3.5-D Integration: A Case Study

Shashikanth Bobba[1], Pierre-Emmanuel Gaillardon[1], Ciprian Seiculescu[1], Vasilis F. Pavlidis[2], and Giovanni De Micheli[1]

[1]Integrated Systems Laboratory, EPFL
CH-1015 Lausanne, Switzerland

[2]Advanced Processor Technologies Group,
University of Manchester, UK

***Abstract*— Two diverse manufacturing techniques for building 3-D integrated systems are vertical integration with *Through-Silicon-Vias* (TSVs), also referred as 3-D TSV integration, and 3-D monolithic integration. In this paper, we present a hybrid integration scheme that combines these two approaches, taking into account their existing technology limits, into a disruptive paradigm called 3.5-D integration. Our novel integration supports circuit-partitioning both at the gate and block level with unprecedented benefits in cost. To demonstrate the effectiveness of 3.5-D integration, we chose as case study a 288-core MPSoC and we made hypothesis on the manufacturing and test cost. We argue a potential 20% decrease in the manufacturing cost and 30% decrease in the test cost when compared to 3-D TSV integration. In order to study the performance improvement of the MPSoC, we benchmarked various blocks of the core and the on-chip interconnection network, connecting all the cores. Our study shows large improvement in performance of the core (average of 11.5%) and latency (average of 24%) of the *Network-on-Chip* (NoC) for the 3.5-D integration when compared to the corresponding 3-D TSV implementation.**

I. INTRODUCTION

Future *Multi-Processors System-on-Chips* (MPSoCs) will integrate multiple layers of active devices on a single 3-D chip [1]. 3-D fabrication technologies include several integration schemes. Two of these schemes are the vertical integration with TSVs [2] and 3-D monolithic integration [3]. Among the different approaches of vertical integration, we consider die-stacking employing TSVs [2]. Similar to authors in [4], we use the term 3-D TSV to denote this technology. Figure 1 illustrates two die-stacking techniques for 3-D TSV circuits, where multiple dies are stacked either by *face-to-face* or *face-to-back* bonding [4]. One of the main challenges of this technology is to reduce the size of the TSVs, which is in the range of few micrometers. Hence, today 3-D TSV technology is limited only to block-level integration [5].

On the other hand, 3-D monolithic integration, though in the early stage, is getting attention from various researchers as it promises to provide high-density 3-D circuits [3, 6]. Figure 2-a illustrates the cross sectional view of a first generation industrial 3-D monolithic technology, in which p- and n-type transistors are grown and optimized sequentially in the same process. Key benefit of this integration scheme is the reduced active footprint due to small vertical contacts in the range of few 100 nm, when compared to TSV sizes in the order of few micrometers. High-density vertical connection is a key feature of 3-D monolithic integration as it enables fine-grain (i.e., gate-level) circuit partitioning. In addition, processing n-type and p-type devices in two different layers, adds flexibility for separate technological optimization to boost their performance.

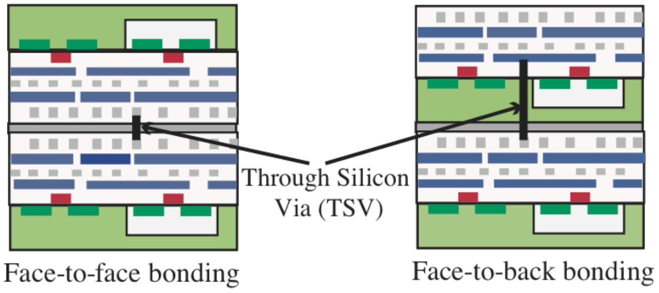
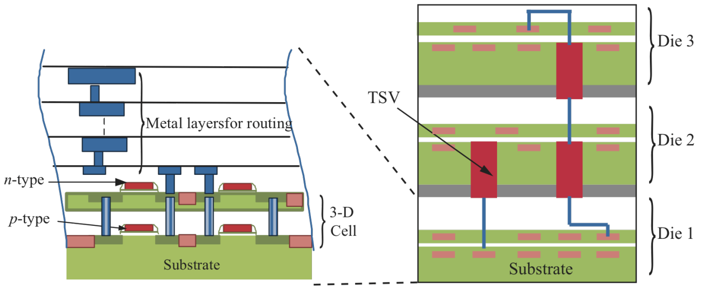


Figure 1. 3-D TSV Integration [4].



