# Die Stacking Is Happening

**Xing Hu, Dylan Stow, and Yuan Xie**
University of California, Santa Barbara

After two decades of research effort, fine-pitched 3D integrated circuits are finally appearing in a growing range of industry products that leverage the benefits of high-bandwidth, high-density circuit integration. This article reflects on the historical development of these 3D technologies, their unique benefits over alternate 3D packaging methods, and the recent industry and research trends in 3D memories and 2.5D integration. Although die stacking is now happening, many potential benefits offered by these technologies remain to be explored, providing an opportunity for researchers and developers to solve these remaining challenges in architecture, methodology, and business.

Attendees of MICRO 2013 may still remember the keynote speech "Die Stacking Is Happening" by AMD's Bryan Black,[1] who predicted that die-stacking technology would soon be coming to mainstream products, emerging from its sole niche as an academic research curiosity. Less than two years after that keynote, Bryan's prediction became a reality: In June 2015, AMD released the world's first commercial GPU, named Fury X, with 4 Gbytes of integrated 3D-stacked high-bandwidth memory (HBM), leading to excitement in the market for both consumers and investors. More recently, AMD and Nvidia have announced their latest GPU architectures—Vega and Volta, respectively—which both integrate 16 Gbytes of HBM2, the second generation of 3D-stacked HBM. Indeed, die stacking is happening.

## WHAT IS DIE-STACKING TECHNOLOGY?

Die-stacking technology, also called *3D integrated circuit* (3D IC) technology, is the concept of vertically stacking multiple IC layers and then connecting them with fine-pitch vertical interconnections. 3D die-stacking integration technologies offer many benefits for IC designs, including:

- reduced interconnect wire length, which results in improved performance and reduced power consumption;[2]
- improved memory bandwidth through on-chip memory integration;
- enablement of heterogeneous process integration for optimized circuit tuning, novel architecture designs, and improved analog/mixed-signal circuitry;
- smaller form factors due to the addition of a third dimension to the conventional 2D layout, resulting in higher packaging density and smaller system footprint;
- potential for lower manufacturing costs and improved binning due to the yield improvement of smaller dies;[3] and
- opportunity for design reuse and system customization.[4]


Published by the IEEE Computer Society
0272-1732/18/$33.00©2018 IEEE

Consequently, 3D integration technology is a promising solution to overcome the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology. Recent research has explored several 3D stacking technologies, including wire bond, contactless (capacitive or inductive), and through-silicon via (TSV) vertical interconnects. Among these integration approaches, TSV-based 3D integration has the potential to offer the greatest vertical interconnect density, and therefore is the most promising choice among all the vertical interconnect technologies.

## Packaging Technologies

There are several categories of 3D packaging technology, as shown in Figure 1. System-in-package (SiP) methodology vertically stacks separate IC chips together on a substrate in a single package with internal wire-bond connections that are bonded to the package, whereas package-on-package (PoP) methodology stacks separate packages vertically through inter-substrate ball grid array (BGA) connections, as shown in Figure 1b. The Apple A8 processor (used in the iPhone 6) integrates a processor package (with dual-core CPU and quad-core GPU) with a 1 Gbyte LPDDR3 DRAM package using PoP technology. 3D chip stacking based on SiP or PoP are packaging-level technologies with the design goal of saving space, but the chips in the stack still communicate with long-distance off-chip signaling with reduced energy efficiency and longer latency. However, these techniques do not require a significant change in processor architecture or design methodologies. On the contrary, true 3D ICs, like the TSV-based system in Figure 1c, can provide a large number of short connections (and improved latency, bandwidth, and power) between two stacked dies, enabling revolutionary architecture innovations but also requiring fundamental changes to design methodologies.

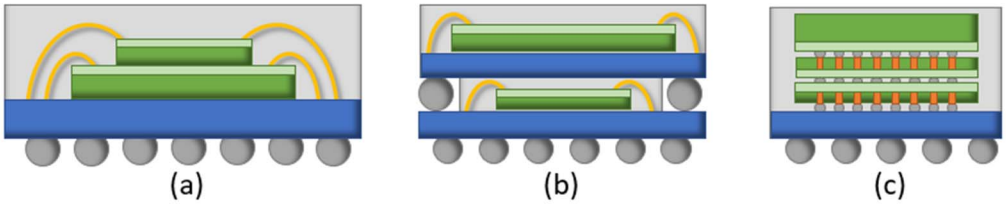


Figure 1. A comparison of 3D packaging technologies with varying vertical-integration densities: (a) 3D system-in-package (SiP) technology; (b) 3D package-on-package (PoP) technology; and (c) face-to-face and face-to-back 3D bonding with TSVs and microbumps.

