Design for 3D Integration and Applications

Invited Paper

Paul D. Franzon, William Rhett Davis, Michael B. Steer, Hua Hao, Steven Lipa,
Sonali Luniya, Christopher Mineo, Julie Oh, Ambirish Sule, Thor Thorolfsson,
Department of Electrical and Computer Engineering,
North Carolina State University, Raleigh, NC, USA
paulf@ncsu.edu , 919.515.7351

***Abstract*—3D stacking and integration can provide system advantages equivalent to up to two technology nodes of scaling. This paper explores application drivers and computer aided design (CAD) for 3D ICs.**

***Index Terms*—3D IC, CAD, VLSI**

I. Introduction

An exciting technological proposition that is becoming very real is the 3D stacking and integration of dies using vertical vias. An SEM showing the cross-section of such a process is shown in Figure 1. This is the 3D 0.18 μm SOI process developed at Lincoln Labs [2]. This process permits wafer stacking and integration of three SOI "tiers" of devices. Though the figure shows two CMOS tiers and one photo-diode tier, Lincoln Labs has also made available a process that incorporates three CMOS tiers. The tiers are integrated with sub-micron 3D vias. The purpose of this paper is not to explore 3D technology (see [4] for a summary) but to illustrate what applications are good drivers for 3D and how are they designed.

II. Why 3D?

When is it advantageous to go vertical and when is it not? Stacking two wafers together and integrating them with vertical vias is not cheap. As a rough rule of thumb, the additional processing cost is about equivalent to that of adding two additional layers of metallization. This cost is even higher if individual die are stacked. This cost must be justified through performance gains or cost savings elsewhere in the system. This cost is much greater than simply that of even "high-end" sophisticated packaging. When might it possibly be justified?

Fortunately, there is a growing consensus that there are several, main-stream, circumstances which justify 3D integration. A list these of potential drivers for 3D integration is provided in Table 1. These will be discussed in turn.

The first, and most obvious, potential motivation is miniaturization. However, through-silicon 3D integration is rarely justified by the desire for miniaturization alone. For most circumstances, if volume reduction is the only goal, then it is much more cost effective to stack and wire-bond. This technology is already in wide-spread use in cell phones, and continues to grow in sophistication. However, one exception that is being widely explored is for memories. Wire-bonding can not be easily used to stack identical memory chips, as they are all the same size. In addition, there are system advantages to thinning and stacking multiple memory die such that the aggregate memory has the same end form factor as one memory package. For example, this technology could enable a credit card sized video storage and viewing device containing 100s of hours of video.

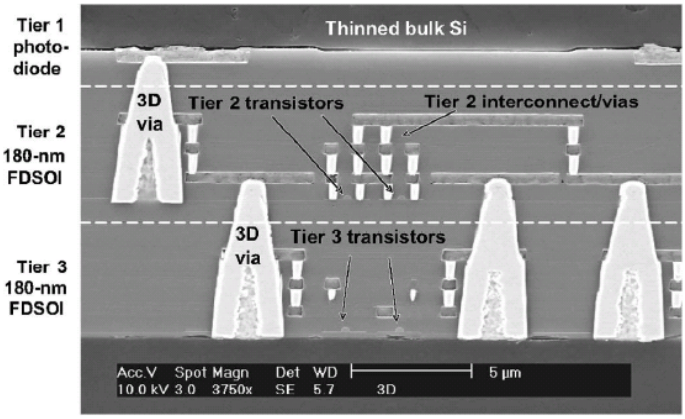

Figure 1. Cross Section of the Lincoln Labs 3D SOI Process, as used in a photo-diode application. The applications described here-in were built using three 180 nm tiers. Taken from [2].