(a) (b)

Figure 2. (a) Transistor stacking with 3-D monolithic.
(b) Potential 3.5-D Integration.

In this work, we propose an idea of a possible 3.5-D integration, which leverages on the key features of both 3-D monolithic and TSV integrations. Figure 2-b illustrates the synergy between 3-D monolithic and TSV integration. 3.5-D supports fewer layers as compared to TSV-based multi-layer systems. Based on existing cost models, we conjecture that the overall manufacturing cost can be reduced by 20% and the test cost can be reduced by 30% for a case study, 288-core MPSoC, with 3.5-D technology when compared to a 3-D TSV implementation. From our simulations we also show an impact on the system level performance of the 288-core MPSoC, with an 11.5% decrease in average delay of the cores and 24% decrease in the latency of the on chip network.

The organization of the paper is as follows. Section II presents our idea about 3.5-D integration and highlights the related advantages. Cost analysis for various integrations schemes is studied in Section III. Section IV presents the performance improvement of an MPSoC by custom technology mapping of the cores and the communication network. Section V concludes the paper. In the rest of this paper, we refer to the various integration approaches using the nomenclature listed in Table I.

TABLE I. NOMENCLATURE

Integration scheme	Details
2-D	Planar technology
3-D TSV [4]	Die-stacking with TSVs
3-D n/p [3]	3-D monolithic integration with *n*-active layer over *p*-active layer
3.5-D	Synergy of 3-D n/p and 3-D TSV

II. 3.5-D INTEGRATION FOR MPSOCS

Technology mapping from planar 2-D to 3-D TSV, 3-D n/p, and 3.5-D integration schemes on an MPSoC is presented in Figure 3. For the sake of scalability, all the processing cores of the MPSoC are interconnected by a homogeneous *Network-on-Chip* (NoC) [7]. With 3-D TSV integration, the planar multi-core chip is integrated into K_1 layers (dies) employing TSVs. A vertical extension of the NoC, a 3-D NoC [8], is required for connecting all the cores across various layers. Alternatively, 3-D n/p integration "folds" the cores by fine-grain stacking of the n-type on top of the p-type transistors, thereby reducing the overall area of the core by 30% (see Section IV). In 3.5-D integration, we exploit both schemes. The 3-D n/p integration increases the integration density, thereby reducing the footprint of the processing cores. Consequently, more cores can be placed for a given die area. Several of the 3-D n/p dies are stacked with TSVs to produce a 3.5-D multi-core SoC. As depicted in Figure 3, with the same die size for 3-D TSV and 3.5-D, we observe 9-cores per die for 3.5-D when compared to 4-cores with 3-D TSV. This situation results in vertical stacking of only K_2 dies (where $K_2 < K_1$). Reduction in the number of layers affects both cost and system performance and are subsequently studied in the following sections. It has to be noted that the thermal issues for 3-D n/p integration is still under study, hence the impact of thermal issues on 3.5-D is not within the scope of this paper.

III. COST ANALYSIS

To demonstrate the effectiveness of 3.5-D integration, the manufacturing cost is analyzed for different technologies. As a case study, we investigate a 288-core homogeneous MPSoC (suited for a telecommunication system) for various integration schemes presented in Table I.

