INT. J. ELECTRONICS, 1995, VOL. 79, NO. 2, 151–162

Digital VLSI circuit design and simulation of an adaptive resonance theory neural network

C. S. HO† and J. J. LIOU†

A digital VLSI circuit design for an adaptive resonance theory (ART) neural network architecture, called the augmented ART-1 neural network (AART1-NN) is presented. An 'axon-synapse-tree' structure is used to realize the activities of the short-term memories and reset subsystem. The long-term memory traces are implemented using NMOS transmission gates. PSpice circuit simulation was carried out to verify the design of a prototype, seven-node AART1-NN. A clock-controlled delay element is included in the simulation to illustrate the functionality of the AART1-NN. It is shown that the AART1-NN node selection activities simulated from the circuit designed are identical to those described by the coupled differential equations governing the AART1-NN.

1. Introduction

Modern information processing has become increasingly complicated and thus requires systems which are potentially robust to noise and incomplete information, adaptive to a changing environment, and intrinsically and massively parallel. Neural networks are capable of meeting such requirements. Most neural networks reported in the literature have been implemented via computer simulation of their corresponding mathematical models. For real-life applications, however, neural networks need to be realized through analogue, digital or hybrid (analogue digital) VLSI circuits.

This work concentrates on one kind of neural network, the binary-input adaptive resonance theory neural network (ART1-NN), which was developed by Carpenter and Grossberg (1987). Some past efforts have been devoted to the design of the ART1-NN. Several analogue realizations have been introduced (Nahet *et al.* 1989, Tsay and Newcomb 1991, Tsay *et al.* 1990, Schneider and Card 1991 a and b), but no attempt has been made to fully implement a complete ART1-NN architecture. A digital implementation of ART1-NN was introduced by Rao *et al.* (1989) based on the pipelined associative memory. However, such an approach does not fully incorporate the parallel mechanism in the ART1-NN.

The question that often arises is whether analogue or digital type circuits are more suitable for neural network implementation. The answer is that both are very desirable. Each is theoretically capable of solving any kind of mathematical relationship. Each, however, has certain practical limitations and advantages over the other. Since digital circuits operate with binary signals, problems involving large amounts in discrete number form would probably be handled most readily by a digital implementation. On the other hand, problems concerning the integration of continuous data may be solved more readily with an analogue implementation. Since the digital approach is much more precise than its analogue counterpart,

Received 30 November 1994; accepted 9 December 1994.

†Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, U.S.A.

problems with precisely defined input data, such as the input of the ART1 neural network, would best be solved by the digital circuit. In addition, in comparison with the analogue circuit, the digital circuit is less sensitive to device parameter changes, is less susceptible to variations of the environment (e.g. temperature change), and it is easier to identify the functionality of the digital circuit. The main disadvantage of the digital approach is that the precise relationship between the input and output signals cannot be found due to the stepwise process.

This paper presents a digital VLSI circuit design based on an augmented AART1-NN model. The AART1-NN, which was developed by Heileman *et al.* (1991), exhibits the same behaviour as the ART1-NN. The major difference between the two types of neural network is that the dynamics of the AART1-NN are completely described by a set of coupled differential equations, whereas the ART1-NN involves additional algorithmic components in its description (Heilema *et al.* 1991).

It should be mentioned that the number of nodes required in the neural network varies greatly with applications. For the AART1-NN, a real system of 64 nodes is capable of classifying a 21-by-21 data array in pattern recognition applications. The same number of nodes has also been implemented in standard CMOS technology for a motion detection neural network (Mead 1989). For a more complicated application, such as a SeeHear visual system, the actual chip will contain 32×36 nodes (Mead 1989).

2. Basic structure of AART1-NN

A schematic of the AART1-NN is shown in Fig. 1. It consists of two subsystems—the attentional subsystem and the orienting subsystem. The attentional subsystem consists of two fields of nodes designated as the fields F_1 and F_2. The nodes in F_1 (e.g. nodes $v_i, 1 \leq i \leq M$) and the nodes in F_2 (e.g. nodes v_j, $M+1 \leq j \leq N$) are used to encode patterns of the short-term memory (STM) activity. Each node in F_1 and in the first layer of F_2 is connected via a bottom-up weighted connection, called the long-term memory (LTM) trace. The pathway from F_1 to F_2 is called the bottom-up LTM trace (denoted z_{ij}) and, likewise, the pathway from F_2 to F_1 is called the top-down LTM trace (denoted as z_{ji}). The F_1 field is often referred to as the input field because the input patterns are presented to it. The first layer of F_2 is often referred to as the category representation layer, because it is the layer that indicates the category to which the input pattern belongs to. The objective of the second layer of F_2 (e.g. nodes $\hat{v}_j, M+1 \leq j \leq N$) is to deactivate the erroneous category representation in the first layer of F_2, whenever such an erroneous representation occurs, and to keep this erroneous category deactivated for as long as the same input pattern is present at F_1. The orienting subsystem in the AART1-NN architecture consists of a single node designated as v_r. The primary purpose of the orienting subsystem is to generate a reset wave to F_2 whenever the category representation in the first layer of F_2 is not a good match with the input pattern.

3. Circuit design of AART1-NN

Throughout the paper, all activities in the AART1-NN are realized by binary values 1(5 V) and 0(0 V). Also, a node is denoted as '1' if it is activated; otherwise, it is denoted as '0'.

