Three DIMENSIONAL-CHIPS

[1]Kumar.Keshamoni, [2]Mr. M. Harikrishna

[1]*Asst.Professor in ECE Dept of RVR Institute of Engineering and Technology, Hyderabad*
[2] *Professor, Head of the Dept.of ECE RVR Institute of Engineering and Technology, Hyderabad*

Abstract: This paper illustrates the performance advantages of 3D integrated circuits with two specific examples, namely 3D-FPGA and 3D-SRAM. Three-dimensional Chip (3D IC, 3D-IC, or 3-D IC) is a chip in which two or more layers of active electronic components are integrated both vertically and horizontally into a single circuit. The semiconductor industry is pursuing this promising technology in many different forms, but it is not yet widely used. Through strategic modification of the architectures to take advantage of 3D, significant improvement in speed and reduction in power consumption can be achieved. The unprecedented growth of the computer and the Information technology industry is demanding Very Large Scale Integrated (VLSI) circuits with increasing functionality and performance at minimum cost and power dissipation. VLSI circuits are being aggressively scaled to meet this Demand, which in turn has some serious problems for the semiconductor industry. Additionally heterogeneous integration of different technologies in one single chip (SoC) is becoming increasingly desirable, for which planar (2-D) ICs may not be suitable.3-D ICs are an attractive chip architecture that can alleviate the interconnect related problems such as delay and power dissipation and can also facilitate integration of heterogeneous technologies in one chip (SoC). The multi-layer chip industry opens up a whole new world of design. With the Introduction of 3-D ICs, the world of chips may never look the same again

Keywords: *SoC, SRAM, FPGA, isolating, Stacking, LBL, architecture, 2D-IC, inter-layer, ASIC*

I.   Introduction

Three-dimensional integrated circuits (3D-ICs) have been studied since the 1980s. This paper explores 3D integration as a supplement to scaling. 3D-IC promises to offer multiple advantages over conventional 2D-IC, including alleviating the communication bottle- neck, integration of heterogeneous materials, and enabling novel architectures. If the 3D-IC is simply a stacking of the 2D circuit blocks with no significant modification in architecture, the gain in performance will be very limited, if any. A strategy in architecture and function partitioning across layers must be developed to take advantage of the third dimension while managing the overall complexity. The performance advantages of 3D architectures will be illustrated with two examples: 3D-FPGA and 3D-SRAM. There are at least three approaches to realize 3D IC's: chip stacking, wafer stacking, and full monolithic integration. Each approach is at different level of maturity and offers various degrees of improvement. An important difference between these approaches is the size of the inter-layer via, ranging from tens of microns for chip stacking to tens of nano meters for monolithic integration. The density of inter-layer via directly impacts the choice of 3D-architecture.

II.   3D-FPGA

The design and prototyping costs of cell-based ASIC have become prohibitive, making FPGAs increasingly popular. However, FPGAs, when compared with cell-based ASICs, have 10−40 time's lower logic density, 3−4 time's higher delay, and 5−12 times higher dynamic power dissipation. 3D integration can help close this performance gap in several ways.

A) Heterogeneous Stacking

A significant fraction of the area in a modern FPGA is occupied by hard IPs, such as memory blocks, microprocessors, and DSP blocks. Using wafer stacking, this IPs can be stacked on top of the FPGA fabric. This reduces the FPGA footprint, resulting in shorter interconnects, lower delay, and lower power. The performance benefits of this approach, however, are limited by the fraction of FPGA area occupied by IPs.

B) Homogeneous Stacking

In this approach, multiple identical FPGA fabrics are stacked using wafer stacking and their switch boxes are vertically connected using through silicon via (TSVs) [5]. Simple analysis of this scenario shows that with TSV pitch of 3−5 times that of inter-metal via in the base CMOS technology, 4−6 fabric layers can be stacked with small area overhead. How- ever, power dissipation in the intermediate layers and TSV parasitic may limit the potential performance benefits of this approach. This approach also entails significant modifications to the FPGA routing architecture and associated CAD tools.