## Stacking Strategies

With 3D integration technologies that reduce the wire length and provide high connectivity between layers, it's possible to partition the structure of a traditional planar processor and stack the resulting dies to improve system performance. There are two strategies for chip stacking:

- In *fine-granularity stacking*, functional units of the processor are separated across the stacked layers, and some components may be internally partitioned and implemented across layers to improve circuit density or to leverage heterogeneous processes. Such a strategy is frequently employed for "logic-on-logic" stacking.
- In *coarse-granularity stacking*, frequently embodied as the "memory + logic" strategy, some on-chip memories are separately partitioned and then stacked above the layers containing the logic components. Coarse-granularity stacking may also be leveraged for modular scaling of top-level units like core clusters and caches.

## EARLY DIE-STACKING ARCHITECTURE EFFORTS

3D integration technology has been an active research topic since the late 1990s. In particular, IBM was one of the pioneering companies to study the required process technologies for 3D integration between semiconductor dies. After many academic efforts, Intel later became one of the first companies to thoroughly explore the architecture of their microprocessors for die-stacking technology. They first demonstrated a logic-on-logic design in 2004, with a single-core processor partitioned across two layers.[5] The design eliminated 25 percent of all pipe stages in the microarchitecture simply due to the new 3D floorplan's reduction in metal routes, resulting in a net 15 percent improvement in performance. Three years later, in 2007, Intel demonstrated another design with memory-on-logic: a prototype two-layer many-core processor with 20 Mbytes of SRAM stacked on top of the 80-core layer. With memory stacking, the bandwidth between the memory and the logic layer could reach 1 Tbytes per second, which was far beyond what the state-of-the-art DDR3 could provide.[6] The research community was excited to see Intel's burgeoning efforts and looked forward to the commercial products that were expected to be available soon after.

## 3D-STACKED PROCESSOR: ARE WE THERE YET?

Despite the volume of active die-stacking architecture research, as well as Intel's successful demonstration with prototype 3D CPUs on both logic-on-logic in 2004 and memory-on-logic in 2007, there was no commercially available Intel microprocessor in the market even several years later. A 2010 *IEEE Micro* article attempted to solve the puzzle by investigating the challenges of making 3D-stacked processors.[7] The conclusion was interesting: Even though an emerging technology such as die stacking can be proven to be technically feasible and beneficial, it might not be adopted by product teams due to other challenges related to cost, risk, demand, or business decisions. For example, stacking DRAM on top of processor layers may result in thermal dissipation and voltage droop challenges, and customizing the DRAM dies for each processor design would introduce design and supply-chain complexity that would increase the overall cost. Additionally, marketable killer applications that can fully leverage the benefits of high inter-layer memory bandwidth are needed to justify the risks of transitioning to a die-stacked architecture.

Further, the momentum of business model decisions can also affect the adoption of radical new technologies. For example, in 2007, Intel adopted the "tick-tock" model to drive the development of their microprocessor designs, with every tick representing a shrinking of the previous microarchitecture's process technology, and every tock designating a new microarchitecture. If 3D technology is to be adopted, should it be used in a tick year or a tock year? It seems to be a technology change best suited for a tick year, but without architectural changes (in a tock year), the benefits cannot be fully realized, and the risk of transition cannot be justified.

## THE RISE OF 3D-STACKED MEMORY

While the 3D-stacked processor design with either logic-on-logic or memory-on-logic approaches remained in the prototyping stage, 3D integration technology caught the attention of the memory industry. Memory has a more regular structure and a simpler circuitry design than processors, and the 3D-stacked memory design, as a separate component, itself does not require close collaboration between the CPU and memory vendors. Further, memory systems require high bandwidth and consume significant portions of the system area, thus naturally benefiting from die-stacking technologies. Major memory vendors such as Micron, Hynix, and Samsung recently developed commercially available stacked DRAM memory technologies. For example, Micro announced its Hybrid Memory Cube (HMC) in 2011 and promised 15 times performance improvement over DDR3, winning the Best New Technology Award from *Microprocessor Report* magazine that year. Several stacked memory standards (HMC, HBM, and Wide I/O) were developed to target different markets with different application scenarios.

