Winner/Loser-Take-All Circuits on SOI Technology for Neural Network Classification

T. A. Duong, C. Saunders†, T. Ngo, and T. Daud

Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109
†Irvine Sensors Corporation, Irvine, CA 92626

ABSTRACT

High connectivity of artificial neural network chip-embodiments combined with currently emerging 3-dimensionally stacked multichip modules for real-time applications of target classification require a scrutiny for low power technology insertion. Conventional CMOS high power consumption limits the allowable density of synapse/neuron elements. However Silicon-On-Insulator (SOI) technology has the potential for successful implementation of-high density neural network because of the following unique features: (a) Operating voltage is reduced 3-fold from 5 to 1.5 volts, reducing power requirements by 9-fold; (b) Reduced substrate offers reduced capacitance and power and an increased speed; and, (c) Latch-up phenomenon is eliminated. Here we describe two practical winner/loser-take-all (W/LTA) circuits fabricated with 0.25 μm fully depleted SOI technology that are useful for neural networks and as compared to other such circuits offer considerable advantage of speed and performance. SPICE circuit simulations show that up to 9-bit resolution can be obtained between a winner and a loser input and with two cascaded circuits. Final characterization tests prove that constructing circuit elements from SOI technology would allow us to build large size neural networks for practical applications.

I. INTRODUCTION

In many artificial neural network architectures, winner/loser-take-all (W/LTA) circuits play an important role in implementing learning [1,2] for high speed classification problems, and in pattern recognition requiring intensive image data computation. Several publications describe W/LTA circuits[3,4] implemented in VLSI hardware; however, these circuits are unsuitable for our application in terms of resolution, speed, and power consumption.

The Ultra Low Power Electronics (ULPE) program using silicon-on-insulator fabrication facilities has provided an opportunity, for the first time, to fabricate neural processing and W/LTA circuits using the foundry at the Massachusetts Institute of Technology/Lincoln Laboratory (MIT/LL) in their 0.25 μm, 1.5 volts, fully depleted SOI technology.

In a nutshell, the advantages of SOI technology are as follows:

1. The operating voltage is reduced to 1.0-1.5 volts which reduces the power by a factor of $(Vdd_{1.5}/Vdd_{5.0})^2$ or 0.09.
2. Power consumption is further reduced because with no substrate capacitance, the total capacitance formed by the reduced feature size and parasitic capacitance is very small.
3. The insulator that replaces silicon substrate does not form a back to back npn or a pnp junction; therefore, no latch-up can occur.

Earlier, using a conventional VLSI (5 volts) technology with 0.8 μm feature size, JPL has developed a 64 IC, 3-dimensional artificial neural network (3DANN) image processing cube with each IC consisting of a 64x64 synapse array circuit. The cube is capable of performing 10^{12} analog multiply-accumulate operations/s with 2.5 watt power dissipation in a 0.5X0.5X0.5-in.(0.125 in^3) volume. Other circuit functionality requirements similarly may impose higher power budget on the cube. From this it is evident that high power density with inherent cooling difficulties will inhibit further circuit size expansion or incçrporation of other required functionalities. However, by using the SOI technology for IC fabrication, it is estimated (considering voltage and capacitancereductions) that the 3DANN power dissipation would have been reduced to mere 0.15 watts. This reduction would be crucial in some cases such as a cube mated to an infrared (IR) sensor array requiring the module cooling to 77K[5,6]. Similarly, there are gains to be made for many complex space-borne system architectures required by DOD and NASA missions (including the NASA initiative of spacecraft-on-a-chip)where high speed combined with low power consumption is generally a key requirement.

In this paper, we describe an overall system configuration in which the W/LTA plays the role of a decision maker (as the final network) to enhance the system speed performance. Secondly, we discuss the WTA and LTA designs which correlate with previous work by Gilbert[7] but offerdistinct advantages in speed and performanceas W/LTA. Thirdly, W/LTA simulation results are provided and obtainable bit resolution is discussed. In addition, a chip has been fabricated using SOI technology with the designed circuits and the hardware results are presented. Lastly, based on these results some conclusions are drawn.

II. SYSTEMARCHITECTURE

A. 3-D Artificial Neural Network:

Algorithms for the solution of spatio-temporal recognition and classification problems are known to require intensive image data computation. To develop a viable hardware solution for such a task, our work at JPL during the past several years has focused on the development of the 3DANN architecture. Because of the importance of low power circuits, considerable attention was paid to low power analog design development. Therefore, the module consists of a set of 64 low power analog chips packaged together as a 3-dimensional "sugar cube" with the surfaceformed by the chip edges mated and electrically connected using 64x64 indium bumps to an infrared (IR) imager to conform to highly parallel processing of image data at very high speeds. For example, an inner-product processing (multiply-accumulate) of a 64x64 image frame with 64 stored templates as weights in the 64 synapse arrays could be achieved with an unprecedented speed of 10^{12} operations per second (Tera-ops) consuming about 2.5 watts of power[8,9].

