Precise Analog Synapse for Kohonen Feature Maps

P. Heim and E. A. Vittoz

Abstract— **A plastic medium-term analog synapse is presented that fulfils the stringent specifications necessary for the Kohonen algorithm. The principle is based on a switched capacitor-like technique implementing a variable time-constant integrator. The memory leakage standard deviation is 2 mV/s for a voltage range of 2 V at room temperature and the learning gain can be varied over two decades. Its differential structure leads to good CMRR, PSRR, and charge injection cancellation. The total synapse area is 1/16 mm^2 using a 3-μm self-aligned contact single-metal CMOS technology. Measurement results of a test chip are also presented.**

I. Introduction

While many papers have been published on the applications of the Kohonen network, as for example in [1], very few hardware implementations have been proposed yet. This may be because the collective functions of the network are difficult to implement with digital techniques, and classical accurate analog circuits need excessive area. Although there are efficient analog implementations for the collective functions such as the winner-take-all (WTA) [2] or the nonlinear network for neighborhood generation [3], only few implementations of a synapse that might fit the requirements of the Kohonen algorithm have been reported recently [4]–[6].

Before introducing the synapse, the impact of some typical inaccuracies on the network's behavior are illustrated to show the level of accuracy required for the synapses. Then the circuit principle is described and some aspects of the implementation are detailed. Finally, measurements of a test chip demonstrate the validity of the design.

II. The Kohonen Map

The Kohonen map [7] is a large one- or two-dimensional ordered array of cells that exhibits very interesting pattern classification and clustering properties. Each cell memorizes a vector $\boldsymbol{m}$ whose components are called synaptic weights. The algorithm is briefly described as follows: a sequence of input vectors $\boldsymbol{x}$ from a data base are presented to all the cells. For each vector the network finds the cell whose weight $\boldsymbol{m}$ best matches $\boldsymbol{x}$ and defines a learning neighborhood around it, in which the cells will be updated according to the following rule (for each component i):

$$m_i(t_{k+1}) = m_i(t_k) + \alpha(t_k)[x_i(t_k) - m_i(t_k)] \qquad (1)$$

Manuscript received December 17, 1993; revised April 2, 1994. This work has been supported by the Fondation Suisse pour la Recherche en Microtechnique (FSRM).

P. Heim was with Ecole Polytechnique Federale de Lausanne (EPFL), Laboratoire d'electronique generale, ELB Ecublens, 1015 Lausanne, Switzerland. He is now with SEDAL, Department of Electrical Engineering, University of Sydney, NSW–2006, Australia.

E. A. Vittoz is with Ecole Polytechnique Federale de Lausanne (EPFL), Laboratoire d'electronique generale, ELB Ecublens, 1015 Lausanne, Switzerland.

IEEE Log Number 9402470.

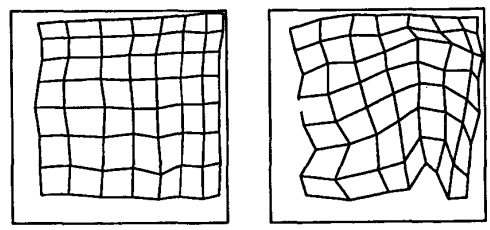

Fig. 1. Effects of inaccuracies on network's behavior.

where $\alpha(t_k)$ is the time dependent learning gain, close to 1 at the beginning or *organization phase* of the process and decreasing to an arbitrary low value during the *convergence phase* of the process. The learning neighborhood is also a decreasing function of time, reduced to the nearest neighbors of the winning cell in the convergence phase.

Although this process is intrinsically collective, small alterations of (1) can drive the network to unpredictable states, and there is no evidence of a simple solution apart from reducing the circuit inaccuracies. Among the various possible inaccuracies of the synapse, two are inherent to storage on capacitors: charge leakage and charge injection. The first is continuous and permanently affects all the cells of the network. The second affects only the cells being updated. To illustrate this, simulations have been made with a two-dimensional database representing a square with uniform distribution. The size of the square corresponds to the full range of the synapses. Fig. 1 shows the effects of leakage and charge injection using the usual representation [7] of the network's mapping: the nodes of the lattice represent the $\boldsymbol{m}_i$ weight vectors in the same coordinate system as the database and the lines show the topological disposition of the neurons. In the absence of inaccuracies, the mapping is regular and evenly distributed. The effect of a leakage which coresponds to 200 mV/s for a data voltage range of 1 V at a learning rate of 10000 iterations/s is shown in Fig. 1(a).