c) Programming Overhead Stacking

FPGAs have lower performance than cell-based ASICs because around 80% of the FPGA area is occupied by programming overhead, including the configuration memory, the interconnect multiplexers and switches. This approach requires monolithic 3D integration because of the high density of inter-layer via required. The study in [6] has shown that such stacking cans achieve3.2 times higher logic density, 1.7 times lower delay, and 1.7 times lower dynamic power consumption than a baseline2D-FPGA implemented in 65nm CMOS technology. These improvements are achieved with appropriate optimization of buffer and transistor sizes, but without any change to the FPGA architecture. In a subsequent study [7], we explored architectural modification to take advantage of 3D. In particular, by merging the connections boxes and switch boxes into a switch layer on top of the logic boxes, as illustrated in Fig.1. The memory layer is split into two layers to provide better local vertical connectivity and relax the requirement on memory cell size. As shown in Fig. 2, the benchmark circuits achieve improvement of 1.7 to 2.9 times in critical path delay, and reduction of 2.5 to 3.2 times in dynamic power.
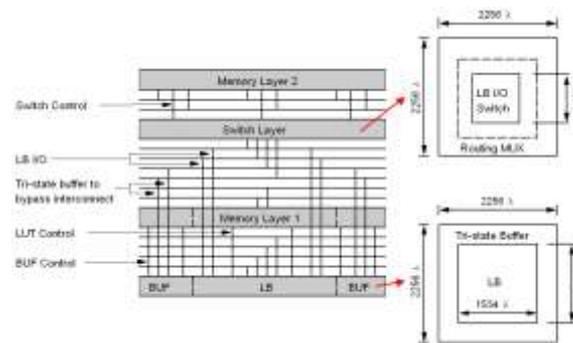

Fig. 1. Monolithically-stacked 3D-FPGA.
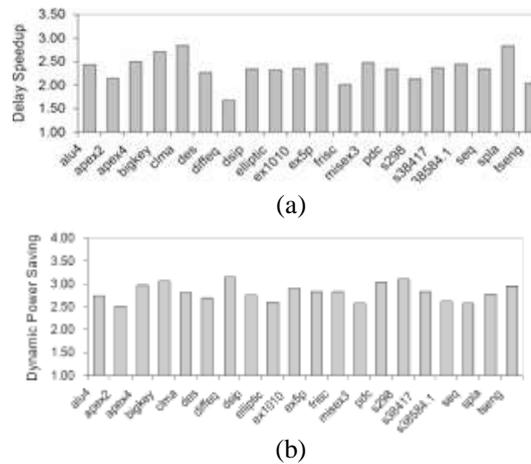

(a)

(b)

Fig. 2. (a) Speed up of critical path delay and (b) reduction of power consumption of 3D-FPGA over 2D baseline for various benchmark circuits.

The stacking approaches discussed above can be combined to achieve performance that approaches that of 2D cell-based ASICs. The idea is to use monolithic stacking to reduce the programming overhead, and to use homogeneous and heterogeneous wafer stacking to achieve further reduction in interconnect length. Additional benefits can be obtained by changing the basic fabric architecture to take full advantage of 3D.

III. 3D-SRAM

The bit-line delay usually constitutes the majority of the total access time of SRAM. A major portion of the active power dissipation is also associated with the bit-line because a large number of bit lines discharge every time a word-line is asserted. Hierarchical bit-line architecture [8] reduces the total bit-line capacitance by isolating the cell junction capacitances from the global bit-lines. However, the overall reduction in bit-line capacitance is limited because a larger portion of the bit-line capacitance comes from the metal coupling, which is proportional to the length and hence is virtually unchanged for the same number of cells per bit-line.