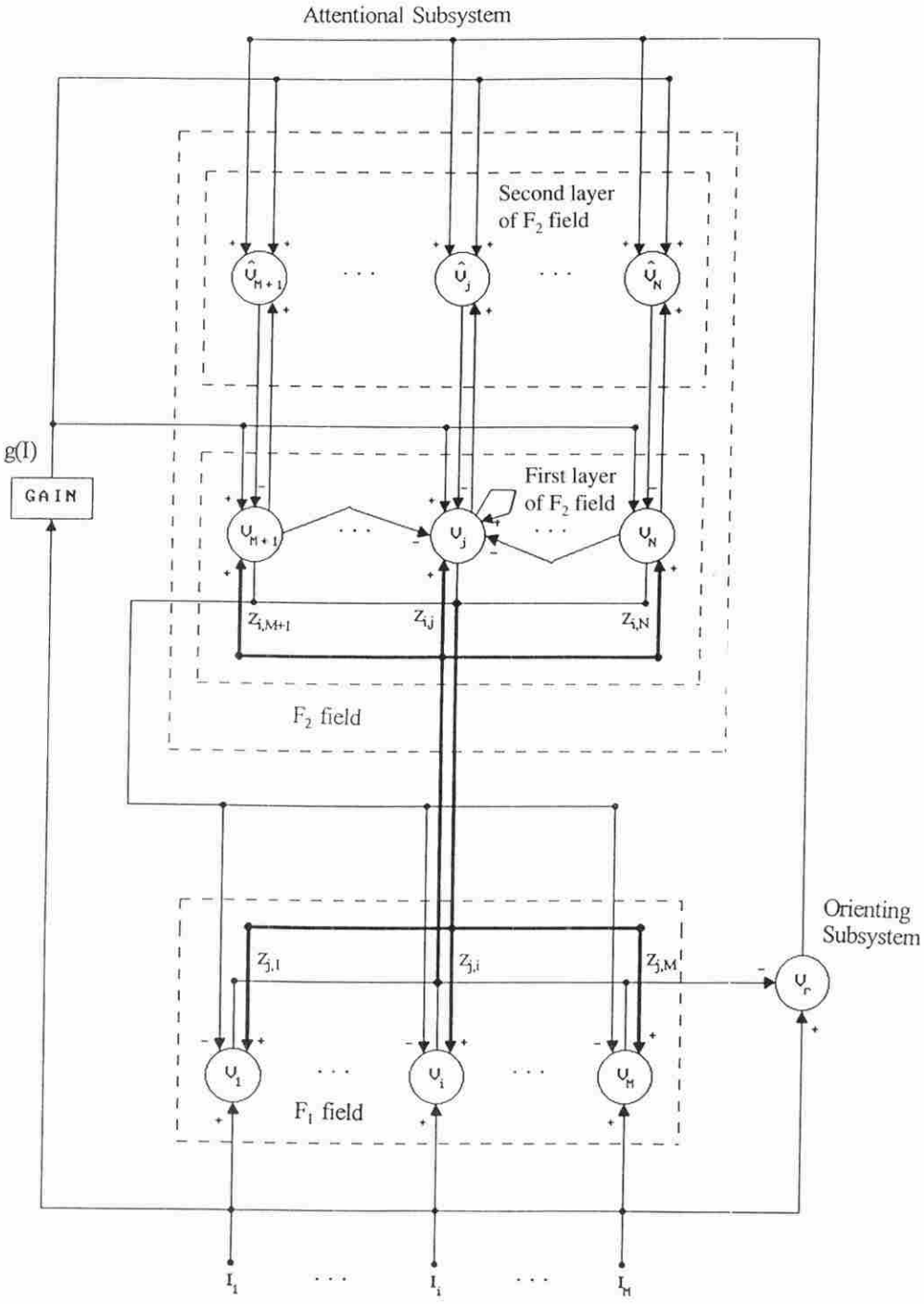


Figure 1. Structure of the augmented ART-1 neural network (AART1-NN).

3.1. *Short-term memory*

In the AART1-NN model, the STM activity of a node in F_1 and the first layer of F_2 can be described by

$$\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} x = -x + (1 - Ax)\left(\sum_{i=0}^{M} S_i\right) - (B + Cx)\left(\sum_{j=1}^{M} T_j\right) \tag{1}$$

where ε, A, B, C and D are arbitrary constants, and the sigmoidal functions S_i and T_j are the total excitatory and inhibitory inputs, respectively. For simplicity, $A = B = C = 1$. Then, the steady-state STM activity x_{ss} can be obtained from (1) as

$$x_{ss} = \frac{\sum_{i=0}^{M} S_i - \sum_{j=1}^{M} T_j}{1 + \sum_{i=0}^{M} S_i + \sum_{j=1}^{M} T_j} \tag{2}$$

