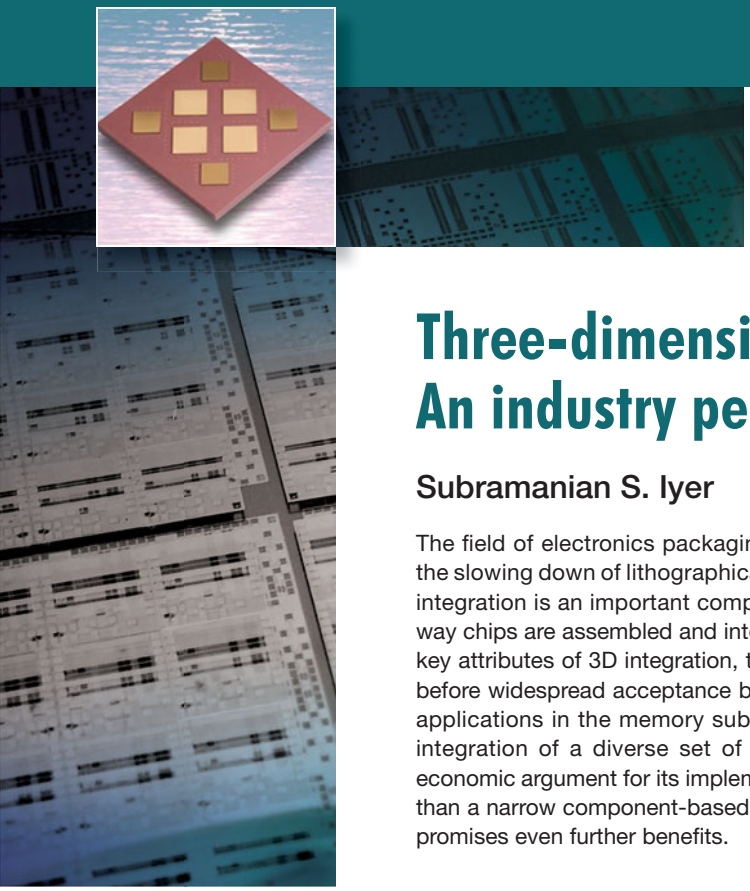


Three-dimensional integration: An industry perspective

Subramanian S. Iyer

The field of electronics packaging is undergoing a significant transition to accommodate the slowing down of lithographically driven semiconductor scaling. Three-dimensional (3D) integration is an important component of this transition and promises to revolutionize the way chips are assembled and interconnected in a subsystem. In this article, we develop the key attributes of 3D integration, the enablers and the challenges that need to be overcome before widespread acceptance by industry. While we are already seeing the proliferation of applications in the memory subsystem, the best is yet to come with the heterogeneous integration of a diverse set of technologies, the mixing of lithographic nodes and an economic argument for its implementation based on overall system function, and cost rather than a narrow component-based analysis. Finally, an extension to monolithic 3D integration promises even further benefits.

Introduction

A combination of lithographic scaling, device scaling,* and the introduction of innovative materials has produced an incredibly sustained gain in silicon productivity as measured by the cost per transistor and power-performance per transistor, and thus, cost per function. This is often popularly interpreted as Moore's Law. More recently, indications show this trend is slowing down. For example, as pointed out by Or-Bach,[1] so-called 14 nm static random access memory cell sizes are approximately 60% larger than what would be predicted by straightforward scaling—although the move to vertically oriented fin-shaped field-effect transistors has provided a significant one-time bump in many of these metrics. It is important to point out that it is not physics, materials, or even technology that is preventing further scaling but rather the effective manufacturing cost per transistor. This is driven primarily by an ever-increasing bill for manufacturing equipment, complexity (especially in lithography and patterning), and the amortization of the cost of development. Non-recurring engineering is another component of cost that includes chip design under restrictive rules, change orders, and mask costs, which are also increasing at unsustainable rates.

However, there are other aspects of system scaling that can come into play. We turn our attention to the package and the board. Interestingly, referring to **Figure 1**, while silicon has scaled by about a factor of 1000 since the 1970s, almost every key parameter in the package and the board has scaled by a mere factor of three to four during the same time frame. There are sound reasons for this neglect in scaling packaging features. Silicon scaling was delivering on cost, performance, and density in a very predictable manner (see the Introductory article in this issue). As a result, more circuitry could be assembled on a single die, leading to the system-on-a-chip (SoC) era. This allowed for more functional integration; consequently, different functional blocks communicated on the same die without needing to go off-chip. This meant there was more return on investment to scale silicon versus improving the package and board level interconnects, allowing systems to operate with the lowest latency and highest bandwidth.†

However, this on-chip interconnection comes at a cost: The die sizes were getting larger and the processes were getting more complex, both of which result in reduced yield.

*A recent retrospective of Dennard's scaling theory by M. Bohr is available at http://www.eng.auburn.edu/~agrawvd/COURSE/READING/LOWP/Boh07.pdf and also in the January 2007 issue of the *IEEE Journal of Solid-State Circuits Society*.

†Latency refers to the time it takes for data to reach its destination after being requested, a measure of delay that includes logic operations and transit time, while bandwidth refers to how much data may be transferred in a unit of time—a measure of how many interconnects exist between two system nodes.