## TAKING A STEP BACK: FROM 3D TO 2.5D

Despite the benefits of true 3D-stacked processors, such as the system in Figure 2a, these designs are plagued with the challenges of thermal and power density and supply-chain complexity. To address these challenges, while also leveraging the benefits of thriving 3D-stacked memory designs (HMC, HBM, and Wide I/O), one possible solution is to step back and adopt an interposer-based 2.5D approach.[8] In such an approach, the design of the 3D-stacked DRAM and the design of logic die are decoupled. Memory vendors (such as Hynix) focus on designing the many-layered 3D-stacked DRAM with an industry standard (such as JEDEC's HBM), whereas the processor vendors (such as AMD or Nvidia) focus on the design of the logic dies. The 3D-stacked DRAM die and the CPU/GPU die are placed side by side on a silicon interposer, as shown in Figure 2b, which provides fine-grained interconnect for high-bandwidth, low-power integration.

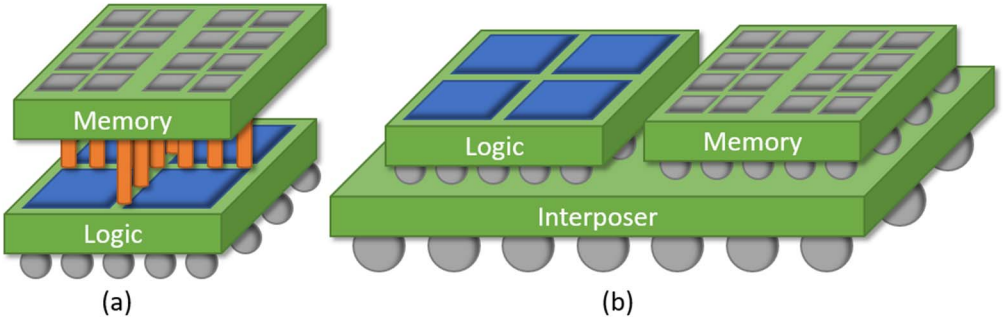


Figure 2. Memory and logic dies can be tightly integrated using vertical or horizontal approaches: (a) memory-on-logic 3D IC or (b) interposer-based 2.5D integration.

AMD became the pioneer to take this die-stacking architecture approach to mainstream computing, offering the world's first commercial GPU product with 3D stacked DRAM in 2015 (the Fury X GPU integrated 4 Gbytes of HBM). Since then, other processor companies have followed up with several variations for different application domains. Nvidia's Pascal/Volta GPU integrated 16 Gbytes of HBM2 for AI acceleration, and Intel's Knights Landing Xeon Phi CPU was packed with 16 Gbytes of Multi-Channel DRAM (MCDRAM), a variant of the HBM design targeted for high-performance computing applications. Beyond CPU/GPU architectures, in 2016, Xilinx also announced that the Virtex UltraScale+ FPGAs are also packaged with HBM. The integration of a large capacity of stacked DRAM with processors has also inspired the architecture community to investigate its usage.[9] For example, Intel Xeon Phi's MCDRAM can be used as a part of main memory in a NUMA fashion, or as a last-level cache, or somewhere in between.

AMD became the pioneer to take this die-stacking architecture approach to mainstream computing, offering the world's first commercial GPU product with 3D stacked DRAM in 2015.

## TECHNOLOGY VERSUS ARCHITECTURE: AN EVOLVING INTERACTION

Die-stacking technology was first investigated more than two decades ago and inspired architects to explore various possible processor architectures (such as fine-granularity logic-on-logic stacking, memory-on-logic stacking, and stacked-memory with logic on interposer), before

eventually becoming a mainstream architecture. Such an evolution reminds us of the classic paper by John Hennessy and Norm Jouppi: "Computer Technology and Architecture: An Evolving Interaction."[10] In it, the authors claimed that "the interaction between computer architecture and IC technology is complex and bidirectional." The characteristics of technologies affect decisions that architects make by influencing performance, cost, and other system attributes, and the developments in computer architecture also impact the viability of different technologies.

