# Retinomorphic System Design in Three Dimensional SOI-CMOS

Miriam Adlerstein Marwick and Andreas G. Andreou
Electrical and Computer Engineering, Johns Hopkins University
Baltimore, MD 21218 USA
{mir,andreou}@jhu.edu

***Abstract*— Three dimensional (3D) Silicon On Insulator (SOI)-CMOS technology offers opportunities for integration of truly complex neuromorphic systems that do not suffer from the limitations that hinder neuron-like local connectivity in 2D CMOS technologies. In this paper, we outline the rationale for morphing neural structures into 3D SOI-CMOS systems. We discuss design challenges for mixed signal neuromorphic circuits in single tier and 3D SOI-CMOS. We also report on a SOI-CMOS compatible photodetector with photosensitivity of 30,000 (A/W), which is the highest ever reported in the literature.**

## I. Neuromorphic Electronic Systems: The Formative Years

Two decades have passed since the first draft of the seminal book "Analog VLSI and Neural Systems" was put in print [1]. The book had a dual objective; i) to create a new design discipline for collective computational systems and ii) to promote a synthetic approach in the understanding of biology and the human brain. In 1986, Mead's group at Caltech was employing bulk CMOS technology with $\lambda$ between 2.5 and 0.7 microns (page 59 of [1]). A quick review of our own publications and laboratory books from that period, reveals that we were fabricating chips in 4 micron Silicon On Sapphire (SOS)-CMOS technology and in 3 micron p-well bulk CMOS. Alas! two decades later, with foundry CMOS technologies at the 90nm and 180nm node, engineers and scientists in the neuromorphic field have not been able to capitalize on the benefits of the $(\times 100)$ improvements in digital MOS transistor area density. Many of the analog VLSI neuromorphic systems rely on analog devices and as such scaling the density these components (mostly MOS transistors and capacitors) did not follow Moore's law.

The power dissipation of neuromorphic systems did not benefit from technology scaling either and our best circuits today hover between 10nW and 100nW per computational cell. Each cell has typically one or two single pole circuit with two or three current branches biased in the nano-ampere current level. Even though one could argue the power dissipation is manageable locally, the energy cost to send the digital representation of the state from one chip to another on the same board is high. Estimates of the energy costs to transmit one bit of information as a function of distance are plotted in Figure (1). It takes less than a femtojoule of energy to move one bit worth of charge through the source to the drain of an MOS transistor, in deep sub-micron CMOS technology, one picojoule to move one bit of information across a $1cm$ die and almost one hundred picojoules to move it from one die to another! This is a poor utilization of energy, and a direct result of the limitations of two dimensional integration and the use of macroscopic components to interconnect chips. Optical interconnects while efficient at distances measured in kilometers are not very helpful at short distances from a power dissipation point of view. In a two dimensional array of cells that is typical in neuromorphic electronic systems (feature maps), additional energy costs are accrued in going from the two dimensional representation of data to a one dimensional stream on the periphery of the die for interchip transmission [2]. This latter is responsible for an energy shift upwards by one to two log energy units in Figure (1).

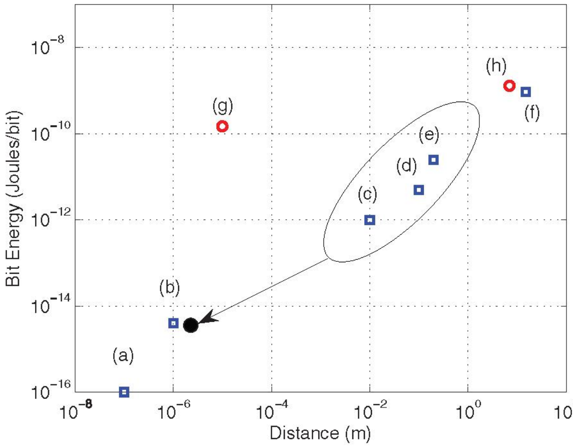

Fig. 1. Energy cost of communication as a function of distance; (a) and (b) switching of a 100nm and 1000nm CMOS inverter [3], (c) switching of metal line across 1cm die [C=1pF, V=1.5V], (d) electrical chip-to-chip link [5pF package capacitance at V=2.5V] (e) optical chip-to-chip link ( [4], [5]), (f) Firewire link, (g) wireless near-field (capacitive coupling) communication between chips [6], [7], (h) Ultra Wide Band radio link. The Firewire and Ultra Wide Band radio link energy budgets were taken from technical specifications in the standards that define the physical layer of the links.

