

This paper was recipient of an OUTSTANDING PAPER AWARD at MIXDES99

**VLSI DESIGN OF A FLEXIBLE-STRUCTURE SEQUENTIAL
MIXED-SIGNAL NEURAL PROCESSOR**

contact author: madrenas@eel.upc.es

*Proceedings of the 6th International Conference Mixed-Signal Design of
Integrated Circuits and Systems (MIXDES'99), Kraków (Poland), June 1999, pp 259.*

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reported without the explicit permission of the copyright holder.

VLSI DESIGN OF A FLEXIBLE-STRUCTURE SEQUENTIAL MIXED-SIGNAL NEURAL PROCESSOR

J. MADRENAS, E. ALARCÓN, J. COSP AND J.M. MORENO

UNIVERSITAT POLITÈCNICA DE CATALUNYA , BARCELONA, SPAIN

KEYWORDS: Mixed-signal Design, Neural Networks, Current-Mode, Charge Injection, Analog VLSI

ABSTRACT: A mixed-signal circuit that can emulate multi-layer perceptron neural networks is presented. The sequential processing provides a tradeoff between speed, power consumption and circuit area, allowing a flexible accommodation of different network structures. The design of the three most critical analog basic blocks of the processor is presented. a) Dynamic analog memories: A technique of switching error cancellation is proposed. b) Semialgorithmic DAC multipliers. They perform an analog-digital product., so the network parameters are stored in digital form. c) Sigmoid circuit. A winner-take-all based sigmoid circuit is used to fulfill the system requirements. The common-mode currents are cancelled by means of a common-mode feedback. Concerning the digital part, a general-purpose programmable digital control circuit with a high degree of circuit control flexibility is presented. Finally, the results and ongoing work are discussed.

1 INTRODUCTION

The good performance of artificial neural networks in solving complex tasks such as classification, prediction, interpolation, information compression or recognition is well-known [1]. Although the most popular model is clearly the Multi-Layer Perceptron (MLP), a large amount of different neural models have been extensively proposed, especially in the recent literature. Given the large parallelism that neural models exhibit, their hardware implementation provides important benefits in comparison to a software emulation in a group of applications. This is the case of applications requiring one or more of the following conditions: fast execution, portability, low-power consumption, reduced physical volume and weight, in many cases in connection with embedded systems. Both digital and analog implementations have been reported, each one presenting specific tradeoffs.

Analog neural network implementations are attractive because of the compact and low-power realization of the processing elements [2]. However, they suffer from the following drawbacks: 1. Limited dynamic range, and therefore, signal-to-noise ratio. 2. Analog memories, which are difficult to implement. 3. Poor repeatability because of manufacturing process tolerances and temperature-dependent parameters. 4. Lack of flexibility in front of the digital programmability.

The first limitation is especially critical for the learning phase. During execution, the required precision is more relaxed in general and can be performed by analog processing. The second limitation can be overcome using either digital memory or CMOS-compatible floating-gate analog memories for weight storage, and current copiers for intermediate data storage [3,4]. Concerning to the third limitation, appropriate design

techniques and chip-in-the-loop learning can reduce its effect. To overcome the last limitation, a systolic-ring analog architecture was reported [5]. Later, the architecture was modified to a sequential one to minimize the analog signal degradation [6].

The architecture principle is based in the sequential emulation of each neuron in a layer-by-layer basis. Briefly explained, first the input data vector is stored in a row of Analog Transfer Cells (ATC). Then, all the synapses of one neuron and its activation function are computed in one clock cycle. The neuron outputs are stored in a second row of ATCs. After emulating all the neurons on the first neuron layer, the computed outputs are used as inputs to emulate the next layer. The procedure is repeated until the last layer is emulated. Two rows of n ATCs, n synapses and one activation function are enough to emulate a network of an arbitrary number of layers, each one consisting of up to n neurons.

Benefits of this architecture are: reduced routing overhead, compact circuitry, unlimited number of layer emulation capability, direct scalability and very flexible network structure. The maximum number of neurons in a single layer is only limited by the implemented number of synapses.

