Rebooting Computing, computing, Moore's law, 3D chip manufacture, chip stacking, monolithic 3D, neural network, machine learning

Editor: Erik P. DeBenedictis, Sandia National Laboratories; epdeben@sandia.gov

Rebooting Computing

Sustaining Moore's Law with 3D Chips[1]

Erik P. DeBenedictis, Sandia National Laboratories
Mustafa Badaroglu, Qualcomm
An Chen, SRC
Thomas M. Conte, Georgia Tech
Paolo Gargini, IRDS

Rather than continue the expensive and time-consuming quest for transistor replacement, the authors argue that 3D chips coupled with new computer architectures can keep Moore's law on its scaling path.

The integrated circuit concept that drove the information revolution was based on downscaling the size of features on the 2D surface of a silicon die. Although the original scaling process is nearing its limits, tantalizing applications in, for example, AI could expand the economy, given more capable computers. The crucial question is whether a new scaling path can be developed to support these applications.

Here, we discuss recent 3D semiconductor manufacturing processes that permit faster, more complex, and more energy-efficient computers, provided they're accompanied by new computer architectures. We preview how the 2D-to-3D shift will support these applications in the next International Roadmap for Devices and Systems (IRDS).[1]

History of the Integrated Circuit

In his famous 1965 article,[2] Gordon Moore projected that the number of devices on integrated circuits, such as the one in Figure 1a, would increase exponentially,

---

[1] Sandia National Laboratories approved for unlimited unclassified release SAND2017-9177 J Published as DeBenedictis, Erik P., et al. "Sustaining Moore's Law with 3D Chips." *Computer* 50.8 (2017): 69-73. **DOI:** 10.1109/MC.2017.3001236

driven by lithographic line width decreasing to about 0.7× its previous value with every generation of one to two years. Decreasing linear dimensions would cause the number of components to increase with the inverse square of this ratio – or about 2× per generation – and the energy on each transistor and wire would decrease proportionally to the size of the devices and wires. As Moore predicted, computers have become exponentially more capable over time, even with no change in their purchase price and energy consumption, thus enabling the information revolution. Unfortunately, this scaling process is reaching its physical limits.

Current 3D Options

The original integrated circuit included a single device layer, but several processes now in high-volume production make use of the third dimension, which allows more components on each integrated structure, shortens wire lengths, and reduces chip-to-chip interconnect bottlenecks.

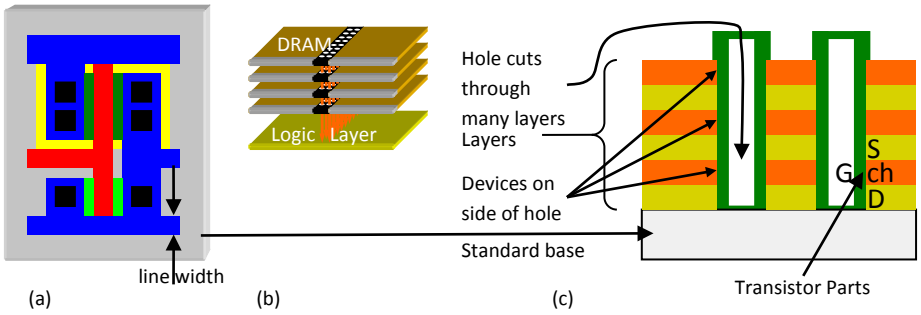

Figure 1. 2D and 3D scenarios. (a) A standard 2D integrated circuit defined as lithographic features with a shrinking line width on a silicon die. (b) Die stacking (High Bandwidth Memory illustrated): four dynamic random-access memory (DRAM) die assembled in layers mechanically and electrically above a logic layer. (c) Monolithic 3D (flash solid-state drive illustrated), manufactured as layers with devices chemically formed on the sides of holes. ch: channel; D: drain; G: gate; S: source.

Die stacking, as Figure 1b illustrates, is one of these 3D processes. It involves manufacturing each layer as a separate die and then assembling these layers into a stack, including electrical connections along the large face between layers. For readers familiar with the inside of a computer, die are silicon rectangles up to 1 cm × 1 cm and a few microns thick within the much larger packages that users can see. About a dozen die can be stacked before imperfect connections cause the yield to drop, meaning that, to the human eye, a whole die stack looks like a single die. Unfortunately, the stacked die must be paid for, so *n* stacked die cost somewhat more

than the *n* die did to start with. Some high-end graphics cards[3] available for gamers contain stacked High Bandwidth Memory (see Figure 1b).

