ESSCIRC'88, 21-23/09/88, Manchester (UK) 08-C10

"A new VLSI architecture for large Hopfield's neural networks"

M. Verleysen (*), B. Sirletti (*) and P. Jespers.

(*) Michel Verleysen and Bruno Sirletti are currently working with an I.R.S.I.A. fellowship in the field of neural networks. The authors are with the laboratory of microelectronics, Université Catholique de Louvain, 3 pl. du Levant, 1348 Louvain-la-Neuve, Belgium.

Abstract

A new CMOS architecture for Hopfield's neural networks is proposed. The use of differential amplifiers and active synapses allows the implementation of hundreds of neurons on a single chip. Since it is fully programmable, the circuit can be used as a content-addressable memory as well as in optimization problems.

Introduction

While conventional sequential computers are particularly suited for executing tasks consisting in clearly formulated sequences of instructions, biological brains are more adapted to tasks such as image or speech processing and pattern recognition. The ability of the human brain to deal with incomplete or noisy data lead to the idea of emulating its structure in order to build machines which share some of those abilities: this idea constitutes the base of connectionnist architectures and artificial neural networks [1].

Neural networks do not need any more to prove their interest and their power in the field of pattern recognition, multi-parameters optimization and np-complete problems resolution has been demonstrated [2]. Indeed they offer a new kind of solution to these problems, with a resolution speed which can be 100 or 1000 times faster than the one obtained with the classical artificial intelligence approach. Nevertheless, practical solutions cannot be realized without considering the implementation of very large arrays with hundreds of neurons; large networks are indeed necessary to preserve the intrinsic parallelism and speed properties of such architectures.

Limitations of conventional neural networks

In artificial neural networks, the elementary processing elements (called neurons by analogy with biological models) are connected through a programmable coupling network. In the model proposed here, each neuron receives information about the state of all the other neurons, weighed by connection strengths [3]. A connection between two neurons (called synapse) can be either excitatory (positive) or inhibitory (negative). Each synapse can source or sink current to the input line of the connected neuron; the direction of the current is determined by a combination between the output of the neuron to which the synapse is connected and a weight value memorized into each synapse. In our model, the output of each neuron is boolean: the logical function of a neuron is thus only a weighed sum of the other neurons with a sign discrimination. Furthermore, only three different connection values are allowed in each synapse: +1, 0 and -1. The direction of the current results from the product of these two values (the neuron output and the connection strength): a product equal to 1 means a sourced current, to -1 a sunk one and to 0 no connection. The architecture proposed here can be generalized if more than three different states are required.

The CMOS solution proposed by De Vegvar and Graf [4] where the synaptic currents are sourced by a P-type transistor and sunk by a N-type one (figure 1) is the simplest from a logical point of view, but it has a major drawback: the matching of currents through the N- and P-type transistors is impaired by the mobility differences between electrons and holes. Since in our model, the logical function of each neuron is to detect the sign of the sum of the sourced and sunk currents, the total mismatching between the P- and N-type transistors on each input line must not exceed one single synaptic current. As the synaptic current itself has to be small enough to allow several single currents to be summed on the same line, mismatching constitutes a strong limitation to the maximum number of neurons that can be put together on the same chip.

Matched current sources

The essential feature of our solution is to sum the sourced and sunk currents on two different lines (figure 2). The value stored in mem1 determines the existence of a connection between neurons i and j. According to the value of the connected neuron output (outj) and the value memorized into the synapse (mem2), the current is switched either to the V_+ or the V_- line. All the elementary sourced and sunk currents are summed separately on the two lines. The neuron (figure 3) determines whether the total sourced current is greater than the total sunk one or not. Those currents are therefore converted into voltages by transistors T_1 and T_2. These voltages are themselves compared by means of the reflector formed by transistors T_3 to T_7. Because of the two-stage comparator embedded in the neuron, its gain is very important and the output (out) is always saturated, either to 5V if the current in neuron i- is greater than the one in neuron i+, or to 0V in the opposite case.

The last problem we have to solve is the equality of the currents in the different synapses when increasing the number of active ones. With the neuron presented in figure 3, the voltages V_+ and V_- are indeed reduced when the number of active synapses is increased. To avoid V_+ and V_- to decrease too much, we insert on the i+ and i- lines a negative feedback loop (figure 4) which forces V_+ and V_- to V_{ref}. No high gain is needed for the feedback loop, hence the amplifier of figure 4 can be reduced to one single transistor.

