VLSI Design of Neural Networks

Ulrich Ramacher
Ulrich Rückert eds.

Kluwer Academic Publishers 1991
(Boston)

PRECISION OF COMPUTATIONS IN ANALOG NEURAL NETWORKS

M. VERLEYSEN, P. JESPERS

INTRODUCTION

VLSI implementations of analog neural networks have been strongly investigated during the last five years. Except some specific realizations where the precision and the adaptation rule are more important than the size of the network [1] [2], most applications of neural networks require large arrays of neurons and synapses. The fan-out of the neuron is not the crucial point: digital or analog neuron can be easily designed so that they can drive a large number of synapse inputs (in the next layer in the case of multi-layered networks, in the same layer in the case of feedback networks). Fan-in is more important: whatever is the transmission mode of information between synapses and neurons (voltage, current, pulses,...) the neuron input must have a large dynamics if it is connected to hundreds of synapses. Digital neurons are of course the solution: if the dynamics of the neuron inputs has to be increased, more bits will be used and the required precision will be obtained. However, digital cells are in general much larger than their analog counterpart: for example, a neuron connected to 100 synapses must contain a digital adder with 100 inputs, each of them coded in several bits. The silicium area occupied by the cells and the connections between cells will be incompatible with the integration of a large number of synapses and neurons on a single chip.

ANALOG NEURONS AND SYNAPSES

In order to compensate for such lack of efficiency, analog cells are used in VLSI neural networks. Several techniques have been proposed to transmit information between synapses and neurons. A.Murray [3] proposed a method inspired by biological mechanisms, where the partial sum-of-products are coded by the frequency of a pulse stream. This solution presents several advantages about the precision of computations, but requires in general a complex circuitry in comparison with other methods.

One of the first VLSI analog neural networks was realized by AT&T Bell Labs [4]. This circuit was based on the sum of positive and negative currents for excitatory and inhibitory synapses (figure 1). A simple logic realizes the product of the synapse input by its internal weight; depending on the sign of this product, a current is sourced to or sunk

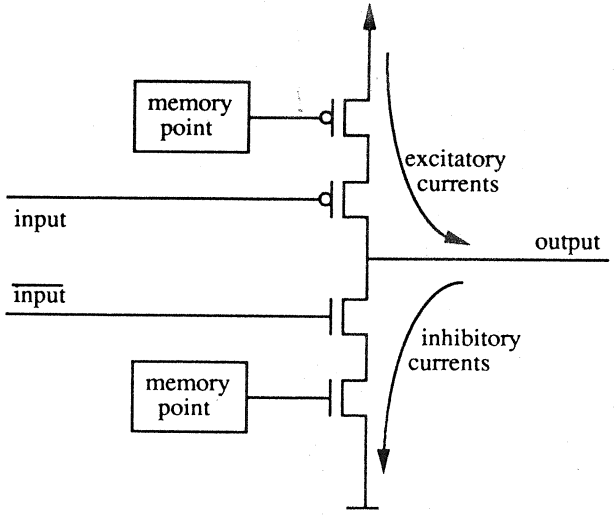

Figure 1 Positive and negative currents

from the input line of the connected neuron. The drawback of this architecture is that one can never assume the excitatory and inhibitory currents to be exactly the same; even with adjustments of the size of the p- and n-type transistors, there is always a risk of mismatching between the sourced and sunk currents because of the technological mobility differences between the two types of charges.

On the other hand, the neuron realizes a non-linear function of its input; in the case of a simple threshold function, the neuron determines the sign of the sum of all synaptic currents. The mismatching between excitatory and inhibitory currents are also summed, and can rapidly become greater than a typical synaptic current. For example, if the error between currents from n-type and p-type transistors is about 20 %, a computation error may occur if only 5 synapses are connected to the same neuron!

In order to compensate for such lack of efficiency, a two-line system can be used: all the excitatory currents are summed on one line, all the inhibitory ones on another line (figure 2). All currents are thus generated by the same type of transistors, and the mismatchings between current sources decrease. In the following, a two-line system is described where the neuron values are binary, and where the synaptic weights can only take three values: + 1, 0 and -1. Such restrictions of course limit the use of the neural network; however, for some architectures, like the Hopfield network, learning algorithms can be found where these restrictions have almost no decreasing effect on the performances of the network [5]. Moreover, the complexity of the synapses is strongly reduced due to the