Subramanian S. Iyer, Microelectronics Division, IBM Systems and Technology Group, USA; ssiyer@us.ibm.com
DOI: 10.1557/mrs.2015.32

© 2015 Materials Research Society

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

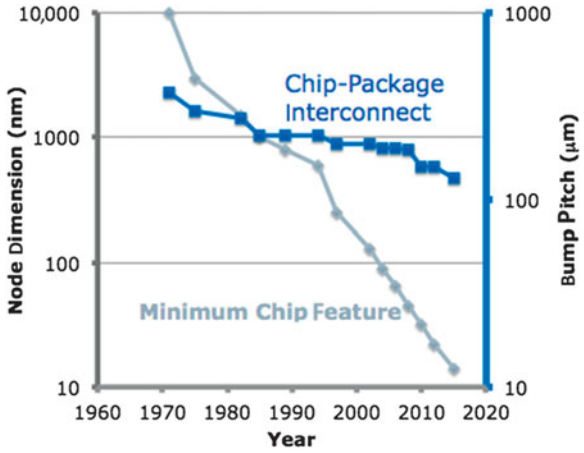


Figure 1. Scaling of minimum chip feature (~1000x) in silicon (left axis) as a function of time, and the scaling of package features (~4x) in the same time frame (right axis). We have used chip-to-package interconnects (bump pitch) in this example, but the results hold true for other package and board-related features as well.

Some parts of the chip do not scale gracefully, for example, the analog (i.e., non-digital) content on a digital chip. To drive off-chip loads that may increase with system complexity, the current drive capability of these devices will need to stay constant or even increase. In other words, analog content, such as input/output (I/O) drivers and radio frequency devices, does not scale. As we migrate to smaller nodes (32 nm to 14 nm), the analog content becomes a larger fraction of the SoC die. To make matters worse, scaling optimizes for digital devices, and the analog devices are suboptimized. As a result, there is no compelling argument to include analog content on an advanced node SoC—provided one can integrate it effectively in some other manner.

Memory integration in systems

One can make similar arguments for other specialized SoC components. Embedded memory is an important example, as sometimes over 60% of the SoC is memory.[2] As clock frequencies have risen, system performance is often gated by the so-called memory wall.[3] Including larger caches allows us to minimize calls to the main memory.[4] The most obvious approach is to use the densest memory available—embedded dynamic random-access memory (eDRAM), which was the approach used in the first generation of the BlueGene systems (class of high performance server).[5] However, that approach could not be used in IBM's higher performance P and Z servers,[6] as the processor (silicon-on-insulator, SOI) and eDRAM (bulk) substrate and technologies were different. These were integrated on either a ceramic or organic multichip module (MCM), as shown in **Figure 2** ("Then"). However, it was not possible to exploit the performance of the eDRAM, as the latencies on the MCM dominated. This issue was addressed by integrating the eDRAM in the same technology as the processor chip, an excellent example of the SoC approach,[7] as shown in Figure 2 ("Now").

However, as technology nodes advance and become more complex, there is a price in overall process complexity. For example, the logic on the chip is burdened by unique processing which DRAM requires—in this case, primarily the deep trench module (used to fabricate the high density capacitor), while DRAM on the chip is burdened by the cost of multiple threshold devices and about six to eight extra wiring levels that it may not need.

Current limitations of integrating multiple chips with low latency and high bandwidth

The main reason to integrate on-chip is to minimize latency and increase bandwidth between different circuit blocks, such as the cache discussed in the previous section. A typical system integrates several SoC chips on a printed circuit board. In principle, this is already being done with one important caveat: The latencies and bandwidths available on the board are significantly worse than what is available on-chip, primarily due to the lack of scaling of the package and board.

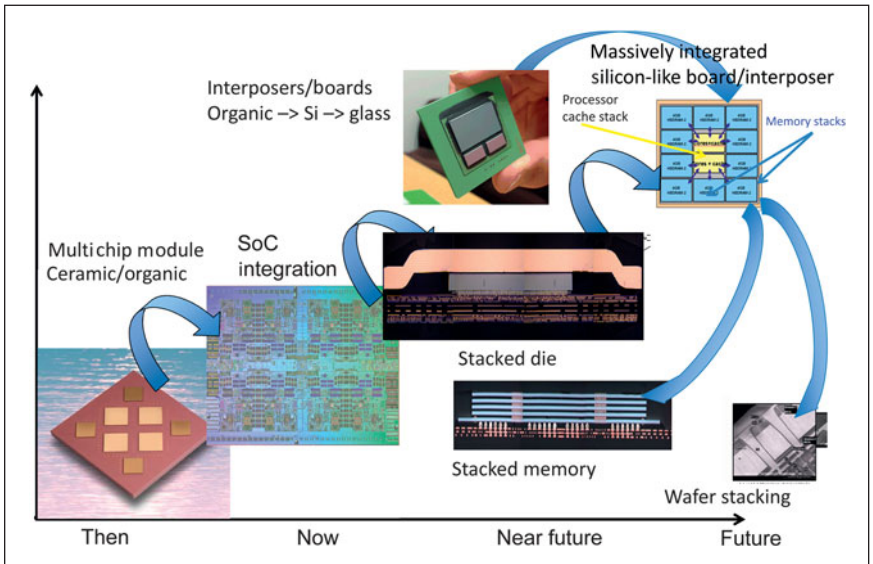


Figure 2. The evolution of the memory subsystem in high-performance processors. The move from multichip modules ("Then") to a single chip solution ("Now") resulted in a ~7x reduction in footprint, decreased latency, and increased bandwidth. Further reduction in these parameters is possible using die stacking and interposers ("Near future"), and finally, in the "Future," all of these may be integrated to produce a compact module on an interposer that can exceed the expectations of Moore's Law. The top inset shows a silicon-like interposer, and the bottom inset shows a stacked memory. The massively integrated board/interposer is shown with a processor cache stack (yellow) surrounded by tightly connected memory stacks (blue). The stacked memory is courtesy of IBM Corp. and Micron Technology, Inc., and the interposer is courtesy of IBM Corp. and Semtech Corp.

