A Simplicial CNN Visual Processor in 3D SOI-CMOS

Pablo S. Mandolesi†
Pedro Julian‡
Departamento de Ingeniería Eléctrica y de Computadoras
Universidad Nacional del Sur, Bahía Blanca, 8000, Argentina
Email: pmandolesi@uns.edu.ar

Andreas G. Andreou
Electrical and Computer Engineering
Johns Hopkins University Baltimore, MD 21218 USA
Email: andreou@jhu.edu

***Abstract*— This paper presents the architecture for a SIMD digital Visual Processor Unit (VPU) that is based on the Simplicial CNN (S-CNN) algorithm. The system is designed for three dimensional CMOS integration in the three tier MITLL 3D SOI-CMOS 0.18μm technology. The architecture includes input/output sub-systems, in the third tier, arithmetic logic units (ALU) and register files on the third and second tiers and instruction cache memory and a timing state machine on the first tier. The partition of the architecture exploits its physical realization in three dimensional CMOS. Parallel optical data input through an array of photodetectors and analog interface circuits in the third tier facilitate testing and characterization.**

I. Introduction

Since the invention of the integrated system in the 1950's, the microelectronics industry has seen a remarkable evolution from the centimeter scale devices created by Jack Kilby [1] to millimeter scale integrated circuits fabricated by Robert Noyce to today's experimental 8nm feature size MOS transistors [2]. During this time, not only have exponential improvements been made in the scaling of devices [3], but the CAD and workstation technologies have advanced at a similar pace enabling the design of complete truly complex systems on a chip that include several memory hierarchies as well as sophisticated high speed input/output interfaces.

The need for ever increased performance and computational power, the microprocessor has driven the scaling of digital CMOS technology. This is true for both general purpose CPU (Central Processor Unit) chips found on the motherboard of computer systems and for GPU (Graphics Processor Unit) chips found in high end graphics cards. These processor chips incorporate multiple levels of metallization (8 to 12) and are fabricated using complex lithographical processes and processing equipment.

However scaling of CMOS necessitates expensive fabrication lines to produce MOS transistors of even finer features. Three dimensional integration is an alternative method to increase the number of transistors, while at the same time preserving locality of reference. The early attempts towards 3D integration were focused on multiple tiers with polycrystallized silicon devices [4].

An alternative approach that has emerged in recent years is based on the three dimensional stacking of wafers fabricated in standard CMOS technologies, augmented with an inter-die via [5]. The latter approach exploits the dramatic advances made in recent years at the back-end CMOS processing i.e. metallization layers and interlayer contacts (vias). Bulk CMOS wafers are first thinned down to about 10μm thickness and then aligned and bonded to form a multi-wafer stack. The whole process poses significant challenges with bulk CMOS wafers especially with the formation of a few micron through the bulk substrate, isolated electrical via. Nonetheless, complete systems have been demonstrated in experimental 3D bulk CMOS integration technology [6], [7].

More recently an alternative approach has been developed using Silicon On Insulator (SOI) CMOS wafers [8]. Three-dimensional (3D) integrated circuits have been demonstrated as viable technology for information processing in high throughput sensor arrays [8], [9] and massively parallel computer architectures that benefit from locality of reference and short interconnects in the third dimension [10], [11]. The first multi-project foundry 3D SOI-CMOS run [12] was held in the Spring of 2005.

Even with SOI-CMOS wafers the whole process is prohibitively expensive for mass production and fabrication yields can be low. However, the recent report by IBM of 10^8 through-wafer-vias per cm^2 in a production SOI-CMOS environment [13] is and indication that 3D SOI-CMOS has the potential for a cost-effective paradigm shift in the design of integrated circuits. It is believed that at the 22nm node, it will be more cost effective to stack four wafers to achieve an (x4) local transistor density than to scale the feature size by a factor of two. Furthermore, wafers need not be of the same technology but one could use optimized wafers for analog circuits, digital microprocessor, digital memory or FLASH memory, with different feature size, metallization layers and

†Comisión de investigaciones Científicas de la Provincia de Buenos Aires.
‡Consejo Nacional de Investigaciones Cientificas y Técnicas, Cap. Fed. 1033, Argentina.

Work supported in part by Office of Naval Research MURI for Intelligent Biomimetic Image Processing and Classification N00014-01-1-0624 , ONR MURI GC183994NGD; *"Desarrollo de tecnología de redes de sensores para aplicaciones en el medio social y productivo"*, PICT 2003 No. 14628, Agencia Nacional de Promoción Científica y Técnica; *"Redes de Sensores"* PGI 24/ZK12, Universidad Nacional del Sur.
Fabrication of the 3D SOI-CMOS chips was provided by MIT Lincoln Labs.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 25, 2009 at 19:00 from IEEE Xplore. Restrictions apply.

power supplies.