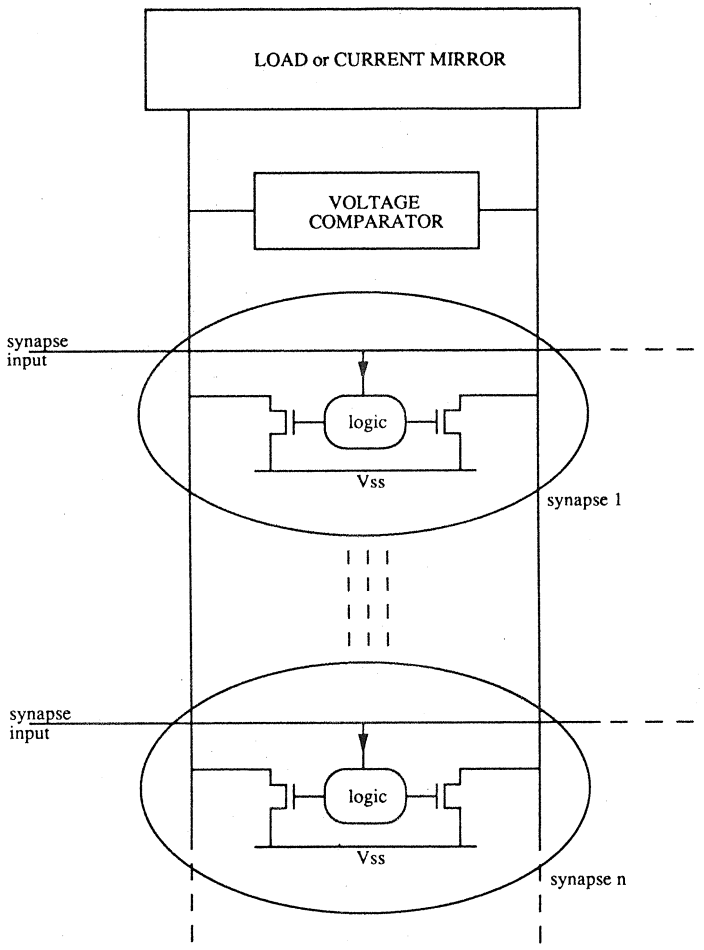

Figure 2 Two-line system

restricted dynamics, and for the same area of silicon a greater number of synapses can thus be integrated.

Each synapse is a programmable current source controlling a differential pair (figure 3). Three connection values are allowed in each synapse. If "mem1" = 1, current is delivered to one of the two lines with the sign of the connection determined by the product of "mem2" and the output of the neuron to which the synapse is connected. If "mem1" = 0, no connection exists between neurons i and j, and no current flows neither to the excitatory and inhibitory lines.

Depending on the state of the XOR function, the current may be sourced either on the line i + or on the line i-. In the neuron, the comparison of the two total currents on the lines i + and i- must be achieved. This is

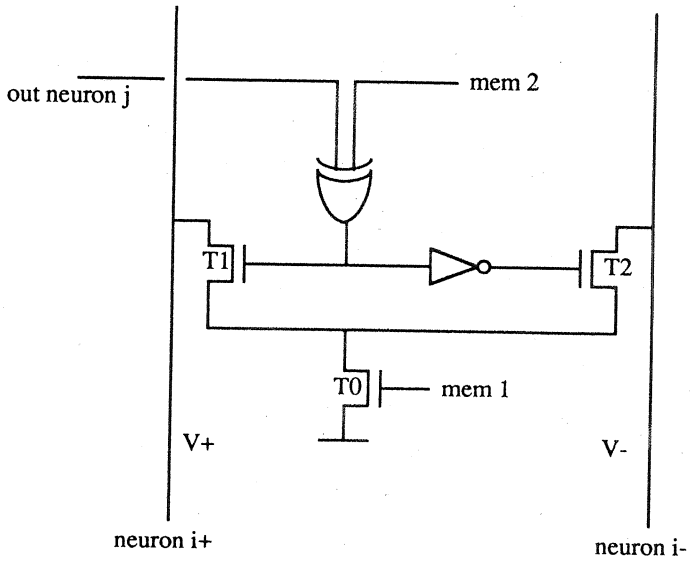

Figure 3 Synapse

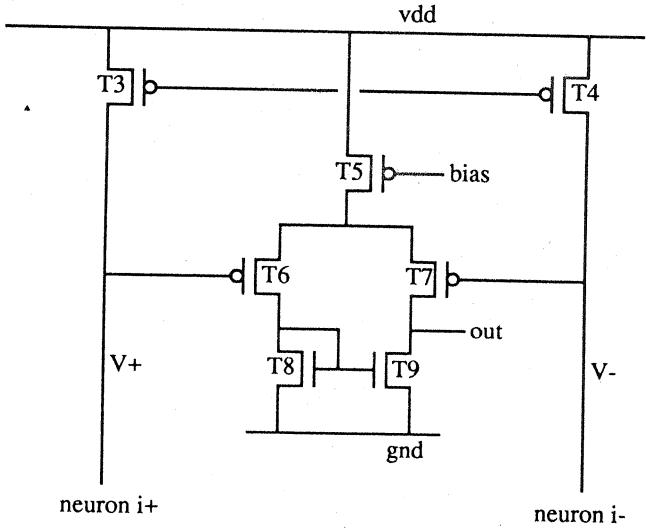

Figure 4 Neuron

done by means of the current reflector shown in figure 4. The currents on the lines are converted into voltages across transistors T3 and T4; these voltages themselves are compared in the differential input reflector formed by transistors T5 to T9. Because of the two-stage architecture of

the neuron, the gain may be very large and the output (out) is either 5V if the current in neuron i- is greater than the one in neuron i +, or OV in the opposite case.