It has been shown that this architecture can emulate a broad number of neural models [7]. In this paper we report the design of an MLP version of the architecture. In the next section, the processor is described at a block level. In sections 3-6, the design of the main analog building blocks and the digital control are discussed.. Finally, a the integration of the blocks to form a processor is briefly explained, and the concluding remarks and ongoing work are pointed out.

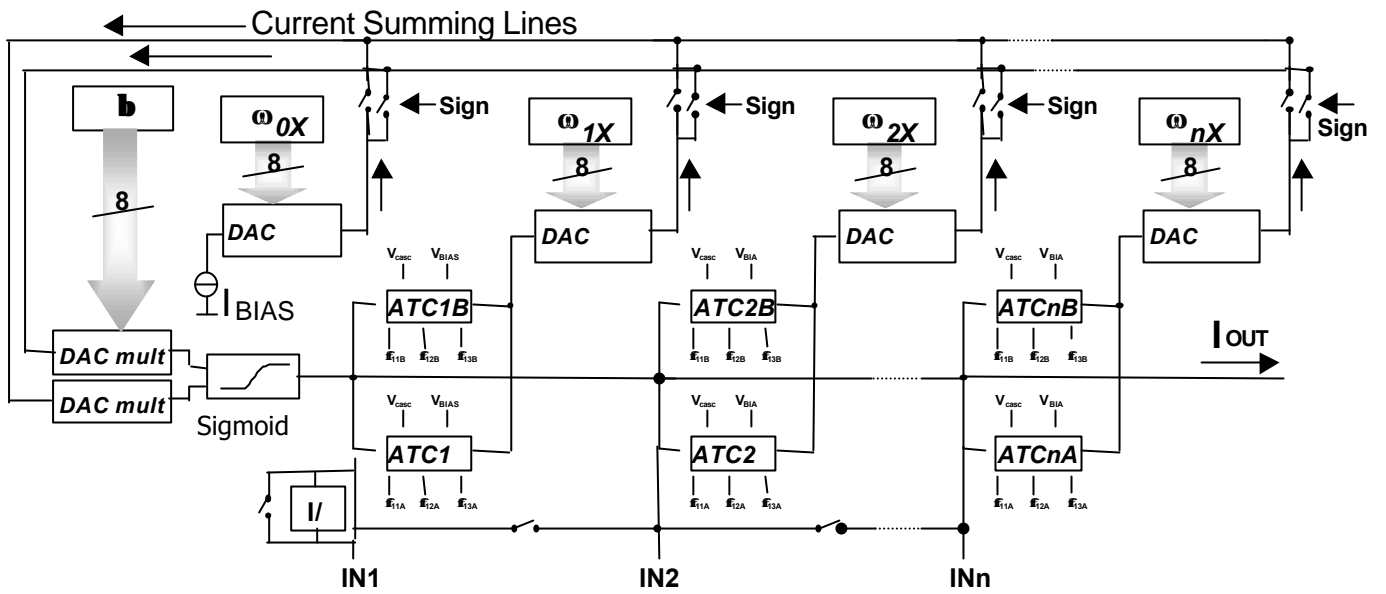


Fig. 1. Analog datapath

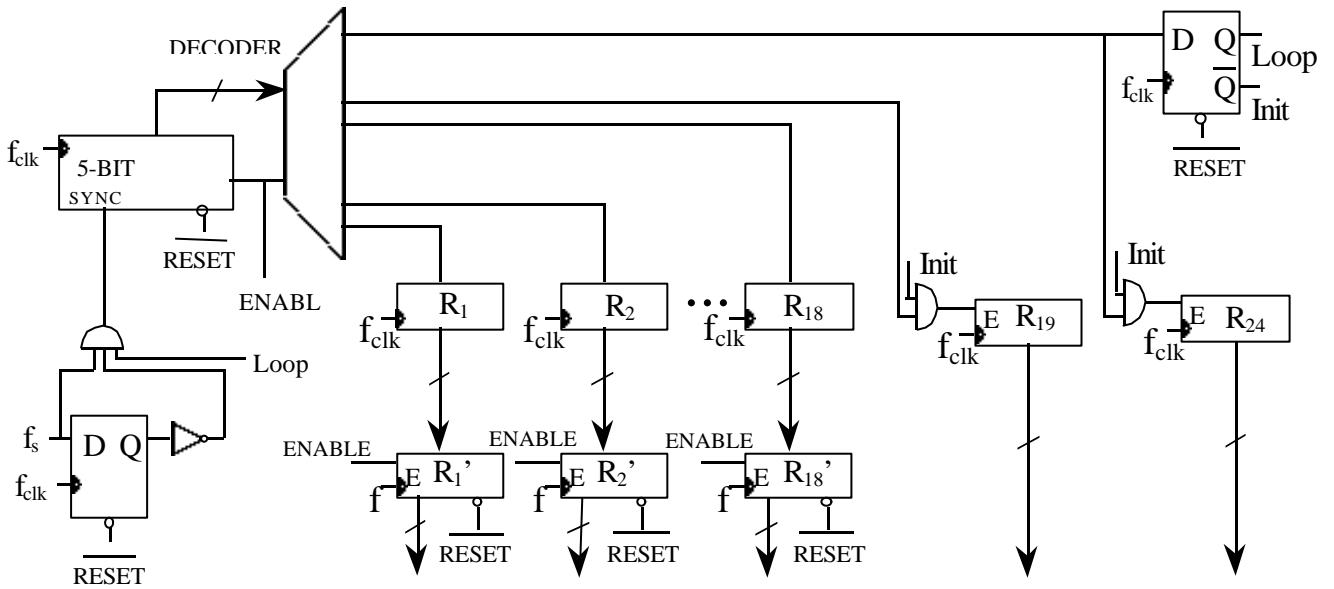


Fig. 2. Digital control

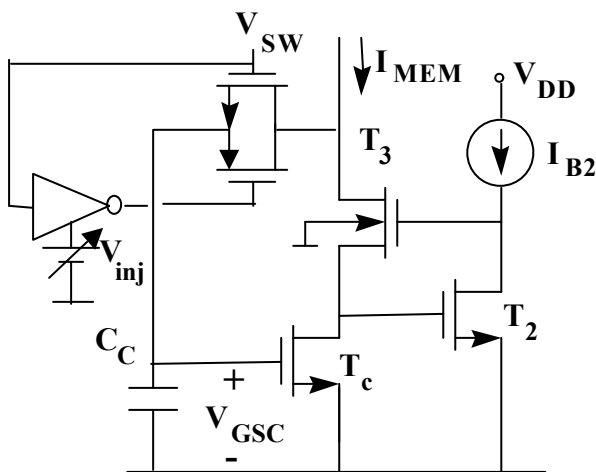


Fig. 3. Regulated-cascode ATC with charge compensation

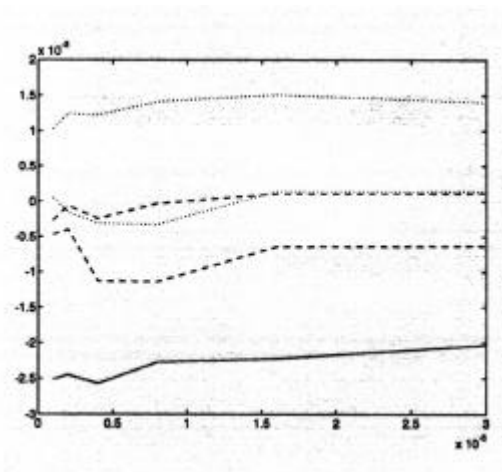


Fig. 4. Simulation of the switch error compensation
X-axis: Ampères. Y-axis: Volts

