1-D Discrete Time CNN with Multiplexed Template Hardware

Gabriele Manganaro¹ and Jose Pineda de Gyvez

Dept. of Electrical Engineering - Texas A&M University - College Station - TX77843-3128 - U.S.A.

tel. +1-409-845-7477, fax +1-409-845-7161

email: gabriele@amesp01.tamu.edu, gyvez@pineda.tamu.edu

ABSTRACT: While VLSI of CNNs has seen significant progress in two-dimensional signal processing little has been done for one-dimensional applications such as audio signal processing and 1-D filtering. This paper presents a new discrete-time programmable Cellular Neural Network suitable for these kind of applications. The proposed VLSI implementation is based on the well-known S^2I technique that among other properties minimizes clock feedthrough effects. This feature renders an accurate signal processing unit. The system's main building blocks are an analog shift register and a switched current multiplier. Yet, the system architecture is novel by itself. Namely, the number of multipliers has been minimized by sharing the multipliers between the $A*y$ and $B*u$ products during the various phases of the controlling clock. The paper presents detailed simulation results of the system architecture

1. Introduction

Although most of the CNN applications and corresponding VLSI implementations regard two-dimensional arrays, one-dimensional Cellular Neural Networks have recently received increasing attention. For the latter ones, quite different applications have been reported, ranging from 1-D signal processing [1-3] to instrumentation and control [4]. In particular a 1-D CNN architecture able to emulate the behavior of FIR filters and to perform the Daubechies Wavelet Transform has been thoroughly discussed in [1-3]. Besides, unlike the case of 2-D CNNs for image processing, the applications reported for 1-D CNNs require a very reduced number of cells [1-4]. This paper presents a VLSI implementation of a new one-dimensional discrete-time programmable CNN suitable for the above mentioned applications. It is based on the well-known S^2I technique. Moreover, the introduction of a time-multiplexing scheme allowed a more efficient use of the hardware.

A block diagram of the proposed architecture is shown in Fig. 1. The main blocks are an analog shift register (i.e. a tapped delay line) and a set of locally connected cells. A cascade of eight delay elements (sr0-sr7) composes the shift register. Input data enters the shift register on the left side (through sr0) and the samples are shifted to the right, passing from sr0 to sr1 and so on. The cells of the CNN receive their inputs from the shift register and from the outputs of the neighbors. The cloning templates are provided externally making the proposed architecture programmable. The state equation describing the behavior of the cell at position c is:

$$x_c(n+1) = \sum_{d \in N(c)} A_{c,d} y_d(n) + \sum_{d \in N(c)} B_{c,d} u_d(n) \quad (1)$$

where $x_c(n+1)$ is the updated state of the cell, $y_c(n) = f(x_c(n))$ is its output, $u_c(n)$ is the output of the shift register srC (e.g. u_1 is the output of sr1), $A_{c,d}$ are the feedback templates while $B_{c,d}$ are the control templates, $N(c)$ is the neighbor set of cell c defined as follows:

$$N(c) = \{d : |d - c| \leq 1\} \quad (2)$$

As shown in Fig. 1, the proposed architecture includes six CNN cells and eight delay cells (the self-feedback is

Fig. 1 : Block diagram of the proposed architecture.

Fig. 2 : S^2I half-delay cell and clock waveforms.

¹ On leave from DEES- University of Catania - Italy

implicit in the cells of Fig.1 and it has not been shown to avoid clutter). In practice, in the real implementation, the first delay stage sr0 is not necessary and the input (*Data In* on Fig.1) can be directly provided in place of it.

2. The Shift Register

A cascade of S²I delay cells [5-7] composes the shift register. A full delay is obtained by cascading two S²I half-delay cells [5-6] whose circuit diagram and corresponding switch control signals are shown in Fig. 2. A replica of the current at the output of the cell is needed for the next delay cell and for any other circuit requiring it as input. Therefore, a current mirror with multiple output branches is placed in-between any full-delay block. In particular, cascode current mirrors have been used in order to obtain very high output impedance. The circuit of a shift register cell is shown in Fig. 3. An impedance is placed between the output and the input node of the half-delay cells in Fig. 3. It represents an NMOST identical to the ones used for the switches with his gate connected to the positive power supply. It compensates for the *ir* drop due to the output switch [6].