To obviate the restrictions of a single IR sensor providing data to the cube, the challenge of delivering highly parallel image data input (64x64 bytes of data in parallel every 250 nanoseconds) has been met by the design of an innovative chip architecture (termed the Column Loading Input Chip or CLIC). The CLIC is mated for its output to the neural processing cube and it replaces the imager on the cube. On the input side, it can be connected to an image frame grabber or a high-speed data formatter through the motherboard thereby being capable of rastering an image of any type (IR, visual, etc.) and large size. This scheme of 3DANN modification, termed 3DANN-M, has been developed and is pictured in Figure 1.

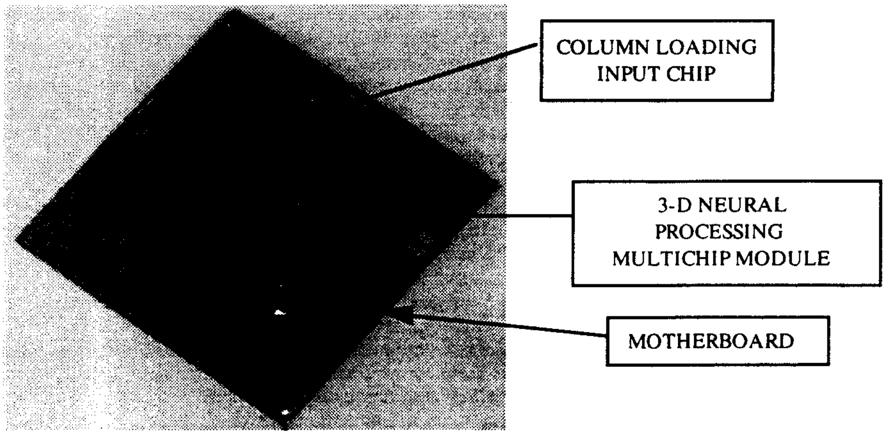
COLUMN LOADING INPUT CHIP
3-D NEURAL PROCESSING MULTICHIP MODULE
MOTHERBOARD

Figure 1:. This photo shows 3-DANN-M with CLIC mated on top of the neural processing module and attached to a motherboard. Learning is off-lineand the weights are then down-loaded into the synaptic array of all the neural processing chips

B. On-Chip Learning:

Even though, the present 3D neural network architectures are only capable of off-line learning whereby simulated weight values can be down loaded to the synaptic array, it is important to study the feasibility of on-chip learning for increased adaptivity and enhanced autonomy. With this motivation, we are now developing an innovative 3D architecture that would integrate neural network on-chip learning with image processing (Figure 2). The purpose of this architecture is to recognize an object with rotation and/or shift invariance. Therefore, on-chip learning on each of the 3D stacked chips provides recognition of an object with respect to a different rotation (or shift)[10]. In the present topology of the 3DANN-M, testing is a very time consuming and tedious undertaking because of the non-availability of some of the critical test points. Our new design makes all the inputs and power lines to the cube fully observable, and at the same time, makes the stacking job simpler because of the straightforward linear metallizations and limited linear indium bump mating.

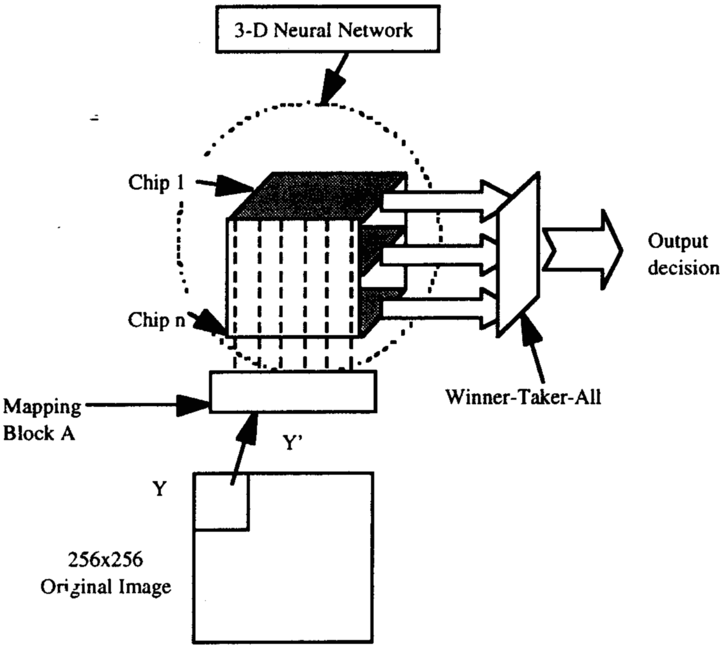

Figure 2: An architecture for a 3-D neural network hardware using on-chip learning along with a decision making WTA plane. A searched window Y that is obtained from a 256x256 image is sent to a mapping block A for conversion to a linear vector before being sent to the 3-D on-chip learning and image processing cube. The WTA is incorporated as the final decision plane in the data processing sequence.