Two-transistor load

In order to convert the two currents into voltages in a two-transistor load, two simple solutions can be considered: two loads (figure 5.a) or a current mirror (figure 5.b).

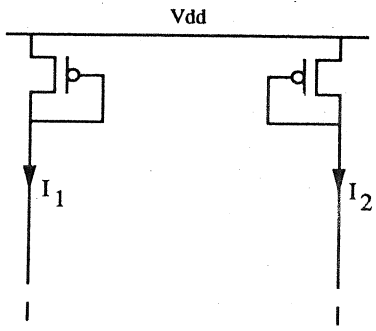

Figure 5a Two loads

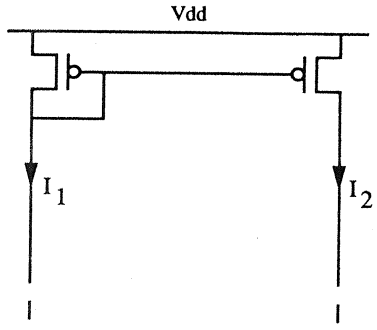

Figure 5b Current mirror

When many current sources are connected to the neuron, the load must be able to discriminate small currents (i.e. one synaptic current) between the two lines, whatever is the common-mode current in these lines (for example, the neuron must have the same behaviour if one excitatory and two inhibitory currents are connected, or with 500 excitatory and 501 inhibitory ones). The ability to discriminate currents in the neuron will of course be enhanced with the differential gain of the load (differential gain is here defined as the voltage difference between the two lines for a given current difference at the input of the load). This gain can easily be computed, assuming the transistors are in saturated mode, and neglecting second-order effects; current can then be expressed by:

$$I = \mu C_{ox} \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2}$$

For the option with the two loads (figure 5.a) we have:

$$\Delta I = \mu C_{ox} \frac{W}{L} \left[\frac{(V_{gs1} - V_t)^2}{2} - \frac{(V_{gs2} - V_t)^2}{2} \right]$$

$$= \mu C_{ox} \frac{W}{L} \left[\frac{(V_{gs1}^2 - V_{gs2}^2)}{2} + V_t (V_{gs2} - V_{gs1}) \right]$$

$$= \mu C_{ox} \frac{W}{L} (V_{gs1} - V_{gs2}) \left[\frac{(V_{gs1} + V_{gs2})}{2} - V_t \right]$$

$$G = \frac{\Delta V}{\Delta I} = \frac{1}{\mu C_{ox} \frac{W}{L}} \frac{1}{\frac{V_{gs1} - V_t}{2} + \frac{V_{gs2} - V_t}{2}} \qquad (1)$$

For the option with the current mirror (figure 5.b), without the Early effect, the same current would flow into the two transistors. The voltage shift will thus be determined only by the Early effect:

$$\Delta V = \Delta I \frac{1}{\frac{I_1}{V_{EAp}} + \frac{I_2}{V_{EAp}}}$$

where V_{EAp} is the Early voltage of p-type transistors, and V_{EAn} the Early voltage of the current sources driving I_2. Gain is thus given by:

$$G = \frac{\Delta V}{\Delta I} = \frac{1}{\frac{I_1}{V_{EAp}} + \frac{I_2}{V_{EAn}}} \frac{1}{\mu C_{ox} \frac{W}{L}} \frac{1}{\frac{(V_{gs1} - V_t)^2}{2V_{EAp}} + \frac{(V_{gs2} - V_t)^2}{2V_{EAn}}} \qquad (2)$$

Comparing (1) and (2), and assuming

$$\frac{V_{gs} - V_t}{V_{EA}} << 1 \ ,$$

the gain in figure 5.b is clearly larger than the gain in figure 5.a.

Once the gain has been computed, the precision of the mirror must be examined. Due to the oxide gradient or other technological imperfections, the threshold voltages of two transistors are never exactly the same; β factors (β = m C_{ox} W/L) can also differ. The impact of the differences in the threshold voltages can be expressed by:

$$I_2 = \frac{\beta}{2} (V_{gs1} - V_t + \Delta V_{tm})^2$$

where V_{tm} is the threshold voltage of the transistors in the current mirror and ΔV_{tm} the possible difference between the V_t of the two transistors. But

$$V_{gs1} = V_t + \left(\frac{2I_1}{\mu C_{ox} \frac{W}{L}} \right)^{\frac{1}{2}}$$

thus

$$I_2 = \mu C_{ox} \frac{W}{L} \frac{1}{2} \left[\left(\frac{2I_1}{\mu C_{ox} \frac{W}{L}} \right)^{\frac{1}{2}} + \Delta V_{tm} \right]^2$$

$$\simeq I_1 + \mu C_{ox} \frac{W}{L} \left(\frac{2I_1}{\mu C_{ox} \frac{W}{L}} \right)^{\frac{1}{2}} \Delta V_{tm}$$

