A 10.35 mW/GFlop Stacked SAR DSP Unit using Fine-Grain Partitioned 3D Integration

Thorlindur Thorolfsson*†, Steve Lipa* and Paul D. Franzon*
*Department of Electrical & Computer Engineering, North Carolina State University, Raleigh, NC 27695
Email: trthorol@ncsu.edu and paulf@ncsu.edu
†Synopsys, Inc.

***Abstract*—In this paper we present a technique for implementing a fine-grain partitioned three-dimensional SAR DSP system using 3D placement of standard cells where only one of the 3D tiers is clocked to reduce clock power. We show how this technique was used to build the first fine-grain partitioned 3D integrated system to be demonstrated with silicon measurements in the literature, which is an ultra efficient floating-point synthetic aperture radar (SAR) DSP processing unit. The processing unit was fabricated in two tiers of GlobalFoundries, 1.5 V 130nm process that were 3D stacked face-to-face by Tezzaron. After fabrication the test chip was measured to consume 4.14 mW of power while running at 40 MHz operating for an operating efficiency of 10.35 mW/GFlop.**

I. Introduction and Related Works

New developments in through-silicon via (TSV) fabrication, wafer alignment, thinning and bonding, allow the 3D integration of stacked dies, enabling the design of 3D integrated systems of various levels of integration. The different levels of integration can be roughly divided into three categories.

First, system level 3D integration. At this level of integration, systems typically integrate dies manufactured in different process technologies (often logic and memory process) and take advantage of the fact that 3D integration virtually allows the use of two different manufacturing processes in one IC, along with more input/output pins and the reduction of interconnect parasitics. A good example of this level of 3D integration is the logic-on-memory 3D integration presented by Zhang et al.[1], which uses 3D integration to provide data at a very high data rate (4.25GB/s) from the DRAM to the logic.

The second level of integration is block-level integration. At this level of integration blocks are placed in different 3D tiers using 3D floorplanning techniques to build a more tightly integrated system, such as the block level 3D IC design presented by Kim et al. [2] or the 3D aware floorplanning with fixed outline constraints presented by Xiao et al. [3] or several others [4], [5]. Examples of systems using this level of integration include the 3D-Maps system[6], [7], which features tiles of processor and SRAM blocks tightly integrated and the NoC 3D system presented by Mineo et al. [8].

The final level of integration is fine-grain partitioned 3D integration[9], [10], also known as intra-block level integration. At this level of integration individual blocks exist in more than one tier of silicon. Although, there has been a significant amount of research work done on the tool side in 3D standard cell placement[11], [12], [13], [14], [15], [16], the work presented in this paper demonstrates the first working fine-grain partitioned 3D integrated system with silicon measurements in the litterature.

The remainder of the paper is organized in the following manner. Section II describes the 3D standard placement technique. Section III describes the architecture of the test chip. Section IV details the measurement results of the test chip. Section V contains the results of thermal simulation of the test chip and Section VI concludes the paper.

