HYBRID-CONTROL OF SYNAPSE CIRCUITS FOR PROGRAMMABLE CELLULAR NEURAL NETWORKS

S. Espejo, R. Domínguez-Castro, R. Carmona, and A. Rodríguez-Vázquez
Centro Nacional de Microelectrónica-Universidad de Sevilla
Edificio CICA, Avda. Reina Mercedes s/n, 41012-Sevilla, SPAIN
Phone: +34 - 5 - 423 99 23. Fax: +34 - 5 - 423 18 32. E-mail: espejo@cnm.us.es

ABSTRACT

This paper describes a hybrid weight-control strategy for VLSI realizations of programmable Cellular Neural Networks (CNNs), based on auto-tuning of analog control signals to digitally specified values. The approach merges the advantages of digital and analog programmability, achieving low areas and reduced number of control lines, simplifying the control and storage of weight values, and eliminating their dependency on global process-parameter variations.

1. INTRODUCTION

The implementation of general-purpose programmable CNN systems [1] is a requisite for the application of this computation paradigm in many areas of great interest [2]. In the design of this class of systems, one of the major trends is the optimization (in terms of area and power efficiency, accuracy, and speed) of the programmable synapses. The efficiency of the synapse circuit is strongly influenced by the type of programmability selected: *analog* or *digital*. This contribution analyzes the two possibilities, and proposes a hybrid (analog/digital) approach which combines the advantages of the two alternatives and virtually eliminates their drawbacks. These results can be extended to other types of massively parallel analog processing systems.

2. PROGRAMMABLE SYNAPSES FOR CNNS

The synapse required for programmable CNN implementations can be represented as in Fig. 1. Both x_i and W are input signals, while x_o is an output. We refer to x_i as the input signal or simply *the input*, while W is called the *weight signal*. The input is driven by the signal to be scaled (the output y^c of the cell), which is a function of time during network operation. On the other hand, the weight signal is time-invariant during the CNN process. The ideal behavior of the scaling block can be formulated as follows,

$$x_o = P(W) \cdot S(x_i) \tag{1}$$

where $S(.)$ is a linear and continuous function of x_i in some range around $x_i = 0$. Function $S(.)$ is normalized to the value of its derivative at $x_i = 0$ and hence, $P(W)$ represents a scaling factor within the linear range of $S(.)$, in which $S(x_i) = x_i$. In general $P(.)$ is not required to be linear, and can be either a continuous or discrete function.

Two general classes of synapses can be considered attending to the nature of the weight signal W. *Digitally-programmed* synapses are driven by a digital weight signal, and hence, function $P(W)$ needs to be defined only for a discrete set of values

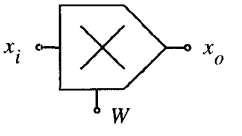


Fig. 1: Representation of a programmable CNN "multiplier"

of its variable. *Analog-programmed* synapses are driven by an analog weight signal, and hence, function $P(W)$ needs to be defined within a continuous set of values of its variable. The use of either class of synapse presents important advantages and drawbacks.

3. DIGITALLY-CONTROLLED SYNAPSES

The weight signal of a digitally-programmed synapse is commonly represented by a binary number of several bits. We will use here an N-bits-plus-sign representation,

$$W = (1 - 2w_s) \sum_{i=0}^{N-1} w_i 2^i \tag{2}$$

In order to achieve a weight range of $[-P_{\max}, P_{\max}]$, we must use the following weight increment

$$\Delta P = P_{\max} / \left(2^N - 1\right) \tag{3}$$

and define the weight function $P(.)$ as

$$P(W) = \Delta P \cdot W = \frac{P_{\max}}{2^N - 1} (1 - 2w_s) \sum_{i=0}^{N-1} w_i 2^i \tag{4}$$

An schematic implementation of a digitally-programmed synapse using this codification is shown in Fig. 2, where a symbolic analog block with a transfer characteristic $S(.)$ and a fixed scaling factor ΔP is used. Clearly, the output signal x_o is related to the input x_i by (1) with $P(.)$ given by (4).