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

The challenge we have going forward is to scale the footprint of the package and the board cost effectively, which has been accomplished with silicon in the last several decades, such that we can continue to pack more function per unit volume and reduce cost per function in a manner that adheres to historical rates. There are three important components needed to make this happen: (1) design infrastructure, (2) technology, and (3) testability and reparability. This article will focus on technology because of the materials implications, but we emphasize that items (1) and (3) are just as important. From a technology perspective, we will examine two complementary approaches: 3D stacking at both the die and wafer level, and implementing interposers that allow us to integrate a few to several dies with high bandwidth. Interposers are not conceptually different from the MCM concept described in Figure 1, but technological innovations allow for lower latency and higher bandwidth.

Technologies for 3D die stacking and interposers

Three-dimensional stacking of, for example, a cache and processor chip allows us to accomplish what is shown in the "Near future" section of Figure 2. This requires the development of through-silicon via (TSV) technology (see the Jiang et al. article in this issue). In the two die high face-to-face stack shown in **Figure 3**, the only way for either die to communicate with the outside world is through these TSVs. There is another scheme that stacks dies face-to-back. Typically all dies are mounted face (device side) down toward the laminate. In this case, connections from the bottom-most die (only) to the laminate do not need to go through TSVs. This method is used when multiple die stacks are needed or when it is difficult for off-chip drivers to go through TSVs due to some design constraint.

TSV technology borrows heavily from the micromachining world, and the Bosch process[8] is used almost exclusively to etch these vias typically as blind holes into the silicon.

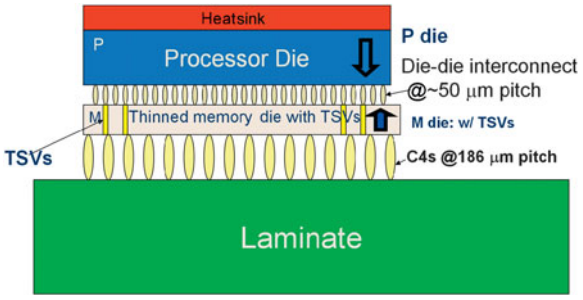


Figure 3. A prototypical two-high stack in a face-to-face configuration. Such a scheme would allow for very low latency and high bandwidth connection between two high-performance dies. All connections to the package go through through-silicon vias (TSVs). While there is excellent interconnectivity between the two dies, connectivity to the outside world is constrained by the laminate. The arrowheads indicate the device side of the die. C4 stands for controlled collapse chip connection. The M die in this case refers to a cache die, while the P die refers to a processor die.

Sides of the via are insulated by a deposited oxide layer whose thickness dictates the low frequency capacitance of the via (at very high frequencies, the capacitance drops significantly, as the free carrier relaxation in the silicon dominates). Typical thicknesses used are in the range of 0.5–1 μm to ensure adequate coverage, especially near the bottom of the TSV. This is followed by the deposition of a barrier layer (typically TaN) and a copper seed layer. The TSV is then filled with copper by electroplating. The aspect ratio of the TSV is limited by the capabilities of these different processes to approximately 5–10. TSV diameters have shrunk from about 25 μm to about 5 μm over the last few years. This results in TSV depths of approximately 100 μm to 50 μm.

The TSV depth is limited by the thinnest wafer manageable, usually >30–50 μm. Most TSV filling processes are similar to those used in the backend interconnects in silicon technology, with one notable exception. Conventional plating of a via results in Cu plating on the top surface, referred to as overburden. Typically, this is about half the via diameter. This overburden needs to be polished off, which is an expensive process. So-called bottom-up plating processes have been developed specifically for TSVs to minimize the overburden. The liners and seeds also need to be significantly thicker than in conventional wiring levels to minimize discontinuities and pinholes. Discussions on other material, design, and reliability considerations of TSVs will be included after details on the stacking process have been discussed.

Figure 4 shows a schematic flow of the die stacking process. Briefly, the wafers exit the conventional fabrication process and are then thinned from the back side by a combination of mechanical thinning, polishing, and etching to expose the metal in the TSV. Usually, the thinning is done at the wafer level with a support wafer for handling as the wafer is thinned. Some authors have reported reconstituting the wafer on another support wafer from a known good die.[9] After exposure, the supported thinned wafer is further processed to make connections to the exposed TSVs as well as to add additional wiring levels for signal and power redistribution. These layers are called redistribution layers, which provide valuable connectivity, and through judicious design optimization, they relieve wiring congestion on the device side of the die.

The thinned die is then mounted on a laminate (or interposer). The challenge here is doing this reliably with thinned dies, which are invariably warped. A number of schemes to manage this warpage have been developed and usually involve warpage control or compliancy along with some form of rigid support during the assembly process. Warpage is highly temperature dependent and arises because of the heterogeneous materials used in silicon, interposers, and packages, which expand at different rates. Warpage is the biggest detractor in all die-to-die, die-to-interposer, and die-to-package assembly operations. In the case of face-to-back assembly, it may be necessary to complete an additional flip of the die/wafer prior to mounting. This thinned die mounted on a laminate is often called the subassembly.