The effect of charge injection is shown in Fig. 1(b) for a distribution of mean value $|c|$ and standard deviation σ (with respect to full range). This simulation shows a qualitative result for which the perturbating term corresponds to about 15% of the average update value. The effect depends on various parameters such as the neighborhood size, the learning gain, and, more constraining, on the number of neurons and the database structure, which fixes the average distance between neurons, as reported in [8].

III. Synapse With Learning Capability

The rule (1) used to update the synaptic weights describes a discrete-time low-pass filter that could be implemented with a

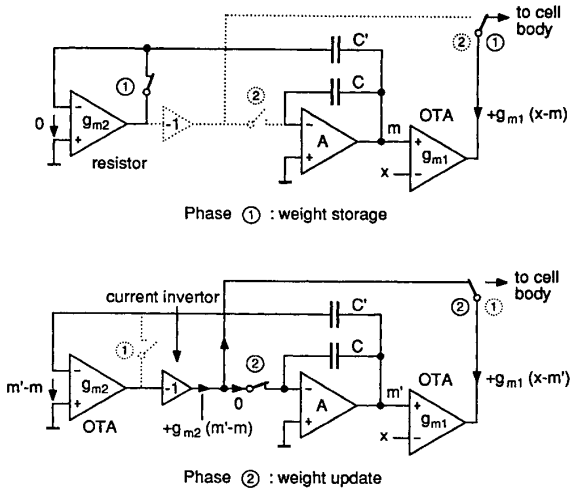

Fig. 2. Principle of the proposed synapse.

classical switched-capacitor (SC) technique if α were constant. To make α variable, a modified technique uses current equalization, instead of the usual charge redistribution, to modify the charge on a capacitor. To implement (1) with this technique, the actual weight value $m = m(t_k)$ must be maintained while computing the new weight value $m' = m(t_{k+1})$, which needs a second capacitor C'. The principle of the synapse is shown in Fig. 2, emphasizing its two steady states. The synapse is made of two transconductance amplifiers (OTA), a high-gain amplifier A, a current invertor, two capacitors and four switches.

During phase 1 (weight storage), the weight m is held on capacitor C and copied to capacitor C'. The updating phase starts with the opening of switches 1, holding m on C'. When switches 2 close, the output current of OTA g_{m1} charges C. The resulting variation of m is applied at the input of g_{m2} by means of C', which generates an output current $-g_{m2}(m' - m)$. This current is then inverted and subtracted from the other charging current. When equalization occurs, the new weight value $m(t_{k+1})$ is fixed according to:

$$g_{m2}[m(t_{k+1}) - m(t_k)] = g_{m1}[x(t_k) - m(t_{k+1})]. \quad (2)$$

Identification with (1) results in the following gain:

$$\alpha = \frac{g_{m1}}{g_{m1} + g_{m2}} \quad (3)$$

which can be controlled by varying either g_{m1}, g_{m2} or both.

Fig. 3 shows a differential version of the synapse. The principle, which is commonplace in SC implementations, further simplifies the circuit because its crossing branches replace the current invertor and its inherent rejection of common-mode signals acts as a charge injection compensation technique. However, it requires 4 capacitors and the amplifiers need output common-mode feedback (CMFB). The eventual mismatch between the two g_{m1} values only affects the internal value of m but does not affect the network behavior.

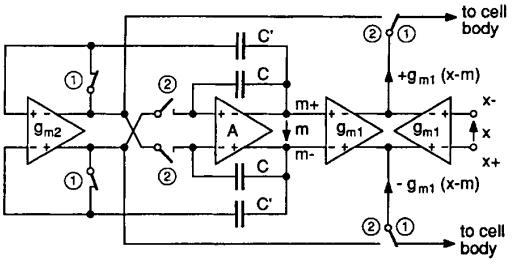

Fig. 3. Differential version of the synapse.

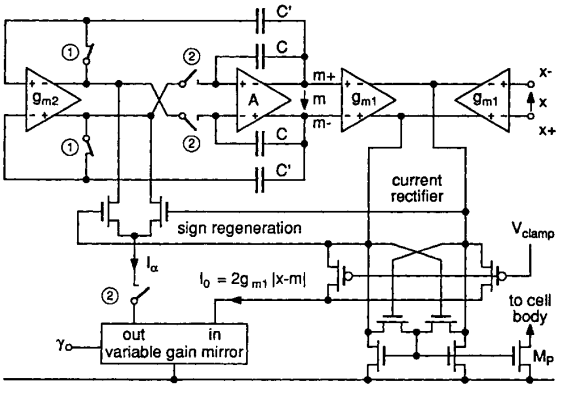

Fig. 4. Synapse version with fixed transconductances.