Networks of electronic components such as switches, capacitors and short wires in VLSI integrated circuits are only weakly connected to each other i.e. each component is connected to only a few other components, often in a pipeline structure. Architectures of modern processors and memories are designed specifically to meet the constraints of limited connectivity in modern two dimensional integrated circuits,

0-7803-9390-2/06/$20.00 ©2006 IEEE

that features a dozen or so metallization layers. Biological connectivity on the other hand is capable of generating very strongly connected, scale-free networks [8]. A typical pyramidal neuron in the folded cortical structures of the brain connects to over $10,000$ other cortical neurons while a typical transistor gate is connected to only a handful of other gates in an integrated circuit.

We believe that the network structure of the brain architecture in three dimensional space contributes significantly to its *effectiveness* and *energy efficiency*. At all levels of the central nervous system, from retina to the cortex, the tissue is organized in a hierarchy of layers. In certain layers there is an abundance of axons, the physical structures in neurons responsible for "communication" while others are densely packed with cell bodies and dendrites, or what one will consider as the "computational" structures in the tissue. Furthermore, the layers are tightly coupled vertically through what is termed in biology a "column". It is only recently that we begun to think quantitatively of neural structures as communication network components (channels) [9] and begun to formulate theories of how fundamental physical wiring and energy constraints at the network level, have guided the evolution of neural tissue [8]. It appears that our efforts to create "brain like" machines have been hindered to some extent by "local wiring" limitations and hence the ability to exploit "locality of reference" in the silicon abstractions of neural computations. The emerging technology of 3D CMOS is posed to change this, and is likely to have a profound impact in the future engineering of neuromorphic systems.

## II. 3D CMOS Integration: A Paradigm Shift

Three dimensional integration through wafer stacking and assembly is an alternative to technology scaling that achieves increase in the number of transistors and short range interconnect per unit area. The recent report by IBM of $10^8$ through-wafer-vias per $cm^2$ in a production SOI-CMOS environment [10] is and indication that 3D has the potential for a cost-effective paradigm shift in the design of integrated circuits. This change is driven primarily by two reasons. Firstly, the costs of scaling the manufacturing from one node to the other is increasing at a higher rate than the rate of market increase for the products. At the system level, there is need for heterogeneous technologies with analog or radio frequency components, that do not take advantage of aggressive technology scaling. It is believed now, that at the 22nm node, it will be more cost effective to stack four wafers to achieve an ($\times 4$) 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 processor, DRAM or FLASH memory, with different feature size, metallization layers and power supplies. Technology scaling will likely to continue, perhaps at a somewhat slower pace, nonetheless three dimensional assembly at the wafer or even at the die level is a promising complementary direction.

The early attempts towards 3D integration were focused on multiple tiers with polycrystallized silicon devices [11]. An alternative approach employs the three dimensional stacking of wafers fabricated in standard CMOS technologies, augmented with an inter-die via [12]. 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 $10\mu m$ thickness and then aligned and bonded to form a multi-wafer stack. Complete systems have been demonstrated in experimental 3D bulk CMOS technology [13], including prototype structures for neuromorphic processing [14]. More recently an alternative approach has been developed using SOI-CMOS wafers [15] and three-dimensional (3D) integrated circuits have been demonstrated as viable technology for information processing in high throughput sensor arrays [15], [16].

The first multi-project foundry 3D SOI-CMOS run using the MITLL 0.18um technology [17] was taped out in the late Spring of 2005. The fabrication of the wafers has been completed and at the time of this writing the third tier in the wafer stack is being integrated. In this paper we discuss the design of neuromorphic circuits in the MITLL SOI-CMOS technology.

## III. Retinomorphic Circuits in 3D SOI-CMOS

Putting aside the difficulty in designing of systems in the third dimension, the design of mixed signal systems in deep submicron SOI-CMOS technology is a challenge by itself. The MITLL technology used in the multi-project run is **three tier**, Fully Depleted Silicon On Insulator (FDSOI) process, with a single poly and three metal layers. A tier denotes a single wafer in the integration stack. The design rules for the back-end processes (metallization and vias) are rather conservative and more typical of a 0.35 micron CMOS technology. BSIMSOI3.2 [18] model parameters that were provided by the foundry were used for simulation. LTspice (available freely) from Linear Technology [19] was also employed for simulation of the circuits as it is robust, very fast and free of bugs.