As shown in Figure 2, the training images would be sent through a mapping block 'A' where the image can be linearized and sent to the chips where the image pixels can be transformed into different configurations[10] such as rotation with different angles, scaling, etc., at each particular chip before training. In addition, training of the neural networks would be performed on all chips in a cube simultaneously and in parallel. During a testing phase, a sub-image Y which can be rastered from a 256x256 original image would be sent to the cube for testing. All the outputs would be evaluated using a WTA to obtain the best fit. One of the most challenging tasks is the design of ultra low power consumption circuits. To overcome this challenge, we design low power circuits operating at 1.5 volts line voltage and utilize the SOI technology to fabricate the chips. Our present focus for this paper is on WTA. Therefore, the details of the overall architecture including the MAX circuits are not discussed here. Design details of WTA and results of the simulation as well as hardware are provided.

III. , WTA CIRCUIT DESIGN

Our WTA circuit has been designed based on Gilbert's MAX/MIN circuit[7] with a feedback architecture. Independently, a similar feedback circuit has been used by Lazzaro[3]. However, his circuit has been designed for operation in current mode and in the subthreshold region of the transistor. We adapt Gilbert's feedback circuit to be able to use it for the WTA.

The WTA circuit is shown in Figure 3. A differential pair is used to compare between two voltages V_{in} and V_{ref}. If V_{inw} is found to be the winner ($V_{inw} > V_{ref}$.), then the corresponding V_{out} is high ($V_{inw} - V_{ref} > 0$). The transistor m5 provides the feedback to ensure that V_{ref} is less than a V_{inw} (winner) and larger than one (or all) V_{inl} (loser/s). The output voltage is given by:

$$V_{out} = A_v(V_{in} - V_{ref})\begin{cases} V_{in} > V_{ref} & if \quad V_{in} \quad winner \\ V_{in} < V_{ref} & otherwise \end{cases}$$

where A_V is the gain of a differential pair, V_{in} is(are) the voltage(s) of the inputs to be compared, and V_{ref} is a common node voltage for all the cells. Further, transistor m6 provides an enhanced gain for the system. When V_{out} for the winner is high, m6 has no conduction; hence the gain A_{Vw} of that differential pair stays constant. In contrast, V_{outl} (loser) is low and m6 is in conduction mode. Transistor m10 is mirrored to transistor m9 to apply a certain amount of current. Thereby, A_{Vl} is reduced and a loser node voltage that is close to that of winner node is reduced near Vss.

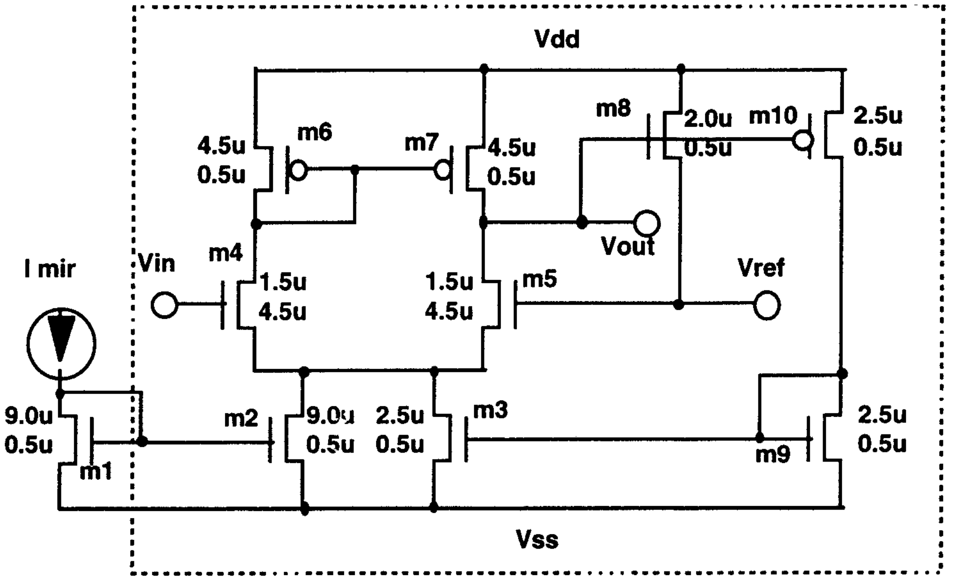

Figure 3: Circuit diagram of a single cell of the Winner-Take-All circuit. In this figure, Vin, Vout, and Vref are brought out. Vref is a common node for all cells in the system. Vout for winner will be high, and for all others (losers) it will be low. The transistor m1 is a global component common to all the cells.

IV. LTA CIRCUIT DESIGN

In contrast with WTA circuit, LTA is used with inverse polarity of the transistors to obtain the complement functionality as shown in Figure 4. Now the equation for V_{out} is modified as:

$$V_{out} = A_v(V_{in} - V_{ref})\begin{cases} V_{in} < V_{ref} & if \quad V_{in} \quad loser \\ V_{in} > V_{ref} & otherwise \end{cases}$$

Again, a differentialpair is used to compare between two voltages V_{in} and V_{ref}. If the voltage V_{inl} is lower than V_{ref}, then the corresponding V_{out} is lower ($V_{inl} - V_{ref} < 0$) compared to all the other V_{out}. The transistor m8 provides the feedback to ensure that V_{ref} is greater than a V_{in} (loser) and less than all other input voltages.