(neglecting second-order effects). The error in the current I_{tm} is thus given by:

$$\Delta I_{tm} \simeq \left(2\mu C_{ox} \frac{W}{L} I_1 \right)^{\frac{1}{2}} \Delta V_{tm}$$

$$\simeq \frac{2}{V_{gs} - V_t} \Delta V_{tm} I$$

The effect of β variations is expressed by:

$$\Delta I_\beta = \frac{\Delta \beta}{\beta} I_1$$

A third error to consider is the mismatching between the two transistors at the input stage of the differential amplifier which will measure the voltage difference between the V_{gs} of the two transistors in the mirror (figure 6). This error is given by:

$$\Delta I_{td} = \frac{\Delta V_{td}}{V_{EA}} I_1$$

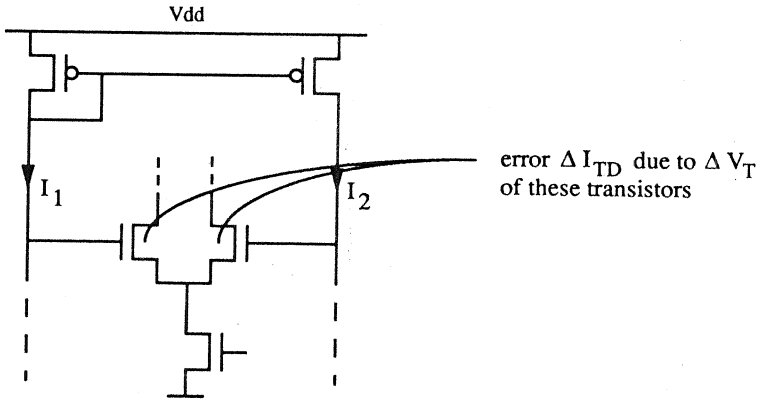

Figure 6 Mismatching of the voltage comparator

In order to compare these three errors, realistic values are chosen:
I: from 0 to 500 µA
$\mu_p C_{ox}$: 1.5 10^{-5} A/V² (standard CMOS process)
W/L: must be chosen to cope with the maximum current (500 µA).
If $V_{gs}-V_t$ can be up to 1.5V, then

$$\frac{W}{L} = \frac{I_{max}}{\mu_p C_{ox} \frac{(V_{gs}-V_t)^2}{2}} \simeq 30$$

V_{EAn}: 20 V
ΔV_{tm}: 10 mV
$\Delta\beta/\beta$: 0.01
ΔV_{td}: 10 mV
(these three last values can be reached with careful design of the mirrors and comparators).

The three currents ΔI_{tm}, ΔI_β and ΔI_{td} are given in figure 7.

The error due to the threshold voltage difference in the current mirror (ΔI_{tm}) is obviously the most important one, especially for small currents. This is due to the fact that this error is proportional to the square root of the size of the transistors in the mirror. Even for small currents, this error

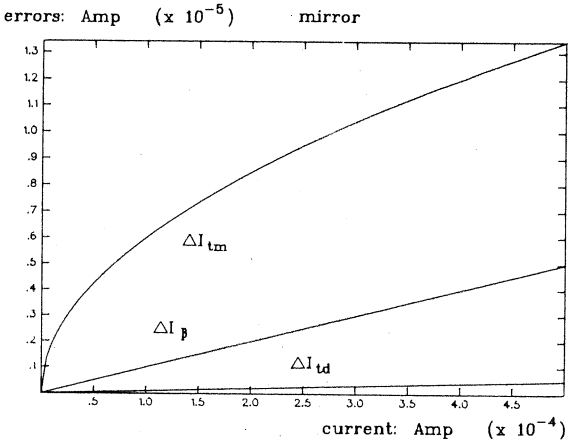

Figure 7 Errors for current mirror

is thus important because these two transistors must be large enough to drive the maximum current (here 500 µA).

One solution to this problem would be to connect several mirrors in parallel, each of them being active only when necessary to drive the total current. This solution is considered in section "multi-transistor load".

Multi-transistor load

A load where several mirrors are connected in parallel through switches is now considered (figure 8). The switches are supposed to be active sequentially, depending on the value of the greatest current among I_1 and I_2; in other words, if this current is I_M, we have:

$0 < I_M \leq I_{ref}$ → 1 load active
$I_{ref} < I_M \leq 2\, I_{ref}$ → 2 loads active
...

where I_{ref} is a given current which will be estimated at the end of this paper.

In order to focus on the advantages of such solution, the same three errors computed in section "two transistor-load" will be estimated, but this time for a load as described in figure 8, with two current mirrors. We suppose first that the switches have no influence on these errors, and that $I_{ref} = I_{max}/2$, where I_{max} is the maximum current in the load (here 500 µA). Since the maximum current I_{max} is supposed to be the same, the size of

