Monolithic 3D-ICs with Single Crystal Silicon Layers

Deepak C. Sekar and Zvi Or-Bach
MonolithIC 3D Inc., San Jose, CA 95124, USA E-mail: deepak@monolithic3d.com *(Invited Paper)*

Abstract

Approaches to obtain monolithic 3D logic and memory ICs are proposed in this paper. 3-4x higher memory density with similar litho cost can be obtained with monolithic 3D memories, while benefits similar to a generation of scaling can be obtained with monolithic 3D logic by doubling the number of device layers. Well-known, manufacturing-friendly materials, process steps and device structures are used.

Introduction

3D-ICs with monolithically stacked transistor layers provide ultra-dense through-silicon connections and short wires. In addition, they often utilize the same litho step to pattern multiple layers of devices, thereby reducing litho cost. A few trends are making monolithic 3D-ICs increasingly attractive. Fig. 1 shows litho tool cost has increased exponentially over the past 40 years. Next-generation litho is now quite expensive and risky. Fig. 2 indicates wire RC delays are a key bottleneck in today's chips. A 28nm GPU, for instance, requires several times more energy for communicating data than for computation. This trend is expected to get worse with scaling (Fig. 2). Down-scaling requires major changes to transistors too, with disruptive innovations such as high k/metal gate and Finfets required every few years.

Ion-Cut Technology

Single crystal silicon for stacked device layers of Monolithic 3D-ICs can be obtained with ion-cut technology, which involves hydrogen implantation, wafer bonding and cleave (Fig. 3). Ion-cut has been demonstrated below 400^{o}C [3], and is well-known due to its two decades of use for SOI wafer manufacturing.

Monolithic 3D Memories

Flash memory, which scales faster than logic and DRAM, may hit the limits of traditional scaling first. Toshiba, Samsung, Micron and Hynix have all invested aggressively in *polysilicon*-based monolithic 3D flash memories, and have them on their roadmaps [4][5]. Fig. 4 shows our proposal for *single crystal silicon* 3D flash memories with litho steps shared among multiple device layers. Single crystal silicon flash memory cells have ~5x higher mobility, better off-characteristics and lower variability compared with poly cells, all of which help with multi-bit storage. As shown in Fig. 4, ion-cut is utilized repeatedly to produce multiple layers of single crystal silicon atop peripheral circuits with tungsten wiring, following which shared litho steps are utilized to define NAND flash memory strings and produce contacts [5]. Fig. 5 shows memory capacity estimations based on [5]. For a 140mm^2 die, the structure in Fig. 4 can provide 256Gbit chips compared with 64Gbit and 128Gbit for conventional scaled NAND and polysilicon-based 3D vertical NAND flash respectively [5]. This is due to the cost advantages of sharing litho steps across multiple memory layers as well as the multi-bit storage capacity of single crystal silicon cells. Maximum aspect ratios for etch and deposition in Fig. 4 are 16:1 compared to higher than 50:1 for poly-based 3D vertical NAND flash, allowing easier manufacturing. Note that ion-cut could be applied to any 3D flash memory with horizontal channels, and is not limited to the exemplary architecture [6] shown in Fig. 4.

Monolithic 3D architectures with shared litho steps have been developed by the authors for floating-body DRAM and resistive memories too [7]. Single crystal silicon is an enabler for these applications. The primary concern with applying ion-cut to memories is cost. Our analysis reveals that with re-use of substrates, ion-cut could cost as low as $60 per layer, making it affordable. Encouragingly, some companies in the cost-sensitive solar industry are adopting ion-cut today.

Monolithic 3D Logic

Much of today's 3D logic stacks utilize chips processed separately with high temperatures that are then thinned, bonded and connected to each other. Silicon thickness of thinned die and misalignment during bonding are concerns with this approach. The ITRS predicts minimum TSV diameter around 1um between 2009 and 2015, indicating difficulties scaling to small TSV sizes needed for many applications. Monolithic 3D-ICs could be a solution. Fig. 6 indicates the main barrier to creating high-quality transistors at Cu/low k compatible temperatures (sub-400^{o}C) is dopant activation. Fig. 7 describes one approach to overcome this problem, which utilizes recessed channel transistors. These have been used in DRAM manufacturing since the 90nm node, and are known to be competitive with standard planar transistors [9]. As can be seen in Fig. 7, high temperature dopant activation steps are conducted before transferring bilayer n+/p silicon layers atop Cu/low k using ion-cut. The transferred layers are unpatterned, therefore no misalignment issues occur while bonding. Following bonding, sub-400^{o}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 alignment marks of the bottom Cu/low k stack since transferred silicon films are thin (usually sub-100nm) and transparent. Sub-50nm through-silicon connections can be produced due to the excellent alignment.