The most explored advantage of 3D is to use it to reduce the interconnect distance between chip functions. Many researchers justify 3D from an interconnect delay, and interconnect power, perspective. From a theoretical viewpoint, the advantages can be substantial. Several studies have presented a Rent's rule style of analysis that presents significant advantages [4, Banerjee01, Rahman01, 5]. The basic argument relies on the fact that with each additional added layer of transistors, there is a similar increase in the number of circuit functions that can be interconnected within a fixed wire length. This leads to a 25% or more decrease in worst case wire length [3] [11], a similar decrease in interconnect power [5], and a decrease in chip area. However, experience shows that many designs do not realize this in practice. Fortunately, with careful choice appropriate design applications can be found. For example, FPGAs are very interconnect bound and can achieve substantial performance and power improvements when recast in 3D [12]. Another interconnect bound circuit is the Ternary

Content Adressable Memory (TCAM). Remapping a TCAM onto 3D can provide a 23% power improvement [10]. Results obtained using two practical examples explored at NCSU are summarized in Figure 2 [Hao06]. This Figure compares data taken from two designs executed in the Lincoln Labs 3-tier process. One is a Fast Fourier Transform (FFT) [Hao06]. The other a dual core process – an Open Risc Processor System On a Chip (ORPSOC) [13]. In this study, the performance benefits of 3D integration were compared with those of technology scaling. In these examples, 3D integration provided about the same performance advantage of two generations of technology scaling – a very compelling case.

Table 1. Potential drivers for 3D integration.

Driving Issue	Case for 3D	Caveats
Miniaturization	Stacked memories. "Smart dust" sensors.	For many cases, stacking and wire-bonding is sufficient
Interconnect Delay	When delay in critical paths can be substantially reduced through 3D integration.	Not all applications will have a substantial advantage
Memory Bandwidth	Logic on memory can dramatically improve memory bandwidth	While memory bandwidth can be improved dramatically, memory size can only be improved linearly
Power Consumption	In certain cases, a 3D architecture might have substantially lower power over a 2D	Limited domain. In many cases, it does not.
Mixed Technology (Heterogeneous) Integration	Tightly integrated mixed technology (e.g. GaAs on silicon, or analog on digital) can bring many system advantages.	Though might justify 3D integration, this driver might not justify vertical vias., except for the case of imaging arrays

Stacking memory die to create a new "super-memory" chip is not the only 3D application involving memory. An interesting and little explored area is logic-on-memory. That is creating a high bandwidth memory interface to the logic. For many end applications, the demand for memory bandwidth is growing rapidly. In many cases, this is due to the increased use of multi-core processors. With the addition of each processor, comes a similar requirement for increasing memory bandwidth. It is predicted that by 2010, a 32-core CPU will require 1 TBps of off-chip memory bandwidth [6]. This, by itself, gives a fairly natural case for 3D, that has been only lightly explored, and then mainly in the context of general purpose computer architecture. For example, 3D caches can lead to 10% - 50% reductions in cache latency, depending on the benchmark used [9] [8]. Other applications that are likely to benefit form logic-on-memory include digital signal processing, graphics and networking. Some application-specific examples will be given below. At least one company is exploring customized 3D memories for logic-on-memory applications [14].

The potential for power reduction comes from two directions. The prospective for interconnect power reduction was discussed above. An area that has been little explored is the potential for trading area for power. Given the relief provided by 3D integration on interconnect issues, this potential exists. Another obvious potential advantageous route to explore is to use 3D memory integration to reduce memory power. This is explored a little in one of the case studies below.

Finally, a compelling driver for 3D technology is mixed technology or heterogeneous integration. The main application explored to date has been imaging arrays. The advantage of using 3-D is that no area has to be sacrificed on the imaging layer for circuitry (making it a more efficient photon collector), and considerable circuitry can be placed right underneath each pixel for pixel-level processing. For example, this approach has been used to produce a laser radar receiver array [1].