This paper presents a 3D SOI-CMOS architecture for a SIMD digital Visual Processor Unit (VPU) that is based on the Simplicial CNN (S-CNN) algorithm [14] and [15]. The (S-CNN) algorithm is derived from the cellular neural network (CNN) paradigm for parallel computation introduced in [16]. The (S-CNN) algorithm offers an efficient way of implementing a parallel search through computations that are done in parallel and collectively by an array of identical programmable units. The mapping of the S-CNN algorithm into digital circuits for a single tier CMOS technology was reported in a previous paper [17], and hence here we focus on the aspects of the architecture that pertain to a 3D CMOS implementation.

The paper is organized as follows: Section II presents the processor architecture and floor planning at the high level. Section III discusses the circuits in each cell of the array and the partition of the layout in the three tiers. Section IV concludes the paper.

II. Processor Architecture and Floor Plan

In this section we briefly summarize the salient features for the architecture of the S-CNN processor presented in [17]. The basic computation in the S-CNN SIMD architecture is done sequentially in time on each cell over the entire array of cells in parallel. The processor functionality is determined by programming a set of memories describing the input and the local cell state relationships, G and F respectively. This is done for a group of cells called the sphere of influence (formed by a cell and its neighbors) and a simple logic relation between them. Relation $F : \Re^m \rightarrow \Re$ and $G : \Re^m \rightarrow \Re$ are piecewise linear functions defined over a simplicial domain (please see [17] [15]). During a program cycle, data encoded sequentially in time, and the evaluation of the composition of functions F and G, namely FoG, evolve. The output of each cell is the integration of the binary value FoG during the complete program cycle. The input/state relationship of the sphere of influence is determined by the mapping that is stored in the G/F lookup table. Each lookup table consists of 512 bits of memory. In the most general case of this processor architecture, each cell could have its own look up table. In the SIMD architecture describe here, all cells share the same lookup table. The value of G/F is extracted by addressing the memory with an ordered set of binary signals, the time coded input/state values of the sphere of influence that form a nine-bit digital word. At completion of the program cycle, an eight-bit value is stored in the state register and can be transferred outside the chip if desired. For a more detailed description of the underlying architecture and its programming the reader is referred to [17].

The S-CNN processor consists basically of three main parts: the cell array, the state machine/ memory units and the I/O interface unit [see Figure (1a)].

Figure (1b) shows a sketch of the floorplan for the architecture mapped onto a three tier CMOS technology. Tier one is used for service circuitry including the processor state machine, the instruction cache memory and part of the I/O subsystems on the sides of the array. Tier two and three are occupied by arrays of ALU and register files as well as input/output functions. Figure (2) shows the overall layouts of the individual tiers. On Tiers two and three the cell array can be easily seen; two instruction cache memory banks and the state machine on the side is evident in Tier 1. Each memory banks stores the information to run the programs and the blank space was left intentionally to add special test structures.

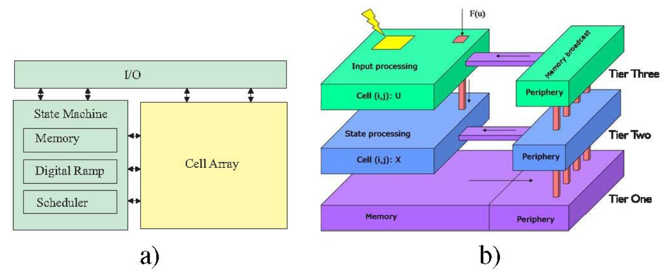

Fig. 1. a) Floorplan for the S-CNN processor mapped onto a single standard CMOS chip; b) Three dimensional floor-plan of the chip

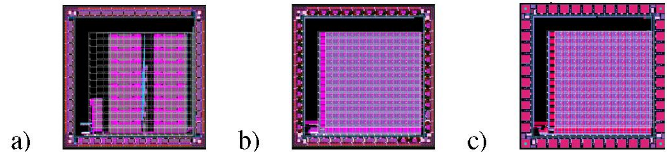

Fig. 2. a) Tier 1 layout; b) Tier 2 layout; b) Tier 3 layout

In designing and partitioning the architecture in the three tiers, care is taken to minimize the number of vias between tiers (wafers). The vias in the targeted technology have dimensions of 1.75 um and pitch of approximately 4 um. While this is a remarkable achievement in an experimental technology, their size and spacing is not commensurate with the transistor feature size in this technology (0.18μm). We have partition the system architecture ensuring that **only one** through wafer via is used for each cell in the array of processing units. Linear arrays of vias are however employed for data and control buses on the sides of the array.