V. WTAARRAYARCHITECTURE

In our proposed architecture, the array contains 64 WTA cells and has a global node V_{mir} to provide a current I_{mir} to the differentialpairs and a common node of V_{ref} which is fedbackto track the maximum input V_{inw}. The array configuration is shown in Figure 5. In our application, as shown in Figure 2, the output from the neural network processing is an analog voltage (V_{in} in Figs. 3, 4, & 5) with its range from 0.3 to 1.35 volts. It is feasible to extract bit-resolution from the circuit simulation results. The intended operational speed of the WTA is ~250 ns per set of 64 inputs. However, presently our focus is not on speed of operation.

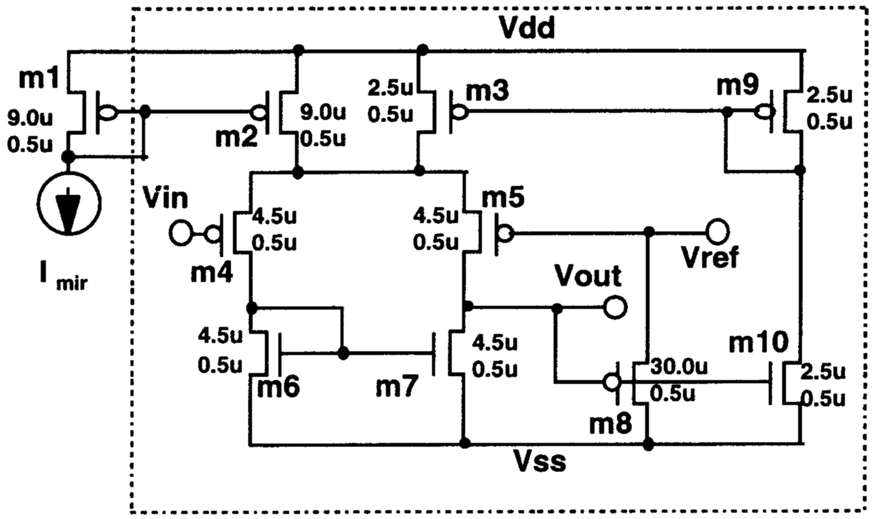

Figure 4: Circuit diagram of a single cell of the Loser-Take-All circuit. In this figure, Vin, Vout, and Vref are brought out. Vref is a common node for all cells in the system. Vout for loser will be low, and for all others it will be high. The transistor ml is a global component common to all the cells.

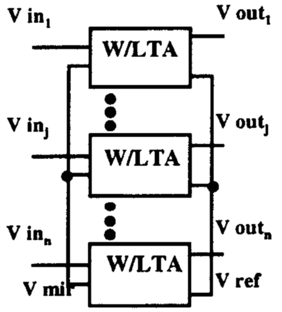

Figure 5: In this figure, the WTA array contains n analog inputs which vary from 0.2 to 1.3 volts, and n digital outputs from which the winner is 1.5 volts (Vdd). Vref is common to all cells, and Vmir is set by a global transistor which injects a constant current Imir.

VI. SIMULATION RESULTS

We have used HSPICE circuit simulator to simulate our designs. We have 4 analog inputs with a range from 0.3 to 1.35 volts. We randomly assign 3 channels with different voltage inputs, and one voltage is ramped up to pass voltages at each channel. We demonstrate in simulation that during the ramping up, when one passes beyond the largest channel voltage, that corresponding output will become a loser. In other words, the output of the ramped up input will be switched to high to be a winner.

We are defining the bit resolution of WTA based on the assumptions as follows:

a) The threshold level will have to be obtained from the maximum among all inputs.

b) The bit resolution is defined as the ratio between the range of operational input over the narrow range of input values around the threshold voltage of the corresponding winner and loser. E.g., input range of operation is from 0.33-1.35 volts then the range of operational inputs is about 1.0 volts and Vin is ramped up, then Vout (ramp, loser) becomes a winner at Vin (ramp) of 0.999 volts and Vout (winner) becomes a loser at the ramp input of 1.001 volts. Thus, the effective input step is 2 millivolts. From these figures, the resolution of WTA (1000mV/2mV) is about 9 bits.

(i) WTA: In our simulation we have used two stages of WTA and at the first stage, we have three of the four channels at constant voltages and the fourth channel voltage is ramped with a 0.1 millivolt step. These four inputs are shown in Figure 6 as Vin1 ramping from 0.26 to 0.42 volts (bottom plot) Vin2 = 0.40 volts, Vin3 = 0.275 volts, and Vin4 = 0.27 volts. In the middle plot the four corresponding outputs are shown. The outputs Vout13 and Vout14 corresponding to Vin3 and Vin4 of the first stage are very low. Initially, Vout12 is a winner (0.61 volts). However, the decision for a winner between Vout11 and Vout12 is a narrow one when Vin1 is close to Vin2. As soon as Vin1 just passes Vin2, the decision for the winner Vout11 is easier to make for the circuit. Finally, the top plot shows the corresponding second stage outputs Vout21 to Vout24 with the inputs Vout11 to Vout14 coming from the first stage. From the output of the cascading 2nd. stage we observe that the step is ~2.0 millivolts in a 1 volt operational range from which it can be roughly estimated that we have ~9 bit resolution.