In our proposed 3D-SRAM architecture [9] shown in Fig. 3, the local bit-lines extend upward, through an inter-layer via that connects SRAM cells vertically. The local bit-line connects through a select transistor to the global bit-line routed on the bottom layer. The overall bit-line capacitance can be reduced significantly because the length of the global bit-line is reduced by a factor of the number of layers. Thus this SRAM array will be denser, faster, and lower power than a conventional design.
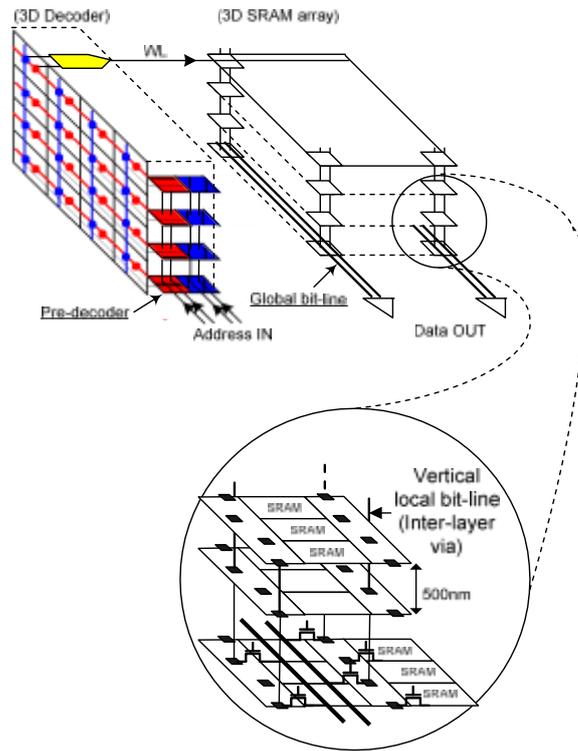

Fig. 3. 3D-SRAM architecture.

The SRAM cell is similar to a conventional 2D-SRAM cell and hence does not require sophisticated 3D technology [10]. Select transistors that connect the vertical local-bit-line and global bit-line are located on the first layer between SRAM cells. If the inter-layer vias are assumed to be similar in size to that of inter-metal vias, the area overhead will only be 18% per cell. Area efficiency is maximized by reusing this area for inter-layer vias in the upper layers. Fig. 4 compares the total bit-line capacitance versus the number of active layers for the 3D architecture, and the corresponding number of cells per local bit-line (LBL) in the 2D-hierarchical (2D-H) architecture. Although the simulation shows that using 8 layers yields the best result, using 4 layers achieves the best tradeoff between performance and complexity. With 4 layers, there is 3.4 times reduction in capacitance compared to 2D-SRAM, and 2.4 times reduction compared to 2D-hierarchical SRAM. Bit-line delay also follows the trend of the bit-line capacitance.   In 3D, the delay is reduced by about 1.8 times compared to 2D-SRAM, and 1.7 times com- pared to a corresponding 2D-hierarchical SRAM.
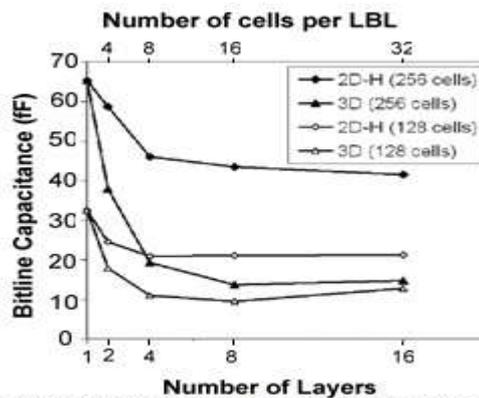

Fig. 4. Comparison of bit-line capacitances in 2D, 2D-hierarchical (2D-H) and 3D-SRAM.