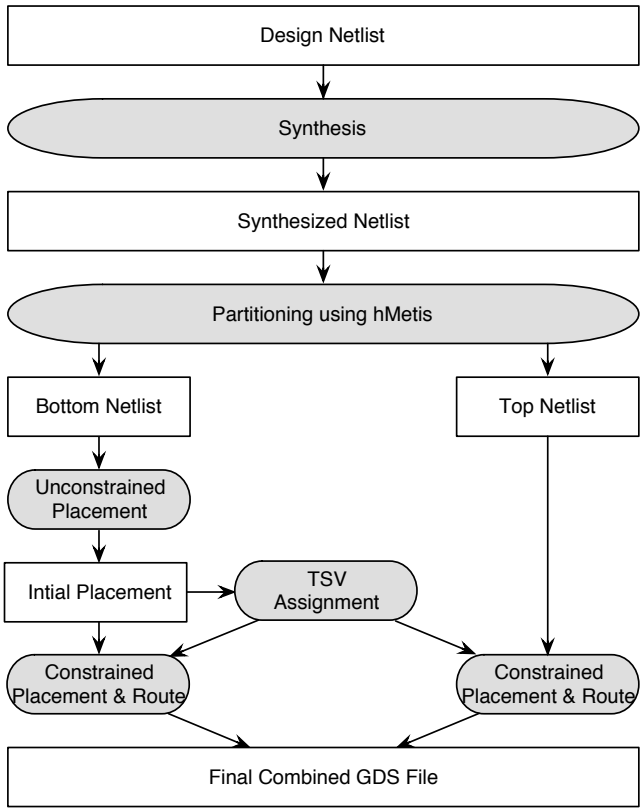

Fig. 1: The design flow for 3D standard cell placement.

II. 3D Standard Cell Placement

The 3D standard cell placement technique used to implement the test chip assumes a 3D stack-up of two tiers face-to-face. In this case the connectivity between the two tiers is through microbumps and the off-chip connectivity is through TSVs.

A significant advantage of using microbumps is that unlike TSVs they do not require a keep out region for logic cells allowing a greater interconnect density. The technique works in the following manner. First, a hypergraph representation of the synthesized netlist is generated. This representation is then partitioned into two groups that have a similar total cell area and a minimum number of signals crossing between the two groups (one group is for the top tier, the other for the bottom tier). In this partitioning, all the standard cells that use the clock are placed in the bottom partition. This serves two purposes. First, it reduces the area that the clock grid has to cover which in turn reduces the total clock power. Second, it decreases the effects of process variation between the two stacked dies on the clock tree.

After partitioning, 3D placement is completed using a series of three discrete placements. First, a rough placement or "unconstrained" placement is generated for the bottom tier. This placement is considered "unconstrained" because it does not consider the location of the input and output pins as constraints during placement. This placement is only used to determine which signal is assigned to which microbump. The actual assignment is then completed using an assignment algorithm[17] that minimizes the sum of the distances from the standard cells that drive inter-tier signals to the microbumps that carry the signals. Final placement of both the top and bottom tiers is then performed using the microbump to signal assignment being used to constrain the input and output locations for placement of the top tier and the final placement of the bottom tier. The diagram in Figure 1 shows an overview of the approach.

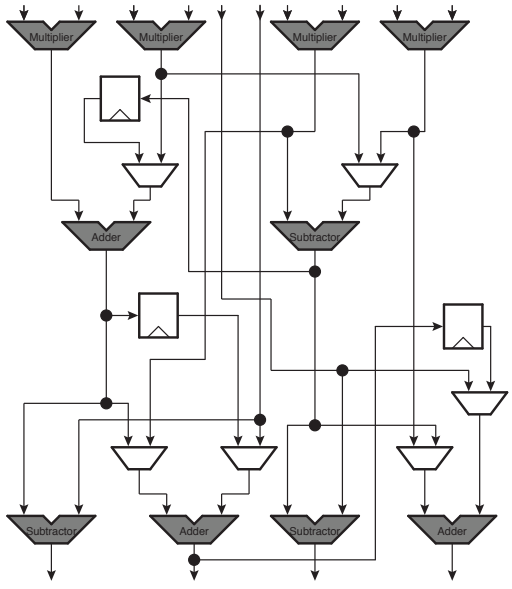

Fig. 2: Architecture of the SAR DSP processing unit.

III. SAR Processor Unit Architecture

The circuit used to demonstrate the fine-grain partitioned 3D integration is a synthetic aperture radar (SAR) processing unit. In the application of a SAR system it is most important to minimize the of number milliwatts required per GFlop of processing power. Figure 2 shows the architecture in inside the SAR processing unit (the flip-flops shown are FIR filter taps). Overall, the SAR DSP processing unit contains 10 basic 32-bit floating-point arithmetic units (4 multipliers, 3 adders and 3 subtractors) and a reconfigurable data-path between them. By reconfiguring the data-path the 10 basic units can be used to implement the four DSP operations (FFT, IFFT, FIR filtering and complex multiplication) that are required for SAR image formation.