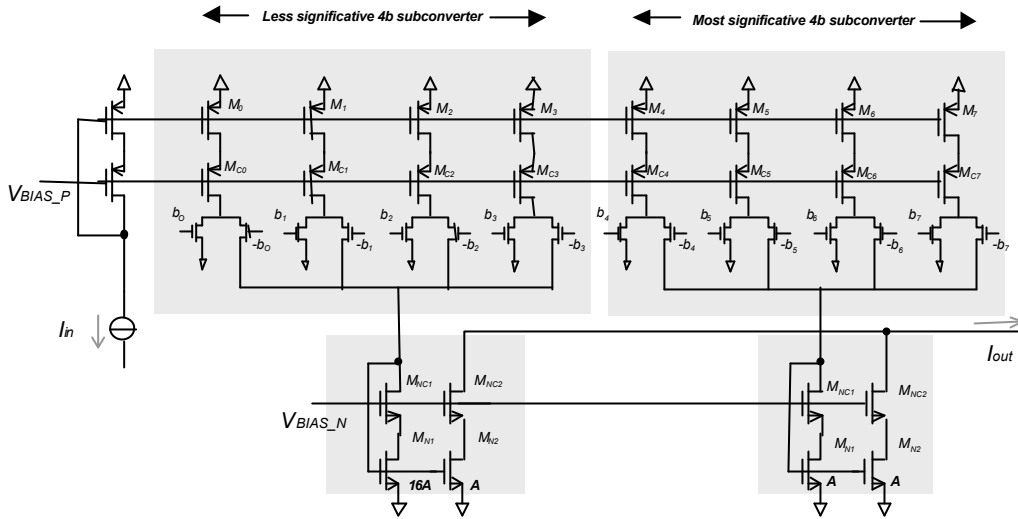


Fig. 5. Mixed-signal semialgorithmic D/A multiplier

2 PROCESSOR DESCRIPTION

In an MLP model, synapses are multipliers while the activation function is a sigmoid. Since the current-mode approach has been adopted, synapse addition is obtained simply by connecting synapse outputs. Concerning network parameters, they are stored in a conventional digital memory, and retrieved as they are required.

The processor implementation is divided into the analog datapath, where processing takes place, and the digital control and registers, in charge of delivering the routing data and the overall synchronization.

In Fig. 1, the analog datapath of the circuit is shown at a block level for a 3-synapse processor. Extension to an arbitrary number of synapses is straightforward. The electrical information carrier is current.

The voltage or current input data can be loaded either serially or in parallel from the bottom analog inputs and are stored in a row of ATCs. The two rows of ATCs store input and output activations, exchanging their role at each new layer emulation (since outputs of one layer become inputs of the next one). The row of ATCs storing the analog input data are multiplied with the digital synapse weights by means of digital-to-analog converters (DACs). Depending on the weight sign, the output current is driven to a positive or negative current aggregation line. After a programmable scaling are used to apply the β (slope) parameter, both currents enter a sigmoid circuit. The sigmoid output is stored into one of the output row ATCs.

Other important features included in the processor are: Digitally programmable offset and gain compensations to reduce processing tolerances., and a delay control in ATC switches to guarantee a correct operation.

To generate the sequence of operations, a very simple digital control is enough. A flexible digital circuitry has been specified, in order to allow any variation in the data flow. The circuit is shown in Fig. 2. The parallel registers that control the datapath routing and weight values are updated at each clock period. The clock cycle has been established in 1 MHz, and the weight

resolution 8 bits plus a sign bit for the first test circuit. This resolution has been shown to be enough for the execution phase of many real applications [8].

In the following sections, the key issues in the design of the main building blocks are discussed.

3 ANALOG TRANSFER CELL

Systems that use analog memory require a careful design of this critical element. The selected dynamic memory cell is a regulated-cascode current copier (Fig. 3) because of its small area overhead and good properties [4,9,10].

A critical error source in analog memories is the charge injection and clock feedthrough produced by turning off the switch, in our case an NMOS transistor, that controls the copier NMOS gate. Modeling of the error mechanisms is complicated, because is very dependent on the control signal slope, signal level, and copier transconductance, among others [10].