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

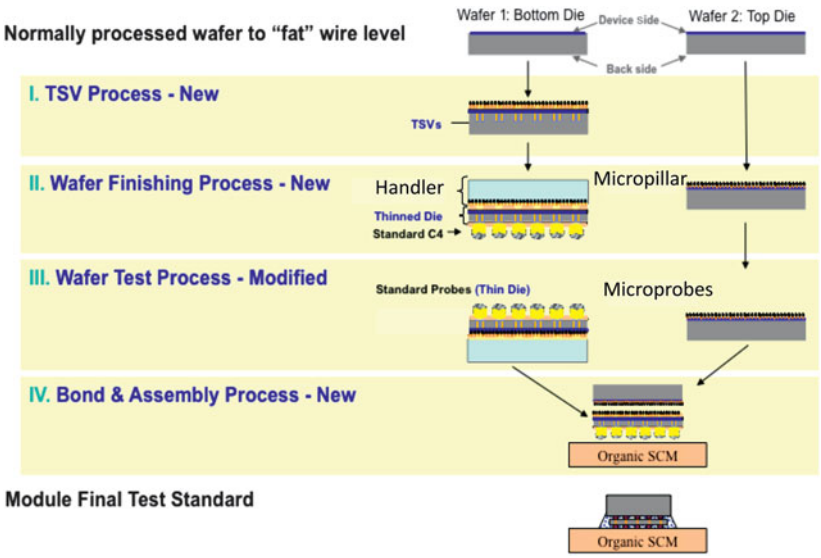


Figure 4. Schematic of wafer finishing and assembly process for a two-high face-to-face stacked die process. Through-silicon vias (TSVs) are introduced in the upper layers of the wiring hierarchy to minimize process disruptions. The top die is unmodified. Only the bottom die has TSVs. The wafer finishing process involves supporting the wafer on a stiff handler such as a silicon wafer, thinning and revealing the TSVs without contaminating the silicon with Cu. Additional wiring levels may be fabricated on the grind side to provide additional wirability, power distribution, and passives such as inductors. New test points may be introduced to guarantee functionality, and the assembly process consists of first attaching the thinned die to the single chip module (SCM) laminate—often called the subassembly—and then mounting the unthinned die to the subassembly followed by the module final test and burn-in, as appropriate. Note: C4, controlled collapse chip connection.

At this point, additional dies are assembled onto the first die serially. In the case of a two-high stack, this die is relatively thick and mounted face-to-face. In the case of face-to-back assembly of multiple dies, the dies are individually assembled onto the subassembly. At multiple points during the assembly process, testing is performed to ensure the integrity of the process. Many things can go wrong. The primary defect is the lack of good connectivity between the dies, usually because of a poor inter-die connection.

The inter-die connection is variously referred to as a controlled collapse chip connection (C4),[10] bump, or pillar. It is perhaps the most important part of any 3D die assembly process, and shrinking it is what makes 3D a valuable proposition. Smaller bumps allow us to increase bandwidth between dies and design a more uniform power and clocking grid. **Figure 5** shows the evolution of this interconnect pitch. Large bumps are joined primarily with a low-temperature melting solder—a process called mass reflow. In recent years, environmental considerations have eliminated the use of lead, and the solders are primarily tin-based. These require higher temperatures to flow and are more brittle than Pb-based solders.

As the bump sizes are reduced, the bump structure evolves into what is often referred to as a pillar, which is composed of a copper standoff pillar topped with solder. The amount of solder is reduced primarily because there is less space laterally for the solder to squeeze out, and there is danger in electrical shorting to neighboring pillars. Since there is much less mass flow, the joining process relies more on thermal compression bonding (TCB).[11] TCB processes require physical proximity between the parts to be joined, significantly more height uniformity of the bumps, higher-pressure contact (i.e., compression), higher temperature, and longer time to complete the bond. It is currently a significant detractor in both time and yield in a manufacturing environment and is thus an active area of research. Planarity of the different components (dies, laminates) is also crucial, but the application of a compliant force can force contact. Even so, warpage is a major concern, as was mentioned earlier.

An example of multi-die stacking face-to-back is shown in the inset second from the left ("Now") in Figure 2. Multiple DRAM strata are assembled sequentially onto logic strata on an organic laminate. The interconnect pitch between the silicon strata is about 50 μm, while the final pitch to the laminate is at approximately 150 μm. Currently, four to eight DRAM layers can be assembled and, in principle, this can go up to 16 layers. The advantage of this configuration is that instead of the main processor communicating separately with each DRAM chip (more independent connections mean more power), multiple connections may be replaced by a single connection to the base logic chip. The base logic communicates with the stack using TSVs, which can be done by straightforward signaling similar to on-chip communication. This can reduce the power budget by almost one-third and, equally important, reduce the footprint of the DRAM by over an order of magnitude. A number of protocols are being developed for stacking and communication, and the reader may consult the references for more detailed information.[12,13]

Returning to the issue of TSV design and placement, **Figure 6** shows different configurations for the introduction of TSVs in a typical hierarchical wiring system used in modern integrated circuit technology.[‡] TSVs are some of the largest features in the chip, and their integration into a silicon process flow can be quite disruptive, especially when one

[‡]Hierarchical wiring refers to a progressive increase in both wiring pitch and thickness while moving away from the silicon surface. The levels near the silicon are fine pitch and used for local interconnects. While moving up the hierarchy, one can wire longer distances. Finally, at the uppermost levels, the pitch and thickness are greatest, and these levels are typically used for power bussing, I/O to the chip, and the clock distribution mesh.