Monolithic 3D chips, shown in Figure 1c, are manufactured by a process that starts with a single, conventionally designed silicon chip. Many layers of material are deposited on top of the base chip without any lithographic patterning, thus creating a plywood-like structure on top of the chip. Next, lithographic patterning is applied that goes through all the layers at once, as if cutting plywood, and elaborate chemistry creates devices on the edges of the layers exposed by the lithographic step. This leads to a 3D array with as many devices in the *x* and *y* dimensions as would be possible using 2D fabrication, but with one device for every two layers in the *z* dimension. The single lithographic step cuts cost but constrains all the layers to the same pattern—so the process only works for memory at the moment. Some currently available solid-state disk drives have 72 layers of devices in the *z* dimension, with this number projected to grow to 200.[4]

3D Roadmap

The 2017 IRDS road map will project the state-of-the-art integrated-circuit technology for each year or so into the future for the next 15 years. Figure 2 offers a preview.[1] As you can see, effective 2D density scaling in accordance with Moore's law is predicted until the mid-2020s, after which advances will come from the 2D density multiplied by the number of layers, or the projection of the 3D density onto a 2D surface. This scaling will reduce die purchase and lifetime energy costs and deliver system-level benefits.[5] This evolution demonstrates that the ultimate point of scaling is 3D VLSI that offers the possibility to stack devices that enable high-density contacts at the device level (up to 100 million "vias" per square millimeter), as Table 1 shows. 3D packaging has been available for decades in primitive forms, with connections, performance, and energy efficiency growing as technology has progressed. Fully integrated 3D is expected to be available in less than a decade.

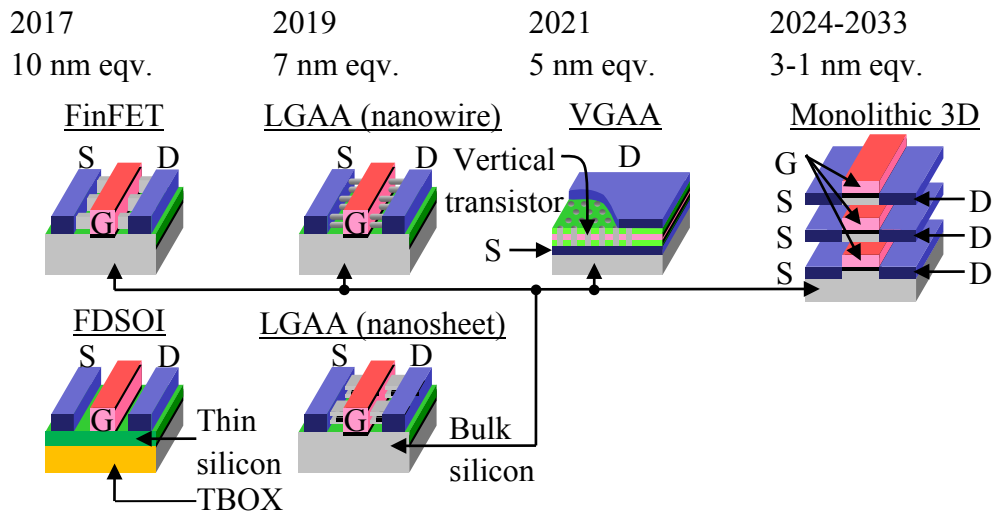

Figure 2. Preview of the 2017 International Roadmap for Devices and Systems' predictions for 3D structures. D: drain, FDSOI: Fully depleted silicon on insulator, G: gate, LGAA: Lateral Gate-All-Around, S: source, TBOX: thin buried oxide, VGAA: Vertical Gate-All-Around.

Options	Links	Bandwidth	Latency	Power	Timeframe
Wire-bond stack	100s	Low	High	High	Available for 30 y
Through Silicon Via (TSV) or microbump stack	1,000s	Medium	Medium	Medium	Available for 10 y
3D VLSI stack	100,000s	High	Low	Low	< 10 y

Table I. Evolution of 3D integration options toward 3D VLSI. 3D packaging has been available for decades in primitive forms, with the number of connections, performance, and energy efficiency growing as the technology progresses. Fully integrated 3D is expected to be available in less than a decade.

The IRDS road map includes heterogeneous integration of devices that are functionally different from the logic gates and memory, such as sensors, new storage devices, and alternative logic and memory devices or circuits. The IRDS projects five generations of functional change in the transition from 2D to 3D VLSI

1. Two stacked layers with non-scaled components such as analog, I/O, and power management in one layer and high-performance logic and memory in the other.

2. Monolithic integration of two layers, where each layer contains one of the two fundamental transistor types in logic, NMOS and PMOS (*n*-channel and *p*-channel metal-oxide semiconductors), stacked on top of each other for increased logic density.