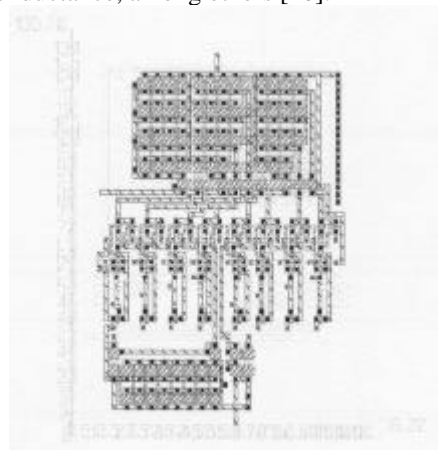


Fig. 6. Layout of the semialgorithmic D/A multiplier

Compensation by a dummy switch has been proposed, but it has been shown that its compensating effect is not clear. Since a PMOS transistor has an opposite sign charge injection, we propose the use of a complementary switch. By controlling the voltage swing of the PMOS gate during switching, charge

injection is adjusted so the error can be virtually eliminated.

Fig. 4 depicts the voltage error for a single NMOS switch (solid line) and for 2 different CMOS switches. The dashed lines correspond to a PMOS width of 1.4 times the NMOS switch width. The dotted lines to a PMOS width 2 times larger than that of the NMOS switch. The two lines represent the error change with a 3 V swing of the control voltage V_{inj} (Fig. 3). For PMOS switch dimensions within 1.4 and 2 times the NMOS switch, the switch-induced voltage (and thus current) error can be compensated by varying the V_{SS} value. The compensation can be done either externally or by means of an internal feedback loop [11].

4 CURRENT-STEERING MIXED-SIGNAL MULTIPLIER

Synaptic-weight product can be implemented with continuous-time circuits *-i.e.* with multiplier blocks, as open-loop transconductors [12] or bipolar Gilbert cells or with mixed-signal multipliers based on digital-to-analog converters. In general, the aggregation operation needed at the output end of the synaptic cells dictates the use of current-domain signal representation. In addition, for the architecture being considered, input signals to the synaptic multipliers are the currents delivered by the current-copier-based ATCs. In this situation, a current-steering switched current mirror digital-analogue converter is used.

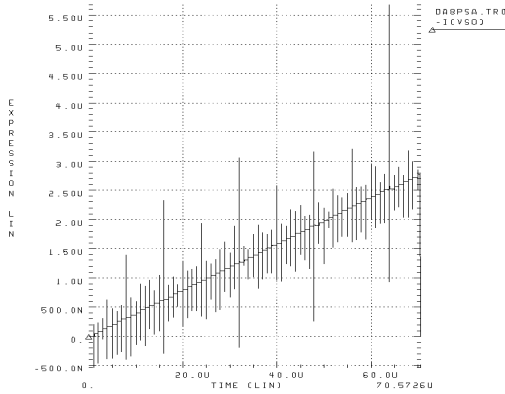


Fig. 7. Transient waveform for the central codes of the semialgorithmic D/A multiplier

The signal-scaling action in current-steering DACs is obtained by means of active-device scaling and, thus, no need for passive components appears, avoiding the use of linear capacitors (which, in turn, require one extra polysilicon layer process step). The current-steering DAC, being in fact a switched current mirror, achieves high-frequency operation because of the inherent reduction of voltage swings, thus minimising sensitivities to parasitic capacitances [13].

The designed DAC (Fig. 5) is composed of a matrix-arranged array of unit cascoded current sources. The use of high-compliance low-voltage cascode stages within the converter assures the reduction of channel length modulation effect and consequently improves

integral linearity and monotonicity, while keeping supply-voltage requirements low.

The use of semialgorithmic techniques [14] reduces matching requirements (imposed by fully algorithmic DACs) while maintaining the linear relation between active chip area and resolution. In order to limit current mismatch to $\frac{1}{2}$ LSB, the maximum allowable mismatch [15] between current sources is given by the relationship

$$\sigma_N^2 \left(\frac{\Delta i}{i} \right) < \frac{1}{2^{n+2}} \xrightarrow{n=8 \text{ bits}} \sigma_N^2 \left(\frac{\Delta i}{i} \right) \leq 0.19\% \quad (1)$$

For an n -bit semialgorithmic DAC, the transfer function given in (2) for I_{in} input current and D^w digital code illustrates the use of fine and coarse $n/2$ -bit subconversion blocks.

$$I_{OUT} = D^w \cdot I_{IN} = I_{IN} \left[\left(\frac{b_{n-1}}{2} + \dots + \frac{b_{n+1}}{2^{n+1}} \right) + \frac{1}{2^2} \left(\frac{b_{n-1}}{2} + \dots + \frac{b_0}{2^2} \right) \right] \quad (2)$$

In each four-bit subconverter in Fig. 5, digital codes are applied to differential current switches (consisting of p-type fully saturated differential pairs) which steer the current sourced by the current generator matrix to either the output summing node or a dummy node.