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

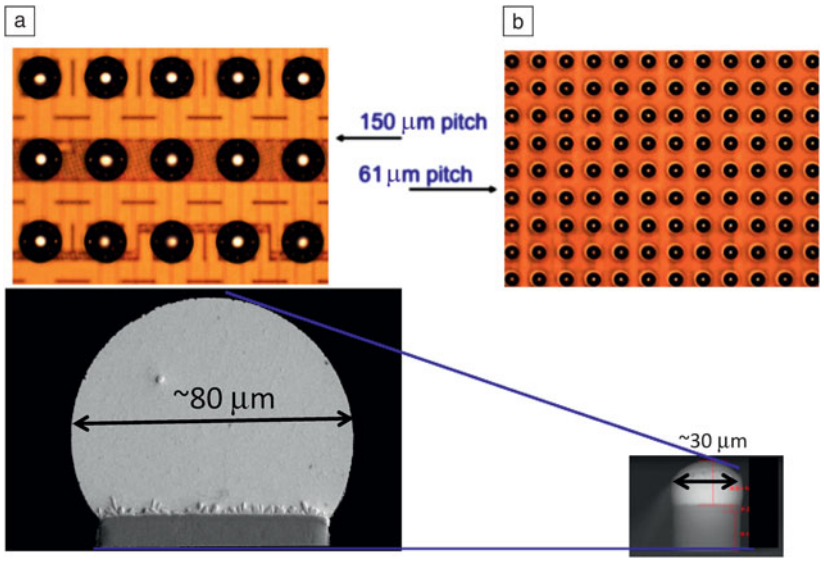


Figure 5. The evolution of bump pitch and structure from (a) ~200 μm pitch solder-dominated controlled collapse chip connection technology (mass reflow) to (b) 50 μm pitch Cu pillar-based technology that requires thermal compression bonding.

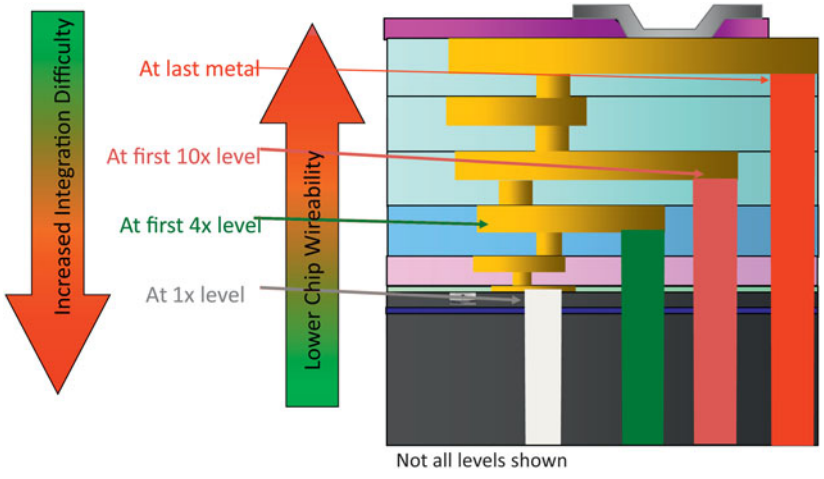


Figure 6. Through-silicon vias (TSVs) may be introduced in a hierarchical wiring system. At lower levels, there is less wiring blockage but more integration challenges. Today's relatively large TSVs are optimally integrated at wiring levels where dimensional discrepancies between the wiring level and the TSV are minimal. A couple (1–2) of wiring levels above the TSV can relieve a significant amount of wiring congestion.

integrates this with finer pitch wiring. It becomes easier to integrate TSVs at higher levels of the hierarchy, where the dimensions are comparable. Also, as we will see later, with the pitches used today, we can only connect 3D dies at the power and I/O levels, so integrating at the higher levels makes more sense. However, since TSVs go right through the back end and silicon, they block wiring and can create routing problems. The right compromise is to integrate TSVs at as high of a level as possible that provides acceptable routing —usually the penultimate wiring level.

Another consideration is the potential mechanical stresses that copper-filled TSVs can exert on silicon (see article by Jiang et al.). These stresses can perturb channel mobilities,[14] and a keep-out zone where devices may not be placed is usually defined to avoid these problems. Finally, there is a significant mismatch in the thermal expansion coefficient between Si and Cu in the TSV. This can cause pistoning of the copper in the TSV during thermal cycles that the chip sees. This effect can be minimized by a combination of appropriate microstructure, additives to the electroplating bath, TSV geometry, and the integration scheme. Today, cost-effective fine pitch (<10 μm) and reliable TSVs are routinely manufactured.

Interposers

Interposers represent a special case of 3D integration. The concept of an interposer goes back several decades and was first introduced in the 1960s to interconnect small numbers of bipolar transistors when backend interconnects had not matured.[15] This particular embodiment was called Solid Logic Technology and has since evolved into a ceramic MCM, where a single ceramic-based interconnect layer makes way for a multilayer ceramic stack where multiple dies could be assembled and interconnected.

This, in turn, has given way to organic MCMs. The goal of an interposer is first and foremost to connect multiple dies in a manner that allows for the lowest latency and the highest bandwidth. This translates to the tightest interconnect pitch between the die and the interposer, and the maximum allowed connections between chips. The latter can most easily be achieved by tightening the wiring and via pitch on the interposer. In both ceramic and organic interposers, this pitch (for the vias that go through the full structure) is usually mechanically defined (by punching for ceramic and drilling for organic laminates through the core), and the wiring pitches are relatively loose. The individual layers of the laminate and the laminate cores are relatively large. This increases the length of interconnects between die (latency). In addition, bump pitches to the ceramic and organic laminates are relatively large and limit the number of connections possible between dies (bandwidth). One can increase the number of wiring layers and compensate for this somewhat, but even so, this basic limitation persists.

The advent of silicon-based 3D integration allows the use of silicon as an interposer, and both the vias and the

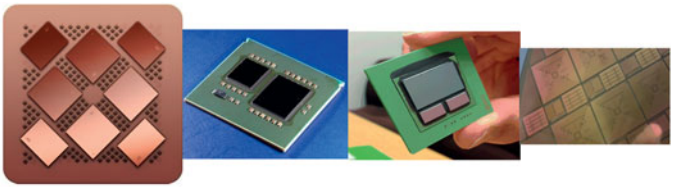
https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

wiring pitches on these silicon interposers are similar to those achieved in silicon technology. Thus, the interconnect density between the die and the interposer limits the degree of interconnectivity between dies. This pitch today can approach approximately <40 μm for silicon on silicon; therefore, silicon interposers offer the highest bandwidth and lowest latency compared to any other option.

