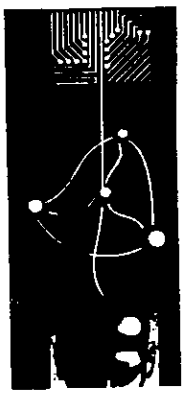

1° Congresso Brasileiro de Redes Neurais

Escola Federal de Engenharia de Itajubá
Itajubá, 24 a 27 de outubro de 1994

LOW-OFFSET NEURAL WINNER-TAKE-ALL NETWORK

Volnei A. Pedroni

California Institute of Technology
Dept. of Electrical Engineering, 128-95
Pasadena, CA 91125 - USA
pedroni@romeo.caltech.edu

CEFET/PR
Depto. de Eletronica e Pos-Graduacao
em Informatica Industrial
Curitiba, PR - Brasil

Abstract - **Winner-take-all (wta) circuits are common building blocks in neural networks, vector quantizers, and other analog parallel signal processing systems. We present a wta circuit that employs a Hopfield-like architecture for the transmission of the positive-feedback coefficients over the 2-D computing array. The properties of this kind of network are further illustrated by means of a 32-input VLSI implementation on a 2.0 μm CMOS chip. Experimental results show a high voltage gain (so digital outputs are immediately available) and very small offsets (under 10mV in the worst-case scenario), with an analog dynamic range resolution of approximately 50 dB.**

I. INTRODUCTION

Analog hardware implementations of vector quantizers, content-addressable-memories, and other *n*-dimensional classifiers lead inevitably to system generation of an also *n*-dimensional set of electric signals which represent the results of some pre-defined vector distance metric computation. The remaining task is to identify which among these signals best satisfies the given metric, that is, to identify the greatest (or smallest) among the resulting signals. A variety of winner-take-all (wta) circuits, which perform this identification function, have been reported recently in the literature [1]-[3]. In this paper, we present a new wta network, which makes use of a neural architecture that closely resembles a Hopfield network [4] (Fig. 1) for the transmission of the positive-feedback coefficients over the 2-D computing array. The network operates in voltage-mode and is capable

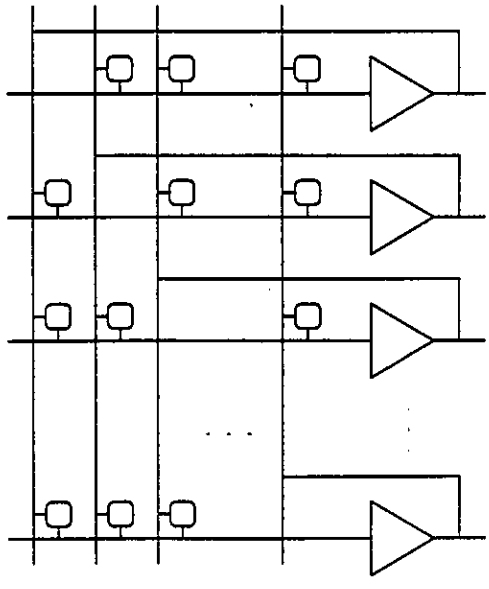

Fig. 1: Neural architecture adopted in the wta circuit.

of detecting very small input voltage differences, limited only by transistor mismatches, therefore overcoming the low detectability (low gain) of conventional $O(n)$ systems [1]-[2], and being also superior to other $O(n^2)$ implementations [3] in the respects that it does not require power supply switches and capacitors, presents higher gain and the inherent higher accuracy of voltage- over charge-mode systems, and also has the internal (computing) nodes completely insulated from the input terminals. The circuit is fully analog, so signals generated by binary processors like Hamming classifiers can be treated simply as a particular and more trivial case in which the input voltages are allowed to take on only certain values rather than any values. Although the interconnect complexity of this kind of circuit is $O(n^2)$, where n is the number of input signals (candidates), it requires just one transistor per synapse, therefore demanding very small silicon area for its implementation.

II. NEURAL WTA CIRCUIT

The neural wta circuit is shown in Fig. 2. It is composed of n amplifying cells (rows), each cell having n-1 active loads, cross-connected in an nx(n-1) array that closely resembles the network of Fig. 1. As can be seen, the circuit has analog inputs. $V_1, V_2, \ldots, V_n$, analog outputs $V_{O1}, V_{O2}, \ldots, V_{On}$, and digital outputs $D_{O1}, D_{O2}, \ldots, D_{On}$. The total current $nI_B = I_1 + I_2 + \ldots + I_n$ is set by the bias voltage V_B and is kept constant thanks to the common bus line V_C, thus providing a high voltage gain and high detectability for small perturbations; it also provides an output voltage which is independent of V_i (1). Preset transistors are also shown.

The principle of operation of the wta of Fig. 2 is based on a positive-feedback loop which only allows one winner in the equilibrium state. The basic operation of the network can be summarized as follows. If we suppose initially that $V_1 = V_2 = \ldots = V_n$, then $V_{O1} = V_{O2} = \ldots = V_{On}$ (theoretically only, due to transistor mismatches).