To investigate the chip-level impact of monolithic 3D-ICs, an open-source 2D/3D chip simulator called IntSim [10] was used. Fig. 8 gives a description of IntSim. Fig. 9 reveals that for a 22nm 600MHz logic core, doubling the number of monolithically stacked device layers can provide similar benefits to a generation of scaling (2x lower power, 2x lower die size). These advantages are due to the shorter wires provided by monolithic 3D, which allow reduced gate (driver) sizes. Note that Rent's Rule based stochastic wire length distributions were used for this analysis, as CAD tools for monolithic 3D design are still immature today.

Conclusions

Several techniques to obtain monolithic 3D chips have been proposed in this paper. Benefits equivalent to several generations of scaling can be obtained with the technology without incurring the cost and risk of next-generation lithography. Wiring and litho limitations of traditional scaling could make monolithic 3D increasingly attractive moving forward.

References

[1] W. Dally, Throughput Computing, Intl. Conf. on Supercomputing, 2010 [2] S. Naffziger, S. on VLSI Technology, 2011 [3] M. Sadaka, et al., Building Blocks for Wafer-Level 3D Integration, Solid State Technology [4] J. Choi, et al., S. on VLSI Technology, 2011 [5] R. Liu, Short Course at S. on VLSI Technology, 2010 [6] H. T. Lue, S. on VLSI Technology, 2010 [7] US Patent # 8,026,521, MonolithIC 3D Inc. [8] Y. Saito, et al., S. on VLSI Technology, 2000 [9] J. Y. Kim et al., S. on VLSI Technology, 2003. [10] D. C. Sekar, et al., ICCAD, 2007.

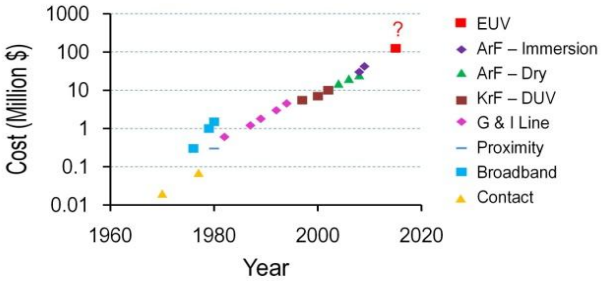

Fig. 1: Exponential increase of litho tool cost. An immersion litho tool costs $40M today while other tools in the fab cost $1M-$5M.

Operation	pJ
Integer Add	1
Fetching operands from	
A register file 1mm away	26
L1/L2/L3 caches	50/256/ 1000

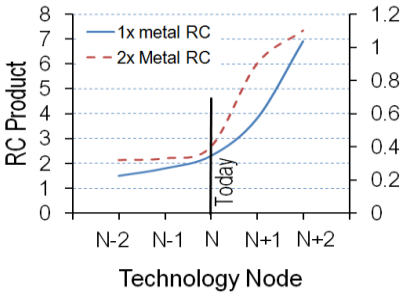

Fig. 2: Wire RC delay and energy trends (Left) nVIDIA 28nm GPUs [1] and (Right) AMD chips [2].

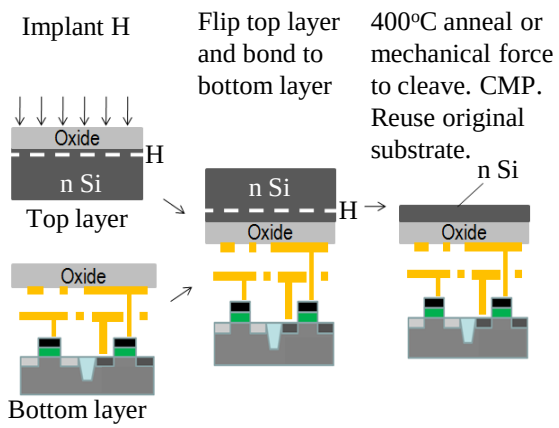

Fig. 3: Ion-Cut process for stacking single crystal silicon (c-Si) layers at less than 400°C.