In addition, using silicon technology can introduce a variety of passive components such as capacitors, inductors, precision resistors, and antennae. These typically occupy large areas, but the interposers can use older (i.e., cheaper) technology nodes. The upper inset in Figure 2 shows an interposer where three dies are integrated,[16] showcasing the power of high bandwidth low latency integration. A 45 nm SOI-based digital signal processor with embedded DRAM fabricated on 300 mm diameter wafers is integrated with two SiGe bipolar complementary metal oxide semiconductor dies in a 130 nm node (node dimensions refer to the lithographic capability used) fabricated on 200 mm wafers. For practical purposes, they behave as if they are on the same die. This heterogeneous integration has tremendous implications as an alternative to putting everything on the same piece of silicon. The simplification of the individual technologies and consequent cost and yield benefits are obvious, but there is some additional complexity in assembly and testing. In spite of this, there is no alternative cost-effective way to achieve this kind of integration. We have not eliminated the need for a package (or laminate); however, the laminate requirements are considerably simpler, and its purpose is mechanical robustness for handling and fanning out to match the board interconnect pitches.

Silicon as a semiconductor is electrically more lossy compared to ceramic or organic dielectrics. Proper shielding can minimize these losses, but this adds cost. In addition, there is a perception that silicon is expensive to process compared to standard packages. This is not necessarily the case, as we may be able to utilize fully depreciated older fabs. Finally, the transition from a passive interposer to an active one (with devices) is easy, and one can, in principle, extend the interposer concept from a passive interposer to an active one where one can accomplish additional functionality such as power management and full 3D functionality.

Glass has been proposed as an alternative material[17] and has most of the advantages of silicon interposers, with the additional advantage of it being a low loss material. In addition, glass can be fabricated in panels (used today in displays) and in rolls, and this provides the possibility of further manufacturing cost reductions. Today, however, there are many challenges, including handling thin glass (glass has lower elastic modulus compared to silicon), especially after processing, cracking of glass after dicing, and making through-glass vias. Nevertheless, glass continues to be an extremely attractive material on account of its lower cost expectation. **Figure 7** compares the properties of different interposer materials.



	Ceramic MCM	Organic MCM	Si Interposer	Glass Interposer
Dielectric Properties	Adequate	Good	Lossy	Excellent
Feature Dimensions	Mechanically defined	Down to ~10 μm L/S	Si-like Lithography	Display-like
CTE-Induced Stress	V. Good	Mod. High	Excellent	Tailorable
Cost	High	Moderate	Moderate	TBD
Available	Available	Available	Available	Development

Figure 7. Comparison of interposer materials. Note: CTE, coefficient of thermal expansion; MCM, multichip module; L/S, lines and spaces.

Barriers to 3D acceptance

We have shown that it is possible to implement the concept of SoC using a specialized die either by stacking on a die or by stacking on an interposer, and we can either emulate the SoC function or reduce the die footprint by stacking. But neither of these approaches allows for the same level of integration as a large chip SoC. In other words, monolithic integration will always be a superior solution, provided it is economically feasible, but that economic feasibility is being called into question because of the realities of scaling. Even so, die stacking, whether die-on-die or dies-on-interposers, is at present a serial process and requires that dies and interposers be all known good through exhaustive testing—an expensive and time-consuming process that is also a supply chain nightmare. After assembly, the assembly needs to be tested and repaired. Usually there is some repair capability but seldom a rework capability, as we cannot disassemble the stacks without damaging the die. Losing a fully assembled 3D module is indeed an expensive proposition and together with the known good die problem,[§] it has presented a formidable barrier to widespread use of 3D technologies. However, for integrated device manufacturers and foundries that can complete a full process, this presents an unprecedented growth opportunity.

[§]The known good die problem is where one needs to assemble fully tested dies that meet all the specifications of the chip, but some of these specifications may not be verified in die form.

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

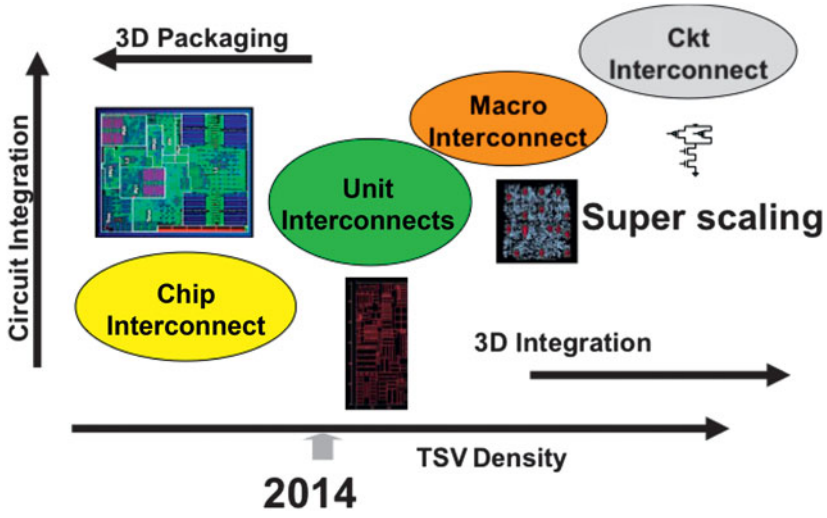


Figure 8. Evolution of three-dimensional (3D) technology from a packaging variant to silicon integration. Note: TSV, through-silicon via.

