A Survey of 3D Circuit Integration

Stephen Tarzia
March 14, 2008

***Abstract*—This paper surveys recent publications on the new class of layered circuit integration techniques termed *3D integration*. We describe both the potential benefits and major pitfalls of 3D integration. Several competing layering approaches are described and compared. We also forecast the impact that a move to 3D integration would have on CAD tools and circuit design flows.**

I. Introduction: why 3D?

In this section we motivate (A) higher levels of integration, (B) the shortening of interconnect, (C) heterogeneous integration, and (D) fine grained testing. 3D integration facilitates each of these tasks.

A. Systems on Chip

The System on Chip (SoC) has become an attractive option due to technology scaling. An SoC is a is a complete electronic system, including digital logic, memory, and analog circuitry, in a single chip. The reasons for this development are clear. In the fabrication of integrated circuits, yield drops off dramatically with increased die area. For this reason, die areas have only slowly increased over the years. Transistor density, on the other hand, has maintained an exponential growth trajectory for decades. As VLSI transistor sizes decrease, more functionality can be integrated onto a given die area. At the same time, pads and PCB traces outside of a chip become larger relative to those smaller transistors within chip; both the power and latency of off-chip communication increase relative to on-chip communication.

In addition. growth in the number of pins available on chip packages is relatively slow; this limits off-chip communication bandwidth. Therefore, there is both the capability and motivation to integrate more system components onto a single die.

B. Interconnect shortening

With increasing per-die circuit size, both in SoCs and in complex monolithic instruction processors, interconnect delay has become the dominant factor for circuit performance. Larger circuit sizes mean global wires connecting opposite ends of the die are larger relative to transistors; their delay now dominates gate delay.

The dominance of interconnect delay has also created a *timing closure* problem for CAD tools [1, p. 608]. Physical layout and routing decisions must be fed back into the tools that synthesize logic and size gates in order to verify that timing deadlines are indeed met; as interconnect delay becomes more dominant, it becomes more difficult to predict path delays in the early stages and thus the chances of meeting deadlines after physical design are lowered. Infinite iteration between high and low level design can result.

Repeater insertion is another headache for hierarchical design flows [2, fig. 9]. Delay along global wires can be reduced by breaking them up with repeaters. However, these repeaters must be squeezed into valuable silicon area underneath the wire. Routing is usually done after placement, so we have a feedback situation akin to timing closure: we may invalidate a placement in the routing stage after realizing that a repeater is needed where there is no open silicon.

It would be highly advantageous to reduce the delay penalties and design complexities due to long interconnects. We can see that 3D integration does this by simple geometry. A given square area A has maximum Manhattan wirelength $2\sqrt{A}$. The same area split into two layers reduces the wirelength to $\sqrt{2}\sqrt{A} + l_v$ where l_v is the length of a via between layers. In general, n layers gives a maximum Manhattan wirelength of $2\sqrt{\frac{A}{n}} + (n-1)l_v$.

C. Heterogeneous integration

In a mixed design, like an SoC, we prefer to fabricate each type of circuit in its own ideal technology; a 3D system allows layers built with different processes to be combined into a single chip. It is possible to fabricate digital logic, memories, DSPs, analog and RF devices on a single die using one technology but this is suboptimal in terms of performance, area, and power. Putting the components on different dies also allows us to better isolate sensitive analog circuitry. Even within the same process, it may be desirable to have layers with different voltage and performance requirements or clocking domains. Looking ahead, heterogeneous layering also would allow the upper layer to be dedicated to optical I/O and low-skew optical clock distribution [1, p. 610].

D. Commodity dies

For all but the largest production runs, tremendous cost savings might be realized by assembling systems from a collection of commodity "prefab" dies rather than creating a new mask set. Mask prices for cutting-edge processes have been increasing steadily, so mask reuse is critical. Prefabbed dies for certain components, especially analog devices, could be used for many years while timing-critical digital logic dies are continually updated for newer processes. Deng and Maly analyse the cost benefits of layered integration in detail [4].

E. Component testing and replacement

Yield would be significantly improved if chip components could be individually tested and repaired prior to packaging.

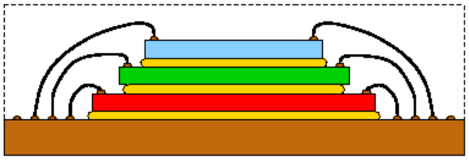


Fig. 1. Die stacking using wire bonding. Image is taken from [10].