Additionally, in order to achieve the lowest mWatt per GFlop and to decrease the power consumed by the clock tree and flip-flops minimal pipelining is used. Although, this reduces the maximum operating frequency because it makes the critical path longer it significantly reduces the number of milliwatts required per GFlop of computation.

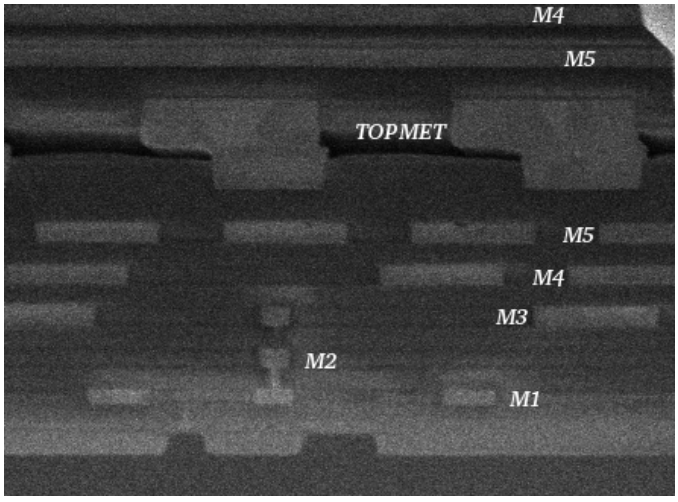

Fig. 3: A cross section of the 3D stack-up showing the microbumps (labeled TOPMET) and metal layers of the face-to-face stacking.

IV. Test Chip Measurement Results

The architecture described in Section III and implemented using the 3D standard cell placement technique described in Section II was fabricated on a test chip. The test chip was implemented in two tiers of Global Foundries 130nm process, stacked together face-to-face using Tezzaron's Copper-to-Copper thermo-compression bonding technique[18]. Each tier contains 5 metal layers and one layer of poly-silicon. The connectivity between the two tiers consists of 4.4 μm by 4.4 μm copper micro-bumps (labeled TOPMET in Figure 3) that are fixed on a 5.0 μm grid. Off-chip connectivity to a given signal is then accomplished by using a bundle of 23 individual TSVs (1.2 by 1.2 μm each) that connect to a copper backmetal bond pads, the bottom of which is shown in Figure 5.

The core power consumption of the processing unit was measured to be 4.14 mW when running at 40 MHz, with a supply voltage of 1.5 V, which translates to an overall efficiency

TABLE I: Comparisons to other works.

Metric	This Work	Vangal [19]	Oh [20]	Arakawa [21]	Aoki [22]	Nam [23]
Efficiency (mW/GFlop)	10.35	194	43.75	89.28	81.58	10.18
Performance (GFlops)	0.4	6.2	32.0	2.8	3.2	2.8
Power (mW)	4.41	1200	1400	250	261	28.5
Frequency (MHz)	40	3100	5600	400	400	200
Process (nm)	130	90	90	130	90	180

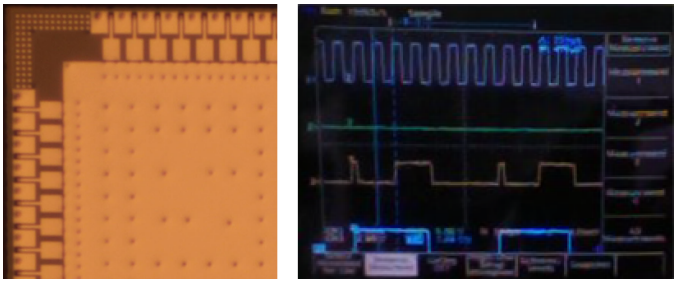

Fig. 4: Die photo and measurement waveform photograph.