It is apparent in (2) that x_{ss} increases with the sigmoidal function ΣS_i and decreases with the sigmoidal function ΣT_j, respectively. Based on these properties, a digital STM circuit was designed by utilizing the axon synapse-tree (AST) structure introduced by Tsay and Newcomb (1991), as shown in Fig. 2. In the circuit, the functionality is realized using the voltage-controlled conductances generated by NMOS devices operated in the linear (or ohmic) region. Note that S_i and T_j are the voltages of binary control signals to the synapse transistors of the AST structure (see Fig. 2). The conductance G_n of an NMOS transistor in its ohmic region is given by

$$G_n = \frac{I_{DS}}{V_{DS}} = \beta_n \left[V_{GS} - V_{TN}) - \frac{V_{DS}}{2} \right] \tag{3}$$

where V_{DS} and V_{GS} are the drain and gate voltages related to the source, V_{TN} is the threshold voltage, $V_{GS} - V_{TN} > V_{DS}$, and $\beta_n = \mu_n C_{ox}(W/L)$ (μ_n is the electron mobility, C_{ox} is the oxide capacitance, and W and L are the MOS channel width and channel

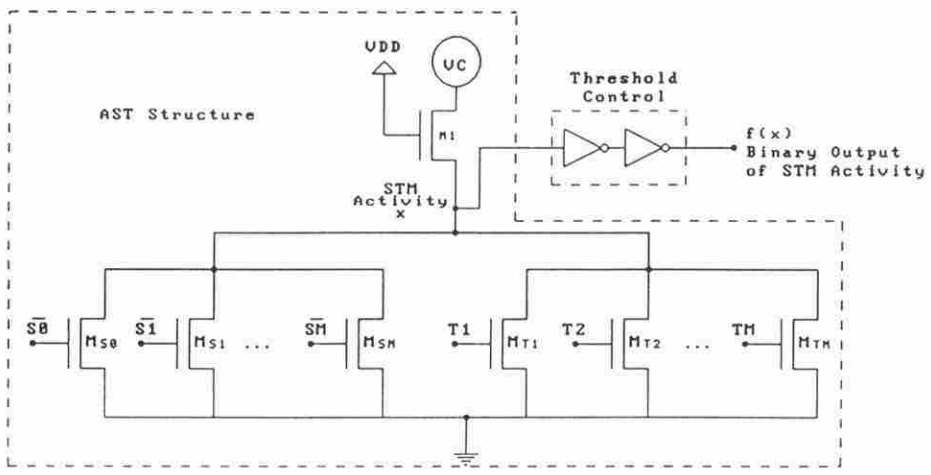


Figure 2. Circuit diagram of the AART-NN short-term memory (STM).

length, respectively). Assuming all synapse transistors have the same device parameter β_n, then they can be lumped into a single equivalent resistor with a conductance $G_1=(a+b)\beta_n(V_H-V_{TN}-X/2)$. Here, V_H is the control signal at 'high' voltage ($V_H=5\,V$ is used throughout the paper), a is the number of synapse transistors $M_{S0}\sim M_{SM}$ that are turned on, and b is the number of synapse transistors $M_{T1}\sim M_{TM}$ that are turned on. Thus, the STM activity X is equal to $V_C G_2/(G_1+G_2)$, where $G_2=\beta'(V_H-V_{TN}-X/2-V_C/2)$ is the total conductance of the axon transistor. Here, β is the device parameter for the axon transistor and V_C is the constant voltage source applied to the circuit.

For proper signal transmission in the NMOS devices, the following conditions have to be met: $S_i-V_T>X$, $T_i-V_T>X$, and $V_{DD}-V_{TN}>X$. Since $V_{DD}=S_{i(high)}=T_{i(high)}=5\,V$ was chosen (V_{DD} is the bias voltage, see Fig. 2(*a*), V_C and X should be less than 4 V when $V_{TN}=1\,V$. It is found that X decreases with increasing a and b, and X increases with the total excitatory input and decreases with the total excitatory input. Therefore, the AST structure shown in Fig. 2 performs a trend that is analogous to the STM activity of AART1-NN.

To generate the binary STM activity (1 or 0), a CMOS inverter circuit with a threshold voltage V_{th} is implemented following the AST structure (Fig. 2). The output activity is '1' when $X>V_{th}$ and is '0' otherwise. According to the characteristics of a CMOS inverter, V_{th} can be determined by

$$V_{th}=\frac{V_{TN}+\left(\frac{\beta_p}{\beta_n}\right)^{1/2}(V_{DD}+V_{TP})}{1+\left(\frac{\beta_p}{\beta_n}\right)^{1/2}} \tag{4}$$

where β_p and V_{TP} are the device parameter and threshold voltage for the PMOS transistor.

To illustrate the characteristics of the STM, PSpice simulation was carried out for node v_1 (with activity x_1) of the AART1-NN. The circuit shown in Fig. 3. In the simulation, $V_{TN}=1\,V$, $V_{TP}=-1\,V$ and $V_C=3{\cdot}2\,V$ were chosen. The input pulses (five gate voltages) of the circuit are: $I_1, f_2(x_3)z_{31}, f_2(x_4)z_{41}, f_2(x_3)$ and $f_2(x_4)$. Their waveforms are shown in Fig. 4(*a*). The corresponding output voltage x_1 and its binary value $f_1(x_1)$ are illustrated in Fig. 4(*b*). During the time period $t=[0,40]$ (in microseconds), it is found that all that inhibitory inputs are 0 and the activity x_1 increases with increasing excitatory inputs. After that, all the total excitatory inputs are high, and the activity of x_1 decreases with increasing inhibitory inputs. The simulation results agree with the STM activity described by the AART1-NN algorithm.