For accuracy reasons, fixed scaling factors equal to a power-of-2 multiple of ΔP are obtained by connecting several identical blocks with voltage-mode input and current-mode output in parallel. The design of these type of synapses is

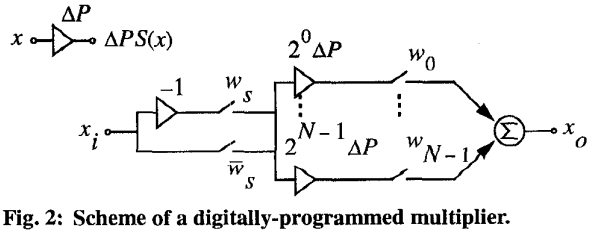


Fig. 2: Scheme of a digitally-programmed multiplier.

0-7803-3073-0/96/$5.00 ©1996 IEEE

straightforward after the unitary element with ΔP scaling factor has been designed. Linearity is easy to achieve because every unitary element is designed with a fixed scaling factor. Also, linearity is independent of the weight value, and the linear output signal-range scales linearly with the weight. Concerning weight accuracy, digitally-programmed synapses are inherently robust against wafer-level process parameter variations, since the weight is given by the number of identical devices being used. Only severe variations affecting the behavior of the basic unitary block will result in wrong performance. The binary codification of the weight signal allows an easy external control of the scaling factors, since the relationship between the weight signal and the actual weight is clearly known a priory and robust against process variations. Weight values can be stored on digital memories, avoiding the time-degradation problems associated with analog memories. Finally, the same binary signals used to codify the weight value can be used to "power-off" the parts of the synapse circuitry not being used for some particular weight value. This feature is convenient in many cases, given the high power dissipation expected from large programmable CNN systems.

The discretization of the possible weight values can be seen as a drawback of digitally-programmed synapses. However, a fair comparison must take into account the achievable weight-accuracy of either digitally- or analog-programmed synapses, affected by systematic and random errors (mismatch). In what follows, we use the concept of *effective* resolution, referred to the minimum relative weight increment $\Delta P/P_{max}$ which can be achieved with a reasonable expectance of accuracy (say 90% probability of weight deviations within the range $\Delta P/2$). Assuming the general figure of about 7 to 8 bits accuracy for common, not calibrated analog circuitry [3], digitally-programmed synapses with eight-bits weight signal will generally result in similar *effective* resolutions than many analog-programmed synapses.

Unfortunately, there are some other relevant disadvantages in the use of digitally-programmed synapses. Area consumption is usually much larger than that of analog synapses, due to the large number of unitary elements and the multiplexing circuitry required and, above this, to the large number of global control lines needed. As an example, a digitally programmable CNN with neighborhood radius of one requires a total of 19 different coefficients, nine for each of the two templates plus the offset term. If each of these coefficients is codified by 7 bits-plus-sign weight-signals, the total number of weight-control lines reaching every cell is of 152, and the number of unitary elements for the MDACs of 2413.

4. ANALOG-PROGRAMMED SYNAPSES

An analog programmable synapse can be characterized, in general, by an expression of the form,

$$x_o = h(W, x_i) \tag{5}$$

where $h(.)$ is assumed to be an approximately linear, continuous function of x_i, at least in some range around $x_i = 0$. If we define

$$P(W) = \left.\frac{\partial h}{\partial x_i}\right|_{x_i = 0} \quad and \quad S(W, x_i) = \frac{h(W, x_i)}{P(W)} \tag{6}$$

then, (5) can be written as

$$x_o = P(W) \cdot S(W, x_i) \tag{7}$$

where the weight function $P(.)$ is now a continuous, generally nonlinear function of the weight signal W, and $S(.)$ can be shown to be normalized to the value of its derivative at $x_i = 0$ for any W. Since for a given W value, $S(.)$ is proportional to $h(.)$, $S(.)$ is an approximately linear, continuous function of x_i in some range around $x_i = 0$. Hence, (7) is similar to the ideal formulation in (1), except that now, $S(.)$ is a function of the weight signal. This accounts for the fact that the linearity and signal range of the normalized output $x_o/P(W)$ of an analog synapse depend, in general, on the particular value of the weight.

The design of area and power efficient analog-programmed synapses with proper performances requires, in general, a significantly higher effort than that required for the design of digitally-programmed synapses. In many cases, the achievable linearity of the synapse is low for extreme weight values, unless high costs are accepted. Also, function $P(W)$ is usually a nonlinear function dependent on process parameters, difficulting the external control of the coefficients. Finally, the on-chip storage of the analog weight values requires analog memories, which lack the robustness of digital memories and present time-degradation problems.