In most 3D technologies, the inter-layer via needs to be enlarged and a landing pad included to accommodate the misalignment between layers. Fig. 5 shows the bit-line capacitance and delay versus the size and alignment accuracy of inter-layer via. As expected, the performance will degrade as the inter-layer via size increase or alignment accuracy worsens. For reference, simulation results of 2D-hierarchical SRAM are shown as horizontal dotted lines. It is important to note that even if the inter-layer via size is relaxed to 0.3μm and alignment accuracy to 0.25μm, the performance remains un- changed.   This is because the area overhead is originally limited by the select transistors, and consequently, slight increase in the inter-layer size or misalignment does not significantly affect the overall capacitance.   However, if the misalignment degrades to more than 0.5 μm, the benefits start to vanish.   Note that inter-layer vias with size of 0.14μm and pitch of 0.40μm have been demonstrated [11]. Thus our 3D-SRAM architecture can be fabricated with currently available most advanced wafer-to-wafer bonding technology and will still achieve great performance improvement. Of course, the inter-layer via needs to scale correspondingly with the technology node to derive similar benefits in the future.
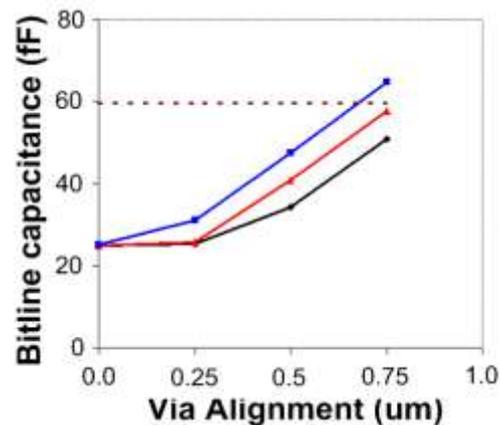
(a)
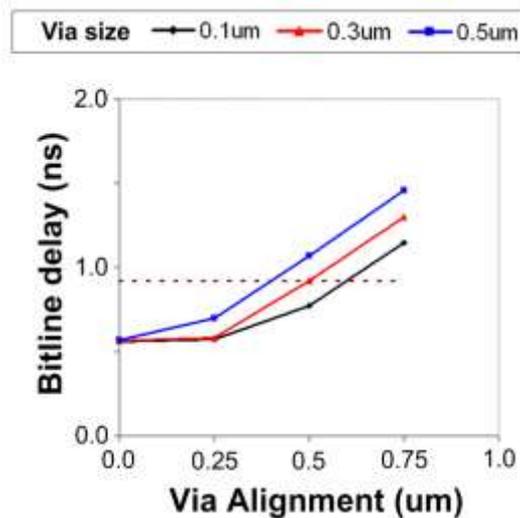
(b)

Fig. 5. Simulated bit-line (a) capacitance and (b)   delay of 3D-SRAM vs. inter-layer via size and alignment accuracy. The horizontal dotted line is the   reference for 2D-hierarchical SRAM.

A 3D row decoder that derives the maximum advantage of 3D is also illustrated in Fig. 3. There are two sets of pre-decoders, one to decode the "layer" information and the other for the "row" information. These pre-decoders are distributed on every layer and operate concurrently. An "AND" logic then drives the chosen word-line in the selected layer. Since only the pre-decoders for the selected layer and row need to be activated, the worst-case capacitance is reduced to 1/n of the corresponding 2D design, where n is the number of active layers used. Speed improvement and power reduction are expected. A 32-kbit proof-of-concept macro using 0.13-μm CMOS technology to emulate the 3D-SRAM architecture has been demonstrated. The 3D-SRAM is

arranged as an array of 4 layers × 32 word-lines × 256 bit-lines. As illustrated in Fig.6, a vertical slice of 4-layer 3D-SRAM cells furthest from the address inputs is transformed into a 2D design and the inter-layer via is emulated by two upper level metal vias. This allows us to evaluate the worst-case delay and power of the 3D array. A conventional 2D-SRAM block with the same capacity has also been fabricated for performance comparison.

The energy-delay characteristics of the SRAM blocks are summarized in Fig. 7.The 3D-SRAM offers about 5 times improvement in energy-delay product over 2D-SRAM. The experimental results verify the 65nm projections discussed earlier.
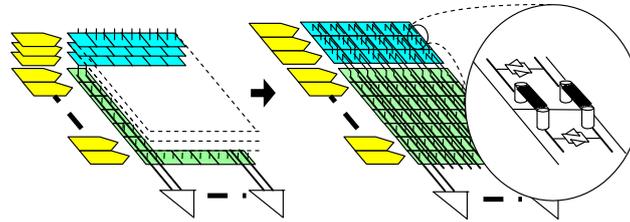

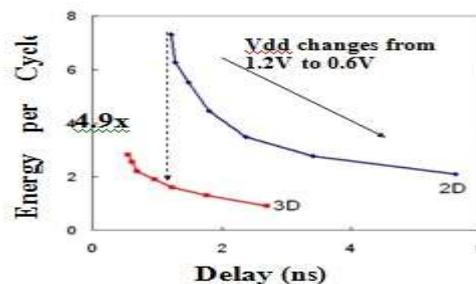Fig. 6. Emulating 3D-SRAM architecture with 2D.