3.2. *Long-term memory*

In the AART1-NN model, the TM traces are described by the following two equations:

$$\varepsilon_z\frac{d}{dt}z_{ij}=-\left[(L-1)f_1(x_i)+\sum_{k=1}^{M}f_1(x_k)\right]f_2(x_j)z_{ij}+Lf_1(x_i)f_2(x_j) \tag{5}$$

$$\varepsilon_z\frac{d}{dt}z_{ji}=-f_2(x_j)z_{ji}+f_1(x_i)f_2(x_j) \tag{6}$$

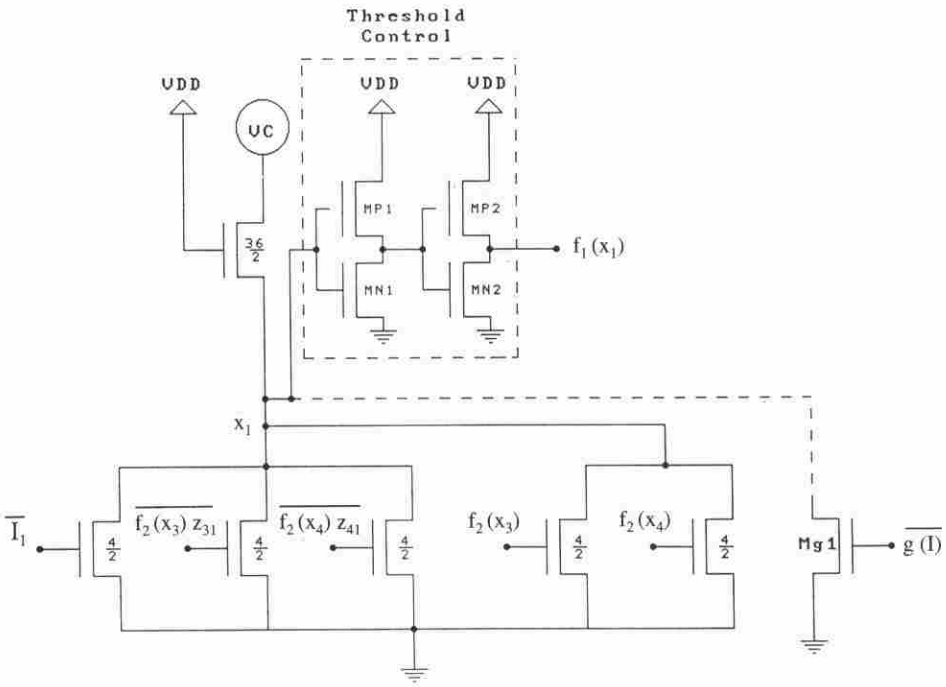


Figure 3. Digital circuit implementation for the short-term memories in the AART1-NN.

Then, the steady-state values of the LTM traces can be obtained from (5) and (6), using $L=1$ and $f_2(x_j)=1$, as

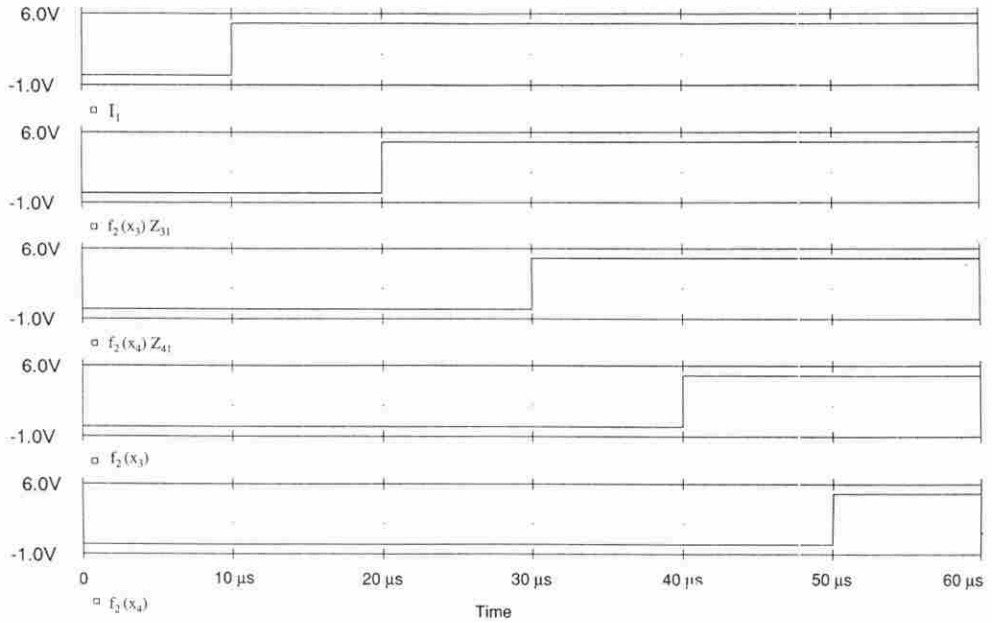
$$z_{ij(\mathrm{ss})}=\frac{f_1(x_i)}{\sum_{k=1}^{M} f_1(x_i)}, \quad z_{ji(\mathrm{ss})}=f_1(x_i) \tag{7}$$