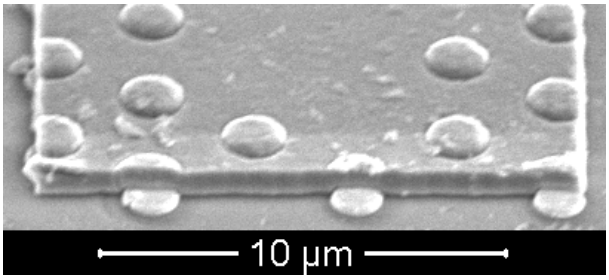

Fig. 5: Close-up of 12 TSVs and backmetal interface.

of 10.35 mW/Gflop (die-photo and wirebonded die shown in Figure 4 and 6). Comparisons to other works are in Table I.

V. Thermal Simulation Results

Thermal issues are exacerbated in 3DICs because the worst case thermal path (from the farthest tier to the heatsink) is more resistive than in a 2D system. To obtain the thermal profile of the 3D stacked processing unit, we perform a thermal simulation of the design using Gradient HeatWave-3DIC. Since this is a low-power, high efficiency design, the temperature rise will be minimal in the simulation results. In order to explore the issue in broader terms, we also simulate the temperature for two additional hypothetical scenarios. In these scenarios the system is running at 10x and 100x the speed of that it was designed for. The junction temperatures of the top tier (the tier that is further away from the heatsink) are shown in Figure 7 for all three scenarios. It is interesting to note how drastically hotter the top tier is due to its distance from the heatsink and the cooling effects of the dummy TSV's, which can be seen in the regularly-spaced circular indentations. Overall, however the simulations shows that at least for low-power systems, high thermal gradients should not be a concern for fine-grain partitioned 3D integrated systems.

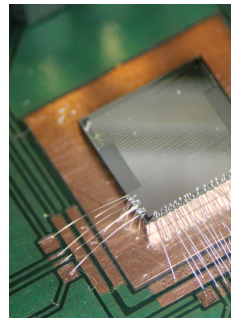

Fig. 6: Photo of wirebonded die

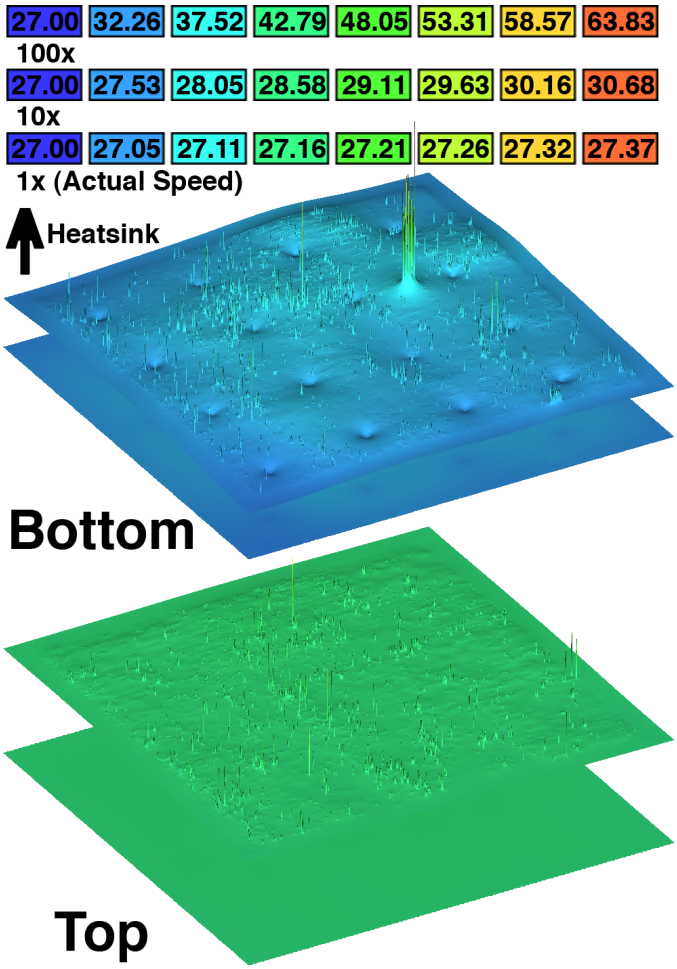

Fig. 7: Thermal simulation of test chip.