For any vertical integration approach, the bonding process contributes significantly to the overall cost. Among the possible bonding approaches *wafer-to-wafer*, *die-to-wafer*, and *die-to-die*, we consider the *die-to-wafer* as a good compromise between manufacturing throughput and yield [9]. In order to estimate the overall cost of the MPSoC, we consider appropriate costs, reported in literature, for TSV *die-to-wafer* process [9] and 3-D n/p integration [3]. In the case of 2-D integration, 5 mask sets are needed for the active layer in which both the n-type and p-type transistors are patterned. Whereas in the case of 3-D n/p integration, 6 mask sets (i.e. 3 mask sets for each active layer), as well as an additional *Silicon-On-Insulator* (SOI) layer, are needed. Taking into account the extra processing steps, the authors of [3] have reported 26% increase in the total cost when compared to planar SOI (8 metal levels 22-nm process), with an assumption of producing 30000 wafers per month. It has to be noted that, with 3-D n/p integration, we reduce the footprint of the active circuit, thereby decreasing the overall cost of the chip.

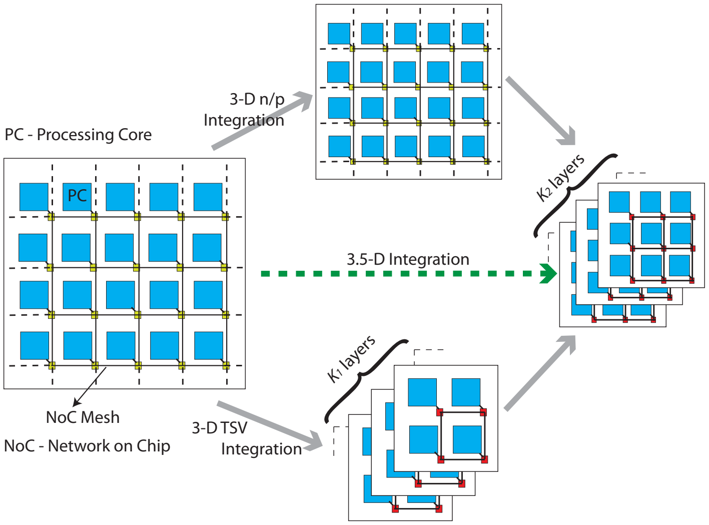


Figure 3. Technology mapping from planar 2-D to 3-D TSV, 3-D n/p and novel 3.5-D integrations.

We target the homogeneous cores presented in [10] as a good vector for scalability. The partition of the 288-core MPSoC with various integration schemes (shown in Figure 3) is reported in Table II.

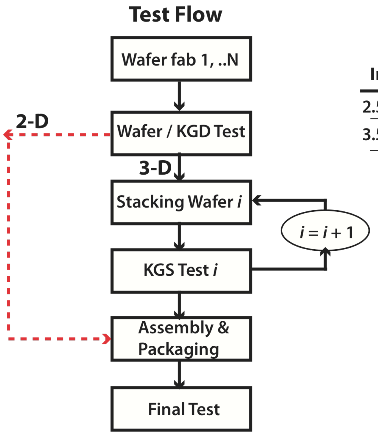
TABLE II. PARTITION OF A 288-CORE MPSOC WITH VARIOUS INTEGRATION SCHEMES

Integration	Cores/Die	Stacked dies	Die area (mm2)
2-D	16 x 18	1	318
3-D TSV	6 x 8	6	59
3-D n/p	16 x 18	1	229
3.5-D	8 x 9	4	64

The increase in the number of cores per die for 3.5-D is supported by the decrease in the active footprint of the core, offered by 3-D n/p integration. From Table II, we can observe for similar die area for 3.5-D and 3-D TSV, we obtain 72 cores/die with 3.5-D when compared to 48 cores in the case of 3-D TSV. Hence by packing more cores onto a given die area, we reduce the number of stacked dies to from 6 to 4. By considering the cost for 3-D n/p integration [3] and the cost of TSV process for Die-to-Wafer stacking [9], we conjecture the manufacturing cost of the MPSoC to be reduced by 20% for 3.5-D integration when compared to a corresponding 3-D TSV implementation.