The peculiarity of the S²I technique is the compensation of many analog errors due to the non-ideal behavior of the MOS transistors [5-6]. Above all, it allows an almost complete correction of the errors due to the charge injection coming from the MOS switches [5].

Fig. 3: Shift register cell

Fig. 4: Nonlinear output function

The transfer function of a S²I half-delay cell, when loaded by another identical one, is approximately [7]:

$$I_0(z) = I_i(z) \frac{H_i(z)}{1 + \frac{2g_0}{\Psi g_m}} + I_{off1}(z) \quad (3)$$

where $H_i(z) = -z^{-1/2}$ is the ideal transfer function of an half-delay cell, $I_{off1}(z)$ represents a constant output offset. $\Psi = g_m/2g_0$ is the voltage gain of an inverter in which M_C acts as driver while M_F acts as current source. However, when the S²I half-delay cell is loaded by a conductance equal to g_m then the transfer function is approximately given by [7]:

$$I_0(z) = I_i(z) \frac{H_i(z)}{1 + \frac{2g_0}{g_m}} + I_{off2}(z) \quad (4)$$

where $I_{off2}(z)$ represents a constant output offset very close to $I_{off1}(z)$. Therefore, due to the sign inversion, when two of these cells are cascaded in order to obtain a full delay, the two offsets practically cancel each other. In our case, the first half-delay cell is loaded by the second one, while the current mirror loads the second one. Therefore the total transfer function of the two blocks in cascade is given by:

$$G(z) = \frac{z^{-1}}{1 + \frac{2g_0}{g_m} \left(1 + \frac{1}{\Psi} \right) + \frac{4g_0^2}{\Psi g_m^2}} \cong \frac{z^{-1}}{1 + \frac{2g_0}{g_m}} \quad (5)$$

This implies that the delay cell slightly attenuate the delayed signal. This attenuation could be minimized making use of cascode transistors in the half-delay cell [8]. However, in the proposed design, it has been preferred to compensate it by choosing a suitable current gain for the current mirror following the delay cell. In fact, performing a Montecarlo analysis, the latter solution has shown a better robustness to parameter variation. Finally, in order to compensate a possible residual output offset current, another delay cell with zero input has been created. This cell's output current is the residual offset. This is copied and subtracted to the output of all the

other delay cells by means of cascode mirrors (it is added to the intermediate node of the current mirrors using the terminal *Off_c* as shown in Fig. 3). This is very similar to the well-known *replica circuit* technique [8].

2.1 CNN Cells

The cell's behavior is characterized by the state equation (1). The output of the cell is:

$$y_c(n) = f(x_c(n)) = \frac{1}{2}(|x_c(n) + 1| - |x_c(n) - 1|) \quad (6)$$

Two cascode current mirrors shown in Fig. 4 easily obtain this. The bias current I_{dd} of the NMOSTs working as cascode current sources fixes the saturation current.

2.1.1 S²I Multiplier

Four quadrants S²I multipliers [7] have been used to implement the programmable templates. The circuit of the multiplier is shown in Fig. 5. The multiplication of any two given currents x and y is accomplished by evaluating their quadratic terms:

$$(x + y)^2 - x^2 - y^2 = 2xy \quad (7)$$

In order to accomplish this, a current squarer circuit is used to obtain the three squares on the left-hand side. The squarer characteristic is:

$$i_0 = I_b + e_1 + (i_i + e_2)^2 / 4I_b \quad (8)$$

where i_0 and i_i are the output and input current respectively, I_b is a constant current related to the bias voltage V_{bias} , e_1 is the output offset error, e_2 is the input offset error. A four-step algorithm (corresponding to the four phases θ_1 - θ_4) is used to evaluate the left-hand side of (7). The control signals for the switches are shown in Fig. 6.

Fig. 5 : Circuit of the S²I multiplier.

Fig. 6 : Control signals for the switches.