VI. Conclusion

In this papers we have demonstrated a technique for 3D standard cell placement that includes a novel approach to keeping all clocked cells one one tier. We have shown how this technique was used to create the first fine-grain partitioned 3D integrated system in the literature with silicon measurements. A SAR DSP chip with a measured operating efficiency of 10.35 mW/Gflop which compares favorably to similar works as demonstrated in Table I and summarized in Table II . Additionally, in this paper we have shown how a low power systems such as this the SAR DSP avoid the thermal pitfalls typically associated with high-performance 3D integrated systems with stacked dies.

Although the benefits of system-level 3D integrated system such as DRAM on logic are more easily attainable and more immediate, we believe that there is a bright future for fine-grain partitioned 3D integrated systems especially for high efficiency and low power systems like the SAR DSP processing unit that was presented. However, this future will depend on advances in microbump and TSV manufacturing, feature size and cost.

TABLE II: 3D Integrated SAR DSP Processor Summary.

Technology	130nm CMOS
Wiring	2 x (1P5M) + BM
Transistors	149,936
Test Circuit Area	0.3104 mm^2
Die Size	5 mm x 5mm
Power Supply	1.5 V
Frequency	40 MHz
Core Power	3.521 mW

Acknowledgments

We thank Matthew Craver and Neil Di Spigna for their help with the assembly of the printed circuit boards and the test setup, Adi Srinivasan at Gradient Design Automation for help with the thermal simulation, along with Gary Yeap at Synopsys for his advice and support. This work was supported by Semiconductor Research Corporation along with the MARCO Musyc Center and GSRC Center, by DARPA under contract FA8650-04-C-7127 and contract FA8650-04-C-7120.

References

[1] T. Zhang, K. Wang, Y. Feng, X. Song, L. Duan, Y. Xie, X. Cheng, and Y.-L. Lin, "A customized design of dram controller for on-chip 3d dram stacking," in *CICC, 2010 IEEE*, sept. 2010, pp. 1 –4.

[2] D. H. Kim, R. Topaloglu, and S. K. Lim, "Block-level 3d ic design with through-silicon-via planning," in *Design Automation Conference (ASP-DAC), 2012*, 30 2012-feb. 2 2012, pp. 335 –340.

[3] L. Xiao, S. Sinha, J. Xu, and E. Young, "Fixed-outline thermal-aware 3d floorplanning," in *Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific*, jan. 2010, pp. 561 –567.

[4] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004*, 2004, pp. 306–313.

[5] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3d architectures," *J. Emerg. Technol. Comput. Syst.*, vol. 2, no. 2, pp. 65–103, 2006.

[6] M. Healy, K. Athikulwongse, R. Goel, M. Hossain, D. Kim, Y.-J. Lee, D. Lewis, T.-W. Lin, C. Liu, M. Jung, B. Ouellette, M. Pathak, H. Sane, G. Shen, D. H. Woo, X. Zhao, G. Loh, H. Lee, and S. K. Lim, "Design and analysis of 3d-maps: A many-core 3d processor with stacked memory," in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, sept. 2010, pp. 1 –4.

[7] D. H. Kim, K. Athikulwongse, M. Healy, M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y.-J. Lee, D. Lewis, T.-W. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. H. Woo, X. Zhao, J. Kim, H. Choi, G. Loh, H.-H. Lee, and S. K. Lim, "3d-maps: 3d massively parallel processor with stacked memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, feb. 2012, pp. 188 –190.

[8] C. Mineo, R. Jenkal, S. Melamed, and W. Davis, "Inter-die signaling in three dimensional integrated circuits," Sept. 2008, pp. 655–658.

[9] Y. Liu, Y. Ma, E. Kursun, G. Reinman, and J. Cong, "Fine grain 3d integration for microarchitecture design through cube packing exploration," in *Computer Design, 2007. ICCD 2007. 25th International Conference on*, oct. 2007, pp. 259 –266.