The use of differential switches ensures current flowing for each cell in the matrix, hence long recovery times are avoided and settling time requirements are fulfilled for low-level currents.

Fig. 6 shows the full-custom layout of the semialgorithmic DAC cell. The validation of this mixed-signal current-mode multiplier is inferred from the post-layout HSPICE one-run Monte Carlo simulation shown in Fig. 7. This simulation is carried out under a constant current input of $I_m = 10 \mu\text{A}$ and cyclic digital stimuli. Device size information is included within statistical simulation parameters to ensure the correct mismatch effect. Only a part of the whole conversion margin is depicted. Some performance evaluation parameters are an active area occupation of about 0.01 mm^2 , and settling times bounded by $\approx 200 \text{ ns}$.

5 WTA-BASED SIGMOID

As was shown in Fig. 1, the positive and negative synaptic currents are accumulated in different lines. These currents should be subtracted, but to avoid the mismatch error introduced by extra current mirror stages, both lines are applied to a differential-input sigmoid circuit.

In Fig. 8a, the sigmoid circuit generator is shown. It is based on a current-mode Winner-Take-All (WTA) circuit [16] and consists of a differential pair with two transistors providing a common-mode feedback.

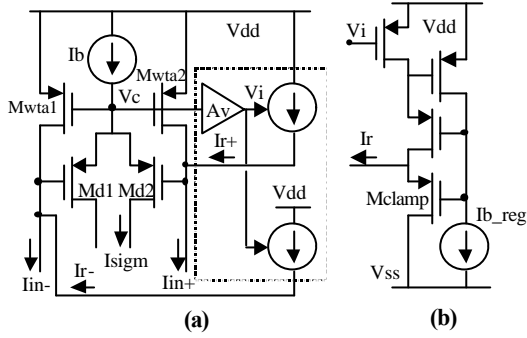


Fig. 8. a) WTA-based sigmoid circuit; b) Rightmost current sources implementation detail.

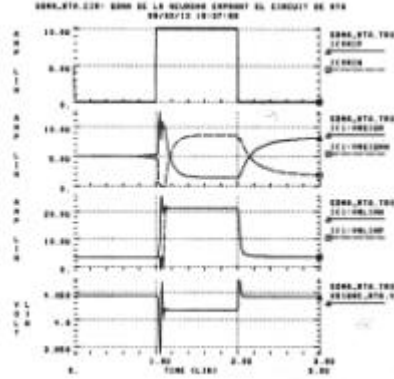


Fig. 9. Simulation of the WTA sigmoid transient response to 10 μA common-mode and 100 nA differential steps.

The sigmoid function is the natural V-I characteristic in a differential pair of exponential devices. For MOS transistors operating in strong inversion, the characteristic reasonably approaches the sigmoid form. The output current I_{sigm} is connected to a common line that can be routed to any ATC or to the circuit output. However, a simple differential pair is not enough since inputs are in current form, so a previous I-V conversion is necessary. Furthermore, a high resistance value is needed to obtain a significant voltage drop at sub- μA levels. For this purpose, the common-mode feedback transistors (M_{wta1} and M_{wta2}) working in saturation region are used as active loads.

Since the MOS transistor small-signal output resistance depends on the bias current, to keep it constant, the common-mode current has to be suppressed. An additional feedback loop performs this task. v_c is amplified and it controls two regulated cascode current sources connected to the WTA inputs (Fig. 8b). This loop keeps almost constant the currents flowing through the active loads, and therefore their resistance, making the sigmoid slope independent of the common-mode signal. In Fig. 9, a transient simulation of this behavior is shown. A reasonably short settling time and common-mode rejection is observed.