In addition to the manufacturing cost, testing cost plays a vital role in determining the overall cost of the 3-D (vertically stacked) ICs. Figure 4 depicts test flows for 2-D and 3-D ICs [11]. A 2-D test flow has two phases, wafer test (performed on the fabricated wafer) and final test (to detect packaging faults).

On the other hand, a 3-D test flow has three phases: the *Known-Good Die* (KGD) test (also called pre-bond test), the *Known-Good Stack* (KGS) test (post-bond test), and the final test [11]. The KGD test is performed after wafer dicing, in order to determine the working dies from the wafer. The KGS test is performed once a die is stacked, in order to detect damages during the stacking process. We can observe from the test flow that both KGD and KGS costs depend on the number of dies (*N* in Figure 4). By applying the 3-D cost model to our 288-core MPSoC, we observe ~30% decrease in test cost for 3.5-D as the number of stacked dies are reduced when moving from 3-D TSV to 3.5-D integration, (6 dies for 3-D TSV and 4 dies for 3.5-D).



Integration	KGD test	KGS test
2.5-D MPSoC	6	5
3.5-D MPSoC	4	3

Figure 4. 2-D and 3-D Test flows [11].

Hence with 3.5-D integration, we conjecture 20% reduction in manufacturing cost and 30% reduction in test cost when compared to 3-D TSV integration. In the following section, we study the performance of the MPSoC when mapped to 3.5-D technology.

IV. Performance Improvement

Performance improvement offered by 3.5-D for MPSoCs is twofold. A performance increase should be noticed at the core level, as well as at the interconnect level. In this section, we first study the performance improvement at the core level and then study the performance improvement of the NoC connecting all the cores.

A. Performance Improvement of the Core

Since 3-D n/p technology offers multiple active layers adjacent to each other, the layout of the standard cell can be folded in two layers thereby forming a 3-D cell. Figure 5 shows the 2-D and the 3-D standard cells. As illustrated in Figure 5b, p-type devices (forming the pull-up network) are realized at the top active layer and n-type devices (forming the pull-down network) at the bottom active layer. By assuming the same design rules of the backend of the line (metal lines), we mapped various existing 2-D standard cell libraries to 3-D cell libraries. One of the primary advantages of this cell transformation is the ease in integration with the conventional design flow, as the design effort consists of developing only the 3-D n/p cell library [12].

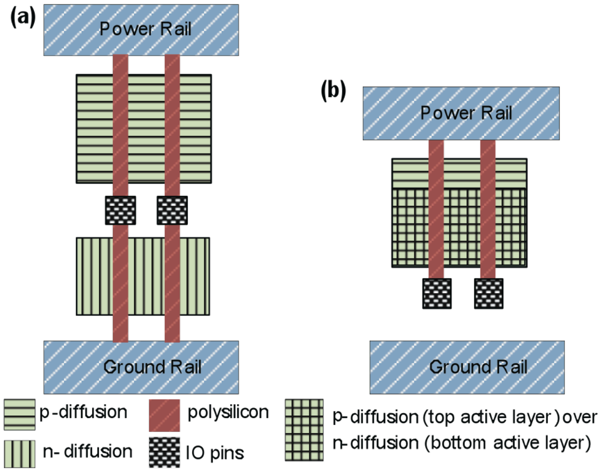


Figure 5. (a) Typical standard cell in 2-D (planar) configuration and (b) Standard cell designed in 3-D n/p by realizing the PUN on the top active layer and the PDN in the bottom active layer.

The standard cells of the 45 nm Nangate Open Cell Library [13] are mapped to the corresponding 3-D equivalent by changing the physical-attributes of the cells. For instance, the height of the cells is reduced by 30% without modifying any width. The size of the I/O pins is retained as in the case of a 2-D cell while the location is altered. Since the drive strength of the gates is not altered, we assume similar delay characteristics as in the planar case. In this study we did not take into account the parasitic extraction of the standard cells. This assumption is valid at the current technology nodes, as the overall delay is dominated by interconnect and transistor delays.