(ii) LTA: With the same terminology as WTA, Figure 7 bottom plot shows that Vin1 ramps from 1.26 volts to 1.28 volts while Vin2, Vin3, and Vin4 are held constant at 1.38, 1.3, and 1.27 volts respectively. Middle plot shows the output of the first stage; Vout11 is loser while the rest are high. As soon as, Vin1 just passes Vin4 at 1.27 volts, the loser is switched from Vout11 to Vout14. As shown in the top plot, the output of the 2nd. cascading stage is enhanced by the gain of the 2nd. stage. Vout21 and Vout24 are switched with a much narrower step as compared to the 1st stage. Using the same definition, it is estimated that roughly 9 bits of resolution for LTA is obtained.

VII. EXPERIMENTAL RESULTS

We now present our experimental results on the circuits fabricated on a silicon-on-insulator (SOI) chip with 0.25 μm feature size and 1.5 volts operating voltage as discussed earlier. There are 8 cells in the WTA circuit. We test the circuit using 2 inputs, Vin1 and Vin2. Figure 8 shows the experimental results; the bottom plot shows Vin1 along the X axis while Vin2 is given along the Y axis. The + sign shows each sampling step of Vin1 while sampling step of Vin2 is given by an * sign. The corresponding WTA outputs are shown in the top plot using the same legend. Vin1 is ramped from 0.5 to 1.5 volts, whereas, Vin2 is held at 0.821 volt. When Vin1 is increasing to 0.82 volts the corresponding outputs are still unchanged. Until Vin1 is increased one more step to 0.894 volt (> Vin2), then Vout1 rises and Vout2 decreases and thus the decision is flipped between Vout1 and Vout2.

We also experimented with Vin1 as a triangular wave. The results are shown in Figure 9. In the two traces on the top, Vin1 (channel 1) has a peak to peak values of 0.4 to 1.3 volts while Vin2 (channel 4) is a reference input of 0.7 volt. These two inputs are applied to WTA circuits, and the resulting outputs are shown in the bottom traces. Here, channel 2 is a corresponding output of the input channel 1, while channel 3 is the output for channel 4 input. Here it demonstrates that the WTAs behave correspondingly when the inputs are changing in time.