It can be seen from (5) that the bottom-up LTM trace z_{ij} can change its value only when node v_j is activated ($f_2(x_j)=$high). Also z_{ij} approaches $z_{ij(\mathrm{ss})}$ when node v_i becomes activated and decays to zero when node v_i is deactivated. Note that the value of $z_{ij(\mathrm{ss})}$ in (7) is not a binary type value. To create binary inputs to the STM circuits, in which all the excitatory and inhibitory inputs are binary values, let $z_{ij(\mathrm{ss})}=1$ (high) when node v_i is activated and $z_{ij(\mathrm{ss})}=0$ (low) when node v_i is deactivated. From these modified properties, a DRAM-type circuit is designed for the bottom-up LTM trace, as shown in Fig. 5. Such a circuit also applies to the top-down LTM trace.

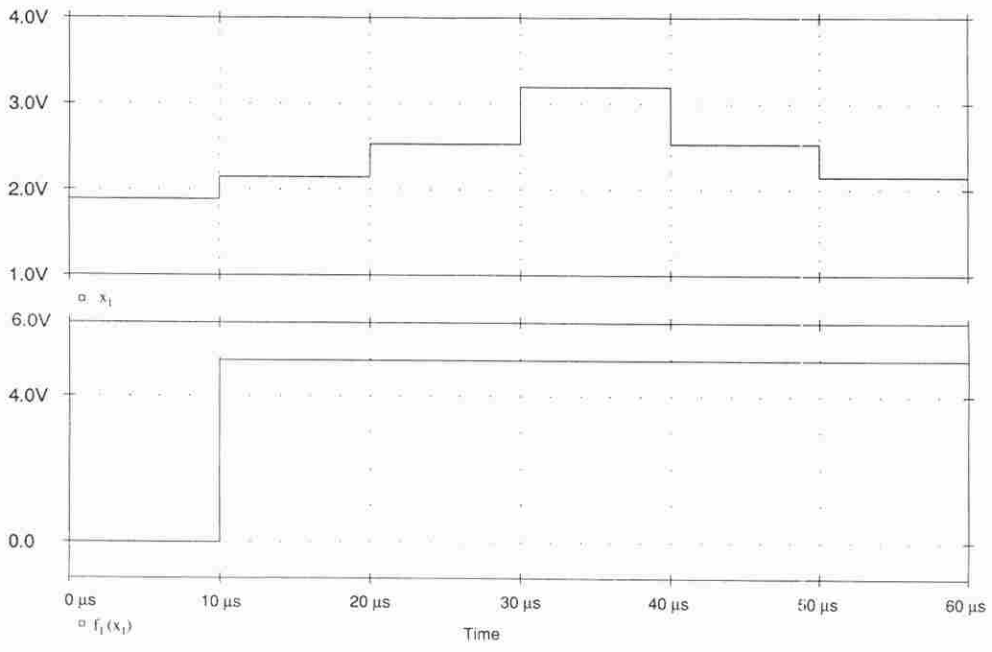
3.3. *Reset subsystem*

The reset subsystem includes the reset node and nodes in the second layer of F_2. In the AART1-NN model, the activity of the reset node x_r is described by

$$\varepsilon_r \frac{\mathrm{d}x_r}{\mathrm{d}t}=-A_r x_r+U\left[P\sum_{i=1}^{M} I_i-Q\sum_{i=1}^{M} f_1(x_i)\right] \tag{8}$$



(*a*)



(*b*)

Figure 4. (*a*) Input waveforms of the STM circuit shown in Fig. 3; (*b*) corresponding output voltage x_1 and its binary value $f_1(x_1)$ obtained from PSpice circuit simulation.

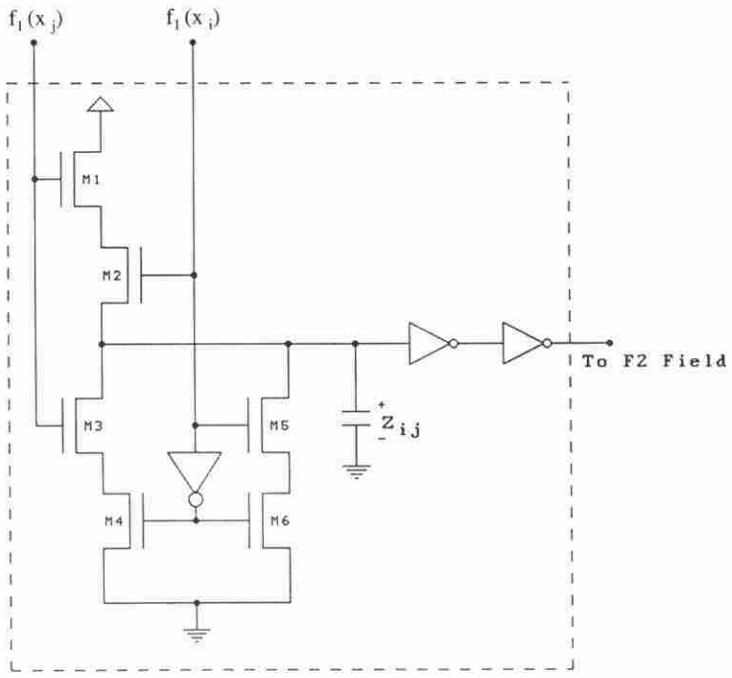


Figure 5. DRAM-type circuit for the bottom-up LTM trace.