1. The cell array Block: The array is produced by tiling identical cells in a regular grid. Each cell is 60μm by 60μm and two tiers tall. The circuits included on each cell handle data input, communicate to its neighbors the input and the state of each cell, and interact with the periphery circuits to execute the selected program cycle, transfer results, or load special inputs.

2. The state machine and memory unit: The state machine includes all the circuits needed to synchronize the processor subsystems and in addition has the capacity to store the programming functions. The state machine signals include the A/D conversion cycle for the Tier 3 diagnostic optical input, controls the I/O operations, the program cycle and the function evaluation cycle control signals. The programs are stored in four different memory banks, arranged in two double banks. Each bank has the state function look up table (F), the

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 25, 2009 at 19:00 from IEEE Xplore. Restrictions apply.

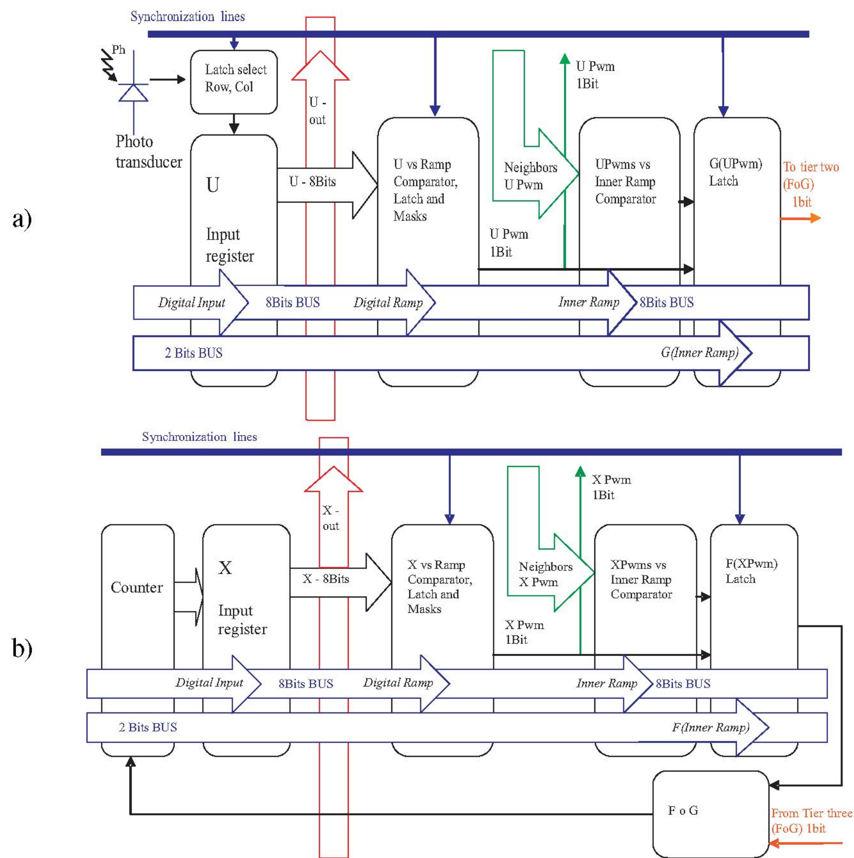

Fig. 3. a) Cell layout for Tier 3 b) Cell layout for Tier 2

input function look up table (G) and the composition logic function (FoG). These circuits are located on Tier one and they communicate with the cell array and the I/O block through 3D vias on the periphery of the chip.

3. The I/O interface: The I/O Interface contains the circuits needed for each tier to interact with the external to the chip components [see Fig.(1)]. The I/O circuitry allows the readout of the data, the loading of data, the programming of the look up table and the configuration of the state machine. These circuits are located on the periphery of the chip and spread out over the three tiers. The chip bonding pads are on the back of tier three, on top of the chip.