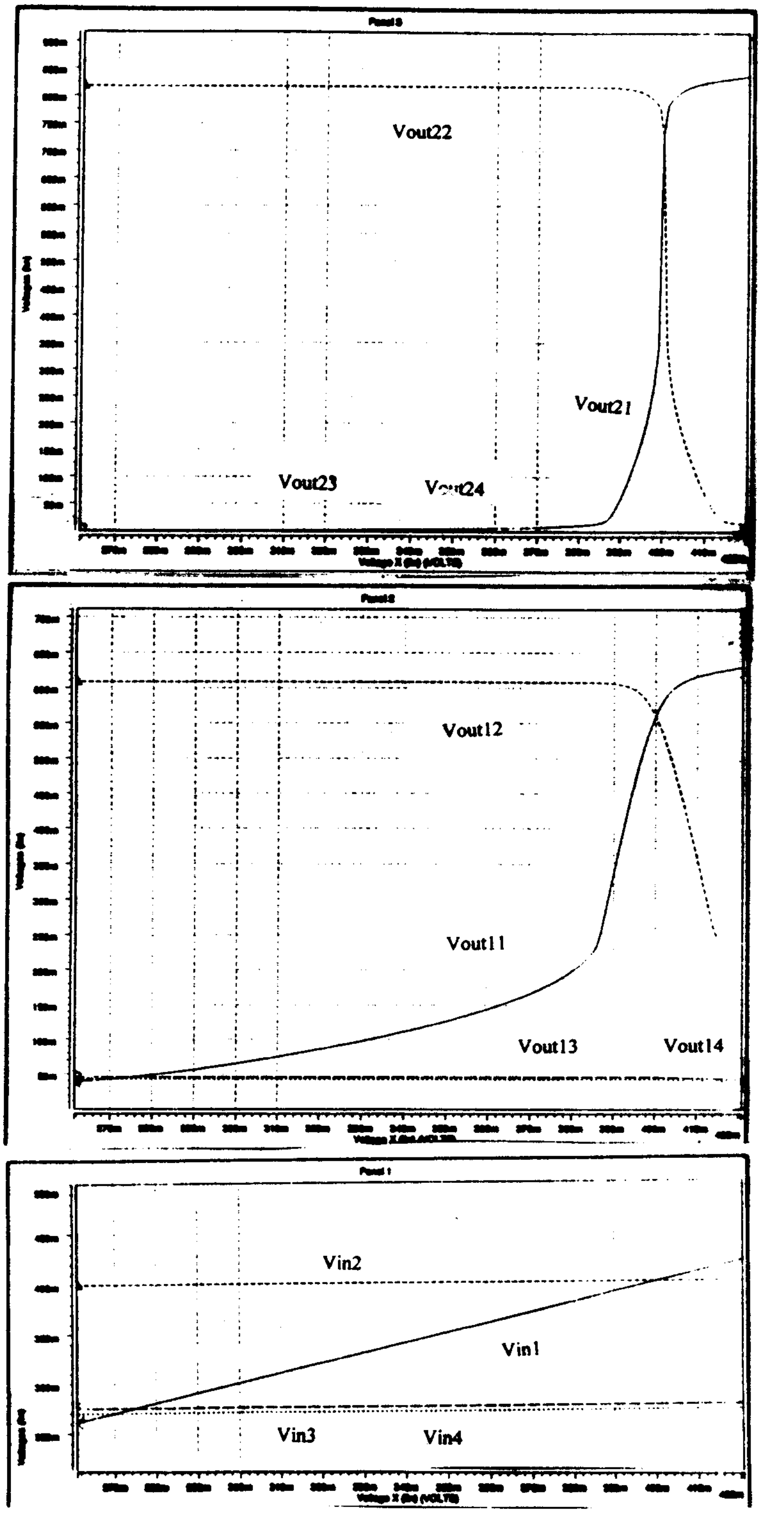

Figure 6. Simulation results of the two stage WTA are shown. Vin1 is ramped, Vin2 through Vin4 are held at respective voltages.

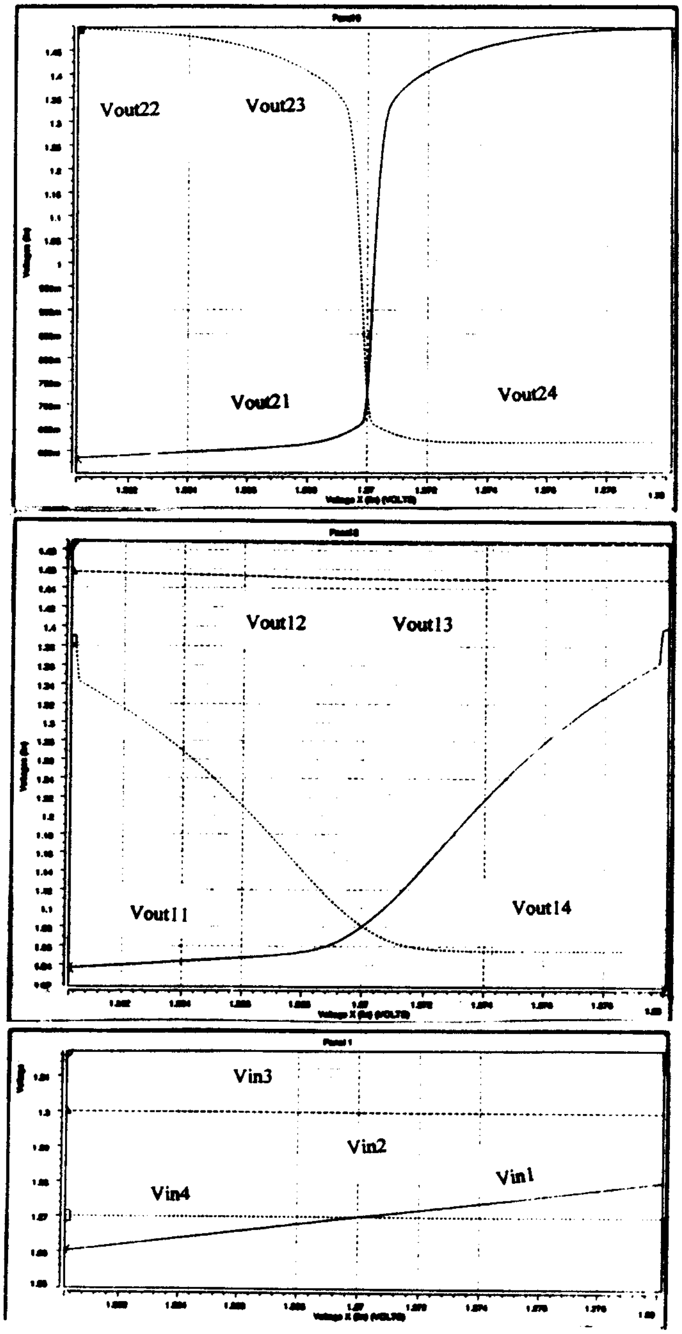

Figure 7 Simulation results of two stage cascading of a loser-take-all circuit array for enhancement of the system gain.