where A_r, P and Q are constants, U denotes the unit step function, and I_i and $f_1(x_i)$ are the binary input and output of the node v_i in F_1, respectively. The reset node becomes activated when $U=1$ and becomes deactivated when $U=0$. Therefore, the binary activity of the reset node can be characterized as: $f_r(x_r)=1$ if $\Sigma I_i > \Sigma f_1(x_i)$, and $f_r(x_r)=0$ otherwise. Following the STM implementation described above, the sigmoidal items shown in (8) can be demonstrated by the AST structure and the digital reset wave can be generated by a threshold control circuit. A general-purpose circuit for the reset node is shown in Fig. 6(*a*).

The activity of node $\hat{v}_j$ in the second layer of F_2 is described by

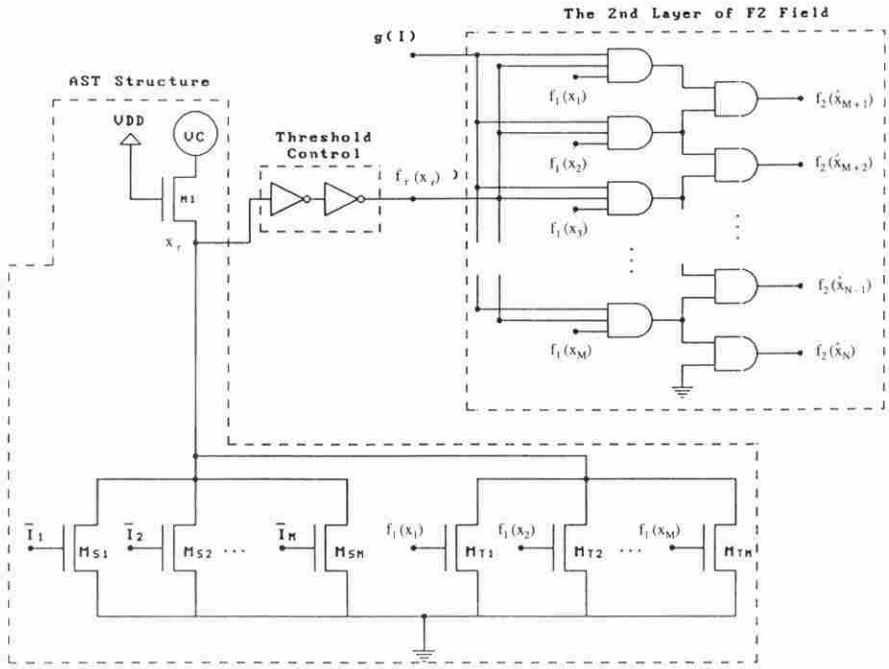
$$\varepsilon_2 \frac{d\hat{x}_j}{dt} = -[1-g(I)]\hat{x}_j + g(I) f_r(x_r) f_2(x_j) \tag{9}$$

where $f_r(x_r)$ and $f_2(x_j)$ are the binary output signals for the reset node and nodes in the first layer of F_2, respectively, and $g(I)$ is defined as a unit function related to the summation of the binary input pattern (Carpenter and Grossberg 1987). According to (9), node $\hat{x}_j$ can become active only when $g(I)f_r(x_r)f_2(x_j)=1$. Consequently, $f_2(\hat{x}_j)=5\,\text{V}$ when $g(I)f_r(x_r)f_2(x_j)=5\,\text{V}$ and $f_2(\hat{x}_j)=0\,\text{V}$ when $g(I)f_r(x_r)f_2(x_j)=0\,\text{V}$. An AND-gate circuit designed to represent this logic function is shown in Fig. 6(*a*).

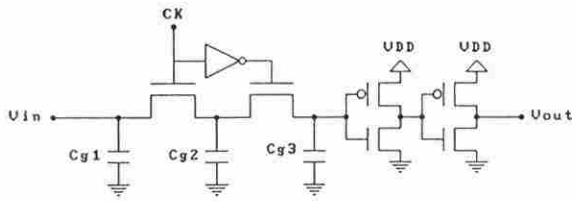
The behaviour of the reset subsystem is characterized as follows:

(*a*) for the case $f_r(x_r)=0$ or $g(I)=0$: $(f_2(\hat{x}_{M+1}), f_2(\hat{x}_{M+2}), \ldots, f_2(\hat{x}_N))=(0,0,\ldots,0)$. Thus, there is no reset activity occurring at this time;

(*b*) for the case $f_r(x_r)=1$ and $g(I)=1$: when $I^1=(I_1, I_2, \ldots, I_M)=(1,0,\ldots,0)$ is presented $(f_1(x_1), f_1(x_2), \ldots, f_1(x_M))=(1,0,\ldots,0)$, and then $(f_2(\hat{x}_{M+1}),$



(*a*)



(*b*)

Figure 6. Circuit diagrams: (*a*) reset subsystem; (*b*) delay element.