The evolution of 3D from die stacking to full integration

Figure 8 shows the evolution of 3D technology to achieve greater integration. It plots the level of integration versus an interconnect metric. This could be the TSV density or the microbump pitch. At coarse pitches (e.g., 150 μm), the level of integration between the dies that are stacked is limited to power delivery and I/O functions—typical of packages. As the pitches decrease or their density increases (e.g., 50 μm pitch), it is possible to integrate units on different dies. For example, we can integrate a cache with a processor and not suffer an inordinate hit in latency or bandwidth. To further integrate macros (large functional blocks) on different strata will require pitches in the 10–15 μm range, which is pushing the capability of die-to-die assembly. To achieve further integration at the circuit or block level will require pitches on the order of a few μm, well beyond the capability of die stacking. There are two approaches to 3D monolithic integration.

Sequential processing

In sequential processing,[18] an amorphous or polycrystalline silicon layer is deposited on a processed wafer with a few backend interconnect layers. If amorphous, this layer will need to be recrystallized, typically with some sort of millisecond annealing using an excimer laser, for instance. The idea here is to create as perfect a layer as possible without perturbing the devices that are already fabricated. New devices are then fabricated in the recrystallized layer with a interconnect layer. Vias are also drilled to connect the wiring in this new layer with those in the initial layer. There is the possibility of connections at the device level between two strata. In principle, the process is repeated as many times as needed. As might be expected, the sequential nature of this process as well as the thermal constraints on annealing the amorphously deposited silicon with devices already present in lower layers will make widespread use of this method difficult. However, a variant of this method is currently employed in the latest generation of NAND (NOT AND) flash,[19] where polysilicon transistors are adequate, and vertical devices do not need additional thermal cycles. Thus, this kind of vertical monolithic integration has its use in some specific but important cases.

Wafer stacking

An alternative approach,[20,21] shown schematically in **Figure 9**, is to directly stack already processed wafers. Several variants have been practiced, but the underlying approach is to use either fusion bonding[22] or TCB followed by wafer thinning to a thickness of a few μm with or without an etch stop layer for thickness control. Since the thinning is performed after bonding, the layers can be made very thin, which allows for much finer vias. The etch stop layer could be the buried oxide layer of an SOI wafer.[23] Global alignment is done at the wafer level. Equipment is now available where this alignment can achieve an overlay tolerance below 2 μm. Vias that connect between the interconnect layers in the two strata are then patterned die by die using die level registration.

Wafer distortion due to internal stresses in the wafer will degrade the overlay, but, even so, it is possible to obtain extremely tight via pitches with excellent alignment, making general purpose 3D circuit level integration possible. This approach is expected to lead to widespread applications, especially in memory scaling and neuromorphic computing, where tens of billions of relatively simple cores and memory blocks need to be interconnected on a massive scale. However, to realize this benefit, system designs need to evolve to be highly

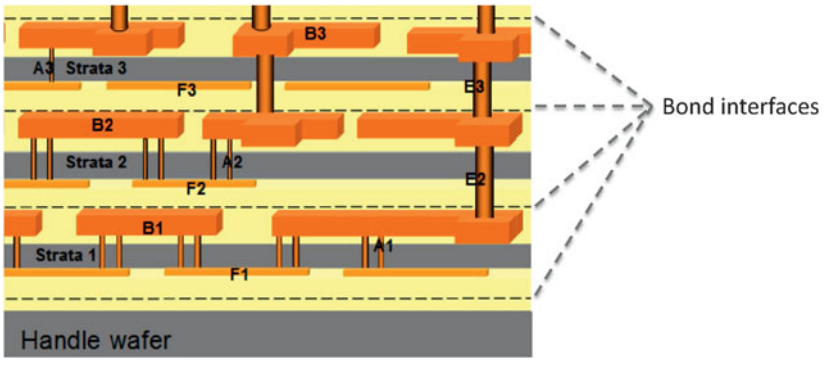


Figure 9. Schematic of wafer-to-wafer stacking that allows for much smaller and tighter interlevel connections between almost completely parallel-processed wafers.

https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press

redundant and extremely fault tolerant. Fortunately, these applications lend themselves to such approaches.

Conclusion

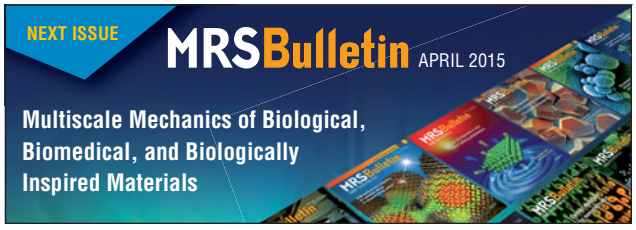
The semiconductor industry is preparing for a major transition into what may be termed "orthogonal scaling." Scaling at the package level has lagged significantly behind scaling at the silicon level, and the industry is well poised to see some radical developments at the package and board levels. Three-dimensional integration is a critical component of this transformation, where we can reduce the footprint between dies by stacking them while simultaneously improving latency, bandwidth, and power in a manner that merely scaling to the next node will not achieve. Monolithic 3D integration, whether by sequential processing or wafer-to-wafer level stacking, can achieve a high degree of interconnection at the circuit level, allowing us to build massive and complex systems. All of this will require concomitant investments in design and test methodologies as well as a significant increase in investment in the packaging development and manufacturing areas.

Acknowledgments

The continued interaction and contributions of all colleagues at IBM, especially Daniel Berger and Gary Patton, are gratefully acknowledged. S.I. would also like to thank Rao Tummala at Georgia Institute of Technology for extensive discussions and insights.