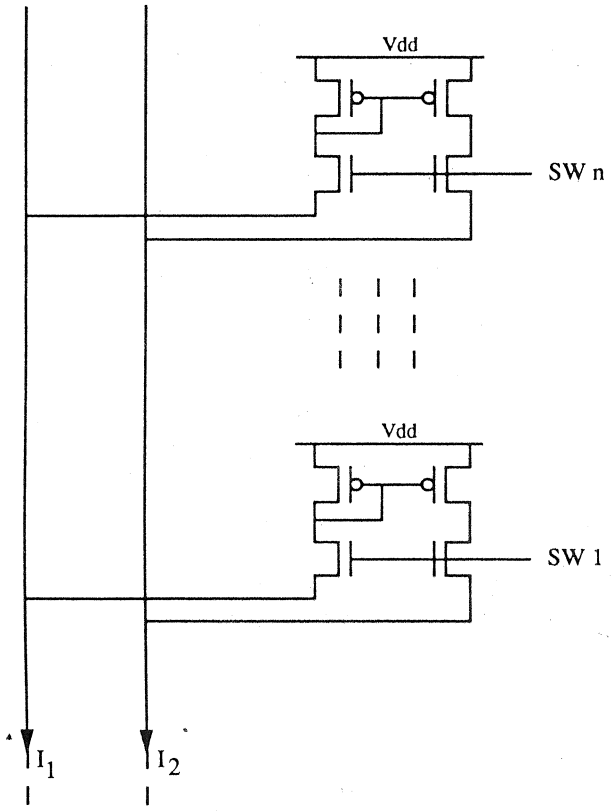

Figure 8 Load with several stages

each current mirror can be reduced to W/L = 15. The values of all other parameters are identical as in section "two transistor-load". Errors ΔI_β and ΔI_{td} do not change; error ΔI_{tm}, however, depends on the W/L of the transistors in the mirrors. As far as only one load is active, ΔI_{tm} is given by:

$$\Delta I_{tm} = \left(2\mu C_{ox}\left(\frac{W}{L}\right)_2 I_1\right)^{\frac{1}{2}} \Delta V_{tm} \quad , where \left(\frac{W}{L}\right)_2 = 15$$

When the two loads are active, error ΔI_{tm} is given by:

$$\Delta I_{tm} = 2\left(2\mu C_{ox}\left(\frac{W}{L}\right)_2 \frac{I_1}{2}\right)^{\frac{1}{2}} \Delta V_{tm}$$

The three errors ΔI_{tm}, ΔI_β and ΔI_{td} are illustrated in figure 9.

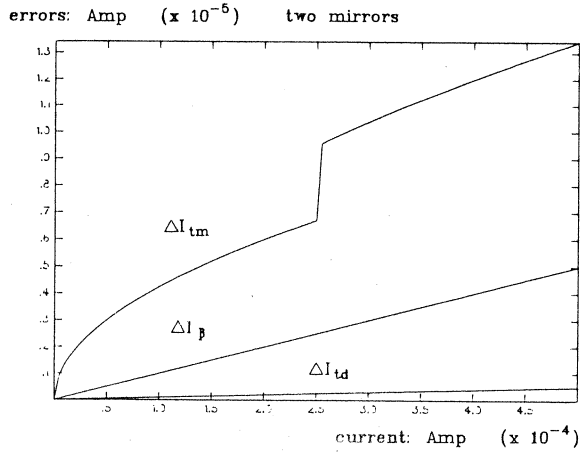

Figure 9 Errors for a load with two mirrors

Two remarks have to be made. First, the error ΔI_{tm} when $I_1 = I_{max}$ does not change between the devices from figure 5.b and figure 8. If the W/L of the transistors in the mirrors are indeed respectively $(W/L)_1$ and $(W/L)_2$, we have for the first case

$$\Delta I_{tm} = \left|2\mu C_{ox}\left(\frac{W}{L}\right)_1 I_{max}\right|^{\frac{1}{2}} \Delta V_{tm}$$

and for the second one

$$\Delta I_{tm} = 2\left|2\mu C_{ox}\left(\frac{W}{L}\right)_2 \frac{I_{max}}{2}\right|^{\frac{1}{2}} \Delta V_{tm}$$

Since $(W/L)_1 = 2\cdot(W/L)_1$, these errors are identical. However, the error ΔI_{tm} for $I_1 < I_{ref}$ is smaller in the second case, due to the fact that the transistors are more efficiently used (a greater current flows in the transistors with respect to their size).

Secondly, the error ΔI_{tm} for $I_1 \geq I_{ref}$ is identical in the two situations, the two devices being equivalent if all the switches are on (the switches are still considered to have no influence on the errors). Furthermore, figure 9 shows a discontinuity in the curve ΔI_{tm} when $I_1 = I_{ref}$. The diminution of the error ΔI_{tm} can thus be improved if this discontinuity is suppressed; a

solution to this problem is presented in section "multi-transistor load with maximum current".

Multi-transistor load with maximum current

The discontinuity in figure 9 can be suppressed if one of the two loads gets his maximum current (I_{ref}) and the other one the remaining current (I_1-I_{ref}) (see figure 10).