### A. Devices

We begin our discussion with the layout of basic devices. MOS transistors in silicon on insulator CMOS technology do not have an explicit body terminal (the equivalent of the substrate contact in bulk CMOS). In analog circuits it is advisable to use transistors that have an explicit body contact. Figure (2) demonstrates the different ways that a body contact can be placed on the otherwise floating silicon island. The SBC-gate device has the a body contact on the source side only and hence it is more compact but asymmetric with respect to drain and source. In contrast, the H-gate device has body contact on both sides of the gate making it a symmetric device for use when symmetry is essential, for example in diffusive circuits [20].

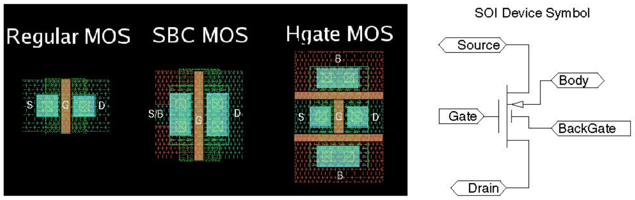

Fig. 2. Layout for three minimal device geometries as allowed by the design rules. Regular transistors, Source-side Body Contact SBC-gate transistor and H-gate transistor (left). There is an area penalty of roughly (×2) in going from a regular "digital" device to a fully body contacted symmetric H-gate "analog" transistor. MOS transistors in SOI technology have five terminals (right). The body contact is the contact to the thin silicon layer between the source and drain. The backgate is the electrical contact to the bulk wafer. In 3D SOI-CMOS there is a possibility of parasitic coupling at the transistor level from one tier to another through the backgate.

Perhaps the most serious challenge in designing retinomorphic systems in 3D SOI-CMOS is the poor photosensitivity of the photodetectors. The photosensitivity of lateral PIN photodiodes fabricated in SOI-CMOS technologies is low, in the order of (0.01 A/W), [21], [22] because the absorbing silicon material is thin (40nm in the MITLL FDSOI process, 100nm in the Peregrine [23] SOS-CMOS technology). The photodetectors in the MITLL technology can be a lateral PIN or PN photodiodes fabricated in the top tier wafer. Even though the wafer is bonded to the lower tiers circuit-side down, the silicon support wafer is removed to the BOX layer thus enabling the illumination of the PIN photodiode from the backside. The photodiode can occupy the entire top level of the three-dimensional circuit and collect the maximum amount of incident light and thus achieve 100% fill factor because additional circuitry is allocated to the lower two tiers of the process.

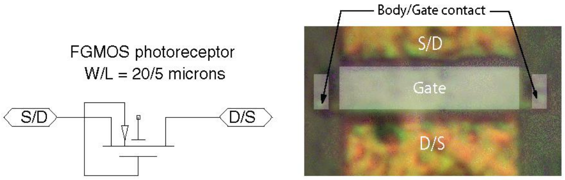

Fig. 3. Floating gate MOS photoreceptor (left) die microphotograph for a fabricated photoreceptor (W/L=20/5 microns). The layout of the device symmetric and two body contacts can be seen on the two sides of the gate.

Here we present an alternative structure suitable for fabrication in SOI CMOS technologies that achieves higher photosensitivity when compared to a PIN photodiode [21], [22], the SOI-CMOS bipolar photodetectors [24] and even the results from a similar structure, the dynamic threshold MOSFET (DTMOS) photodetector [25]. A sensitive photoreceptor can be formed by connecting the front gate of an MOS transistor to its body contact as shown in Figure (3). The body contacts are designed carefully on the sides of the channel as shown in Figure (3). The experimental data presented in this paper [see Figure (4)] are from an NMOS device threshold of $\sim 0.3V$ fabricated in the Peregrine [23] SOS-CMOS technology. The temporal characteristics of the device are shown in Figure (5). The turn–on time of the device is of the order of 1ms but the turn–off time is longer, in the tens of milliseconds, still sufficient for imaging and neuromorphic vision systems. Similar test structures have also been designed in the MITLL FDSOI technology, but their characteristics are not known at the time this paper is submitted for publication.