On the other hand, there are important advantages on the use of analog-programmable synapses. Area requirements are much lower than for digitally-programmed synapses (the required number of transistors is usually around 10 or less). also, the scaling factor of every synapse implementing the same coefficient (one per cell), can be transmitted through one or two global lines.

5. A SYNERGY OF ANALOG AND DIGITAL PROGRAMMABILITY

Table 1 provides a brief comparison of the advantages and drawbacks of digital and analog programmability for the realization of general purpose CNN systems. As can be seen, most of the disadvantages of analog programmability are related to the control and storage of the weight values and their dependency on process parameters. On the other hand, digitally-programmed synapses require large areas and an excessive number of control lines, turning them inefficient for high-density CNN implementations.

	Analog	*Digital*	*Hybrid*
Effective resolution	7-8 bits	7-8 bits	7-8 bits
Area consumption	Low	Very high	Low
Number of control lines	Low	Very high	Low
Power dissipation	Low	High	Low
Sensitivity to process var.	High	Low	Low
Design cost	High	Low	High
External control	Difficult	Simple	Simple
Linearity	Difficult	Simple	Difficult
Weight storage	Difficult	Simple	Simple

Table 1: Simplified comparison of programming alternatives.

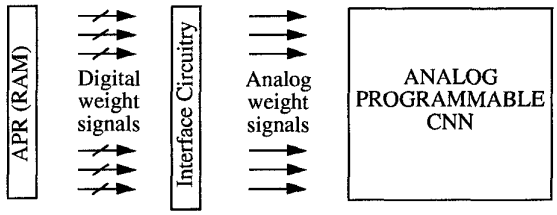


Fig. 3: Symbolic architecture of a hybrid analog-digitally programmable CNN.

We propose a hybrid approach, illustrated in Fig. 3, based on the use of analog-programmed synapses within the cells (which provides high area efficiency and low number of control lines), and digital control from the exterior of the network (which facilitates the control and on-chip storage of the weights). In this manner, the APR [1] can be realized by a digital RAM memory.

As shown in Fig. 3, an interface circuitry is required to generate the internal analog weight-signals from their digitally coded values. This circuitry must be compound by several identical blocks, one for each programmable coefficient in the network. An uniform CNN with unitary neighborhood radius has 19 different coefficients, and hence, 19 of this interface blocks will be required in general. In any case, from a system-area perspective, a reduction in the synapse area is scaled by the number of synapses in every cell and by the number of cells, while the area of the interface circuitry is approximately constant.

The functionality required from the interface blocks is basically that of a nonlinear digital to analog (D/A) converter. If a linear relationship between the digital signal and the programmed weight is desired, the nonlinear characteristic of the converter must cancel out the non-linearity of the weight function of the analog synapse. Assume that the desired weight value is given by

$$p = \Delta P \cdot W_D \tag{8}$$

where W_D and ΔP are defined by (2) and (3), respectively. If the analog synapse being used has a weight function $p = P_A(W_A)$, we need the following transfer characteristic from the interface block

$$W_A = P_A^{-1}(p) = P_A^{-1}(\Delta P \cdot W_D) \tag{9}$$

Since the synapse has two input signals (the weight and the input signal), its inverse function can only be defined for some fixed value of one of the inputs. For our purpose, we must set the input signal x_i to a fixed reference level x_{ref}. The inverse function of the synapse can be obtained using an adaptive architecture involving an analog- and a digitally-programmed synapse, both driven by the same reference signal level x_{ref}, as shown in Fig. 4.

For the analysis of this scheme, we rewrite the transfer characteristics of the analog-programmed synapse (7) as

$$x_o = P_A(W_A)S_A(W_A, x_i) \tag{10}$$

and that of the digitally-programmed synapse (1) as

$$x_o = P_D(W_D)S_D(x_i) \tag{11}$$

with W_D given by (2). The architecture in Fig. 4 can then be described by the following differential equation,

$$\tau_w \frac{dW_A}{dt} = -P_A(W_A)S_A(W_A, x_{\text{ref}}) - P_D(-W_D)S_D(x_{\text{ref}}) \tag{12}$$

which defines a first order dynamical system. If function $P_A(.)$ is a

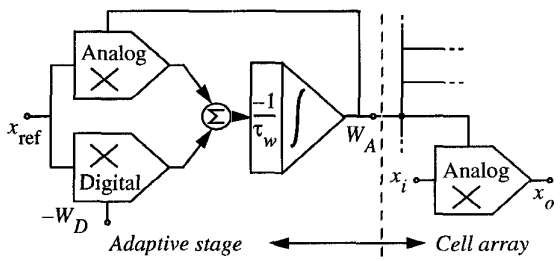


Fig. 4: Architecture of the D/A interface circuitry required for the hybrid-control approach.