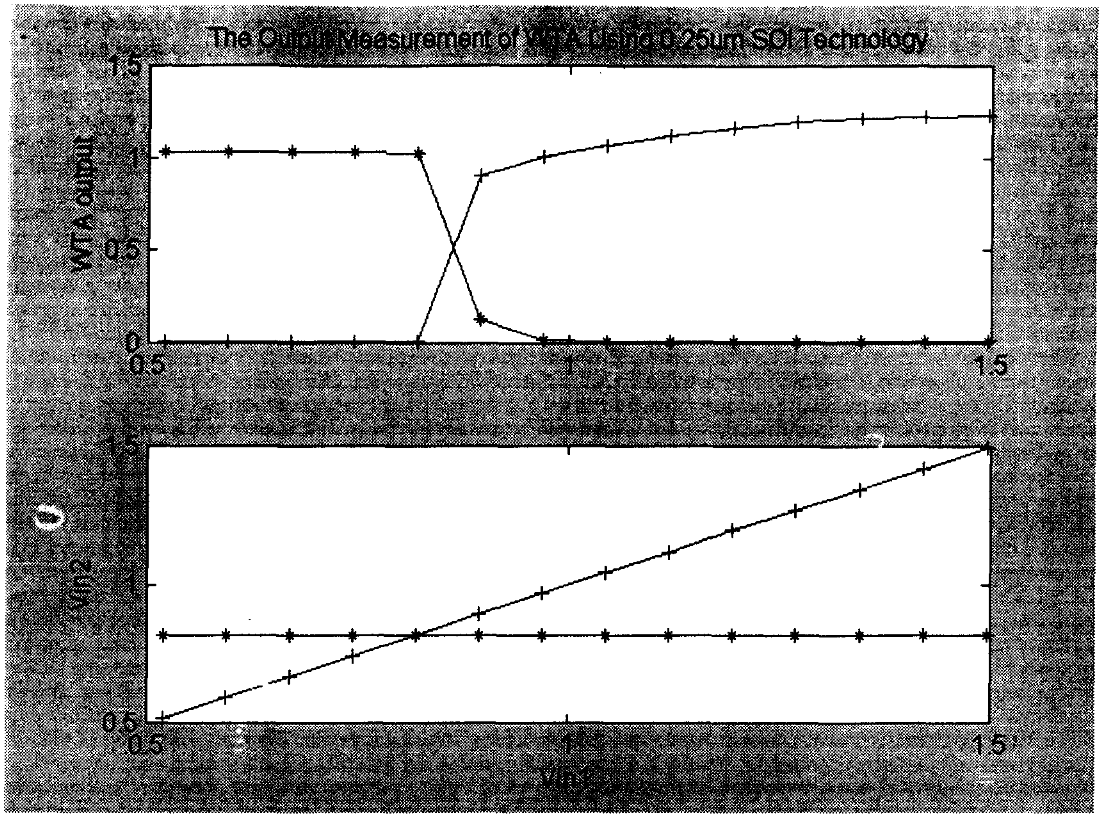

Figure 8. The measurement results of input and output of WTA chip. Vinl and Vin2 shown in the bottom plot are the input into the WTA circuit and WTA outputs are shown in the top plot.

VIII. DISCUSSION

This first low power circuit design experiment with a WTA bears out well the performance in hardware as per the simulated results. It has the potential to provide WTA which is needed in low power, high speed, and latch-up free configuration. The proposed architecture can be fully analyzed theoretically from a control system view point. This will lead to an expansion of its capability for real applications.

The SOI technology offers a capability of a sizable reduction in power consumption. This would enable more compact and power thrift assemblies for military and commercial products, and in particular would impact larger size 3-D architectures. This would also benefit overall response time by a large margin and hence provide a new direction in optimization and speed up of the whole architecture rather than individual components. As is obvious, 3-D electronics would also benefit as a result of reduction in the power.

IX. CONCLUSION

In this paper, we demonstrated that the proposed circuit fabrication with SOI technology would offer benefits as follows:

- Lower power consumption since the voltage source is reduced from 5 to 1.5 volts
- Free of latch-up since no semiconductor substrate exists in the wafer and hence there is no feedback between npn and pnp transistors.
- High speed due to less parasitic as well as component capacitances.

In conclusion, our first ever design and testing of a WTA circuit in SOI technology is very encouraging. Further tests are ongoing, and will lead to better future designs of the WTA. Further study to