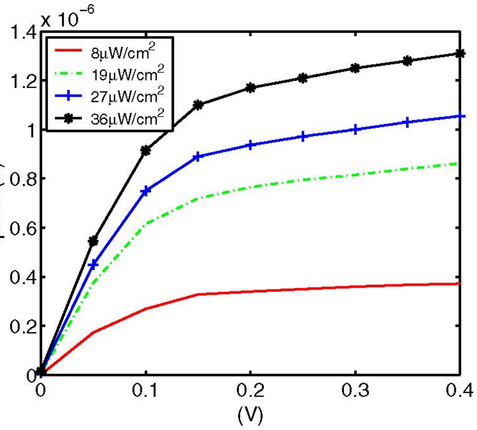

Fig. 4. Current-voltage characteristics of the floating gate MOS photoreceptor. The device was illuminated from the back with a 470nm blue LED.

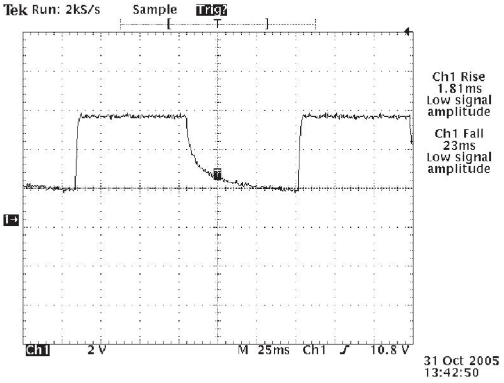

Fig. 5. Temporal characteristics of the FGMOS photoreceptor measured by illuminating the device with a red (≈640nm) 5mW solid-state laser with a bandwidth in the kHz range.

### *B. Adaptive Silicon Photoreceptor*

In what follows is a discussion of one neuromorphic circuit that was incorporated in the MITLL multiproject run, the Tobi-Mead adaptive photoreceptor [26]. In the design of this circuit, the entire adaptive photoreceptor circuit is allocated to Tier B while a PIN photodiode was placed in the top Tier C.

The circuit schematic for the basic photoreceptor is shown in Figure (6). The circuit has low gain for static output from the photodetector, and high gain for a transient optical signal centered about the adaptation point. The PIN photodiode and transistor $M1$ form a simple source-follower whose output voltage is a function of the light intensity as well as the voltage on the gate of transistor $M1$. The output of the source follower is then amplified by the high-gain cascode amplifier represented by transistors $M2$, $M3$, and $M4$ in Figure 6. The output is fed back to the gate of $M1$ through the non-linear adaptive element transistor $M5$ and through the capacitive divider divider of $C1$ and $C2$. The adaptive element is implemented with the compact layout of an SBC-gate transistor. The photoreceptor occupies approximately $20\mu m$ x $20\mu m$ in area. Even though in SOI-CMOS significant area savings can be achieved due to the absence of the well, our layout is rather large in area because of conservative transistor sizing to alleviate potential transistor mismatch problems.

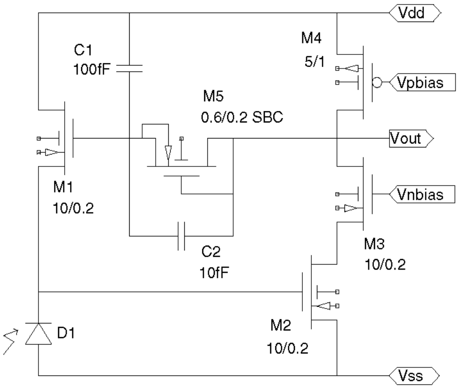

Fig. 6. Adaptive photoreceptor circuit (left). $V_{pbias}$ sets the operating point of the output and hence the response of the circuit and $V_{nbias}$ is the bias for the cascode transistor.

## Acknowledgment

Hiroyuki Kurino (Hiro) a colleague, collaborator and friend of AGA was the inspiration to work in 3D VLSI. It is unfortunate, and sad that he is not with us anymore to witness the progress that we are making.