Yield is poor in large monolithic 2D ICs because just a single fault in that large chip area dooms the entire chip. On the other hand, a 3D IC might be tested in parts. Chips would be built only from Known Good Dies. Yield would increase dramatically since each fault causes only a fraction of the entire system to be discarded. Increased yield would translate into lower cost and higher feasible circuit area.

II. Performance and area estimates

Banerjee *et al.* have a thorough analysis predicting performance, area and thermal characteristics of 3D ICs [1, sec. 3]; we summarize this here. The area of high performance ICs is assumed to be wire pitch limited; therefore the new routing freedom granted by 3D ICs leads to decreased total area, not just decreased footprint. Shorter interconnect causes shorter delays; however, if we allow the same footprint as a 2D chip (and therefore twice the total area) performance can be further improved by widening wires. The above analysis ignores the effects of temperature; we will return to this later.

III. 3D technologies

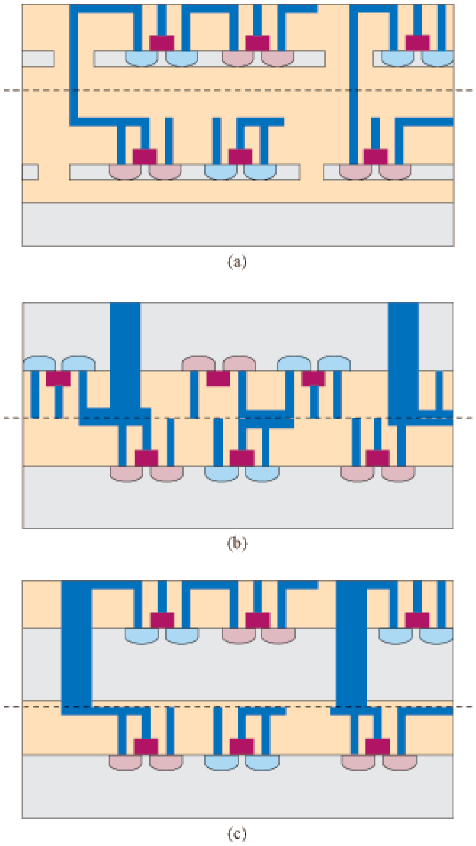
Several options have been proposed for 3D integration. Only die stacking has yet reached production.

A. Die stacking

The simplest option for 3D integration is stacking of successively smaller dies, as shown in figure 1. The product is called a Multiple Chip Module (MCM). In this approach, die alignment requirements are not very precise; wire bonding, tape automated bonding, or controlled-collapse chip connections (C4 solder) are possibilities for connections between layers. Die-stacked chips with wire bonding are already on the market [2]. While simple to produce, the benefits of die stacking are limited. Compared to the discrete chip case, inter-die connections are lower in impedance but they are still limited in number.

B. Wafer bonding

Wafer bonding is the process of joining two or more wafers prior to dicing and packaging. Such bonding must create inter-layer vias while isolating the transistors from adjoining layers. Figure 2 illustrates several of the proposed processes. In all three processes, through-silicon vias are sunk deep into the substrate and later revealed by temporarily attaching the wafer to a glass “handle” and thinning its back end down. In the face-to-back processes (figure 2 a. and c.), thinning the upper layer exposes vias that contact the lower layer. In the face-to-face process (figure 2 b.) thinning the upper layer exposes wire-bonding pads for packaging.



(a)

(b)

(c)

Fig. 2. Closeup view of wafer-bonded layers. Dashed line indicates bonded interface: (a) SOI face-to-back bonding with thin layers, (b) face-to-face bonding, (c) face-to-back bonding with thick layers and deep vias. Image is taken from [9].

The SOI process scales better to many layers since each layer is very thin, which aids in dissipating heat from the lower layers; we will see that heat is an important consideration. However, SOI processes carry their own disadvantages.

Wafer bonding requires very precise alignment of wafers during bonding. Current alignment precision is limited to about $\pm 2\,\mu$m [1, p. 627] [4, sec. 2C]. Through-silicon vias would be relatively easy to drive; one estimate of their RLC electrical characteristics is a few mΩ, less than 1 pH. and a couple of fF. [2]

C. Silicon growth

There is also a class of 3D integration proposals based on growing substrate layers above complete, metalized wafers [1, p. 627] [4, sec. 2C]. These techniques include Beam Recrystallization, Epitaxial Growth, and Solid Phase Crystallization. One drawback of these techniques is that layers must be homogeneous since layers are created in immediate succession on the same equipment. Another problem is that the underlying copper metalization layers are somewhat sensitive; they must be kept below 450 degrees celsius while the higher layers are created.