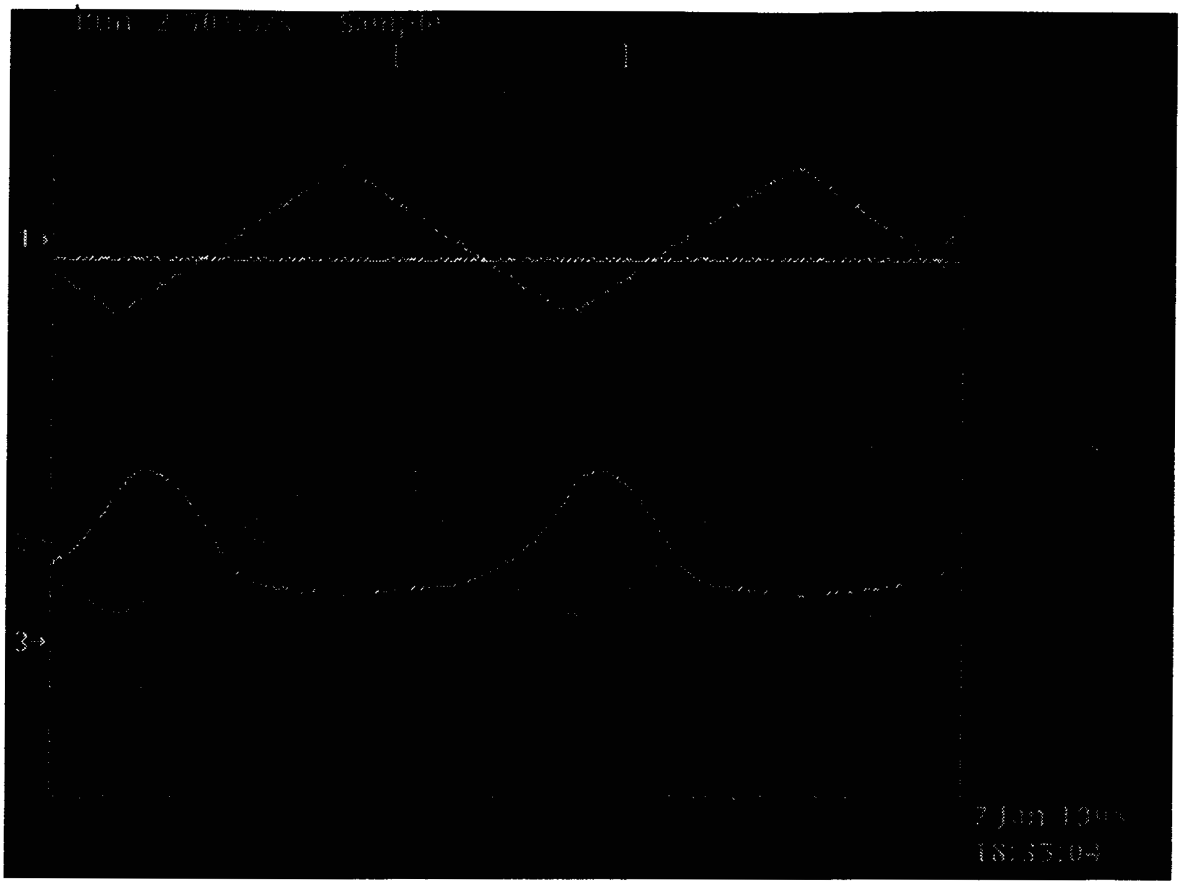

Figure 9. Response (bottom curves) of the 2-input WTA circuit to a triangular wave and a constant voltage inputs (top curves). Separation in the outputs reduces when the two inputs are closer in value.

theoretically analyze the W/LTA circuits will result in an optimization of the design, and hence bring the high speed performance and robustness for real applications.

ACKNOWLEDGMENTS

The research described herein was performed in part by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology and Irvine Sensors Corporation, and was jointly sponsored by the Ballistic Missile Defense Organization/Innovative Science and Technology Office (BMDO/IST) and the National Aeronautics and Space Administration (NASA). The authors would like to thank Drs. Anil Thakoor , John Carson, and Leon Alkalai for useful discussions.

REFERENCES

1. T. Kohonen, "***Self-Organization and Associative Memory***," *Springer-Verlag*, Berlin Heidelberg, 1989.
2. R. Hecht-Nielsen, "***Neurocomputing,***" Addison-Wesley Publishing Company, 1989.
3. J. Lazzaro, S. Reyckebusch, M.A. Mahowald and C.A. Mead, "***Winner-take-all network of O(N) complexity,***" in: "Advances in Neural Information Processing Systems-I," Ed: D. Touretzky, Morgan Kaufmann, 1989, pp. 703-711.
4. J. Choi and B.J. Sheu, "***A high-precision VLSI winner-take-all circuit for self-organization neural networks,***" IEEE J. of Solid State Circuits, Vol. 28, no 5, pp. 576-583, May 1993.

5. T.A. Duong, S. Kemeny, T. Daud, A. Thakoor, C. Saunders, and J. Carson, "***Analog 3-D Neuroprocessor for Fast Frame Focal Plane Image Processing***," *The Industrial Electronics Handbook,* Chap. 73, Ed.-In-Chief: J. D. Irwin, CRC PRESS, 1997.
6. T.A.Duong, T. Daud, A. Thakoor, and B. Lee "***Room and Low temperature performance of high speed neural network circuits***," Electrochemical Society Proceedings, vol. 97-2, May 2-4, 1997, Montreal, Canada, pp. 369-377.
7. B. Gilbert, ***"Nano power nonlinear circuits based on translinear principle,"*** Proc. of the Officeof Naval Research (ONR) Workshop on Hardware Implementation of Neuron Nets and Synapses, Jan. 14-15, 1988, San Diego, CA; pp. 135-170.
8. A. Duong, S. Kemeny, M. Tran, T. Daud, A. Thakoor, "***High speed, low power ASICs for a 3-D neuroprocessor***," Proc. of SPIE Conf. on Non Linear Image Processing VI, San Jose, CA, June 28-July 2, 1995, pp. 470-477.
9. A. Thakoor, T. Daud, C. Padgett, S. Udomkesmalee, and S. Suddarth, "*VIGILANTE:* Multiple Sensors and Neural Network Hardware Based Automatic Target Recognition Experiment," To be presented in Session 9 at the Government Microcircuits and Applications Conference (GOMAC), Arlington, VA, March 16-19, 1998.
10. T.A. Duong, "***An optimal mapping from a two-dimensional analog pixel window to a one-dimensional input vector for fully parallel processing network,***' NASA Tech Briefs, vol. 20, No.3, pp. i1-i2, 1996.
11. T. A. Duong, "***An Analysis of a winner-take-all circuit.***" (In preparation)