IV. Synapse With Fixed Transconductances

Because it is difficult to implement a variable transconductance OTA while keeping a wide input saturation voltage range, we have developed a modified version of the synapse using fixed-transconductance OTA's only, as shown in Fig. 4.

The variable gain is implemented by means of a one-quadrant translinear current multiplier [9], used as a variable gain current mirror. For this purpose, the output current of g_{m1} is rectified [10] and multiplied by a coefficient γ to produce the update current I_α. The sign is regenerated with a source-coupled pair which directs the current I_α to the correct node. The subsequent disequilibrium is compensated by the action of the CMFB of g_{m2}, which mimics complementary symmetrical injected currents of value $I_\alpha/2$.

The current at the output of the rectifier

$$I_0 = 2g_{m1}|x - m| \quad (4)$$

is injected at the outputs of g_{m2} after multiplication by γ and division by 2 through the above-mentioned action of the CMFB. This leads to the following new value for α:

$$\alpha = \frac{\gamma}{g_{m2}/g_{m1} + \gamma} \quad (5)$$

which is a weakly nonlinear function of γ and may therefore not be taken into account when programming the temporal evolution of α.

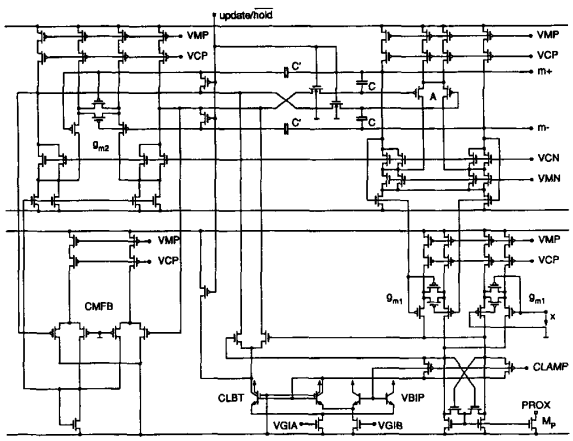

Fig. 5. Complete schematic of the synapse.

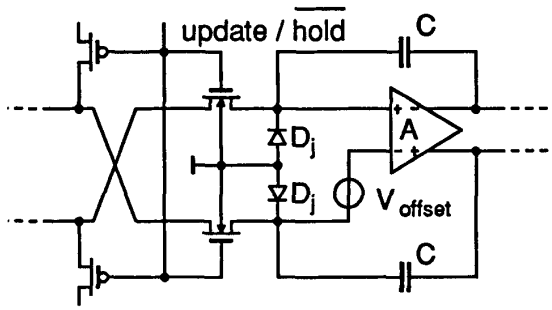

Fig. 6. Implementation of the switches.

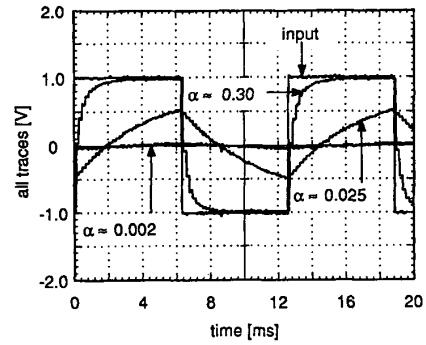

Fig. 7. Synapse working as a variable time-constant.

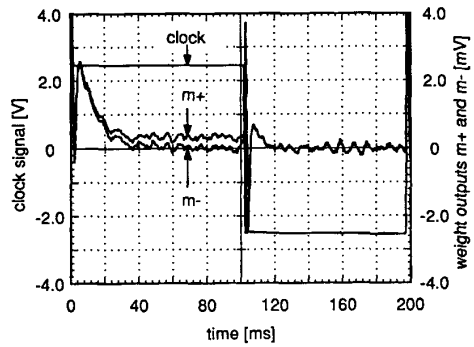

Fig. 8. Measurement of charge injection with a clock rise and fall time of 10 ns.