During θ_1 both x and y are added up at the input node of the current squarer. The squared result is stored in memory 1. During θ_2 only x is fed in; the result of the squaring operation is added to the previous result that was fed (and sign inverted) by memory 1 at node A. This sum is stored in memory 2. During θ_3 no input is presented to the squarer that feeds the residual at node A. At the same time the latter partial sum is fed (and sign inverted) at node A by memory 2 and the total current is stored into memory 1. Finally, during θ_4 the current y is squared and added to the previous sign inverted sub-total given by memory 1. The total current I_{out} is provided at the output. Applying this algorithm to (7) by using (8) and assuming ideal behavior for the circuits and devices (infinite output/input impedance ratio of the building blocks constituting the circuit) the following first-order relationship is obtained:

$$i_{out} = \left\{ I_b + e_1 + \frac{(x + y + e_2)^2}{4I_b} \right\} - \left\{ I_b + e_1 + \frac{(x + e_2)^2}{4I_b} \right\} + \left\{ I_b + e_1 + \frac{(e_2)^2}{4I_b} \right\} - \left\{ I_b + e_1 + \frac{(y + e_2)^2}{4I_b} \right\} = \frac{xy}{2I_b} \quad (9)$$

Thanks to the third term in which no input to the squarer exists, the errors due to the offsets in the squarer are canceled out. And so the final result is proportional to the product of x and y . Due to the fact that the transfer function of any single S²I memory cell is inverted type, the order in which the various terms are evaluated and the intermediate results are retrieved is important. In reality, the samples retrieved from the memory cells are slightly

attenuated. This would imply a not-exact cancellation of the quadratic terms in (7). In practice, most of the error comes from the fact that during θ_4 the output of the squarer is added to the attenuated partial result coming from memory 1. In order to compensate this unbalance, on θ_4 , the current gain of the mirror M_{4a} - M_5 is adjusted by connecting M_{4b} in parallel to M_{4a} . The ratio changes from 1:1 to 1: β with $\beta < 1$. A much more detailed analysis of the multiplier gives the following output relationship:

$$i_0 \cong i\rho \quad (10a)$$

where:

$$i = e_1(\eta^3 - \eta^2 + \eta - \beta) + e_2^2 \frac{1}{4I_b}(\eta^3 - \eta^2 + \eta - \beta) + \frac{x^2}{4I_b}\eta^2(\eta - 1) + \frac{y^2}{4I_b}(\eta^3 - \beta) + \frac{xe_2}{2I_b}\eta^2(\eta - 1) + \frac{ye_2}{2I_b}(\eta^3 - \beta) + \frac{\eta^3}{2I_b}xy; \quad \eta \equiv \frac{\Psi g_m}{\Psi g_m + 4g_{ds}}; \quad \rho \equiv \frac{g_L}{g_L + 4g_{ds}}; \quad (10b)$$

The first two terms of the right-hand side of i in (10b) represent an offset, the last term is the desired result while the remaining terms constitute the nonlinear distortion. Note that η is very close but less than 1. Therefore the third and fifth terms, being multiplied by a factor $\eta^2(\eta - 1)$, can be neglected. Relationship (10b) can be rewritten as:

$$i = i_{off} + \frac{y^2}{4I_b}(\eta^3 - \beta) + \frac{ye_2}{2I_b}(\eta^3 - \beta) + \frac{\eta^3}{2I_b}xy; \quad i_{off} = (\eta^3 - \eta^2 + \eta - \beta)(e_1 + e_2^2/4I_b); \quad (11)$$

where i_{off} is the offset current. The nonlinear error is canceled by setting the current mirror ratio as:

$$\beta = \beta_0 \equiv \eta^3 = \left(1 + 4g_{ds}/\Psi g_m\right)^{-3} \quad (12)$$

Assuming that $(W/L)_5 = (W/L)_{4a}$, the aspect ratio of M_{4b} is obtained as follows: $(W/L)_{4b} = (W/L)_{4a}(1 - \beta)/\beta$.

However, the final design of M_{4b} also depends on the actual voltage drop on the switch in series with M_{4b} . The final expression is therefore:

$$i = e_1(-\eta^2 + \eta) + e_2^2 \frac{1}{4I_b}(-\eta^2 + \eta) + \frac{\eta^3}{2I_b}xy \quad (13)$$

It is worth noting that because $\eta \cong 1$ the offset is almost canceled as well. So finally:

$$i_0 \cong \frac{\eta^3 \rho}{2I_b}xy \quad (14)$$

The experimental characteristic curves of a prototype of the multiplier are shown in Fig. 7. The multiplier's gain is $\eta^3 \rho / 2I_b = 5000A^{-1}$. It is worth noting that the operands have to be kept constant at the inputs of the multiplier while the result of the multiplication is released on θ_4 .