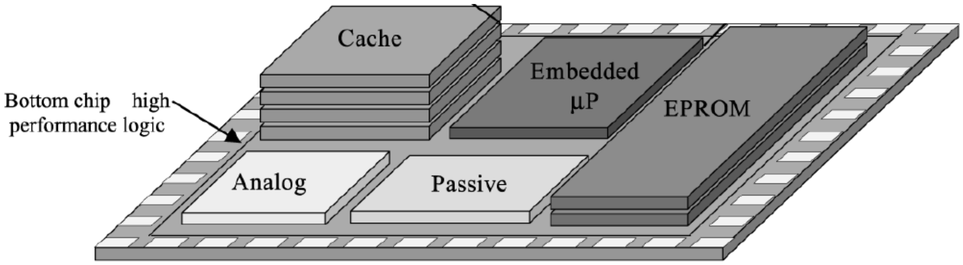


Fig. 3. A 2.5D system. Image is taken from [4].

D. 2.5D integration

It is important to note that there are no yield benefits for any of the previously mentioned 3D processes. This is because testing cannot be done until assembly is completed. Actually, we should expect yield to be lower than for an equivalent 2D chip due to increased processing complexity.

2.5D integration is a revision of die stacking which adds yield benefits [4]. As shown in figure 3, 2.5D integration stacks small dies on top of a large bottom layer containing high-performance logic. The key feature of this process is incremental, hierarchical testing an assembly. First, the bottom layer is partially packaged and tested. It is then used as a "test chassis" to test the dies added on top of it. It is important to note that thorough testing can currently only be done on packaged dies.

The 2.5D process presents several challenges. A die that tests negatively must be removed; this type of die reworking may require some technical advances to be feasible. Also, the dies and the chassis would have to be designed with modular testing in mind. Of course, modular testing is already being developed for future massively multi-core processors, so this requirement should not deter us from considering this technology. However, the cost of fine-grained testing and reworking must be added when doing a cost-benefit analysis.

The 2.5D process as described by its authors calls for die bonding techniques similar to wafer-bonding rather than coarser wire or solder-ball bonding. Performing such bonding with an individual die is much more difficult (or at least costlier) than bonding wafers. The authors suggest that alignment can be aided by high-precision (presumably etched) self-alignment mechanical latches on the surface of the die.

Despite its name, 2.5D integration is certainly not really an intermediate step between 2D and 3D processes. It is quite sophisticated and requires at least two difficult advances in processing technology.

The advantages and disadvantages of the 3D technologies discussed are summarized in figure 4.

IV. CAD

The addition of a third dimension would have require support from more advanced CAD tools. We wish to leverage new design choices but, at the same time, complexity forces tools to adopt more hierarchical design flows, leading to closure problems, as discussed above [1, p. 608].

An alternative to hierarchical design is exploration of design space in its full complexity [4]. This type of approach has already been used to consider thermal effects early in the design process for 2D circuits [6]. Thermal-aware design, in general, will be more important in 3D technologies.

There has been some work specifically targetting 3D routing and placement [3] [5]. I am not sure how useful such work would be. In the 2.5D approach, each die is its own independent system which can be designed independently once an I/O interface is chosen.

V. Thermal challenges

Aside from costs and technological hurdles, the main challenge facing 3D integration is heat dissipation. IC cooling has traditionally been handled by a heatsink attached to the top surface of the chip package. In this model, we can consider the chip to be a one dimensional thermal system; heat flows straight up from the chip to the heatsink. A move to 3D integration decreases the chip footprint and therefore increases power density at the heatsink interface. Another problem is that upper layers insulate lower layers from the heatsink. Silicon has high thermal resistance, so we expect a sharp vertical temperature gradient to develop in the chip.

Weerasekera *et al.* give an example wherein die temperature increases to to over 300 degrees Celsius after moving to 3D integration [10, table 3]. From Banerjee *et al.*'s analysis of 3D ICs we have the temperature rise of the deepest (n_{th}) silicon later [1, eqn. 50]:

$$\Delta T_n = \frac{P}{A}\left[\frac{R}{2}n^2 + \left(R_1 - \frac{R}{2}\right)n\right]$$

where P is the chip's power dissipation, A is the chip's surface area, R is the thermal resistance between layers, and R_1 is the thermal resistance between the top layer and theheatsinkk. If we assume that $R_1 \gg R$, then there is an approximately linear relationship between n and ΔT_n. Their numerical example indicates that when moving a circuit onto a two layer 3D technology, the package thermal resistance must be roughly halved in order to maintain the same temperature while taking advantage of the doubling of frequency that it makes possible.