## TECHNOLOGY-DRIVEN VERSUS APPLICATION-DRIVEN ARCHITECTURE INNOVATION

The emergence of die-stacking architecture is essentially a technology-driven innovation. On the other hand, finding the killer applications that can leverage the technology's benefits is also important in such architectural innovations. Over the last decade, emerging workloads have evolved rapidly from the traditional desktop applications. In the cloud, the explosive volume of big-data applications and the demands of real-time analytics call for the capability to efficiently process large amounts of data. For example, large-scale graph processing and in-memory data analytics are memory-intensive tasks with simple computational intensity, which indicates that workloads are increasingly constrained by memory bandwidth rather than by compute capability. In the PC and gaming market, AR and VR systems demand the high performance of HBM to provide low-latency responses to satisfy strict user experience requirements (which is a contributing factor to the success of AMD's Fury X GPU). The demand will be even larger with the growing adoption of 4K resolution (and even higher future resolutions).

Finally, the recent renaissance in AI motivates architects to design better hardware accelerators for applications such as voice recognition, object detection, scene labeling, and autonomous vehicles, all of which require high memory bandwidth and large memory footprints to train and run deep neural networks. For example, Google's recent Tensor Processing Unit (TPU) paper specifically mentioned that the memory system was the bottleneck to limit the performance of their TPU design.[11] With the trend towards deeper, larger neural network algorithms and the development of compute-focused machine-learning accelerators, the memory system is becoming an increasingly critical bottleneck to the improvement of power and performance.

The recent renaissance in AI motivates architects to design better hardware accelerators for applications such as voice recognition, object detection, scene labeling, and autonomous vehicles.

Consequently, die-stacking architectures with high-bandwidth and large-capacity memories become attractive solutions for all AI hardware accelerators, thus motivating the usage of 3D-stacked memory. Google disclosed that their second-generation Tensor Processing Unit (TPU) accelerator integrates 16 Gbytes of HBM2 to achieve a peak memory bandwidth of 600 GB/s: a 20x improvement over the first-generation's 30 GB/s. Overall, Google includes 4 TB of HBM DRAM capacity in each pod during deployment.[12] Wave Computing, as another example, announced that their Dataflow Processing Unit (DPU) will include four HMC Gen2 interfaces, each equipped with 15 Gbps SerDes links, to provide 0.96 TB/s memory bandwidth and 8 GBytes of memory capacity in a single DPU, or 15 TB/s and 128 Gbytes per DPU cluster. Additionally, Intel's Nervana Neural Network Processor (NNP), which targets both neural network training and inference, employs 32 Gbytes of HBM2 on the shared interposer to achieve 1 TB/s of access bandwidth. The demanding requirements of machine-learning acceleration have created an ideal market for die-stacked memories, ensuring their continued growth and guiding the development of future memory technology standards.

3D integration technology will play an even more important role in future architecture design. In addition to high-volume CPU/GPU processor designs, which can now leverage die-stacking architectures to achieve high-bandwidth and large memory capacity, 3D/2.5 integration can also be

attractive for low-volume applications that desire effective heterogeneous integration and development cost reduction through IP reuse.[4] This goal has become the major motivation for DARPA's recent CHIPS program, which aims to reduce the design cost and effort of low-volume integrated systems through the development of a cross-compatible modular chiplet ecosystem.[13] System integration at the die level can leverage existing hardware across heterogeneous technologies, providing optimized processes for cores, memory, and analog, and can combine these unit–dies into novel custom systems with high efficiency and performance and low engineering overhead. Furthermore, known good die (KGD) testing of the constituent dies before bonding can significantly improve functional manufacturing yield and device binning,[3] further improving the resulting system's cost and efficiency and allowing for larger scale-out systems in a single package.

