# 3D Stacked Microprocessor: Are We There Yet?

**GABRIEL H. LOH**
Georgia Institute of Technology

**YUAN XIE**
Pennsylvania State University

Three-dimensional integration has received considerable attention in the last several years from academic researchers and industry alike. This technology provides multiple layers of devices connected by a high-density, low-latency, layer-to-layer interface that can enable integrated circuits with more devices per unit area and allow the integration of different types of devices within the same 3D chip stack. Academic and industrial researchers continue to explore new ways to exploit 3D integration, but can we really expect to see commercial 3D-integrated microprocessors anytime soon? Is 3D integration yet another promising (or perhaps just overhyped) technology that will eventually and quietly go away?

To try to answer these and related questions about the future of 3D integration technology, we surveyed a variety of researchers, executives, and other technological leaders from a range of institutions, including major semiconductor companies, government agencies, and industry consortia. (Most respondents answered our questions on condition of anonymity, and some chose not to reply at all due to concerns over confidentiality and exposure of proprietary information.) Their responses provide a view of where 3D integration technology for microprocessors currently stands, where it is heading, and what challenges still need to be addressed.

## State of 3D technology

For some domains (such as cell phones and memory), 3D integration in various forms is already in use. For example, many modern cell phone devices use wire-bonded 3D stacking (sometimes referred to as *system-in-package* [SiP], see Figure 1a) or *package-on-package* (PoP) 3D stacking. As Samsung, Tezzaron, and a few other companies have demonstrated, industry has reached the consensus that stacked memory will become mainstream. In this article, we focus on 3D stacking technology based on through-silicon-via (TSV) technology (see Figure 1b), which provides much faster and higher density inter-die connections than SiP or PoP.

The first question that many people are interested in is simply when TSV-based 3D integration technology will be ready for high-volume, mass-market manufacturing in the microprocessor market. Answers from those surveyed varied from the end of 2011 to somewhere in the 2012–2013 timeframe, with many companies currently engaged in serious efforts to develop it into a commercially viable technology. The nearly unanimous consensus was that the first commercial 3D integrated microprocessor products will feature memory stacked on a processor, with the following main benefits:

- a massive increase in bandwidth from using TSVs instead of conventional I/O pins,
- the increase in on-chip storage capacity enabled by 3D integration's ability to mix process technologies (that is, DRAM on CMOS), and
- power savings from reducing the need for the interconnect and I/O drivers associated with off-chip memories.

### Editors' Note

We live in a 3D world. It is hard to imagine a large city, such as New York City, with only single-level structures. There would be no skyscrapers, no mixed-use, no live-work. It would be a long walk (or drive) between everything, especially between dissimilar uses—all in all, very inefficient!

Integrated circuits today are typically designed using single-level Manhattan geometries, nothing like the layout of the real city. In this prolegomenon, Gabriel Loh and Yuan Xie survey 3D integrated circuit technology, demonstrating the virtues, potentials, and challenges of applying three dimensions to future microprocessor designs and exploiting the locality and diversity of real-world Manhattan geometries.



Published by the IEEE Computer Society

A few challenges remain, however, before such a memory-stacked microprocessor becomes a reality. On the technical side, the success of such an architecture will depend on effective thermal management to reduce the impact of heat dissipation from high-power-density processor cores on the leakage-sensitive stacked DRAM. Placing the power-hungry core layer closer to the heatsink could help alleviate the problem. However, such an arrangement implies that the memory layer should be customized for massive TSVs that connect the core layers to the I/Os, which suggests the need for a close collaboration between the processor vendor and the memory vendor, or the creation of an industry-wide processor-memory interface standard.

The collective outlook for 3D integration is that it is coming soon, but various engineering and other issues still need to be worked out. In the short term, we need design tools and electronic design automation to support 3D-integrated designs. In the longer term, designing these tools is a significant challenge that will require heavy investment. However, the immediate needs might be satisfied by evolutionary changes to existing tools because initial 3D-integrated microprocessor products will only use simple topologies such as monolithic memories stacked on conventional processor cores. Even for a conceptually simple application such as 3D memory-on-processor, standards will likely need to be developed. Without standardization, each processor company must partner with a memory vendor and agree to their own processor-memory interface. This situation would leave both parties vulnerable since the processor company's ability to manufacture chips would depend on the memory vendor's ability to deliver a sufficient volume of memory chips, and the memory vendor would depend on the processor company's ability to sell the memory-processor stacked chips. An industry standard processor-memory interface would let the processor companies purchase stackable memory chips from any vendor, and the memory vendors sell their chips to any processor company (or any other IC company needing 3D-stacked memory).