The function of transistor M_P is to provide a first order (or Manhattan) proximity measure to the cell body, to select the cell whose weight $\boldsymbol{m}$ best matches $\boldsymbol{x}$. This proximity measure is a current which is maximum when $x = m$, and perfectly centered on the zero-crossing of the rectifier, i.e., it has its maximum value for $I_0 = 0$ whatever the mismatch of the transistors.

V. Circuit Implementation

The complete schematic of the synapse is shown in Fig. 5. The circuit uses ± 2.5 V supply voltages. The ground (GND) is the reference voltage for the CMFB's and thus fixes the switches' steady state potentials and the virtual ground at the inputs of A. The technique used to reduce the leakage currents is similar to that proposed in [11] and is based on the reduction of the voltage across the drain to bulk junctions of the access switches as shown in Fig. 6. For this purpose the separate well of the switches is tied to GND and the input offset voltage of A appears across the series back-to-back diodes. If there are residual common-mode currents, they will charge the two capacitors until the forward current of one diode equals the reverse current of the other diode.

Since all the nodes are at potentials close to GND, the two phases can be implemented using complementary type transistors driven with the same command signal. This makes the two phases inherently non-overlapping, and the large negative voltage (-2.5 V) on the gates of the n-channel transistors suppresses all residual weak inversion current.

The compensation of charge injection is based on the symmetry of the circuit. The resulting value thus depends on the matching of the switches which is roughly 10% (minimum size transistors). The estimated value using the method described in [12] is less than 0.5 $\mathrm{mV_{RMS}}$, the capacitors C and C' having a value of about 0.8 pF.

The OTA's are implemented around a folded cascode structure, leading to high dc gain and avoiding stability considerations. g_{m1} and g_{m2} use a linearized differential pair [13] with a saturation voltage of 1.25 V, allowing a 2-V differential input voltage range with 25% margin.

The one-quadrant multiplier is made using two compatible lateral bipolar transistors (CLBT) and two vertical bipolar transistors (VBIP), and is controlled by the ratio of two currents. These currents are distributed throughout the chip by the voltages V_{GIA} and V_{GIB}. Its area is 61 μm $\times$ 64 μm.

VI. Circuit Measurements

Fig. 7 shows three measurements on the synapse working as a time constant with very different values to show its wide range of operation. The clock frequency is 5 kHz (continuous

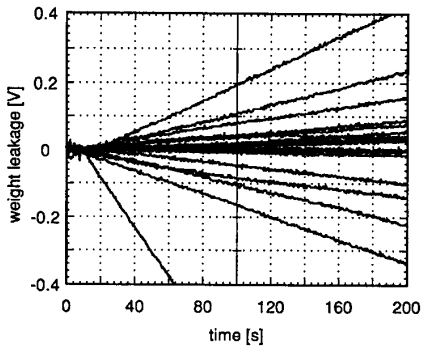

Fig. 9. Leakages of 20 sample synapses (about 3 min elapsed time).

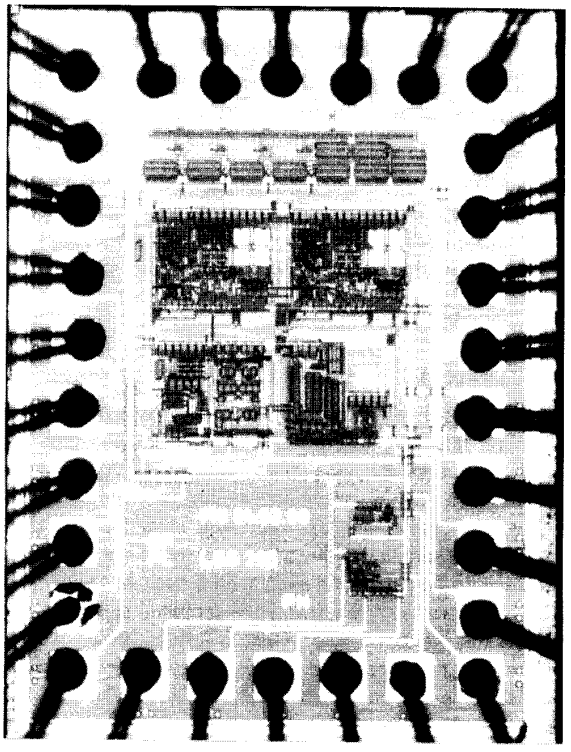

Fig. 10. Microphotograph of the chip (1.2×1.6 mm^2).

learning) and the period of the input signal is 64 times larger than that of the clock. $\alpha = 0.3$ is the largest possible gain for this design, and $\alpha = 0.002$ is the smallest gain value for which the weight m was still maintained within the midrange of the input signal.