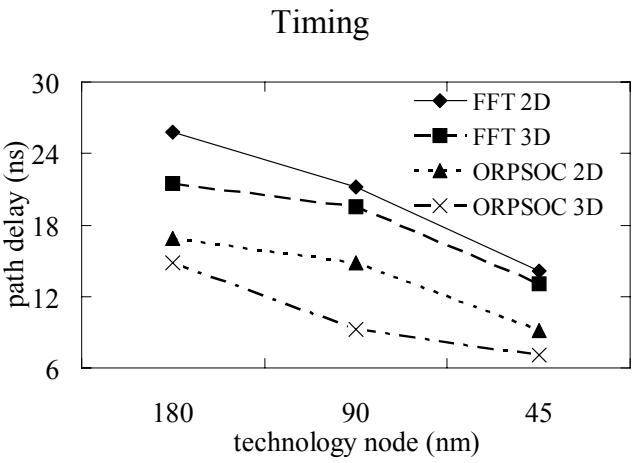

Figure 2. Improvement in path delay improvements achieved using a 3D 180 nm technology, vs. technology scaling alone. 3 metal layers were assumed for each silicon layer. FFT=Fast Fourier Transform. ORPSOC = Open core RisC Processor System on a Chip (the 2 core design presented in Figure 2). Adding two additional silicon layers is roughly equivalent to two generations of technology scaling. Taken from [Hao06].

III. Computer-Aided Design

In support of the 3D design community, the 3D Group at NC State University has developed a complete CAD flow, based on conventional tools provided by Cadence and PTC and with ongoing collaboration of PTC and the University of Minnesota. This flow is shown in Figure 6. Several changes have been implemented to enable 3D designs. Though a first release of the tools is available, the work in this tool flow is ongoing. In the first release, Cadence 2D tools were used, with the addition of a customized partitioning tool. PTC's Mechanica suite was used for thermal analysis, interfaced to the IC design tools via a macromodeling tool. In the second generation, 2D place and route will be replaced by a true 3D place and route tool.

For a designer, the main differences to consider in a 3D flow are as follows:

3D Floorplanning. True 3D floorplanning is needed to enable true 3D optimal designs. Careful consideration of the 3D spatial relationships of different modules in a design leads to short