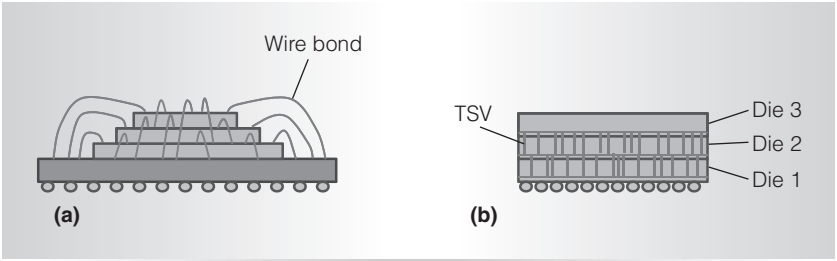


Figure 1. Current state of 3D technology: wire-bonded 3D stacking, or 3D system-in-package (a), and 3D integrated chip using through-silicon-via (TSV) technology (b).

At this point, the exact timing of the adoption of 3D integration for high-volume manufacturing is more a business than a technical decision. The design cycle of a modern high-performance microprocessor typically lasts three to five years, and the risk factors associated with switching to 3D integration differ from those of other past changes to process technology.

For example, Intel follows a ''Tick-Tock'' business model, where the company delivers a new silicon process technology one year and a new processor microarchitecture the next year to drive technology innovation on a reliable and predictable timeline. The introduction of 3D-stacked memory may likely be both a ''tick'' and a ''tock'' as it is a change in the manufacturing technology and may require a new microarchitecture to exploit the new memory interface.

Another risk factor is that the introduction of 3D is an all-or-nothing bet. Compare it to the switch from aluminum interconnects to copper. If the fabrication process engineers had discovered a problem at the last minute, the adoption of copper to the next technology node would have been delayed, but a processor that had been designed with the expectation of copper wiring would still have worked. The cost would simply be that the final product might have achieved a lower clock frequency than initially projected with a copper-based solution. In the case of 3D integration, product planning occurs three to five years before manufacturing, and if the engineers decide to go with 3D integration, it had better be ready when the time comes to manufacture the chip.

There is a question of overall cost to the companies to equip their fabs to support a 3D process, as well as the recurring additional cost-per-die to manufacture each chip due to the required additional processing steps. On the other side of the equation, it still is not clear how much the end user will ultimately be willing to pay for the additional value 3D integration delivers in terms of performance improvement, power benefits, and overall user experience. Other concerns include manufacturing issues related to scaling the fabrication process to high volume, reliability of the 3D-stacked chips in terms of thermals/cooling and power delivery, ensuring high yield, and testing/debugging. The general opinion was that although these issues need to be worked out, there are no fundamental obstacles. Once industry decides that the timing is right for 3D integration, the necessary resources will be invested to accelerate the engineering efforts to overcome any remaining issues.

## Looking forward

Near-term 3D-stacked microprocessors will likely comprise two to five layers of silicon (that is, one processor layer plus one to four layers of memory). Beyond this, the predictions for how far 3D integration can be pushed vary

considerably. At the low end, some respondents believed that no more than four layers would be seen anytime soon, while at the more aggressive end, respondents predicted eight, 12, and even 16 layers. Some respondents observed that as far as technical feasibility, stacks of as many as 100 chips could be possible, but the real question is what one would actually do with that many layers since a usage model for that many layers currently does not exist. A comparison was also drawn with metal routing layers in conventional integrated processors, where adding more layers of metal beyond the existing 10 to 12 rapidly diminishes benefits. Without some radical new 3D architectures, stacking more than a certain number of chips will likely not deliver enough additional value to justify the added expense.