## CONCLUSION

Historically, both technology scaling and architectural innovation have played equally important roles in improving microprocessor performance.[14] However, as technology scaling slows down and more predictions about the end of Moore's law are made (with even Intel forced to change their model to "tick-tock-optimization"), traditional 2D shrinking becomes more difficult and less beneficial. Consequently, going vertical offers a new dimension of scalability in chip design, enabling the integration of more transistors in a single system despite an end of Moore's law.[15] In particular, looking beyond the current 2.5D integration and TSV-based 3D integration, monolithic 3D integration (fabricating transistors layer by layer on a single silicon substrate) with extremely fine vertical interconnections could open a rich new field of architectural possibilities.

## ACKNOWLEDGMENTS

We thank Gabriel Loh from AMD and Norm Jouppi from Google for their valuable feedback on this article.

## REFERENCES

1. B. Black, *Die Stacking Is Happening* (MICRO-46), 2013; microarch.org/micro46/files/keynote1.pdf.
2. Y. Xie, J. Cong, and S. Sapatnekar, *Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*, Springer, 2010.
3. D. Stow et al., "Cost-effective design of scalable high-performance systems using active and passive interposers," *IEEE/ACM International Conference on Computer-Aided Design* (ICCAD), 2017, pp. 728–735; ieeexplore.ieee.org/document/8203849/.
4. D. Stow et al., "Cost analysis and cost-driven IP reuse methodology for SoC design based on 2.5D/3D integration," *IEEE/ACM International Conference on Computer-Aided Design* (ICCAD), 2016, pp. 1–6; ieeexplore.ieee.org/document/7827633/.
5. B. Black et al., "3D processing technology and its impact on iA32 microprocessors," *IEEE International Conference on Computer Design* (ICCD), 2004; ieeexplore.ieee.org/document/1347939/.
6. S. Vangal et al., "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS," *IEEE International Solid-State Circuits Conference* (ISSCC), 2007; ieeexplore.ieee.org/document/4242283/.
7. G.H. Loh and Y. Xie, "3D Stacked Microprocessor: Are We There Yet?," *IEEE Micro*, vol. 30, no. 3, 2010, pp. 60–64; computer.org/csdl/mags/mi/2010/03/mmi2010030060-abs.html.
8. X. Dong et al., "Simple but Effective Heterogeneous Main Memory with On-Chip Memory Controller Support," *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis* (SC '10), 2010, pp. 1–11; ieeexplore.ieee.org/document/5645467/.

9. Y Xie and J Zhao, *Die-stacking Architecture*, Morgan & Claypool Publishers, 2015.
10. J.L. Hennessy and N.P. Jouppi, "Computer Technology and Architecture: An Evolving Interaction," *Computer*, vol. 24, no. 9, 1991, pp. 18–29; computer.org/csdl/mags/co/1991/09/r9018-abs.html.
11. N.P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," *ACM/IEEE 44th Annual International Symposium on Computer Architecture* (ISCA), 2017, pp. 1–12; computer.org/csdl/proceedings/isca/2017/4892/00/08192463-abs.html.
12. J. Dean, "Recent Advances in Artificial Intelligence via Machine Learning and the Implications for Computer System Design," *Hot Chips: A Symposium on High Performance Chips*, 2017.
13. D. Green, *Common Heterogeneous Integration and IP Reuse Strategies (CHIPS)*, DARPA; darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies.
14. A. Danowitz et al., "CPU DB: Recording Microprocessor History," *ACM Queue*, vol. 10, no. 4, 2012; queue.acm.org/detail.cfm?id=2181798.
15. L. Ceze, M.D. Hill, and T.F. Wenisch, *Arch2030: A Vision of Computer Architecture Research over the Next 15 Years*, Computing Community Consortium, 2016; cra.org/ccc/wp-content/uploads/sites/2/2016/12/15447-CCC-ARCH-2030-report-v3-1-1.pdf.

## ABOUT THE AUTHORS

**Xing Hu** is a postdoc researcher at the University of California, Santa Barbara. Contact her at xinghu.cs@gmail.com**.**

**Dylan Stow** is a PhD student at the University of California, Santa Barbara. Contact him at dstow@ucsb.edu.

**Yuan Xie** is a professor in the Department of Electrical and Computer Engineering at the University of California, Santa Barbara. Contact him at yuanxie@ece.ucsb.edu or www.ece.ucsb.edu/~yuanxie.