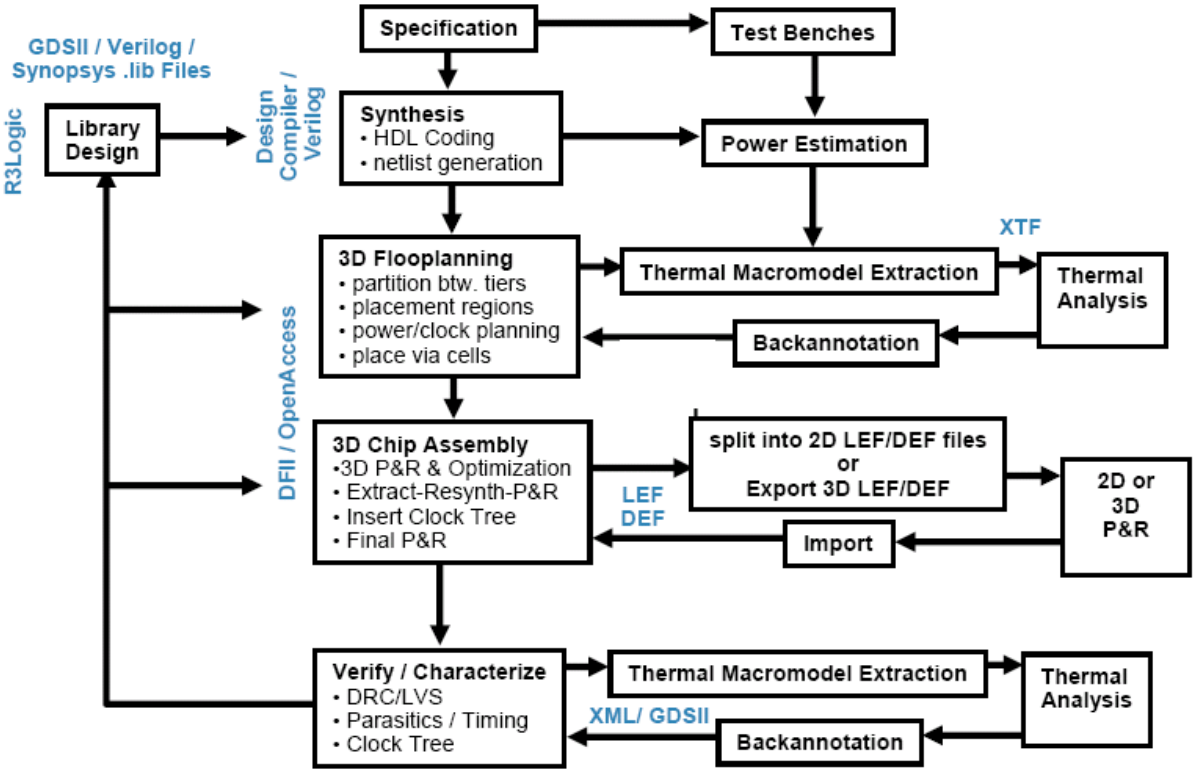

Figure 3. CAD Flow

interconnect paths. The floorplan has a strong impact on thermal operation, as discussed below.

Close integration with thermal analysis. This aspect is very important. Heat density, and thus temperature increases with each additional layer of transistor devices. For example, in all the designs above, the total heat density was almost trebled over that of a 2D design. Thermal analysis is essential at the floorplanning level to determine if modules need to be rearranged in order to control temperature. It might also be used to determine if additional area should be allocated to thermal vias. However, it is important to note that additional vertical vias only provide an incremental improvement when temperatures are too high. At this stage, temperature is better controlled through better choice of the design blocks implemented, and their arrangement in the 3D floorplan. E.g. A "hot" module should be placed on a silicon layer closer to the heat sink. Thermal analysis is also important after detailed design is complete. For example, timing can not be predicted unless device temperature is well known. An important circuit to analyze is the clock circuit. The clock buffers not only produce a lot of heat (10%-20% is typical) but requires detailed timing analysis. Mis-prediction of clock skew can easily lead to chip failure. This is one circumstance where the addition of thermal vias around clock buffers can be used to provide fine-grain temperature control.

IV. DISCUSSION

With careful choice of application, and good design planning, integrated 3D designs can often offer a very superior performance/cost point over their 2D equivalent. However, without both of these, 3D designs are not likely to offer a strong advantage, especially in designs where performance is dominated by the performance of multiple long-range interconnects.

3D design does bring some additional burdens though. The increased complexity and importance of thermal design was discussed above. Additional issues include the increased cost and risk of associated with prototyping, and yield management issues.

It is well known that the masks and wafer processing costs associated with the first prototype run in the multi-million dollar range, and increase rapidly with smaller technology nodes. This is true for one 2D design – imagine taping out multiple 2D designs simultaneously to create a 3D design. The cost and risk associated with a 3D ASIC could be quite daunting. There has not been much public discussion on this issue but possible approaches to alleviating this cost and risk include the following:

- Focusing on logic-on-memory. This separates the risks. Each can be approached independently.
- Make each chip in the stack identical at the physical layer. This means only one mask set and wafer run.

Configure the connections using soft switches after stacking and integration. Alternatively, simply minimize the differences between each layer of silicon, e.g. limit the difference to one mask layer only.

- Multi-project runs.

Yield management is another issue that has been little discussed. If the yield of a single die at wafer level is 90%, then yield of a die from two stacked wafers would be 81%; from three wafers, 73%, etc. This would create a clear financial disincentive for 3D. Fortunate, there are at least two ways to alleviate this issue. One approach would be to use small die that have high yield. Another would be to use a die on wafer 3D integration approach, rather than wafer on wafer. If the separated die were even partially tested and sorted (as is commonly done before dicing), the yield would be that of undiced wafer die. If only good sites on the wafer were populated with mounted die, then the yield impact of 3D integration would be minimal.