Concerning matching of the regulated cascodes, layout techniques have been used. Also, their biasing sources can be separately fine tuned to compensate the unavoidable mismatches. It is interesting to highlight the function of the M_{clamp} transistors. They are normally

off, but when one input voltage rises too much, the corresponding M_{clamp} turns on, fixes that voltage, thus preventing the M_{wta} to operate in the linear region. Moreover, it prevents the feedback source to become out of regulation by sinking the extra current. Finally, the always saturated operation speeds up the processing.

6 DIGITAL CONTROL

The digital control circuitry was shown in Fig. 2. In order to simplify the system, data are loaded from an external host. Two clock signals at different frequencies are used. The signal at f_{CK} is used to load the next set of values. A counter at the same clock frequency and a decoder generate the set of signals that enable loading cyclically one register each time.

When all the necessary registers of the upper row are loaded, a rising edge of signal f_s moves synchronously the loaded values to the lower row of registers, so all the control signals of the analog datapath are updated at the same time. Also, the rising edge of f_s produces a reset at the counter, and the loading cycle of the upper row of registers starts again.

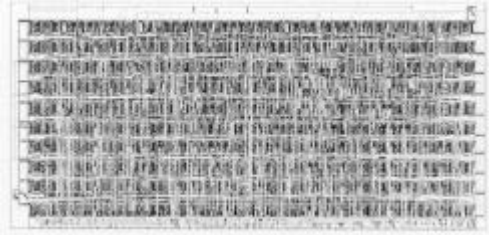


Figure 10. Layout of the digital control circuit

This strategy allows selection of the number of registers to load by means of the f_{CK}/f_s ratio. Additionally, a set of constant value registers is loaded the first time after reset. These are the rightmost registers of the upper row. After reset, the *init* signal enables the loading of that registers. Then, the D-flipflop stores a logic '1', and thus *loop* = '1' and *init* = '0'. This disables further changes in the constant registers, and the loop signal enables counter resetting. In Fig. 10 the layout of the standard cell digital control is shown. Layout area is $1620 \times 770 \mu\text{m}^2$.

7 THE COMPLETE PROCESSOR

The cells have been extensively simulated and optimized for a 0.8 μm CMOS technology. Hspice level 47 models (BSIM3) were used for simulations.

Offset is digitally adjusted by means of an additional constant-input DAC multiplier. Two extra DAC multipliers.

The analog cells have been designed to be compatible and consistent in terms of voltage levels, input and output resistances and to keep error under 0.5% in order to guarantee an 8-bit precision. A conservative 1 MHz clock is used in the first version

The complete system has been functionally simulated and validated using macromodels. Also, the different cells have been connected in a 3-synapse test circuit, and preliminary global simulation results show a promising behavior.

The processor has been designed in such a form that each cell can be separately tested. Furthermore, by programming the appropriate control digital words, the data flow can be varied with a high freedom degree. For instance, data copy between ATCs, and weighted current addition.

8 CONCLUSIONS

The implementation of an analog CMOS neural processor has been presented. This processor has the property to allow the recall emulation of any feedforward network structure provided the largest layer has no more units than the processor number of synapses. Several circuits and techniques suitable for the analog implementation have been presented and discussed.

Preliminary global simulations with all the cells show a correct behavior. The time response and errors are within the estimated values.

Currently the layout of the 3-neuron test processor is being finished. In a near future the circuit will be manufactured.

As an alternative to the digital weight storage, CMOS floating-gate memory for weight storage will be considered in a large-scale next generation processor. After a careful characterization of the cell, the best solution will be adopted. If floating-gate cells are used, this would enable the use of a more compact fully analog multiplier as well.

In this testchip, static CMOS standard logic is used. To minimize digital noise, a careful design of guard rings, and separate power lines and pins has been done. However, if measurements still show a high crosstalk level, a low-noise CVSL logic is to be employed in a future version.

Competitive applications envisaged for the processor described are cost-sensitive and low-power. Especially, when input data is in analog form and analog output is of interest (for instance, in control and embedded systems applications). Although digital processors can perform the described functions, it is at a higher area overhead and power consumption cost.