Various benchmark circuits within the core of the MPSoC are considered [14]. Synopsys Design Compiler is used for mapping the RTL of the benchmarks onto target 3-D standard cell library. Cadence Encounter is used as the physical synthesis engine to generate the virtual seed placement in wirelength driven mode. In Figure 6, we show the improvement in wirelength, delay, and power of various benchmark circuits after placement is performed using the 2-D and 3-D n/p cell libraries. The power numbers include all components of the power dissipation namely leakage and switching power. By partitioning the circuit at the gate-level, with 3-D n/p integration, the active-area footprint is reduced

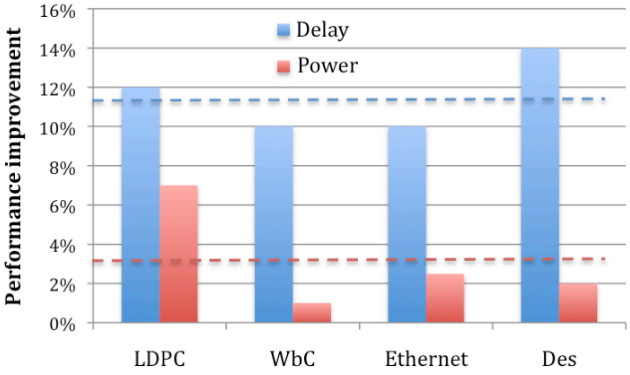


Figure 6. Performance improvement in terms of delay and power of various blocks of the core.

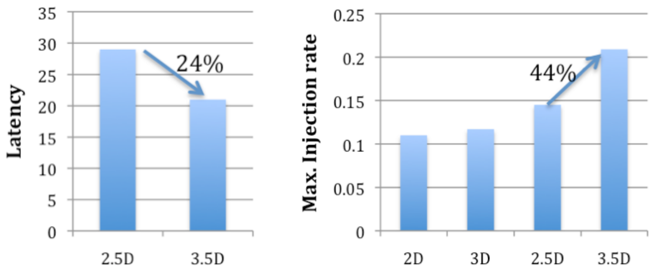
thereby leading to reduction in the average interconnect wirelength of the circuit. This reduction in interconnect wirelength mainly attributes for performance improvement from 2-D to 3-D n/p integration. We observe 11.5% delay improvement and 3.4% power improvement when averaged across various benchmark circuits shown in Figure 6.

B. *Performance Improvement of the NoC*

To assess the system-level performance of the different designs, from the communication point of view, we use a cycle accurate NoC [15]. The simulator assumes best effort NoC architecture similar to that described in the xpipes library [16]. We assume wormhole flow control with input buffered switches, that use round-robin arbitration and ON/OFF flow control [17]. Without loss of generality the arbitration and crossbar switching are done in one cycle. There are no output buffers in the switches, but pipeline stages are placed on long links in order to achieve the required operating frequency. We inject packets that are 10 *flow control units* (flits) long.

For our case study, we generate different mesh and 3D-mesh topologies that correspond to the different partitioning of the cores according to the various integration schemes as presented in Table II. We simulate for different injection rates to assess the latency of the packets and the actual possible injection rates for the different NoC configurations. We inject uniform random traffic and for each configuration we perform simulations for 100000 cycles. First, we study the latency of the NoC at a constant injection rate of 0.1. When compared to planar implementation, with a 2-D NoC (mesh size 16x18), the latency is reduced by 57% and 68% for the 3-D NoC connecting the 288-cores in 3-D TSV configuration (with mesh size 6x8x6) and 3.5-D configuration (with mesh size 8x9x4) respectively. Given a ~60% reduction in latency of a 3-D NoC when compared to a 2-D NoC, in Figure 7a, we only show the latency for the two relevant cases of the 3-D NoC. We observe 24% decrease in latency of the NoC for 3.5-D configuration when compared to 3-D TSV configuration.