Description of the chip

Simulations showed that the circuit described here allows the implementation of hundred of neurons on a single chip. An experimental CMOS 3 micron test chip has been realized in our laboratory with only 14 neurons and 196 synapses for a first evaluation (figure 5). The bottom row of the circuit contains the neurons (1). Each neuron is connected to the others by means of programmable synapses (the synapse network (2) occupies most of the chip area). The 14 input/output pads (3) on the left side of the chip are used to enter data into the network as well as to program the different connection strengths of each synapse; the decoder (4) selects the column of synapses to address when the chip is in programming mode. We also inserted a break point in the feedback loop (between the neuron outputs and the synapse inputs) in order to allow testing of the chip. Since it is fully programmable, the circuit can be used either as a content-addressable memory or to solve optimization problems. A powerful learning algorithm able to program this circuit as a content-addressable memory has been developped recently [5].

References

[1] B. D. Shriver, "Artificial neural systems", Computer review of the IEEE, March 1988 pp. 8-9.

[2] R. P. Lippman, "An introduction to computing with neural nets", IEEE ASSP Magazine, April 1987 pp. 4-22.

[3] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proceedings National Academy of Sciences USA, April 1982.

[4] H. P. Graf and P. De Vegvar, "A CMOS associative memory based on neural networks", Proceedings ISSCC 1987.

[5] B. Sirletti, M. Verleysen, A. Vandemeulebroecke, P. Jespers, "A new learning algorithm for content-addressable memories using Hopfield's neural networks", paper submitted to Electronics Letters.