Fig. 7. Energy-delay characteristic of SRAM blocks

IV.   Conclusions

We show that a monolithically stacked 3D-FPGA can achieve about 2.5 times reduction in critical path delay and 2.9 times reduction in dynamic power consumption over a conventional 2D-FPGA. A 3D-SRAM can achieve about 5 times reduction in power-delay product over conventional 2D-SRAM. Other 3D design examples include 3D-imager, in which a photo-diode layer is bonded with a CMOS electronics layer [12], and 3D one-time programmable (OTP) memory, in which 8 layers of OTP cells are stacked on top of a CMOS Substrate [13]. These designs illustrate that with strategic partitioning and modification of architecture, 3D-ICs can achieve significant performance improvements over conventional 2D-ICs. Heat removal from a 3D stack is a major challenge. The 3D architecture must take this into consideration. In the 3D de- signs illustrated above, the high power electronics are placed in the substrate, whereas the upper layers are limited to memory cells or other low activity devices to minimize heat generation. As the cost for developing and deploying a technology cycle escalates, 3D integration can be an effective supplement to extend the life of the technology node.

References

[1]   For example, M. Yasumoto, H. Hayama, and T. Enomoto, "Promising new fabrication process   developed for stacked LSIs," IEEE Int. Electron Devices Meeting, pp. 816−819, Dec. 1984.

[2]   For example, K. D. Gann, "Neo-stacking technology," HDI Magazine, Dec. 1999.

[3]   For example, H. Kurino et al., "Intelligent image sensor chip with three dimensional structure," IEEE   Int. Electron Devices Meeting, pp. 879−882, Dec. 1999.

[4]   For example, S. Wong et al., "Monolithic 3D integrated circuits," Int. Symp. VLSI Technology, Systems, and Applications, p.66−67, Apr.2007.

[5]   A. Rahman, et al., "Wiring requirement and three-dimensional integra- tion technology for field programmable gate arrays." IEEE Trans. on VLSI Systems, vol. 11, no. 1, pp. 44-54, Feb. 2003.

[6]   M. Lin, et al., "Performance benefits of monolithically stacked 3-D FPGA," IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 26, no. 2, pp. 216−229, Feb. 2007.

[7]   M. Lin and A. El Gamal, "A low-power field-programmable gate array routing fabric," to be published in IEEE Trans. VLSI Systems, 2009.

[8]   K. Osada et al., "A 2ns access, 285 MHz, two-port cache macro using double global bit-line pairs," IEEE Int. Solid-State Circuits Conf., pp. 402−403, Feb. 1997.

[9]   H. H. Nho, M. Horowitz, and S. S. Wong, "A high-speed, low-power 3D-SRAM architecture," IEEE Custom Integrated Circuits Conf., pp.201−204, Sep. 2008.

[10]   S.-M. Jung et al., "The revolutionary and truly 3-dimensional 25F2SRAM technology with the smallest S3   cell, 0.16um2, and SSTFT for ultra high density SRAM," IEEE Symp. VLSI Tech.., pp. 228, Jun. 2004.

[11]   A. W. Topol et al., "Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs)," IEEE Int. Electron Devices Meeting, pp. 352−355, Dec. 2005.

[12]   V. Suntharalingam et al., "Megapixel CMOS image sensor fabricated in three-dimensional integrated circuit technology," IEEE Int. Solid-State Circuits Conf., pp. 356−357, Feb. 2005.

About the Authors:

Kumar.Keshamoni, Asst.Professor
Received B.Tech degree in Electronics and Communication Engineering from the University of JNTU and Pursuing M.Tech degree in VLSI from the University of JNTU-Hyderabad. He is currently Asst.Professor in ECE Department of RVR Institute of Engineering & Technology, Hyderabad, His currently research interests include VLSI Design, Nano Technology and Embedded Systems.
E-mail: kumar.conf@gmail.com
Contact Number: +91-9000080590

Professor **M. HARIKRISHNA**
Received Bachelor of Engineering and M.Tech degrees in ECE from Marathwada University, Aurangabad and JNTU-Hyderabad. He is currently Professor in ECE Department of RVR Institute of Engineering & Technology, Hyderabad and pursuing Ph.D degree at Department of ECE OU-Hyderabad. His currently research interests include Mixed Signal VLSI design, Bio Technology with Signal Processing.
E-mail: yecmhechkay@gmail.com
Contact Number: +91-9985091040