Next, we study the maximum injection rate possible for various configurations. The injection rates are the values that actually affect the end-to-end NoC latency, hence the maximum injection rate gives the best performance of the NoC. In Figure 7b, we show 44% improvement in injection rate from 3-D TSV to 3.5-D MPSoC implementation.



(a) (b)

Figure 7. (a) Latency at constant injection rate of 0.1.
(b) Maximum injection rate possible for various technologies.

V. Conclusion

In this work, we propose a novel vertical integration scheme, called 3.5-D integration, which synergizes existing TSV-based and 3-D monolithic technologies. The feature of gate-level stacking with 3-D n/p integration is leveraged to stack more cores onto a die, when compared to a straightforward 3-D TSV integration. With 3.5-D integration, the number of stacked dies is reduced by 30% when compared to a 3-D TSV implementation. Based on existing cost models for various technologies, we conjecture that the overall manufacturing cost can be reduced by 20% and the test cost can be reduced by 30% for a case study, 288-core MPSoC, with 3.5-D technology when compared to a 3-D TSV implementation. From our simulations, we show performance improvement of 11.5% (on an average) for various benchmarks comprised in the core. In the case of interconnection network, we observe large improvement in the latency of the 3-D NoC (average of 24%) for 3.5-D implementation when compared to 3-D TSV implementation of the MPSoC.

Acknowledgments

The authors would like to thank P. Batude from CEA-LETI for her insight on 3-D monolithic technology. This research was supported in part by the European Research Council grant (ERC-2009-AdG-246810).

References

[1] V. F. Pavlidis and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*. Morgan Kaufmann, 2009.

[2] Jan Van Olmen *et al.*, "3D Stacked IC Demonstration using a Through Silicon Via First Approach," *Proceedings IEEE International Electron Devices Meeting*, pp. 1–4, December 2008.

[3] P. Batude *et al.*, "Advances in 3D CMOS Sequential Integration," *IEDM* 2010.

[4] Y. Deng *et al.*, "2.5D System Integration: A Design Driven System Implementation Schema," *ASP-DAC,* 2004.

[5] G. H. Loh, Y. Xie, and B. Black. "Design in 3D Die Stacking Technologies," *IEEE Micro Magazine*, 27(3), May–June 2007.

[6] www.monolithic3d.com

[7] L. Benini and G. De Micheli, "Networks on Chips: A New SoC Paradigm," *IEEE Computers*, pp. 70-78, Jan. 2002.

[8] C. Seiculescu *et al.*, "SunFloor 3D: A tool for Networks On Chip topology synthesis for 3D systems on chips," *Design, Automation & Test in Europe Conference & Exhibition, 2009,* pp.9-14, April 2009.

[9] X. Dong *et al.*, "System-Level Cost Analysis and Design Exploration for Three-Dimensional Integrated Circuits (3DICs)," *ASP-DAC*, 2009.

[10] C. Jalier *et al.*, "Heterogeneous vs Homogeneous MPSoC Approaches for a Mobile LTE Modem," *DATE*, 2010.

[11] E.J. Marinissen and Y. Zorian, "Testing 3D chips containing through-silicon vias," *International Test Conference, 2009, ITC 2009*, pp.1-11, Nov. 2009.

[12] S. Bobba *et al.*, "CELONCEL: Effective Design Technique for 3-D Monolithic Integration targeting High Performance Integrated Circuits," *ASP-DAC*, 2011.

[13] www.nangate.com

[14] www.opencores.org

[15] C. Seiculescu *et al.*, "CCNoC: On-Chip Interconnects for Cache-Coherent Manycore Server Chips", *WEED*, 2011.

[16] S. Stergiou *et al.*, "×pipesLite: a Synthesis Oriented Design Library for Networks on Chips," pp. 1188-1193, Proc. DATE 2005.

[17] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann, 2004.