3. Two layers: logic and memory.

4. Analog, I/O, and RF connectivity as an extra layer, giving more freedom to include special devices in the design.

5. 3D VLSI with fine-pitch logic-on-logic as well as special function layers exploiting new architectures (discussed in the next section).

Evolving Architectural Requirements

AI is a new and exciting application area that can benefit from the shift to 3D chips. IBM Watson's *Jeopardy!* win[6] and Google AlphaGo's win over the best human Go player[7] both ran on server clusters too large for wide-spread use. Today's self-driving cars use more modest processors, but the cars aren't necessarily safe. Making them safer will require more sophisticated control systems that are likely to increase computational requirements. The von Neumann architecture divides computers into a processor and memory. The resulting serialization of memory access – affectionately called the "von Neumann bottleneck" – makes programming easy for humans but limits performance. AI applications learn in lieu of being programmed, further decreasing the motivation to maintain architecture and programming traditions.

A new generation of AI architectures is emerging that more tightly integrates processor and memory. For example, IBM's TrueNorth (www.research.ibm.com/articles/brain -chip.shtml) neural network chip is mostly memory, whereas Google TensorFlow (www.tensorflow.org) and new versions of the GPUs used in self-driving cars have very wide DDR5 memories. Analog neural network arrays, such as resistive random-access memory (RRAM) crossbars, use a single device to store data and perform logic on it, and arguably represent the ultimate in logic–memory integration. The DARPA Hierarchical Identify Verify Exploit (HIVE) program is developing what might be this progression's ultimate endpoint: a processor that's tuned to sparse data representations and operations found in many AI applications but isn't focused on a specific application.[8, 9] The example below highlights this data-merging capability.

As Figure 3a illustrates, a traditional von Neumann computer system is actually the heterogeneous integration of logic chips (yellow), memory chips (orange), and chip-to-chip interconnects (gray). Just as cell phones use a specialized semiconductor process for RF electronics and another for camera sensors, the

dynamic random-access memory (DRAM) process produces slower devices with 10×higher memory density than the logic process of the same generation. Thus, algorithms in which information dependencies alternate between logic and a large memory (that is, memory too large for implementation by cache) must inevitably move data across the slow, high-energy interface between logic and memory, twice per repetition (as shown by the green line.

Figure 3b shows the equivalent data-dependency diagram for one merge stage of a bitonic sort,[10] a very fast sorting or merging algorithm that can process sparse data representations efficiently. The yellow logic layer on top shows the merging of two sorted lists of eight data records, which occurs in four stages. The first stage does pairwise comparison of all 16 values, swapping the records if needed to put the largest on the upper left. The second and later stages do pairwise comparisons as well, but on more groups of fewer records. When implemented with the structure in Figure 3b, the comparisons and swaps can be laid out on the surface of the logic layer. Data to be processed (Data A) and a place to store the result (Data B) are collocated in the orange memory immediately below and connected via short wires crossing the large 2D surface, rather than an edge.

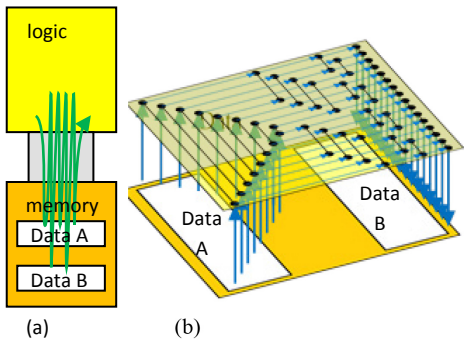

Figure 3. Advantage of 3D for interconnections. (a) 2D systems comprise logic and memory chips, with the green curve illustrating a mixed logic-memory calculation that's inefficient precisely because of this logic and memory partitioning. (b) 3D systems with tight coupling between logic and memory avoid high latency paths, bandwidth bottlenecks, and conversion of signals to high energy levels for off-chip interconnects. The blue curve shows a representative data movement step in sorting.

The sorting algorithm's speed varies depending on whether it's implemented using Figure 3a's or 3b's physical structure. For the 3D structure in Figure 3b, the data can be read all at once from physically collocated memory, and compared and swapped in hardware and in parallel. In contrast, a von Neumann computer would need to

move all the data from memory to the processor sequentially, taking n times as long for n units of data. For small n values, the data could be cached in memory, but this isn't an option for large n values; thus, the bottleneck becomes a principal issue when processing big data graphs and other low-level kernels in AI.