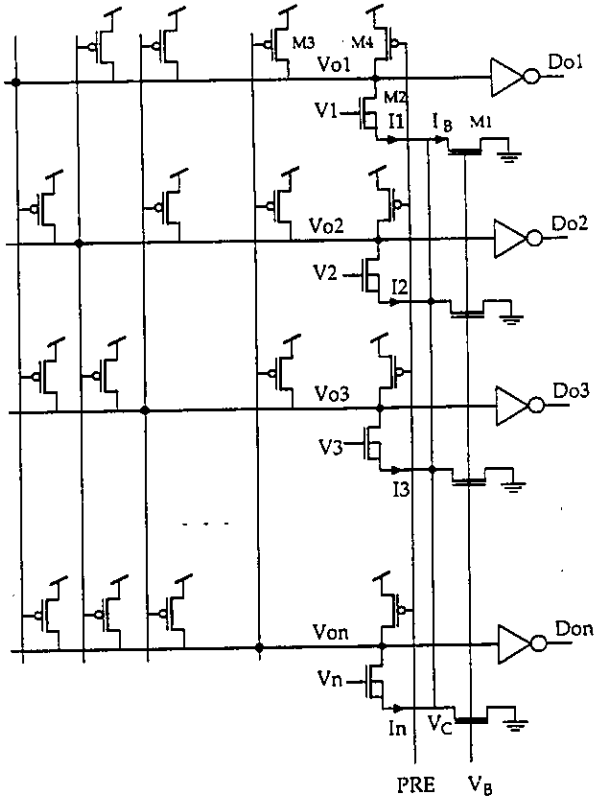

Fig. 2: Neural winner-take-all network.

If now we let one of the inputs, say V_1, increase, V_{O1} has to decrease in order to bring M2 of row 1 into linear mode, needed to keep I_1 constant ($V_{O2}, \ldots, V_{On}$ initially unchanged), so increasing the current driven by the leftmost P-transistor of rows 2 to n. However, the total current must remain the same, what forces the gate voltages of all the other P-transistors in rows 2 to n to increase, thus making the total current driven by the P-transistors in cell 1 smaller, what forces V_{O1} to become even lower.

Quantitatively, the behavior of the circuit can be summarized in the following way. Consider initially the state $V_1 = V_2 = \ldots = V_n \equiv V$ once again, in which case $V_{O1} = V_{O2} = \ldots = V_{On} \equiv V_{O=}$ (theoretically) and $V_C \equiv V_{C=}$. With all transistors in saturation and $\beta_j = (\mu C_{ox} W / L)_j$, we obtain that

$$V_{C=} = V_i - V_T - \sqrt{\lambda_1}(V_B - V_T) \quad \text{and}$$

$$V_{O=} = V_{DD} - V_T - \sqrt{\frac{\lambda_2}{n-1}}(V_B - V_T) \qquad (1)$$

where $\lambda_1 = \beta_1 / \beta_2$ (bias/input transistor) and $\lambda_2 = \beta_1 / \beta_3$ (bias/load transistor). From (1) we verify that, for the saturation condition to be satisfied for any input level, the transistor parameters must obey

$$\sqrt{\lambda_1} \le \frac{V_{i\min} - V_B}{V_B - V_T} \quad \text{and}$$

$$\sqrt{\frac{\lambda_2}{n-1}} \le \frac{V_{DD} - V_{i\max}}{V_B - V_T} \qquad (2)$$

If we let now one of the inputs (say V_1) grow, with all the others still alike (worst case), then V_{O1}, the winning output, decreases, while $V_{O2} = V_{O3} = \ldots = V_{On}$ increase. Call these two voltages levels V_L (low) and V_H (high), respectively. We know that, since M2 of row 1 and M3 of all the other rows are now in triode mode,

$$V_L \approx V_C \approx V_{C=} \quad \text{and} \quad V_H > V_{O=} \qquad (3)$$

A final consideration refers to the transistor parameter ratios λ_1, λ_2. If they are small enough, they may cause $V_H > V_{DD} - V_T$. If so, all of the P-transistors of row 1 (the winner) will tend to the cutoff state, in which case a preset mechanism (M4 in Fig. 2) is necessary in order to remove the circuit from this monostable state before executing the next computation. This situation exists if

$$\lambda_2 < \frac{(2V_{DD} - 2V_C - 3V_T)V_T}{(V_B - V_T)^2} \qquad (4)$$

which is independent from n, as expected, since in the monostable state each active cell is reduced to just one active load. The value of V_C in (4) can be obtained from (1), since $V_C \approx V_{C=}$.