2.1.1.1 Ancillary circuits

In order to bias the current squarer, the multiplier's input node (B in Fig. 6) must be kept at $V_{dd}/2 = 1.5V$. Therefore a common gate amplifier has been used as level shifter. Moreover, in order to have the same order of magnitude for the two operands of the multiplier, a current amplifier is placed at the signal input. Finally, on-chip linear I-V converters allow the external observation of the data in the delay line.

2.1.2 Cell Behavior and Hardware Multiplexing

From what was discussed in the above sections two remarks can be done. First of all, the shift register cells allow new data to enter and retrieve the stored data only during ϕ_1 . Data is internally exchanged between the two half-delay cells on ϕ_2 . The full delay cell is completely isolated by the rest of the system on this phase. During this second phase the rest of the hardware would be essentially idle. These two phases determine the time synchronization represented by the time index of the state equation (1). Secondly, the result of the multiplication process is required before the end of a clock cycle. Therefore, a strategy that permits to exploit the available hardware during the idle phase and to save in terms of area is outlined next.

Let us refer to the block diagram shown in Fig. 8 depicting a CNN cell. The programmable synaptic connections using the multipliers are drawn on the left-hand side of the figure. Those multipliers accept the outputs of the shift register u_d on ϕ_1 with the corresponding weights (namely the control templates $B_{c,d}$) provided by off chip currents. On ϕ_2 , instead, the neighbor outputs y_d are fed to the inputs of the multipliers in place of u_d , while the feedback template $A_{c,d}$ is fed instead of the control template $B_{c,d}$.

Fig. 7 : Measured transfer characteristic. Output current I_0 versus input current I_x . The original scales are in volts. The corresponding currents are reported

Fig. 8 : Cell block diagram.

In other words, during ϕ_1 the weighted sum of the shift register outputs $\sum_{d \in N(c)} B_{c,d} u_d(n+1)$ is present on the summing node at the output of the multipliers. This sum enters the full delay cell. At the same time, the previous value of this sum is present at the output of the delay cell (because it entered the delay cell during the previous period n). Moreover, let us assume that the output of the half-delay cell, depicted below the full delay cell, is providing the weighted sum $\sum_{d \in N(c)} A_{c,d} y_d(n)$ at its output. (this assumption will soon be proven). Therefore, the

state $x_c(n+1)$ is obtained at the summing node of the two previous outputs, in accordance to the state equation (1). This one enters the second half-delay cell. Its output instead is zero and so is the output of the nonlinearity. On ϕ_2 the weighted sum $\sum_{d \in N(c)} B_{c,d} u_d(n+1)$ is stored in the full delay cell. It is passing from the first half-delay

cell to the second half-delay cell that constitutes the full delay cell itself. It is therefore completely isolated from the rest of the system. The cells outputs y_d and the feedback templates $A_{c,d}$ are fed to the synapses. So the weighted sum $\sum_{d \in N(c)} A_{c,d} y_d(n+1)$ is fed to the input of the half-delay cell on the left-hand side of the figure. This

value is available at its output on ϕ_1 of the next clock period (namely on period $n+2$). This proves the above assumption on the output of this half-delay cell. On ϕ_2 there are no currents at the summing node on the right of the full delay cell. The half-delay cell on the left hand side releases its stored value (namely $x_c(n+1)$) and so, due to the nonlinear block, the output $y_c(n+1)$ of the cell is available. This is consistent with the fact that the outputs $y_d(n+1)$ are provided as inputs of the cells during this phase. The above approach allow us to i) exploit the rest of the hardware during the idle phase ϕ_2 of the shift register and ii) use only three multipliers instead of six, saving in area and power. To obtain all of this, only two half-delay cells have been added to the classical CNN cell architecture. From this scheme it can be seen that while the outputs of the delay cells are available during the whole corresponding phases, instead, the sampling is accomplished only during $\phi_{1a}-\phi_{1b}$ (or $\phi_{2a}-\phi_{2b}$ for other half-delay cells) corresponding to θ_i (namely when the multiplication is completed). A whole multiplication cycle is performed during ϕ_1 and another one during ϕ_2 .

3. Simulation results

HSPICE simulation results, at transistor and functional level, are here presented. Let us first consider some transistor level simulation results. In a first case a sinusoidal input is fed at the input of the tapped delay line and shifted along it. The outputs of the on-chip linear I-V converters corresponding to the delay stages sr1, sr2, sr3 and sr7 respectively, are shown in Fig. 9a. The initial values present in the delays before the first sample reaches the stage can also be noticed.