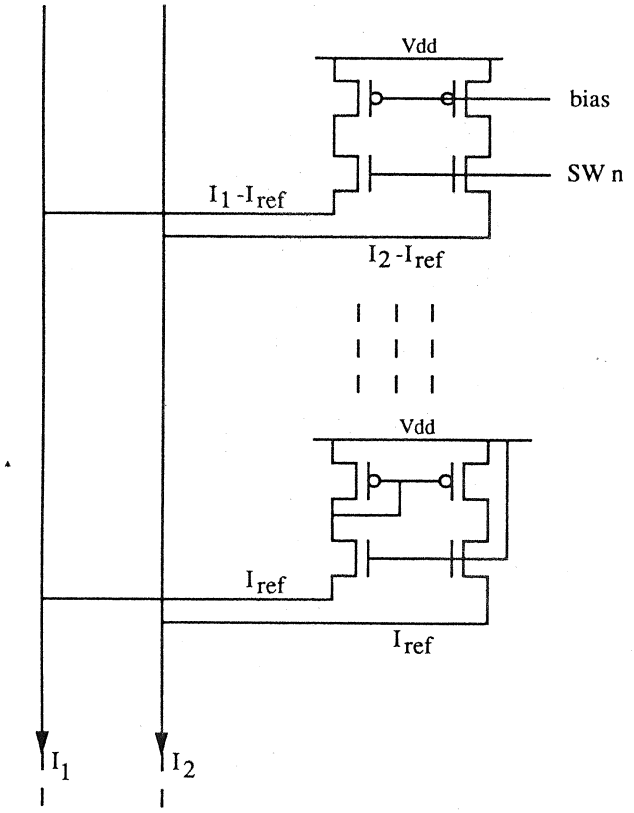

Figure 10 Loads with maximum currents

Error ΔI_{tm} is then given by:

$$\Delta I_{tm} = \left[2\mu C_{ox} \left(\frac{W}{L} \right)_2 I_1 \right]^{\frac{1}{2}} \Delta V_{tm} \qquad if I_1 \leq I_{ref}$$

$$\Delta I_{tm} = \left[2\mu C_{ox} \left(\frac{W}{L} \right)_2 I_{ref} \right]^{\frac{1}{2}} \Delta V_{tm} + \left[2\mu C_{ox} \left(\frac{W}{L} \right)_2 (I_1 - I_{ref}) \right]^{\frac{1}{2}} \Delta V_{tm}$$

$$if I_2 \geq I_{ref}$$

The three errors ΔI_{tm}, $\Delta I\beta$ and ΔI_{td} are illustrated in figure 11.

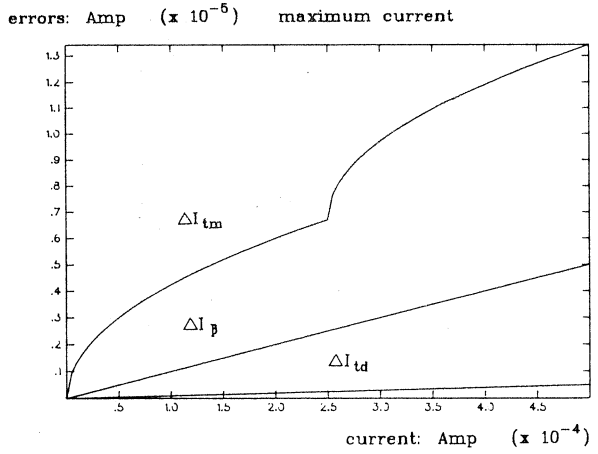

Figure 11 Errors for a load with maximum currents

VLSI NEURON

The solution of section "multi-transistor load with maximum current" can be used to implement a VLSI neuron for an artificial neural network where the number of synapses connected to a single neuron is important. In order to avoid changes in the current flowing through the synaptic current sources, an operational amplifier in a feedback loop is introduced as shown in figure 12. By this way, the drain voltage of the current sources is kept fixed, and the synaptic currents remain identical whatever is the total current in the load. Furthermore, with this feedback loop, the current in the two lines is directly determined by the synapses, and parameter variations in the switches have thus only second-order effect.

The principle explained in section "multi-transistor load with maximum

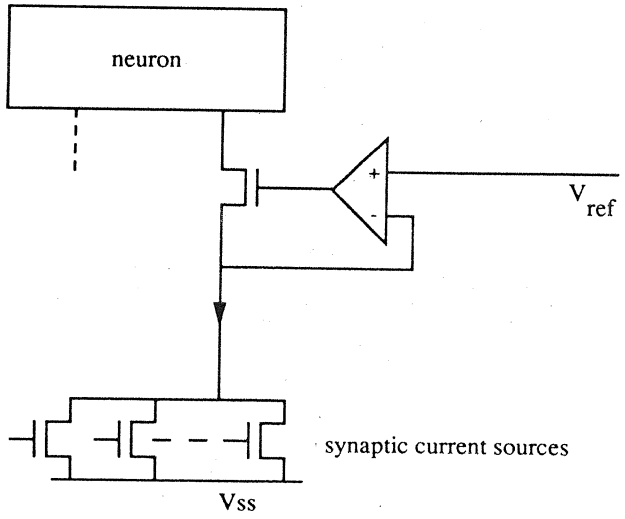

Figure 12 Fixed synaptic current

first stage will be always connected to the sources, while the second one will only be connected when $I_M = \max(I_1, I_2)$ is greater than I_{ref}, the third one when I_M is greater than $2.I_{ref}$, and so on. An efficient I_{ref} will be computed in section "Number of stages in the neuron".