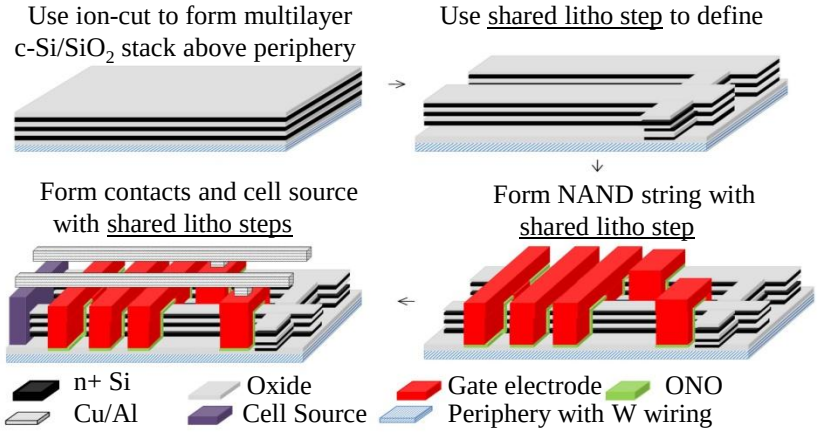

Fig. 4: Process Flow for monolithic 3D NAND flash with junction-free charge-trap flash cells made of single crystal silicon.

140 sq. mm die	2D NAND @22nm	Poly Vertical 3D NAND (BiCS) 32 layers @45nm	3D Single Crystal Silicon NAND 8 layers @22nm
Density	64Gbit (3 bits/cell)	128Gbit (1 bit/cell)	256Gbit (2 bits/cell)
Aspect ratio		60:1	16:1

Fig. 5: Comparison with conventional NAND flash and poly 3D NAND.

	Sub-400°C possible?	Method
Single Crystal Silicon	Yes	Ion-Cut
Shallow Trench Isolation	Yes	Radical Oxidation [8], HDP
High k/Metal Gate	Yes	ALD/CVD
S-D Dopant activation	No	>750°C anneal
Contacts	Yes	Nickel Silicide

Fig. 6: Monolithic 3D with conventional logic transistors would need temperatures higher than 400°C.

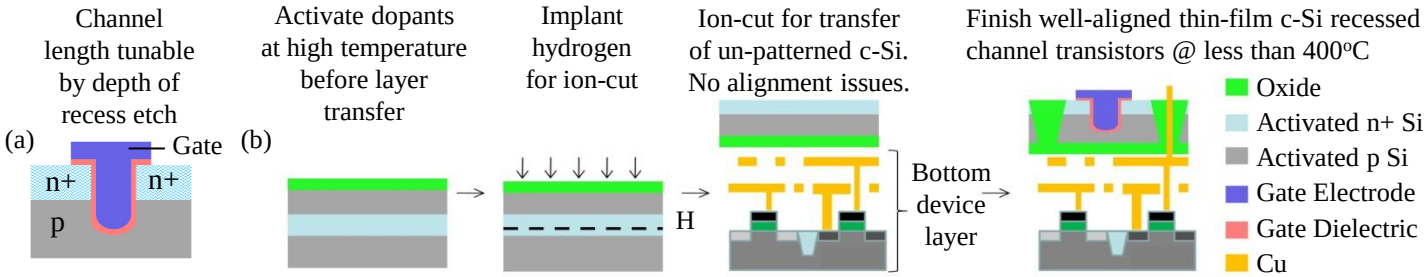

Fig. 7: (a) A recessed channel transistor (b) Process flow for monolithic 3D logic. Bottom device layer with Cu/low k does not see more than 400°C. Through-silicon connections can be close to minimum feature size due to the thin-film process.

Inputs:
- Gate count
- Die area
- Frequency
- Rent's parameters
- Number of device layers

IntSim v2.0

Contains models for: Stochastic wire length distributions of 2D/3D-ICs, logic gates, repeaters, chip power as well as power, clock and thermal interconnect networks

Outputs:
- Chip power
- Metal level count
- Wire pitches

Fig. 8: IntSim, an open-source 2D/3D chip simulator.

10 metal layers per device layer	2D @ 22nm	2 layer 3D @ 22nm	Comment
Average wire length	6um	3.1um	Since 3D and optimal die size less
Average gate size	6W/L	3W/L	Since less wire cap.
Optimal total silicon area (Footprint)	50mm² (50mm²)	24mm² (12mm²)	Since smaller gates, shorter wires
Power	1.6W	0.8W	Since gate, wire, repeater area less

Fig. 9: IntSim's results for a 22nm 600MHz logic core.