Microchannels etched into the substrate can form channels through which cooling liquid could be pumped. This form of cooling would be effective in cooling even the hottest 3D chips. However, it would be imprudent for researchers to expect salvation from an exotic research technology such as this. Dummy thermal vias might also help; these are electrically isolated vias whose only purpose is to conduct heat vertically toward the heatsink

	heterogeneous	component testing (high-yield)	interlayer via density	fabrication difficulty
die stacking (MCM)	yes	no	low	**low**
wafer bonding	yes	no	**high**	medium
silicon growth	*no*	no	**high**	?
2.5D	yes	**yes**	**high**	*high*

Fig. 4. Comparison of 3D technologies

Thermal issues are less important in certain lower power devices like DRAMs. These chips have power density of about $0.01\,{}^{\mathrm{W}}\!/_{mm^2}$; compare this with $2\,{}^{\mathrm{W}}\!/_{mm^2}$ that you'd see in the hottest parts of a high speed logic [2]. Integrating DRAMs directly above CPUs may be a very attractive option since the DRAMs will not add much power [8].

Traditionally high-power devices like instruction processors might be stepped-down for implementation in 3D ICs by using a lower supply voltage. The associated performance drop might be made-up by increased parallelism. However, low-voltage design is more sensitive to process variability which affects threshold voltage. And of course, parallelism is not always useful.

VI. Conclusion

3D integration was considered as far back as 1979 [1, p. 626]. However, there was no reason to give it much consideration while 2D technology was scaling so well. It is natural that it is being reconsidered now given the myriad problems in deep-submicron billion-transistor circuits. Clearly, not all of the proposals mentioned above will be adopted. Of those discussed, I suspect that a second layer of low-power DRAM will become popular. I think that it is also worth mentioning that 3D integration would allow multicore processors to adopt interconnection topologies more closely resembling those used in supercomputing, rather than just simple grids [7].

References

[1] K. Banerjee, S.J. Souri, P. Kapur, and K.C. Saraswat. 3-d ics: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. *Proceedings of the IEEE*, 89(5):602–633, May 2001.

[2] Kerry Bernstein, Paul Andry, Jerome Cann, Phil Emma, David Greenberg, Wilfried Haensch, Mike Ignatowski, Steve Koester, John Magerlein, Ruchir Puri, and Albert Young. Interconnects in the third dimension: Design challenges for 3d ics. *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*, pages 562–567, 4-8 June 2007.

[3] Jason Cong and Yan Zhang. Thermal-driven multilevel routing for 3-d ics. In *ASP-DAC '05: Proceedings of the 2005 conference on Asia South Pacific design automation*, pages 121–126, New York, NY, USA, 2005. ACM.

[4] Yangdong Deng and W.P. Maly. 2.5-dimensional vlsi system integration. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 13(6):668–677, June 2005.

[5] Brent Goplen and Sachin Sapatnekar. Thermal via placement in 3d ics. In *ISPD '05: Proceedings of the 2005 international symposium on Physical design*, pages 167–174, New York, NY, USA, 2005. ACM.

[6] Zhenyu Gu, Jia Wang, R.P. Dick, and Hai Zhou. Incremental exploration of the combined physical and behavioral design space. In *Proceedings of 42nd Design Automation Conference*, 2005.

[7] V Kumar, A Grama, A Gupta, and G Karypis. *Introduction to parallel computing: design and analysis of algorithms*. Benjamin-Cummings Publishing Co., Redwood City, CA, 1994.

[8] C.C. Liu, I. Ganusov, M. Burtscher, and Sandip Tiwari. Bridging the processor-memory performance gap with 3d ic technology. *Design & Test of Computers, IEEE*, 22(6):556–564, Nov.-Dec. 2005.

[9] A. W. Topol, B. K. Furman, K. W. Guarini, L. Shi, G. M. Cohen, and G. F. Walker. Enabling technologies for wafer-level bonding of 3d mems and integrated circuit structures. In *Proceedings of the 54th Electronic Components and Technology Conference (ECTC)*, page 931, 2004.

[10] Roshan Weerasekera, Li-Rong Zheng, Dinesh Pamunuwa, and Hannu Tenhunen. Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs. In *ICCAD '07: Proceedings of the 2007 IEEE/ACM international conference on Computer-aided design*, pages 212–219, Piscataway, NJ, USA, 2007. IEEE Press.