monotonically increasing function of W_A, there is a unique equilibrium point which is locally stable, and hence, the system is globally asymptotically stable. The value of the equilibrium point is easily obtained,

$$W'_A = P_A^{-1}(-P_D(-W_D)\frac{S_D(x_{\text{ref}})}{S_A(W'_A, x_{\text{ref}})}) \tag{13}$$

If $P_D(.)$ is of the form given in (4), which is an odd function of W_D,

$$W'_A = P_A^{-1}(\Delta P \cdot W_D\frac{S_D(x_{\text{ref}})}{S_A(W'_A, x_{\text{ref}})}) \tag{14}$$

and if x_{ref} is within the linear range of $S_D(x)$ and $S_A(W_A,x)$,

$$S_D(x_{\text{ref}}) = S_A(W'_A, x_{\text{ref}}) = x_{\text{ref}} \tag{15}$$

from where (14) results in

$$W'_A = P_A^{-1}(\Delta P \cdot W_D) \tag{16}$$

which is the desired relationship formulated in (9).

Because this adaptation is achieved for a fixed input signal value x_{ref}, and since analog synapses within the network will be driven by variable signals x_i, synapses offset and linearity are of extreme relevance. In particular, the obtention of (16) using the simplification in (15) must be examined carefully, attending to the real forms of S_D and S_A, including their linear ranges and random variations.

Note that offsets in either function will result in large errors in the adapted analog weight signal, if x_{ref} is not much larger than average offset values (for instance measured from the standard deviation). On the other hand, large absolute values of x_{ref} may produce errors as well, unless the analog synapse is highly linear. Further analysis of these error sources on the adapted analog weight signal can be carried out from (14).

The digitally-programmed synapse (a linear multiplying DAC) and the rest of the circuitry in the adaptive stages can be implemented using relative large areas (and hence with low errors [3]), including the analog-programmed synapse, which can be compound of several identical blocks in parallel.

Another alternative is to use a precalibration step to cancel the offsets in the adaptive stages. This alternative requires the offsets to be approximately independent of the weight value. This property of some analog synapses allows the use of a precalibration step to cancel the offset of the processing circuitry in the cells [4].

An important feature of the proposed adaptive approach is that global variation on the transfer characteristic of the synapse have

no effect on the weight values, since the adaptive scheme settles the value of the weight to that specified by the digital signal, for any analog synapse transfer characteristic satisfying a few weak conditions.

6. EXPERIMENTAL DEMONSTRATOR

A particular adaptive stage has been designed and tested as part of a larger programmable CNN chip [4]. The prototype was fabricated in a 0.8μm, n-well, one poly, two metals, 5v, CMOS technology. Fig. 5 shows a simplified schematic of the weight adaptation circuitry. It consists of an analog multiplier, a digitally-controlled multiplier, a fully differential class-two current-conveyor, and some reference circuitry. The digitally-controlled multiplier is realized by a multiplying DAC (a binary-weighted array of current sources) and a single to differential converter with sign control. The realization of the other blocks is shown in Fig. 6.

Fig. 7 shows the layout and microphotograph of the stage, and Fig. 8 contains some experimental measurements. Worst-case adaptation time was below 1μs, and the error of the adapted weight in the range of the quantization of the MDAC (8 bits).