The two input currents of one of the multipliers are shown in Fig.9c. One of the two inputs is fed alternatively with B and A while the other one is fed with u and y . The different signals on the distinct phases can be distinguished. It is seen that on phase ϕ_1 the corresponding B template coefficient is provided at one of the inputs while the output current coming from the delay line (u) is fed at the other one. Besides, the 4 phases $\theta_1-\theta_4$ of the multiplier are distinguished from the fact that one of the inputs (the template coefficient) enters on $\theta_1+\theta_2$ while the other input is accepted on θ_1 and θ_4 . Conversely, the template coefficient A is provided on phase ϕ_2 both with the corresponding output of the cell. A practical example of the functionality of the proposed architecture is now given. The voice of one of the authors has been sampled at 8KHz. White noise has been added obtaining a (4000

samples) noisy signal with signal-to-noise ratio $SNR=1.1965$. A SPICE macromodel of the chip has been simulated to process a wavelet decomposition of this signal (top of Fig.9b) according to the algorithm described in [1-2]. The wavelet coefficients are $\phi = [d_0 \ d_1 \ d_2 \ d_3]$ with $d_0 = (1+\sqrt{3})/4\sqrt{2}$, $d_1 = (3+\sqrt{3})/4\sqrt{2}$, $d_2 = (3-\sqrt{3})/4\sqrt{2}$, $d_3 = (1-\sqrt{3})/4\sqrt{2}$. The corresponding control templates for the algorithm are $B_1 = [d_0 \ d_1 \ 0]$ and $B_2 = [d_3 \ d_4 \ 0]$ (see [1] for the detailed description of the algorithm). The filtered signal (bottom of Fig.9b) has a $SNR=6.4288$, an improvement of 14.6043dB. In Fig.9b the time shift among the two signals cannot be seen at this level of magnification. The noise reduction, instead, is visible, particularly where the low frequency components of the vocal signal are dominant (at the beginning, in the middle and at the end).

Fig.9 : (a) tapped delay line outputs for sr1,sr2, sr3 and sr7; (b) noisy vocal signal (top) and filtered output (bottom); (c) inputs of one of the multipliers

4. Conclusions

The VLSI implementation of a discrete-time one-dimensional Cellular Neural Network has been discussed. One of the peculiarities of the proposed architecture is a hardware-multiplexing strategy. This allows to efficiently use the hardware halving the number of multipliers and storing the intermediate results into temporary memories. Simulation results at transistor and functional level have been reported. A CMOS N-well MOSIS Orbit 2μ chip is currently in fabrication.

5. References

- [1] O.Moreira-Tamayo and J.Pineda de Gyvez, "Filtering and Spectral Processing of 1-D Signals using Cellular Neural Networks", Proc. of International Symposium on Circuits and Systems, Atlanta (USA), vol.3, pp.76-79, 1996
- [2] O.Moreira-Tamayo "Analog Systems for Spectral Analysis and Signal Processing", Ph.D. Dissertation, Texas A&M University, December 1996
- [3] S.H.Bang and B.Sheu, "A Neural Network for detection of signals in communication", IEEE Trans. Circuits & Systems - Part I, vol.43, pp.644-655, 1996
- [4] B. Andò, S. Baglio, S. Graziani, N. Pitrone, "A novel analog modular sensor fusion architecture for developing smart structures", Proc. Of IEEE IMTC97, May 19-21, 1997, Ottawa, Canada.
- [5] J.B. Hughes, K.W. Moulding, "S²I: a switched-current technique for high performance", *Electronics Letters*, vol. 29, no.16, 5th August 1993, pp.1400-1401
- [6] J.B. Hughes, K.W. Moulding, "Enhanced S²I switched-current cells", *Proc. of Int. Symp. On Circuits and Systems*, Atlanta, USA, 1996
- [7] G.Manganaro and J.Pineda de Gyvez, "A Four Quadrant S²I Switched-current Multiplier", submitted for publication
- [8] C. Toumazou, J.B. Hughes, N.C. Battersby (editors), *Switched-currents: an analogue technique for digital technology*, IEE Circuits and Systems series, London: Peter Peregrinus, 1993