As 3D integration technology matures, increasingly heterogeneous mixes of technologies will open up new and exciting system design possibilities. Beyond the simple combination of memory and processor cores, one could also integrate on-stack photonics, RF devices, power regulators, mixed-signal processing, nonvolatile memory technologies, charge-coupled device (CCD) arrays, and other technologies into the 3D stack. Even just considering CMOS logic processes, different layers could be implemented in different technology generations to provide a range of device characteristics with different performance, leakage power, and reliability tradeoffs. One respondent suggested that standardizing a common die size and TSV interface between layers would let system designers easily exploit these different technologies. They could then design modular IP blocks to conform to this standard, and assemble systems by choosing a mix of IP blocks to stack together. Many nonstandard process technologies could then be made widely available.

Stacking multiple layers of chips together effectively increases the number of devices per unit area. Combining this stacking with the conventional scaling of device feature sizes, 3D integration could accelerate the Moore's law curve to more than double the device density per generation. Most respondents, however, believed that as device scaling becomes more difficult and the end of the silicon roadmap approaches, 3D stacking will more likely be a means of simply staying on track. Even just using 3D integration to keep Moore's law going for a few more generations is not without challenges. Increasing the number of transistors by stacking is fundamentally different from increasing the number of transistors by shrinking device geometries. For a given feature size, the power per transistor is the same whether or not you use 3D stacking. That is, in the same technology node, using 3D to double the number of transistors per unit area also doubles the amount of power consumed per unit area (assuming identical activity factors for both layers). Similarly, even though stacking two chips doubles the effective device density, the manufacturer must still pay for twice the total silicon area, so 3D does not provide the conventional economic scaling (that is, an approximate halving of the price per device per generation).

One pertinent observation was that at the end of the day, consumers do not really care about Moore's law and the absolute number of transistors in the products they purchase. The end user expectation is simply that each successive technology generation provides continued scaling of value (for example, performance per dollar for processors and bits per dollar for DRAM), and this added value is ultimately what the consumer is willing to pay for. One respondent noted that many past technology enhancements (such as copper interconnects, low-k dielectrics, and metal gates) did not explicitly help increase device density. However, they were nevertheless critical to ensuring that the additional device density was sufficiently useful so that products manufactured in each successive technology generation continued to provide the expected scaling of value to the end user. The commercial success of 3D integration will also hinge on whether industry can translate the technology's potential benefits (mixed technology integration, extreme layer-to-layer bandwidth, reduction of interconnect power, and so on) into actual benefits to the consumer.

## Opportunities for innovation

Based on our respondents' many comments, it appears that the near-term plan for 3D microprocessors is reasonably well understood and that these microprocessors will likely be commercialized within a few years. What happens after the first few iterations of memory-on-processor 3D stacks is an open question, and the respondents strongly encouraged researchers to explore and invent innovative processor and system architectures that can exploit this technology. One simple observation is that today's processor designs cannot come close to using all of the available bandwidth that a TSV-based 3D-stacked memory would provide. There is a wide-open space of new processor and overall system architectures that could be devised to better exploit this massive memory bandwidth. A few respondents even suggested enabling extreme compute density by building aggressive processor architectures using the 3D stacking of multiple logic/CMOS layers. On the other side, many opportunities exist to create new memory organizations, structures, and interfaces to better match the processor needs as well as power and thermal constraints. Although radically new architectures would present too much risk for early-generation 3D processor designs, once 3D integration hits high volume, the tools and other design support will rapidly develop, thereby making more aggressive 3D architectures much less risky and more feasible.

In many ways, the continued inheritance of 3D integration across technology generations will be necessary to amortize the development costs of 3D CAD tools, testing methodologies and

debug equipment, standards setting, advanced packaging, and everything that will be needed in a 3D integrated ecosystem. All of this speaks to the likely permanent adoption of 3D integration, but it is predicated on whether processor architects and system designers will be able to transform these opportunities into real products that people will buy. Finding innovative ways to provide the mapping from potential to reality is a huge opportunity, and should be a high priority for processor architects and systems researchers alike.

**Gabriel Loh** is an assistant professor in the College of Computing at the Georgia Institute of Technology. His research interests include computer architecture, processor microarchitecture, emerging technologies and 3D die stacking. Loh has a PhD in computer science from Yale University. He is a senior member of IEEE and the ACM.

**Yuan Xie** is an associate professor in the Computer Science and Engineering Department at Pennsylvania State University. His research interests include computer architecture, electronic design automation, VLSI design, and embedded system design. Xie has a PhD in electrical engineering from Princeton University. He is a member of IEEE and the ACM.

Direct questions or comments about this article to Gabriel Loh, loh@cc.gatech.edu.