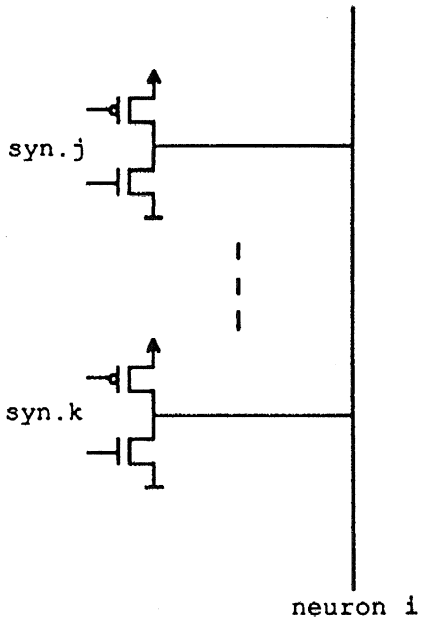


fig.1

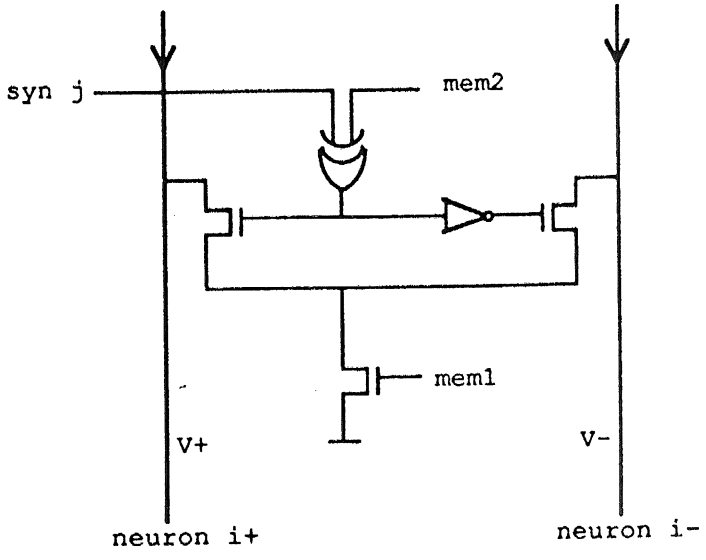


fig.2

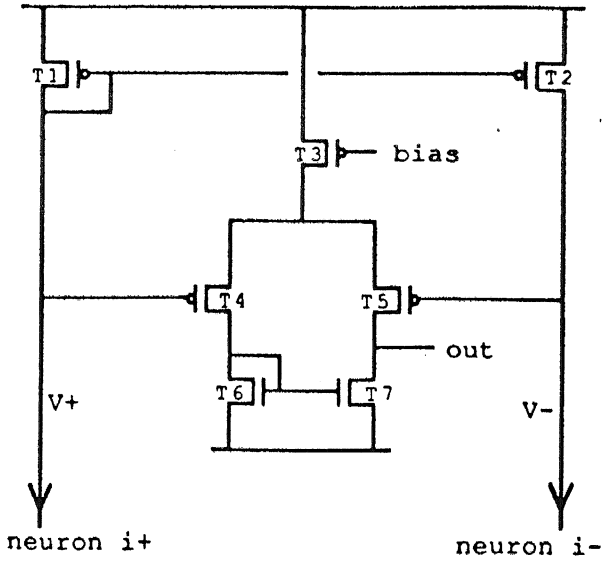


fig.3

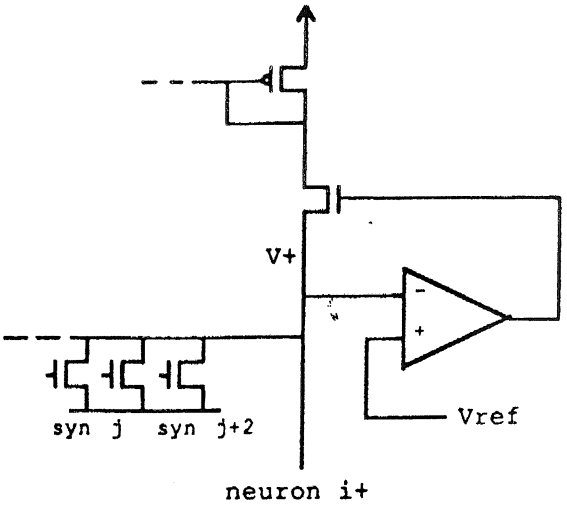
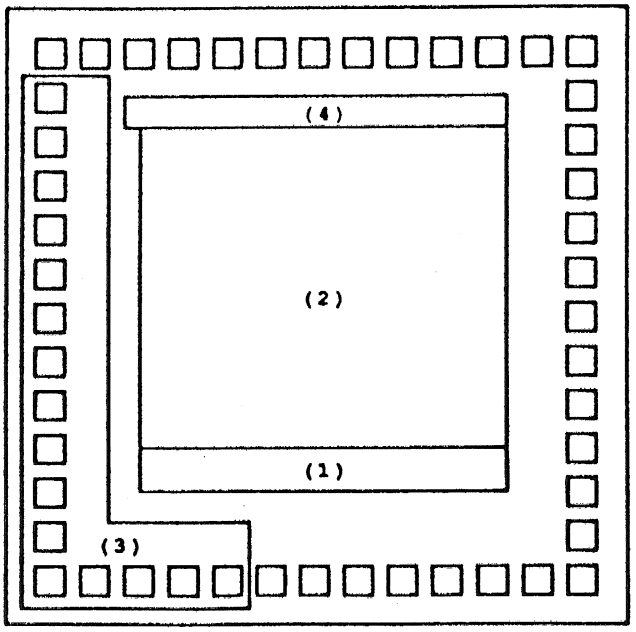


fig.4



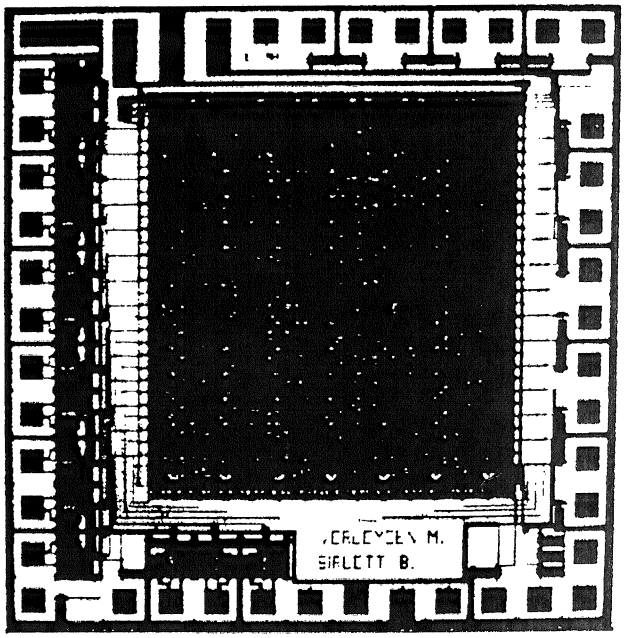


fig.5