References

1. Z. Or-Bach, "Intel vs. Intel," available at http://www.monolithic3d.com/blog/intel-vs-intel (August 13, 2014).
2. S.S. Iyer, *Proc. IEEE IEDM* 33.1 (2012).
3. J.L. Hennessy, D.A. Patterson, *Computer Architecture—A Quantitative Approach* (Morgan Kauffman, New York, 1996), p. 374.
4. J. Handy, *Cache Memory Book* (Academic Press, San Diego, 1993).
5. S.S. Iyer, J.E. Barth Jr., P.C. Parries, J.P. Norum, J.P. Rice, L.R. Logan, D. Hoyniak, *IBM J. Res. Dev.* **49** (2), 333 (2005).
6. H.Q. Le, W.J. Starke, J.S. Fields, F.P. O'Connell, D.Q. Nguyen, B.J. Ronchetti, W.M. Sauer, E.M. Schwarz, M.T. Vaden, *IBM J. Res. Dev.* **51** (6), 639 (2007).
7. S.S. Iyer, G. Freeman, C. Brodsky, A.I. Chou, D. Corliss, S.H. Jain, N. Lustig, V. McGahay, S. Narasimha, J. Norum, K.A. Nummy, P. Parries, S. Sankaran, C.D. Sheraw, P.R. Varanasi, G. Wang, M.E. Weybright, X. Yu, E. Crabbe, P. Agnello, *IBM J. Res. Dev.* **55** (3), 1 (2011).
8. C.J.D. Craigie, T. Sheehan, V.N. Johnson, S.L. Burkett, A.J. Moll, W.B. Knowlton, *J. Vac. Sci. Technol. B* **20** (6), 2229 (2002).
9. W.S. Liao, H.N. Chen, K.K. Yen, E.H. Yeh, F.W. Kuo, T.J. Yeh, F. Kuo, C.P. Jou, S. Liu, F.L. Hsueh, H.C. Lin, C.N. Peng, M.J. Wang, W.C. Wu, S.P. Hu, M.F. Chen, S.Y. Hou, S.P. Jeng, C.H. Yu, K.C. Yee, D. Yu, *Proc. IEEE Symp. VLSI Technol.* C18–19 (2013).
10. L.F. Miller, *IBM J. Res. Dev.* **13**, 239 (1969).
11. O.L. Anderson, H. Christensen, P. Andreatch, *J. Appl. Phys.* **28**, 923 (1959).
12. Hybrid Memory Cube Consortium, available at http://www.hybridmemorycube.org.
13. JEDEC, *High Bandwidth Memory DRAM*, available at http://bit.ly/1gWBE5E.
14. G. Van der Plas, P. Limaye, I. Loi, A. Mercha, H. Oprins, C. Torregiani, S. Thijs, D. Linten, M. Stucchi, G. Katti, D. Velenis, V. Cherman, B. Vandevelde, V. Simons, I. De Wolf, R. Labie, D. Perry, S. Bronckers, N. Minas, M. Cupac, W. Ruythooren, J. Van Olmen, A. Phommahaxay, M. de Potter de ten Broeck, A. Opdebeeck, M. Rakowski, B. De Wachter, M. Dehan, M. Nelis, R. Agarwal, A. Pullini, F. Angiolini, L. Benini, W. Dehaene, Y. Travaly, E. Beyne, P. Marchal, *IEEE J. Solid-State Circuits* **46** (1), 293 (2011).
15. A. Totta, R.P. Sopher, *IBM J. Res. Dev.* **13**, 226 (1969).
16. Semtech, "Semtech and IBM Join Forces to Develop High-Performance Integrated ADC/DSP Platform Using 3D TSV Technology," available at http://bit.ly/1niX1vw.
17. V. Sukumaran, Q. Chen, L. Fuhan, N. Kumbhat, T. Bandyopadhyay, H. Chan, S. Min, C. Nopper, V. Sundaram, R. Tummala, *Proc. IEEE Electron. Compon. Technol. Conf.* 557 (2010).
18. P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J. Hartmann, L. Sanchez, L. Baud, V. Carron, A. Toffoli, F. Allain, V. Mazzocchi, D. Lafond, S. Deleonibus, O. Faynot, *Proc. IEEE Int. Symp. Circuits Syst.* 2233 (2011).
19. H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, R.M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, A. Nitayama, *Proc. IEEE Symp. on VLSI Technol.* 14 (2007).
20. R.S. Patti, *Proc. IEEE* **94** (6), 11214 (2006).
21. P. Batra, D. LaTulipe, S. Skordas, K. Winstel, C. Kothandaraman, B. Himmel, G. Maier, B. He, D.W. Gamage, J. Golz, W. Lin, T. Vo, D. Priyadarshini, A. Hubbard, K. Cauffman, B. Peethala, J. Barth, T. Kirihata, T. Graves-Abe, N. Robson, S.S. Iyer, "Three-Dimensional Wafer Stacking Using Cu TSV Integrated with 45 nm High Performance SOI-CMOS Embedded DRAM Technology," paper presented at the IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference, Monterey, CA, October 2013.
22. S.S. Iyer, A.J. Auberton-Herve, Eds., *Silicon Wafer Binding Technology: For VLSI and MEMS Applications* (INSPEC, The Institution of Electrical Engineers, London, UK, 2001).
23. V. Suntharalingam, R. Berger, J.A. Burns, C.K. Chen, C.L. Keast, J.M. Knecht, R.D. Lambert, K.L. Newcomb, D.M. O'Mara, D.D. Rathman, D.C. Shaver, A.M. Soares, C.N. Stevenson, B.M. Tyrrell, K. Warner, B.D. Wheeler, D.-R.W. Yost, D.J. Young, *IEEE Int. Solid-State Circuits Conf. Tech. Dig.* **48**, 356 (2005). □





https://doi.org/10.1557/mrs.2015.32 Published online by Cambridge University Press