7. SUMMARY

We have proposed the use of a hybrid weight control strategy for the realization of programmable CNNs, based on automatic adaptation of the analog weight signals to values specified by digital words. The weight function of the analog synapse is required to be a monotonically increasing (or decreasing) function of the analog weight signal, and it is convenient that the random offset of the analog synapses be independent of the weight signal. This approach merges most of the advantages of digitally- and

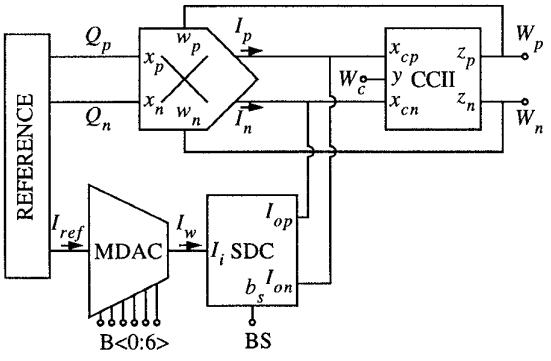


Fig. 5: Simplified diagram of the designed adaptive stage.

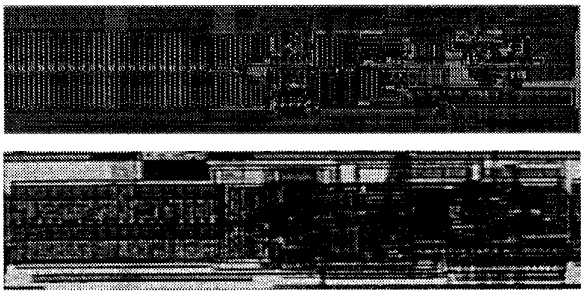


Fig. 7: Layout and microphotograph of the adaptive stage.

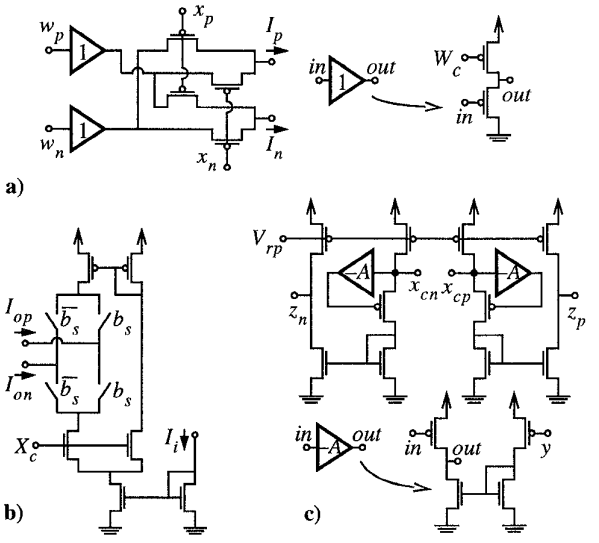


Fig. 6: Adaptive stage blocks: a) multiplier, b) signed single to differential converter, c) differential current conveyor

analog-programmed synapses, achieving low areas and reduced number of control lines, simplifying the control and storage of the weight values, and eliminating their dependency on global process parameter variations. Last column in Table 1 above summarizes the advantages of the proposed hybrid control strategy. Except for the slight increase in design cost, the best of the analog and digital programmability is exploited. A demonstrator has been designed, manufactured in a standard CMOS technology, and successfully tested.

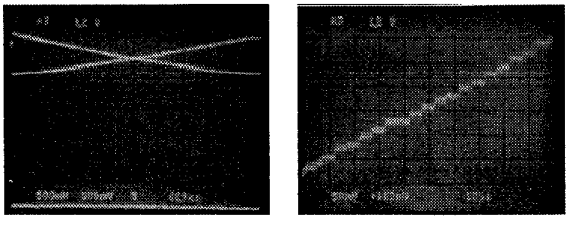
REFERENCES

[1] T. Roska and L.O. Chua: "The CNN Universal Machine: An Analogic Array Computer". *IEEE Trans. Circuits and Systems II*, Vol. 40, pp 163-173, March 1993.

[2] Proceedings of the Third IEEE International Workshop on Cellular Neural Networks and their Applications, Rome, December 1994.

[3] M.J.M Pelgrom, A.C.J. Duinmaijer and A.P.G. Welbers: "Matching Properties of MOS Transistors". *IEEE J. Solid-State Circuits*, Vol. 24, pp 1433-1440, October 1989.

[4] S. Espejo: *"VLSI Design and Modeling of CNNs"*. Ph. Dissertation, University of Sevilla, March 1994.



a) b)

Fig. 8: Experimental results: a) Differential output (W_p,W_n) versus digital input (B<0:6,S), b) enlarged view of the difference W_p-W_n