Scaling Through Tightening the 3D Interface

Combining the ideas just presented leads to a potential new scale-up path. Although this path will have limited duration compared to the line-width scaling that drove Moore's law for 50 years, it will be entirely independent of improvements in device physics. The interface between logic and memory has restricted systems for decades. Removing this bottleneck could let the benefits of past semiconductor advances carry forward to improve system performance. However, doing so will require advances (as shown in Figures 2 and 3) whose engineering processes will take some time to work through. But each advance will reduce the von Neumann bottleneck's impact a little. For applications that would benefit from tighter integration, this could offer generation-to-generation performance increases similar to linewidth scaling.

3D's Relevance to Moore's Law

These ideas are unlikely to move the semiconductor industry into the post–Moore's law era. However, Moore's law is popularly interpreted as applying to everything from semiconductor line width to microprocessor throughput. We see Figure 1c's monolithic 3D as completely compatible with the manufacturing-cost arguments central to Moore's 1965 article.[2] His piece wasn't about microprocessors – which weren't invented until 1971 – but as far as we can tell, it applied just as much to potential future microprocessors as to the bitonic sort–merge network invented in 1968.[10] Moore's article discussed future applications, including "automatic controls for automobiles," although, admittedly, he was probably referring to antilock brakes and not autopilots.

Computer performance isn't improving fast enough to sustain the overall IT sector's expansion at historic rates. The transistor replacement that was expected to save the day remains out of sight, but 3D memory – in the form of stacked DRAM and solid-state disks – is in large-scale production. Computer performance is traditionally viewed as coming from the processor not the memory, and the von Neumann bottleneck between processor and memory is holding up progress.

We've shown how to get around these obstacles to restore industry expansion. Instead of trying to build computers with the von Neumann architecture to satisfy programmers' needs, we suggest that industry build hardware for the big emerging application areas of AI, machine learning, data analytics, and so on. Because such

computers wouldn't use the von Neumann architecture, they wouldn't experience its bottleneck. With 3D chips, some critical algorithms will run better – *n* times better in algorithmic terms, which is quite a bit.

According to the IRDS, the 2D-to-3D shift will take a decade or more. To benefit along the way, industry should develop product families of architectures, systems, and applications that follow the road map with tighter integration in the third dimension.

Acknowledgments

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the US Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

References

1. Online: IEEE International Roadmap for Devices and Systems," website, IEEE, 2017; irds.ieee.org.
2. G.E. Moore, "Cramming More Components onto Integrated Circuits," Reprinted from *Electronics*, Volume 38, Number 8, April 19, 1965, pp. 114 ff, *IEEE J. Solid-State Circuits Newsletter*, vol. 11, no. 5, 2006, pp. 33–35.
3. High *Bandwidth Memory (HBM) DRAM*, JESD235A, Item 1797.99F., JEDEC, Nov. 2015; www.jedec.org/standards-documents/results/HBM.
4. H.S.Yoon et al., "Vertical Cross-Point Resistance Change Memory for Ultra-High Density Non-Volatile Memory Applications," *Proc. Symp. VLSI Technology* (VLSI-Technology 09), 2009, pp. 26-27.
5. M. Badaroglu and J. Xu, "Interconnect-Aware Device Targeting from PPA Perspective," *Proc. IEEE/ACM Int'l Conf. Computer-Aided Design* (ICCAD 16), 2016; doi.org/10.1145/2966986.2980068.
6. D. Ferrucci et al., "Building Watson: An Overview of the DeepQA Project," *AI Mag.*, vol 31, no. 3, 2010, pp 59-79.
7. D. Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, vol. 529, no. 7587, 2016, pp. 484–489.
8. T. Tran, "Hierarchical Identify Verify Exploit," DARPA; www.darpa.mil/program/hierarchical-identify-verify-exploit.
9. J. Kepner and J. Gilbert, eds., *Graph Algorithms in the Language of Linear Algebra*, Soc. for Industrial and Applied Math., 2011.
10. K.E. Batcher, "Sorting Networks and Their Applications," *Proc. Am. Federation of Information Processing Soc.Spring Joint Computer Conf.* (AFIPS 68), 1968, pp. 307–314.

Erik P. DeBenedictis is a technical staff member at Sandia National Laboratories' Center for Computing Research. Contact him at epdeben@sandia.gov.

Mustafa Badaroglu is a staff program manager at Qualcomm. Contact him mustafab@qti.qualcomm.com.

An Chen is the executive director of SRC's Nanoelectronics Research Initiative. Contact him at chen@src.org.

Thomas M. Conte is a professor at Georgia Tech and the director of its Center for Research into Novel Computing Hierarchies (CRNCH). Contact him at conte@gatech.edu.

Paolo Gargini is the chairman of the International Roadmap of Devices and Systems. Contact him at paologargini1@gmail.com.