Acknowledgements. This work is supported by CICYT Project No. TIC96-1195. The author Jordi Cosp holds a CIRIT Research fellowship.

THE AUTHORS

The authors are with the Department of Electronic Engineering, Universitat Politècnica de Catalunya. Address: Gran Capità s/n, mòdul C4. 08034 Barcelona (Spain). E-mail: madrenas@eel.upc.es.

REFERENCES

- [1] Lippmann R. *An Introduction to Computing with Neural Nets*, IEEE Acoustics, Speech and Signal Processing Magazine, 4(2):4-22, April 1987.
- [2] Madrenas J., Moreno J.M., Cabestany J., *Analog and Mixed-Signal Neural VLSI Processors: Taxonomy, Comparison and Performance Evaluation*. MIXDES'96, Proceedings, pp. 389-400, Lodz (Poland).
- [3] Chai Y.Y., Johnson L.G., *A 2×2 Analog Memory Implemented with a Special Layout Injector*, IEEE Jour. Solid-State Circuits, Vol. 31 No. 6, June 1996, pp.856-859.
- [4] Madrenas J., Moreno J.M., Cabestany J., *CMOS Current-mode Analog Basic Blocks for Neural Processing*. Parts 1 and 2. MIXDES'95 Proc., pp. 109-120, Kraków (Poland), 1995.
- [5] Moreno J.M et al., *An Analog Systolic Neural Processor Architecture*, IEEE MICRO, Vol.14, No.3, pp.51-59, Jun. 94.
- [6] Madrenas J. et al., "A Current-Mode Sequential CMOS A/D Variable-Structure Processor for Neural Networks Emulation", Proc. of the Design of Circuits and Integrated Systems Conf., pp. 536-541. Madrid (Spain), 1998.
- [7] Moreno J.M., Madrenas J., Cabestany J., *Systolic Modular VLSI Architecture for Multi-Model Neural Network Implementation*, Proc. of the 4th Int. Conf. on Microelectronics for Neural Networks and Fuzzy Systems, pp. 118-124, 1994.
- [8] Blayo et al., R1-C and R2-C: Hardware Implementations. ELENA ESPRIT BRA 6891: Enhanced Learning for Evolutive Neural Architectures. 1994.
- [9] Daubert J., Vallancourt D., Tsvividis Y.P., *Current Copier Cells*, Electr. Letters, pp. 1560-1562, Dec. 1988.
- [10] Macq D., Jaspers P., *Charge Injection in Current Copier Cells*, Electronics Letters, Vol. 29, No. 9, pp. 780-781, April 1993.
- [11] Espejo S., Domínguez-Castro R., Medeiro F., Rodríguez-Vázquez A., Tunable feedthrough cancellation in switched-current circuits, Electronics Letters, Vol. 30, No. 23, pp. 1912-1914, 10th Nov. 1994.
- [12] S. Satyanarayana, Y. Tsvividis and H.P. Graf, "A Reconfigurable VLSI Neural Network", *IEEE Journal of Solid-State Circuits*, Vol. 27, N°1, January 1992, pp. 67
- [13] C.A.A.Bastiaansen, D.W.J.Groeneveld, H.J. Schouwenaars and H.A.H. Termeer, "A 10-b 40-Mhz 0.8 μ m CMOS current-output D/A converter", *IEEE Journal of Solid-State Circuits*, Vol. 26, N°7, July 1991, pp. 917.
- [14] K.L. Fong and A.T. Salama, "A 10 bit semi-algorithmic current mode DAC", *proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'93)*, pp. 978.
- [15] C. Abel et al. "Characterization of Transistor Mismatch for Statistical CAD Submicron CMOS Analog Circuits", *Proc. ISCAS'93*, pp. 1401, Chicago.
- [16] Lazzaro J. et al., *Winner-Take-All Networks of $O(n)$ Complexity*, Advances in Neural Information Processing Systems, Vol. 1, Touretzky D.S., ed., Morgan Kaufmann Publishers, pp. 703-711, 1989.