-cut-cost1
27. Xie, Y.: BooK: three-dimensional IC Design. Springer, Berlin (2009)
28. Zhou, L.: CASCADE: a standard supercell design methodology with congestion-driven placement for three-dimensional interconnect-heavy very-large-scale integrated circuits. In: IEEE Transactions on CAD of IC and Systems, July 2007
29. Sarhan, H.: 3DCoB: a new design approach for monolithic 3D integrated circuits. ASP-DAC (2014)
30. Bobba, S.: CELONCEL: effective design technique for 3-D monolithic integration targeting high-performance integrated circuits. ASP-DAC (2011)
31. Bobba, S.: Cell transformations and physical design techniques for 3D monolithic integrated circuits. ACM JETCS, Sept. 2013
32. Chang, L.: IBM, technology optimization for high energy efficient computation. Short Course, IEDM (2012)
33. Wei, H.: Cooling three-dimensional integrated circuits using power delivery networks. IEDM (2012)
34. Sekar, D.C.: Semiconductor device and structure. US Patent 8,686,428
35. Tanaka, H.: Bit cost scalable technology with punch and plug process for ultra high density flash memory. VLSI (2007)
36. Fukuzumi, Y.: Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory. IEDM 2007
37. Nitayama, A.: 3D NAND flash memories. Tutorials, S3S (2014)
38. Depak, S.: 3D RRAM. IEEE S3S (2014)
39. Jang, J.: Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra-high-density NAND flash memory. VLSI Symposium (2009)
40. Or-Bach, Z.: Semiconductor device and structure. US Patent 8,379,458
41. Fitzerald, E.: Monolithic 3D integration in a CMOS process flow. IEEE S3S (2014)
42. Ebrahimi, M.: Monolithic 3D integration advances and challenges: from technology to system levels. IEEE S3S (2014)
43. Or-Bach, Z.: FPGAs as ASIC alternatives: past & future. EE Times Apr. 21 (2014)