III. Cell Description and Floor Plan

The cell structure is depicted in Figure (3) where the block diagrams (a) and (b) correspond to Tier three and two respectively. The 8 bit bus, the synchronization lines and the 2 bit bus lines shown in Figure (3) are common to each cell array row. The output bus is shared by a cell array column. The neighbor arrows depict the lines coming from all cells in cell's sphere of influence; similarly, the pulse width modulated digital 1 bit data line (PWM) goes to all the cells in the neighborhood. The superimposed arrows show the information transmitted through the 8-bit bus at different processing stages. On the right of picture in Figure (3a) the one bit line with the time coded information of G connects to the right arrow on the FoG operational block in Figure (3b). What is important, is that this data is carried from one tier to the other through a singe 3D via. The basic parts in each cell are:

1. The pixel or digital input: This circuit includes the photosensitive element and the associated A/D conversion to facilitate testability and diagnostics. In the diagnostics or "imager" mode, a single PIN photodiode (designed as an external ring in tier three, [see Figure (4)] integrates photocurrent in its capacitance to give a voltage output. The voltage is sampled, held and compared to an external analog ramp that runs synchronized with a digital ramp. When the ramp voltage is greater than the photodiode voltage, the comparator changes its output and the value of the digital ramp is latched. This is a single slope parallel A/D converter and the input digital value is stored in a register (the capability of loading an image directly to the input register cell by cell also exists). These circuits are all located on tier three, the top of the wafer stack, allowing direct illumination of the array.

2. The encoder block: includes the circuits required for the time encoding of the input and state values (UPwm and XPwm). Every signal is coded with one bit in such a way that it is zero when the input (or state, respectively) is greater than the cycle ramp and one otherwise. The encoders are two digital comparators, one in Tier 3 [Fig. (3a)] for the input and one in Tier 2 [Fig. (3b)] for the state. They compare the value of the

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 25, 2009 at 19:00 from IEEE Xplore. Restrictions apply.

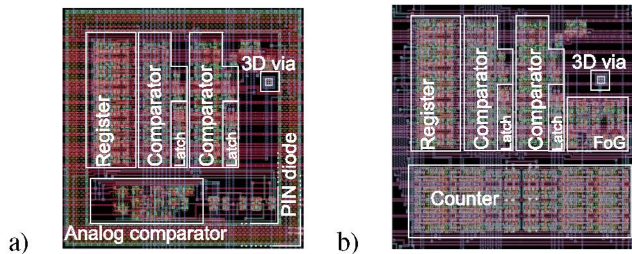

Fig. 4. a) Cell tier three picture b) Cell tier two picture

digital ramp with the stored digital values. As a result, two control signals are obtained that are shared with all neighbors in the sphere of influence (UPwm and XPwm). Every cell collects nine pairs of these encoded signals, one per neighbor. The UPwm signals run on Tier 3 and the XPwm signals run on Tier 2.

3. The processing block: This aggregates the vectors that correspond to the inputs and states of the cells in the sphere of influence in two digital time varying words Wu and Wx for the neighbors UPwm and the neighbors XPwm, respectively. At each step of the cycle the value of function F and G at the vertices indicated by Wx and Wu are obtained. This is done by comparing the broadcast of the memory address with Wu (and Wx) and latching the value of F (and G) when a match between the inner ramp and Wu (Wx) occurs. These tasks are done with two digital comparators (8 bits each) and two multiplexers that are controlled by bit number 9 of Wu and Wx to select the lower or higher memory value. Broadcasting simultaneously two values of the memory reduces the number of cycles in the inner instruction loop (ramp) but adds complexity to the cell. After the two functions are evaluated, the result of the programmed logical operation FoG gives the value of the ramp step S-CNN function. This bit is integrated on a counter (accumulator). All the previous circuits are distributed on Tiers 3 and 2. Tier three has all those components related to u and UPwm, and Tier two has the components related to x and XPwm, plus the FoG circuits and the integration counter. It is important to point out that the partition of the processing blocks as described, **requires only one Tier 2 to Tier 3 via for each cell**. The via communicates the time coded results obtained on Tier 3 to Tier 2 to complete the computation [Fig. (3)]. This is a fundamental characteristic of our architecture that minimizes the number of tier to tier vias in every cell of the array.

IV. Conclusions

The (S-CNN) algorithm is especially suited to a 3D CMOS architecture. The encoding of data in time, allows transfer of information from tier to tier through a minimal number of vias and making more efficient use of space in the 3D floor planning. At the same time, because of pipelining and on chip instruction cache, data and instruction multiplexing on the chip does not reduce the computational throughput in the array. The 3D vias are small but not small enough if compared with metal to metal vias in production CMOS technology with similar feature size. The chip currently in fabrication includes a 14×14 cell array, in a 1.25mm x 1.25mm area with extra test structures in the blank areas. The cellular processor architecture has a density of $\sim 30,000$ through-wafer-vias per cm^2, a rather small number compared to the densities reported in [13], ensuring that the ultimate system could be manufactured with good production yields.