The measurement of charge injection is made under continuous learning with the input $x = 0$. Under these conditions, the weight m will converge to a value for which the update value $\alpha(x-m)$ is just sufficient to compensate for the injected charge. Fig. 8 shows the switches' command (clock) and the differential outputs ($m+$ and $m-$) of a synapse which has injected charge corresponding to 0.4 mV on the synaptic weight. This value is equal to the standard deviation obtained for clock rise time and fall time both equal to 10 ns. It is comparable to the value used for the simulation of Fig. 1(b).

The leakage measurements on 20 synapses have been grouped in Fig. 9, showing a Gaussian-like distribution with nearly zero mean value and a sigma of 2 mV/s at room temperature. This corresponds to a leakage current of 1.6 fA with the 0.8 pF capacitors and is two orders of magnitude smaller than the value used for Fig. 1(a). This is small enough to use the network under continuous training. Although the technology used has quite large junction reverse currents, these results confirm the efficiency of the technique.

Finally, a microphotograph of the test chip is shown in Fig. 10. The chip includes output buffers (on top), two synapses (just below), biasing circuits and a separate one quadrant multiplier for direct measurement. The synapses have a size of 250 μm $\times$ 250 μm, and can be abutted together both horizontally and vertically.

VII. Conclusion

A plastic analog medium-term memory implementing a synapse that can fulfil the requirements for the implementation of a Kohonen map is reported. The use of a differential switched capacitor-like structure together with an efficient leakage reduction technique has led to an accurate cell using a single control signal. Significant area was consumed in achieving such performance, and further investigation is required to simplify the circuit and increase density.

References

[1] T. Kohonen, K. Mäkisara, O. Simula and J. Kangas, Eds., *Artificial Neural Networks,* North-Holland: Elsevier Science Publishers B.V., 1991.

[2] J. Lazzaro *et al.*, "Winner-take-all networks of order *N* complexity," in *Proc. 1988 IEEE Conf. on Neural Information Processing—Natural and Synthetic*, Denver, CO, 1988, pp. 703–711.

[3] P. Heim, B. Hochet and E. Vittoz, "Generation of learning neighbourhood in Kohonen feature maps by means of simple nonlinear network," *Electron. Lett.*, vol. 27, no. 3, Jan. 31, 1991.

[4] O. Landolt, "An analog CMOS implementation of a Kohonen network with learning capability," *Int. Workshop on VLSI for Neural Networks and Artificial Intelligence,* Oxford, 1992.

[5] D. Macq, M. Verleysen, P. Jespers and J. D. Legat, "Analog implementation of a Kohonen map with on-chip learning," *Trans. Neural Networks,* vol. 4, no. 3, May 1993.

[6] Y. He and U. Cilingiroglu, "A charge-based on-chip adaptation Kohonen neural network," *Trans. Neural Networks,* vol. 4, no. 3, May 1993.

[7] T. Kohonen, *Self-Organization and Associative Memory*, Berlin: Springer Verlag, 1988.

[8] P. Heim, "CMOS analog VLSI implementation of a Kohonen map," Ph.D. thesis no. 1174, Swiss Federal Intitute of Technology, Lausanne, Switzerland, 1993.

[9] E. Seevinck, "Analysis and synthesis of translinear integrated circuits," Elsevier, 1988.

[10] P. Heim, F. Krummenacher and E. A. Vittoz, "CMOS full-wave operational transconductance rectifier with improved DC transfer characteristic," *Electron. Lett.*, Jan. 30, 1992, vol. 28, no. 3.

[11] E. A. Vittoz *et al.*, "Analog storage of adjustable synaptic weights," *Proc. of the ITG/IEEE Workshop on Microelectron. for Neural Networks,* Dortmund, Germany, June 25–26, 1990.

[12] G. Wegmann, E. A. Vittoz and F. Rahali, "Charge injection in analog MOS switches," *IEEE J. Solid-State Circuits,* vol. 22, no. 6, p. 1091, Dec. 1989.

[13] F. Krummenacher and N. Joehl, "A 4 MHz CMOS continuous-time filter with on-chip automatic tuning," *IEEE J. Solid-State Circuits,* vol. 23, pp. 750–758, June 1988.