V. CONCLUSIONS

In general, 3D is justified in designs where interconnect resources dominate performance. By permitting a reduction in wire lengths, or an increase in bandwidth, there are many examples where 3D integration can improve performance and/or power consumption by 20% or more. This is equivalent to about two technology nodes of scaling. However, 3D does complicate design. A 3D specific design flow is needed, and thermal design, test and yield management, all require careful attention.

ACKNOWLEDGMENT

This work was funded in part by DARPA under the 3D IC program, and by NSF.

REFERENCES

[1] B. Aull, J. Burns, C. Chen, B. Felton, H. Hanson, C. Keast, J. Knecht, A. Loomis, M. Renzi, A. Soares, V. Suntharalingam, K. Warner, D. Yost, and D. Young, "Laser Radar Imager Based on 3D Integration of Geiger-Mode Avalanche Photodiodes and Two SOI Timing Circuit Laters," Proc. IEEE ISSSC, Feb, 2006, pp. 1179-1188.

[2] J.A. Burns, B.F. Aull, C.K. Chen, C. Chang-Lee, C.L. Keast, J.M. Knecht, V. Suntharalingam, K. Warner, P.W. Wyatt, D. Yost, "A wafer-scale 3-D Circuit Integration Technology," IEEE Trans ED, Vol. 52, No. 10, Oct. 2006, pp. 2507-2516.

[3] K. Banerjee, S. Souri, P. Kapur, K. Saraswat, "3-D ICs: A Novel Chip Design For Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," Proc. IEEE, Vol. 89, No. 5, 2001.

[4] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P.D. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Veritical," IEEE Design and Test of Computers, VOl. 222, No. 6, Nov-Dec, 2005, pp. 498-510.

[5] S. Das, A. Chandrakasan, R. Reif, "Timign, Energy and Tehrmal Performance of Three Dimensional Integratied Circuits," Proc. Great Lakes Symposium on VLSI, pp.338-343, 2004.

[6] H.P Hofstee, "Future Microprocessors and off-chip SOP Interconnect," in IEEE Trans. Advanced Packaging, Vol. 27, No. 2, May 2004, pp. 301-303.

[7] H. Hua, "Design and Verification Methodology for Complex Three-Dimensional Digital Integrated Circuits," Ph.D. Dissertation, NC State University, 2006. Under the direction of Dr. W.R. Davis.

[8] S.A. Kuhn, M.B. Kleiner, P. Ramm, W. Weber, "Performance Modeling of the Interconnect Structfure ofa Three-Dimensional Integrated RISC Processor/Cache System," IEEE Trans. CPMT, Part B, Vol.19, No.4, Nov, 1996, pp. 719-727.

[9] F. Li, C. Nicopoulos, T. Richardson, Y. Xi, V. Narayanan, M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," Proc. ISCA'06, pp.130-141.

[10] E.C. Oh, P.D. Franzon, "Design Considerations and benefits of Three-Dimensional Ternary Content Addressable Memory," Proc. IEEE CICC, Oct., 2007.

[11] A. Rahman, . Fan, R. Reif, "Comparison of Key Performance Metrics in Two and Three Dimensional Interrgated Circuits," Proc. Int. Interconnect Technology Conference, pp. 18-20, 2000.

[12] A. Rahman, S. Das, A. Chandrakasan, R. Reif, « Wiring Requirement and Three-Dimensional Integration Technology for Field Programmable Gate Arrays, » IEEE Trans. VLSI, Vol. 11, No. 1, Feb, 2003, pp. 44-54.

[13] K. Schoenfliess, "Performance Analsysi of System-on-Chip Application of 3D Integrated Circuits," MS. Thesis, NC State University. Under direction of Dr. W.R. Davis.

[14] www.tezzaron.com