The switches which connect the successive loads can be driven by a device as illustrated in figure 13.

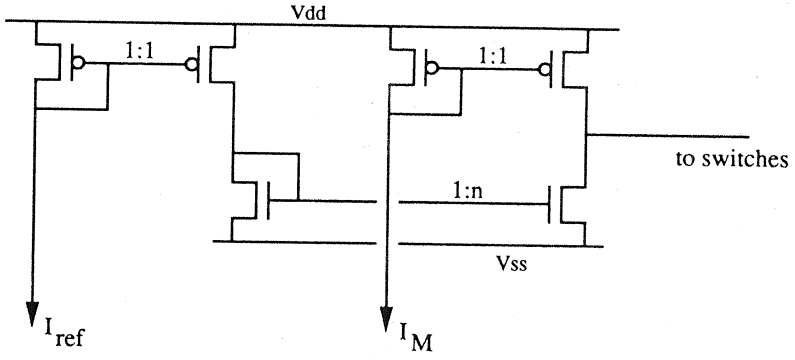

Figure 13 Command for switches

The ratio n:1 used in the N-type current mirror depends on the current n.Iref to which I_M must be compared. This device needs the current I_M as input; this could be done by inserting another cell which generates the

maximum of the currents I_1 and I_2. However, I_M can be replaced by I_1 or I_2 without loss of performance. It is not very important, indeed, if the loads are not activated exactly at I_{ref}, $2I_{ref}$, $3I_{ref}$,... but well at an approximation of these values. If the currents I_1 and I_2 are thus nearly identical, one of these two currents can be used for I_M. If I_1 and I_2 are quite different, the circuit will work properly if, by chance, the current I_1 and I_2 which is chosen to drive the cell of fig.10 is greater than the other. If this is not the case, say $I_M = I_1$ and $I_1 << I_2$, only the number of loads necessary to drive properly I_1 will be active. The transistors driving I_2 will not be sufficient for such a current, and their drain voltage will thus spectacularly decrease. In this case, it will not be difficult to discriminate between I_1 and I_2 with a simple comparator; the current I_M can thus be replaced for example by I_1. An important point is to avoid to duplicate the current I_1 at the output of the synapses; the advantage of the circuit would indeed be lost because of the imperfections in the mirrors used to duplicate the current. A second, less precise, current I_1 has thus to be generated directly in the synapses. The complete circuit is shown in figure 14; all the mirrors have unity ratios, except those indicated in the figure.

NUMBER OF STAGES IN THE NEURON

The last question to solve is the choice of the current I_{ref}, and so to decide the optimum number of stages in the neuron. First, the number n of stages in the neuron and the current I_{ref} are related by:

$$n\, I_{ref} \leq I_{max} < (n+1)\, I_{ref}$$

where I_{max} is the maximum current the neuron has to drive.The link between this architecture to compute analog sum-of-products and neural networks can now be restored. The function to realize is the sum of fixed synaptic currents, and the logic comparison between the total excitatory and inhibitory currents. If I_{syn} is one single synaptic current, the error $\Delta I_{tot} = \Delta I_{tm} + \Delta Ib + \Delta I_{td}$ has no influence on the logic comparison as long as $\Delta I_{tot} < I_{syn}$. This relation can be developed:

$$\left[2\mu C_{ox}\frac{W}{L}I\right]^{\frac{1}{2}}\Delta V_{tm} + \frac{\Delta\beta}{\beta}I + \frac{\Delta V_{td}}{V_{EA}}I < I_{syn}$$

In section "multi-transistor load", the fact was proven that the introduction of several mirrors does not change the error when $I = I_{max}$. It can also be shown that the errors computed at values of the current which switch on a new stage are proportional to this current. The optimum number of stages will thus be:

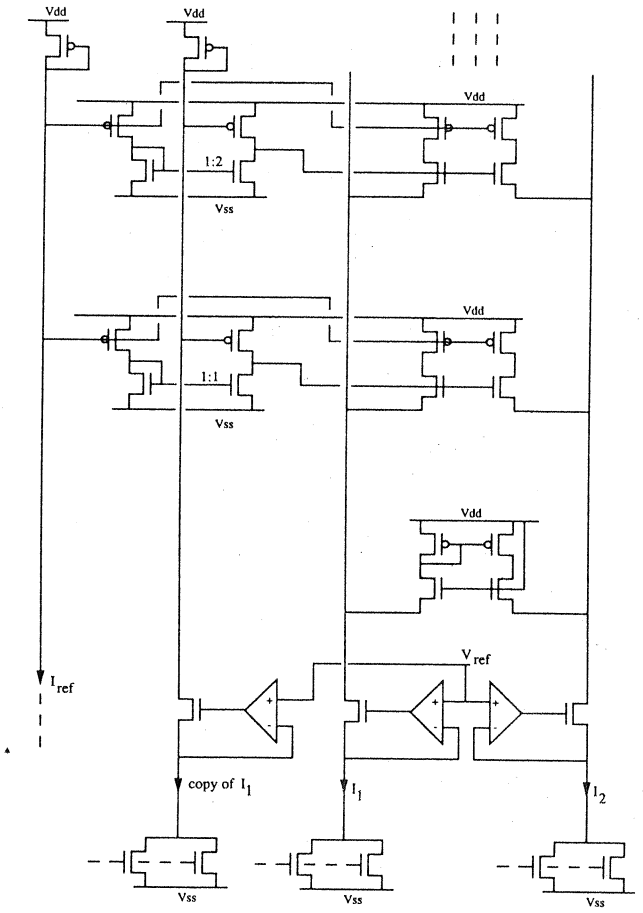

Figure 14 Complete neuron

$$n = int\left\{ \frac{\left[2\mu C_{ox}\left(\frac{W}{L}\right)_{tot} I_{max}\right]^{\frac{1}{2}} \Delta V_{tm} + \frac{\Delta\beta}{\beta} I_{max} + \frac{\Delta V_{td}}{V_{EA}} I_{max}}{I_{syn}} + 1 \right\}$$

where $(W/L)_{tot}$ is the size of a transistor which could drive the total current I_{max}. The value of I_{ref} is then given by:

$$I_{ref} = \frac{I_{max}}{n}$$

This value of I_{ref} corresponds to a maximum error of I_{syn}. It would probably be useful to have a security factor on the allowed error, i.e. to replace I_{syn} by 0.9 I_{syn}. It can easily be verified that if the error with a current I is less than I_{syn}, then the error with a current 2I will be less than $2I_{syn}$, and so on. This value of I_{ref} is optimum, because the current for which the error is less than I_{syn} is maximum (and also for $2I_{syn}$,...). It would be unprofitable to enhance the number of stages, and thus to decrease I_{ref}; the error current indeed would decrease in absolute value, but not in terms of integer multiples of I_{syn}; this would thus have no effect on the logic comparison between the total excitatory and inhibitory currents.

TEST CHIP

A test chip has been realized according to the design of figure 14. The cells implement a 4-stage neuron, with automatic switching of the different stages. A microphotograph of the cell is given in figure 15. The large dimensions of the cell (600 x 886 µm) are due to the fact that it has not been optimized (for example, very large transistors were used in the current comparators).

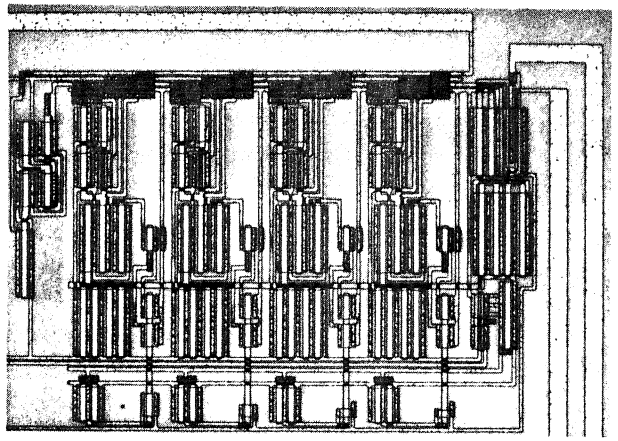

Figure 15 Microphotograph of the cell

CONCLUSION

A method is presented to reduce the errors due to mismatching of components in a VLSI neuron used in a neural network where the information is transmitted by currents. Since the number of neurons in a chip is much less than the number of synapses, the loss of area due to this neuron is not very significant. However, the errors in the decisions taken by the neurons are reduced, especially when relatively few synapses are connected (for example is sparsely-coded memories).

Furthermore, another improvement of this neuron, but which cannot be precisely predicted, is the fact that if the mirror is splitted into several parts, the probability that one mismatching between components will be compensated by another mismatching is enhanced. The neuron must of course be carefully designed, for example by physically inverting half of the mirrors, in order to compensate for oxide gradients.

ACKNOWLEDGEMENTS

All our acknowledgements go to Brigitte Wénin-Dupont, who developed and helped us to use Bananas, a graphical software used to plot the simulations of this paper. This work has been partially financed by the ESPRIT-BRA project NERVES.

REFERENCES

[1] E. Vittoz and X. Arreguit, "CMOS integration of Herault-Jutten cells for separation of sources", Analog implementation of neural systems, C. Mead and M. Ismail eds., Kluwer Academic Publishers, Norwell, MA, 1989.

[2] M. Sivilotti, M. Mahowald and C. Mead, "Real-time visual computation using analog CMOS processing arrays", Proceedings of the 1987 Stanford conference on advanced research in VLSI, P.Losleben ed., MIT Press 1987.

[3] A. F. Murray, "Pulse arithmetic in VLSI neural networks", IEEE Micro, vol.9, n°6, December 1989.

[4] H. P. Graf and P. de Vegvar, "A CMOS implementation of a neural network model", roceedings of the 1987 Stanford conference on advanced research in VLSI, P.Losleben ed., MIT Press 1987.

[5] M. Verleysen, B. Sirletti, A. Vandemeulebroecke and P. Jespers, "Neural networks for high-storage content-addressable memory: VLSI circuit and learning algorithm", IEEE Journal of Solid-State Circuits, vol.24, n°3, 1989.