III. EXPERIMENTAL RESULTS

A n=32 wta circuit was fabricated on a 2.0 μm MOSIS CMOS chip (Fig. 3). The transistors are all L=10 μm long, with $\lambda_1 = 0.5$ and $\lambda_2 = 5$. The experiments agreed consistently with the predictions, and the measured detectability was better than 10mV in the worst case, i.e., all inputs equal but one, with an input dynamic range slightly over 3V (50 dB).

A qualitative view of the experiments is shown in Fig. 4, with the upper trace of the scope showing the preset clock, the second channel showing the winning output, and the last channel showing one of the other outputs. As can be seen, the winning output becomes low when preset is released (PRE=5V), while the others stay high. The measurements show always only one winner present at a time, as imposed by the equilibrium state of the circuit.

The gain of the system is depicted in Fig. 5 by means of two separate measurements, which were performed with $V_B = 1.0\,\text{V}$ and $V_{REF} = 2.0\,\text{V}$, being V_{REF} the voltage applied to all the inputs except one, to which V_{in} was connected. The low system offset can be readily observed.

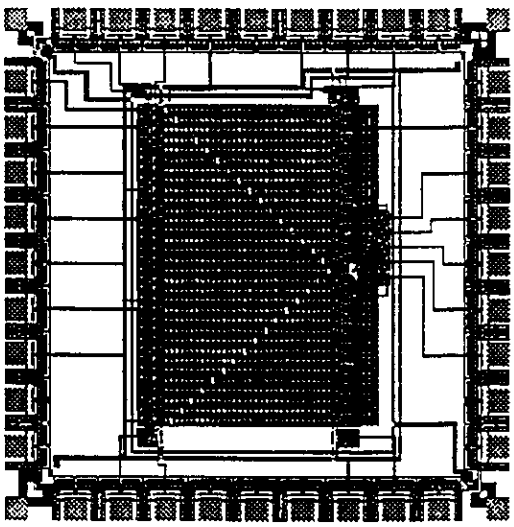

Fig. 3: n=32 wta on a 2.0μm CMOS chip.

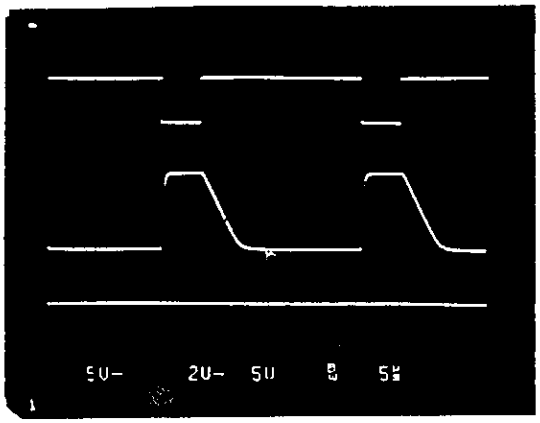

Fig. 4: Winning versus losing outputs.

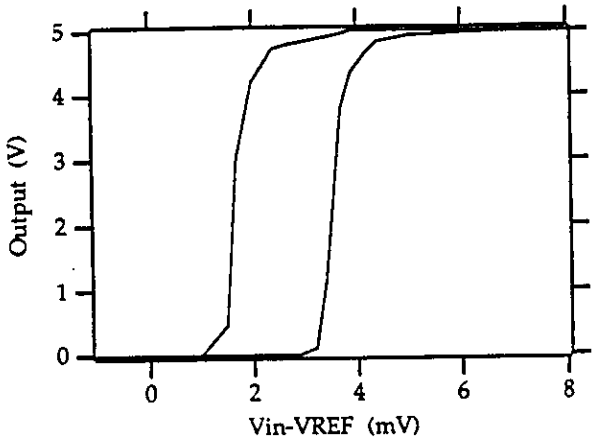

Fig. 5: Plots of two separate gain measurements.

V. CONCLUSION

We have made use of a neural architecture to introduce a new winner-take-all network. Basic properties of the circuit were qualitatively and quantitatively discussed and experimental results, obtained from a prototype chip fabricated using conventional 2.0 μm CMOS technology, were also presented. Positive-feedback, high gain, and small offset make possible the detection of very small perturbations, which are immediately decoded through digital outputs directly available on the network.

REFERENCES

[1] J. Lazzaro, S. Ryckebush, M. Mahowald, C. Mead, "Winner-take-all networks of O(N) complexity", NIPS, vol. 1, 1989.

[2] A. Andreou, K. Boahen, P. Pouliquen, A. Pavasonic, R. Jenkins, "Current-mode subthreshold MOS circuits for analog VLSI neural systems", IEEE Trans. on Neural Networks, vol. 2, 1991, pp. 205-213.

[3] Y. He, U. Cilingiroglu, E. Sanchez-Senencio, "A high-density and low-power charge-based Hamming network", IEEE Trans. on VLSI, 1993, pp. 56-62.

[4] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proc. Nat. Ac. of Sciences, vol. 79, 1982, pp. 2554-2558.