$f_2(\hat{x}_{M+2}),\ldots,f_2(\hat{x}_N))=(0,0,\ldots,0)$. Thus, there will be no reset activity occurring at this time. When $I^3=(I_1,I_2,\ldots,I_M)=(1,1,\ldots,0)$ is presented, $(f_1(x_1),f_1(x_2),\ldots,f_1(x_M))=(1,1,\ldots,0)$, and then $(f_2(\hat{x}_{M+1}),$ $f_2(\hat{x}_{M+2}),\ldots,f_2(\hat{x}_N))=(1,0,\ldots,0)$. Thus, node v_{M+1} will be reset and node v_{M+2} will be chosen to represent the input pattern.

Notice that input pattern I^2 is a zero pattern at which $g(I)=0$.

4. Circuit simulation

PSpice circuit simulation for a seven-node prototype AART1-NN circuit, which has two nodes (v_1,v_2) in F_1, two nodes (v_3,v_4) in the first layer of F_2, two nodes

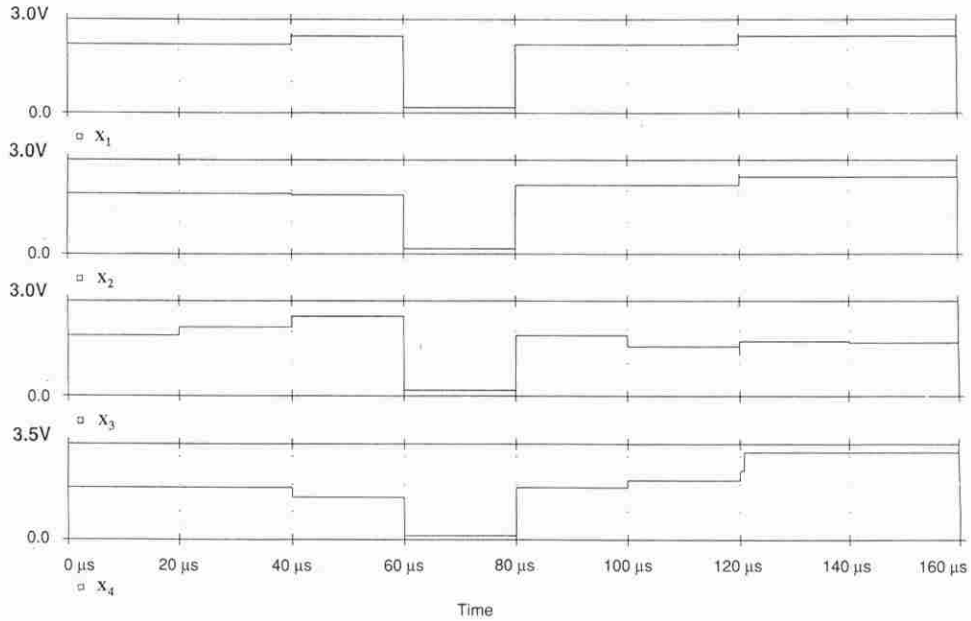


Figure 7. Simulation results for the short-term memory activities x_1, x_2, x_3 and x_4.

(v_3, v_4) in the second layer of F_2, and one reset node v_r, has been carried out on a Sun SPARC workstation. In the simulation, the input pattern was designated as $I^k (1 \leq k \leq 3)$ and $I^i = (I_1, I_2)$, where I_1 and I_2 are the binary input components to nodes v_1 and v_2 in F_1, respectively. Three continuous input patterns, $I^1 = (1, 0)$, $I^2 = (0, 0)$, $I^3 = (1, 1)$, are provided during the time interval [0, 60], [60, 80] and [80, 160] (in microseconds), respectively.

Delay elements were designed and added following every STM node and the reset node in order to maintain a time duration for nodes in the second layer of F_2 to generate deactivate signals to the first layer of F_2 and to improve the glitching problem caused by the undesirable timing mismatch. The delay element circuit is shown in Fig. 6(*b*), where the ratios of Cg1/Cg2 and Cg2/Cg3 are chosen to be larger than 20 to avoid the charge sharing problem. Furthermore, because the zero pattern is always presented between presentations of useful patterns in the AART1-NN model, an additional grounded NMOS device is designed to reset every STM activity during the presentation of the zero input pattern.

The PSpice simulation results for the STM activities are given in Figs 7 and 8. After the first pattern $I^1 = (1, 0)$ is presented to F_1, node v_1 becomes active (x_1 is larger than the threshold $V_{th} = 2\,\text{V}$). After a clock cycle delay, $f_1(x_1)$ goes to 5 V at $t = 20\,\mu\text{s}$ (Fig. 7). Then, node v_3 is selected to present the input pattern ($f_2(x_3)$ goes to 5 V at $t = 40\,\mu\text{s}$) because node v_3 receives more excitatory inputs than node v_4. According to the function of the designed reset subsystem, there is no rest wave generated at this point $((f_2(\hat{x}_3), f_2(\hat{x}_4)) = (0, 0))$. Consequently, node v_3 is deemed by the architecture as the right node in F_2 to represent the input pattern I^1. For the LTM traces, it is found that z_{13} remains at 5 V and z_{31} increases to 5 V once $f_2(x_3) = 5\,\text{V}$ (these results are not shown). After the second pattern $I^2 = (0, 0)$ is presented to F_1, all STM activities decrease below the threshold ($V_{th} = 2\,\text{V}$) and their