Acknowledgments

We thank Dr. Craig Keast for his personal interest and support. This work was done while one of the authors (AGA) was on a sabbatical leave of absence at the Universidad Nacional del Sur, Bahía Blanca.

References

[1] J. S. Kilby, "Turning potential into realities: The invention of the integrated circuit," http:/nobelprize.org/physics/laureates/2000/kilby-lecture.pdf, Dallas, TX, December 2000.
[2] ITRS, "International technology roadmap for semiconductors," http://public.itrs.net, 2005.
[3] G. E. Moore, "Progress in digital integrated electronics," *1975 International Electron Devices Meeting*, vol. 21, pp. 11–13, 1975.
[4] R. Buchner, W. VanDerWel, K. Haberger, S. Seitz, J. Weber, and P. Seegebrecht, "Process technology for 3D-CMOS devices," in *IEEE SOS/SOI Technology Conference*, October 1989, pp. 72 —- 73.
[5] M. Koyanagi, H. Kurino, K. W. Lee, K. Sakuma, N. Miyakawa, and H. Hitano, "Future system-on-silicon LSI chips," in *IEEE Micro*, July-August 1998, pp. 18 —- 22.
[6] H. Kurino, K. Lee, T. Nakamura, K. Sakuma, K. Park, N. Miyakawa, H. Shimazutsu, K. Kim, K. Inamura, and M. Koyangi, "Intelligent image sensor chip with three dimensional structure," *Electron Devices Meeting, 1999. IEDM Technical Digest International*, pp. 879–882, Dec. 1999.
[7] M. Koyanagy, Y. Nakagawa, K. W. Lee, T. Nakamura, Y. Yamada, K. Park, and H. Kurino, "Neuromorphic vision chip fabricated using three-dimensional integration technology," in *IEEE International Solid-State Circuits Conference*, vol. 1, February 2001, pp. 270 – 271.
[8] J. Burns, L. McIlrath, J. Hopwood, C. Keast, D. Vu, K. Warner, and P. Wyatt, "An SOI three-dimensional integrated circuit technology," in *IEEE International SOI Conference*, October 2000, pp. 20 —- 21.
[9] V. Suntharalingam, R. Berger, J. Burns, C. Chen, C. Keast, J. Knecht, R. Lambert, K. Newcomb, D. O'Mara, C. Stevenson, B. Tyrrell, K.Warner, B. Wheeler, D.Yost, and D.Young, "CMOS image sensor fabricated in three-dimensional integrated circuit technology," in *IEEE International Solid-State Circuits Conference*, vol. 1, February 2005, pp. 356 – 357.
[10] K. W. Lee, T. Nakamura, T. Ono, Y. Yamada, T. Mizukusa, H. Hashimoto, K. T. Park, H. Kurino, and M. Koyanagi, "Three-dimensional shared memory fabricated using wafer stacking technology," in *IEEE International Electron Devices Meeting*, 2000, pp. 165 —- 168.
[11] X. Lei, C. C. Liu, H. S. Kim, S. K. Kim, and S. Tiwari, "Three-dimensional integration: technology, use, and issues for mixed-signal applications," *IEEE Transactions on Electron Devices*, vol. 50, pp. 601 – 609, March 2003.
[12] Massachusetts Institute of Technology Lincoln Laboratory, "MITLL low-power FDSOI CMOS process design guide," June 2005.
[13] A. Topol and twenty eight other authors, "Enabling SOI-Based Assembly Technology for Three-Dimensional (3D) Integrated Circuits (ICs)," in *Proceedings IEDM*, 2005, pp. 363–365.
[14] P. Julián, R. Dogaru, and L. O. Chua, "A piecewise-linear simplicial coupling cell for CNN gray-level image processing," *IEEE Trans. Circuits Syst. I*, vol. 49, pp. 904–913, July 2002.
[15] R. Dogaru, P. Julián, and L. O. Chua, "The simplicial neural cell and its mixed-signal circuit implementation: An efficient neural network architecture for intelligent signal processing in portable multimedia applications," *IEEE Trans. Neural Net.*, vol. 13, pp. 995–1008, July 2002.
[16] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst. I*, vol. CAS-35, pp. 1257–1272, October 1988.
[17] P. S. Mandolesi, P. Julián, and A. G. Andreou, "A scalable and programmable simplicial CNN digital pixel processor architecture," *Transactions on Circuits and Systems Part I*, vol. 51, no. 5, pp. 988–996, May 2004.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 25, 2009 at 19:00 from IEEE Xplore. Restrictions apply.