This work supported in part by Office of Naval Research MURI for Intelligent Biomimetic Image Processing and Classification N00014-01-1-0624 , ONR MURI GC183994NGD and by NSF-ECS-0225489 and NSF-ECS-GOALI-0010026. M.A.M is also supported by a Graduate Research Fellowship from the Johns Hopkins University Applied Physics Laboratory (JHUAPL). Fabrication of the 3D SOI-CMOS chips was provided by MIT Lincoln Labs. We thank Dr. Craig Keast for his personal interest and support. Mike Engelhardt of Linear Technology has provided invaluable help and support by incorporating the Berkeley BSIMSOI3.2 model code in LTspice.

## References

[1] C. A. Mead, *Analog VLSI and Neural Systems*. DRAFT 13th October, 1986.

[2] E. Culurciello and A. Andreou, "A comparative study of access topologies for chip-level address-event communication channels," *IEEE Transactions On Neural Networks*, vol. 14, no. 5, pp. 1266–1277, September 2003.

[3] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S. Lo, G. Sai-Halasz, R. Viswanathan, H. C. Wann, S. WInd, and H. Wong, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, vol. 85, no. 486-504, 1997.

[4] A. Apsel and A. G. Andreou, "A 5mw, gigabit/s silicon on sapphire cmos optical receiver," *Electronics Letters*, vol. 37, no. 19, Sept. 13 2001.

[5] J. Liu, Z. Kalayjian, B. Riely, W. Chang, G. Simonis, A. Apsel, and A. Andreou, "Multichannel ultrathin silicon-on-sapphire optical interconnects," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 9, no. 2, pp. 380–386, March/April 2003.

[6] E. Culurciello and A. Andreou, "Capacitive coupling of data and power for 3D silicon-on-insulator VLSI," in *IEEE International Symposium on Circuits and Systems, ISCAS*, Kobe, Japan, May 2005, pp. 4142–4145.

[7] E. Culurciello, P. Pouliquen, A. Andreou, K. Strohbehn, and S. Jaskulek, "A monolithic digital galvanic isolation buffer fabricated in silicon on sapphire CMOS," *IEE Electronics Letters*, vol. 41, no. 9, pp. 526–528, April 2005.

[8] S. B. Laughlin and T. J. Sejnowski, "Communication in neural networks," *Science*, vol. 301, no. 5641, pp. 1870–1874, September 2003.

[9] P. Abshire and A. Andreou, "Capacity and energy cost of information in a bilogical and silicon photoreceptor (Invited Paper)," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1052–1064, July 2001.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Massachusetts Institute of Technology Lincoln Laboratory, "MITLL low-power FDSOI CMOS process design guide," June 2005.

[18] The device group, department of EECS, UC Berkeley , "http://www-device.eecs.berkeley.edu/ bsimsoi/," February 2004.

[19] Linear Technology, "http://www.linear.com/index.jsp," 2004.

[20] K. Boahen and A. Andreou, "A contrast sensitive silicon retina with reciprocal synapses," in *Advances in Neural Information Processing Systems*, vol. 4. San Mateo: Morgan Kafmann Publishers, 1992, pp. 764–772, reprinted in: Analog Vision Chips, C. Koch editor, IEEE Press, 1993.

[21] A. Apsel, E. Culurciello, A. Andreou, and K. Aliberti, "Thin film PIN photodiodes for optoelectronic silicon on sapphire CMOS," vol. 4, pp. 908–911, May 2003.

[22] E. Culurciello and A. Andreou, "16 x 16 pixel silicon on sapphire CMOS digital pixel photosensor array," *IEE Electronics Letter*, vol. 40, no. 1, pp. 66–67, January 2004.

[23] Peregrine Semiconductor Corporation, "http://www.psemi.com," 2004. [Online]. Available: http://www.psemi.com/

[24] W.Zhang, M. Chan, S. Fung, and P. Ko, "Performance of a CMOS compatible lateral bipolar photodetector on SOI Substrate," *IEEE Electron Device Letters*, vol. 19, pp. 435–437, November 1998.

[25] W.Zhang, M. Chan, and P. Ko, "Performance of a Floating Gate/Body Tied NMOSFET Photodetector on SOI Substrate," *IEEE Transactions on Electron Devices*, vol. 47, pp. 1375–1384, July 2000.

[26] T. Delbruck and C. Mead, "Analog vlsi adaptive logarithmic wide-dynamic-range photoreceptor," *IEEE International Symposium on Circuits and Systems*, pp. 339–342, 1994.