[10] Y. Ma, Y. Liu, E. Kursun, G. Reinman, and J. Cong, "Investigating the effects of fine-grain three-dimensional integration on microarchitecture design," *J. Emerg. Technol. Comput. Syst.*, vol. 4, no. 4, pp. 17:1–17:30, Nov. 2008.

[11] R. Hentschke, G. Flach, F. Pinto, and R. Reis, "Quadratic placement for 3d circuits using z-cell shifting, 3d iterative refinement and simulated annealing," in *SBCCI '06: Proceedings of the 19th annual symposium on Integrated circuits and systems design*. New York, NY, USA: ACM, 2006, pp. 220–225.

[12] S. Das, A. Chandrakasan, and R. Reif, "Design tools for 3-d integrated circuits," in *ASPDAC 2003 conference on Asia South Pacific design automation*. New York, NY, USA: ACM, 2003, pp. 53–56.

[13] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3d ics using a force directed approach," *Computer Aided Design, 2003. ICCAD-2003. International Conference on*, pp. 86–89, Nov. 2003.

[14] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3d ic placement via transformation," in *Design Automation Conference, 2007. ASP-DAC '07. Asia and South Pacific*, Jan. 2007, pp. 780–785.

[15] J. Cong and G. Luo, "A multilevel analytical placement for 3d ics," in *ASP-DAC '09: Proceedings of the 2009 Asia and South Pacific Design Automation Conference*. Piscataway, NJ, USA: IEEE Press, Jan. 2009, pp. 361–366.

[16] Y. Deng and W. Maly, "2.5d system integration: a design driven system implementation schema," in *Design Automation Conference, 2004. ASP-DAC 2004. Asia and South Pacific*, jan. 2004, pp. 450 – 455.

[17] T. Thorolfsson, N. Moezzi-Madani, and P. Franzon, "Reconfigurable five-layer three-dimensional integrated memory-on-logic synthetic aperture radar processor," *Computers Digital Techniques, IET*, vol. 5, no. 3, pp. 198 –204, may 2011.

[18] R. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214–1224, June 2006.

[19] S. Vangal, Y. Hoskote, N. Borkar, and A. Alvandpour, "A 6.2-gflops floating-point multiply-accumulator with conditional normalization," *JSSC*, vol. 41, no. 10, pp. 2314–2323, Oct. 2006.

[20] H.-J. Oh, S. Mueller, C. Jacobi, K. Tran, S. Cottier, B. Michael, H. Nishikawa, Y. Totsuka, T. Namatame, N. Yano, T. Machida, and S. Dhong, "A fully pipelined single-precision floating-point unit in the synergistic processor element of a cell processor," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 4, pp. 759–771, April 2006.

[21] F. Arakawa, T. Yoshinaga, T. Hayashi, Y. Kiyoshige, T. Okada, M. Nishibori, T. Hiraoka, M. Ozawa, T. Kodama, T. Irita, T. Kamei, M. Ishikawa, Y. Nitta, O. Nishii, and T. Hattori, "An embedded processor core for consumer appliances with 2.8gflops and 36m polygons/s fpu," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, Feb. 2004, pp. 334–531 Vol.1.

[22] H. Aoki, T. Kawahara, M. Yamaoka, C. Yoshimura, Y. Nagasaka, K. Takayama, N. Sukegawa, Y. Fukumura, M. Nakahata, H. Sawamoto, M. Odaka, T. Sakurai, and K. Kasai, "A powerful yet ecological parallel processing system using execution-based adaptive power-down control and compact quadruple-precision assist fpus," in *VLSI Circuits, 2008 IEEE Symposium on*, june 2008, pp. 186 –187.

[23] B.-G. Nam and H.-J. Yoo, "A 28.5mw 2.8gflops floating-point multifunction unit for handheld 3d graphics processors," in *Solid-State Circuits Conference, 2007. ASSCC '07. IEEE Asian*, Nov. 2007, pp. 376–379.