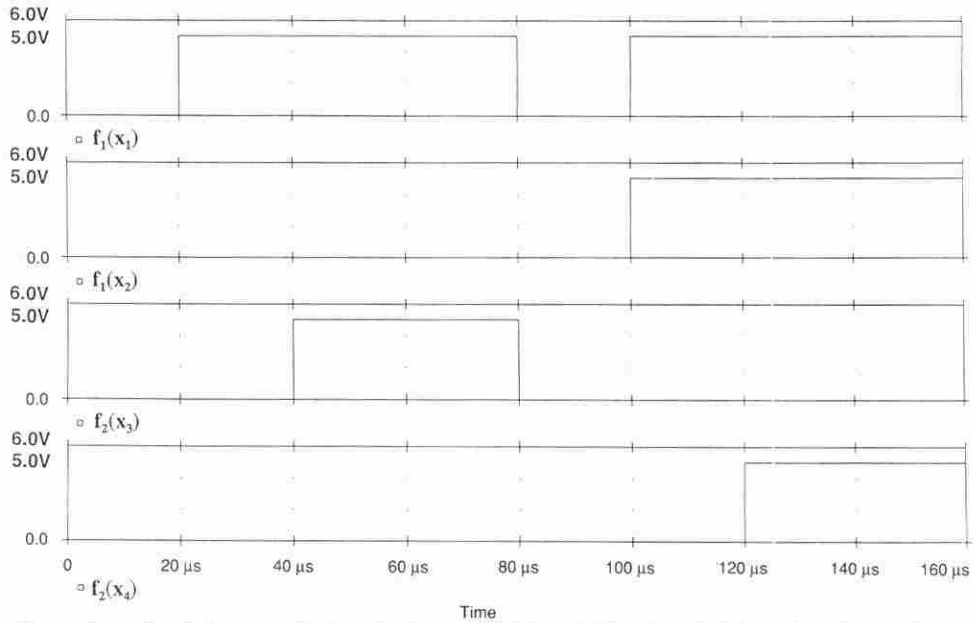


Figure 8. Simulation results for the binary STM activities $f_1(x_1)$,$f_1(x_2)$,$f_2(x_3)$ and $f_2(x_4)$.

binary outputs decay to 0 V at $t=80\,\mu s$ (Figs 7 and 8). After the third pattern $I^3=(1,1)$ is presented to F_1, nodes v_1 and v_2 are activated (Fig. 7) and $f_1(x_1)$ and $f_1(x_2)$ increase to 5 V at $t=100\,\mu s$ (Fig. 8). After one clock time delay, nodes v_3 and v_4 are supposed to become active ($x_3=5$ V and $x_4=5$ V) at the same time. However, the reset wave is generated to deactivate node v_3 at this moment ($f_2(\hat{x}_3)$,$f_2(\hat{x}_4)$ $=(0,1)$ are found in Fig. 8. Eventually, node v_4 is chosen to present the input pattern I^3 (see Fig. 8, $t=[120,160]$). Thus, the LTM traces corresponding to node v_4 approach 5 V ($z_{14}=5$ V and $z_{41}=5$ V) once $x_4=5$ V (not shown).

5. Conclusions

The AART1-NN can cluster in a parallel manner an arbitrary collection of binary input patterns. This capability makes the AART1-NN very attractive for high-speed signal processing applications. In this paper, a prototype seven-node AART1-NN circuit has been successfully designed with a digital VLSI circuit and verified in PSpice circuit simulations. It has been shown that the simulated digital circuit node selection activities are identical to those described by the coupled differential equations governing the AART1-NN. The design procedure suggested in this paper can easily be extended to larger AART1-NN architectures as well as to other types of neural networks.

Acknowledgments

This work was supported in part by the Florida High Technology Council and by the Division of Sponsored Research at the University of Central Florida.

References

Carpenter, G. A., and Grossberg, S., 1987, A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, **37**, 54–115.

HEILEMAN, G. L., GEORGIOPOULOS, M., and ABDALLAH, C., 1991, A dynamical adaptive resonance architecture. *Proceedings of the IEEE International Joint Conference on Neural Networks*, Singapore, Vol. 3, pp. 2658–2663.

MEAD, C., 1989, *Analog VLSI and Neural Systems* (Reading, Mass: Addison-Wesley).

NAHET, B., DARLING, R. B., and PINTER, R. B., 1989, Analog implementation of shunting neural networks. *Proceedings of the 1988 Neural Information Processing Systems*, p. 695.

RAO, A., WALKER, M. R., CLARK, L. T., and AKERS, L. A., 1989, Integrated circuit emulation of ART1 networks. *Proceedings of the first IEEE Conference on Artificial Neural Networks*, pp. 37–41.

SCHNEIDER, C., and CARD, H., 1991 a, CMOS implementation of analog Hebbian synaptic learning circuits. *Proceedings of the 1991 IEEE International Joint Conference on Neural Networks*, Seattle, Vol. 1, pp. 437–442. 1991 b, Analog CMOS synaptic learning circuits adapted from invertebrate biology. *IEEE Transactions on Circuits and Systems*, **38,** 1430–1438.

TSAY, S. W., EL-LEITHY, N., and NEWCOMB, R., 1990, CMOS realization of a class of Hartline neural pools. *Proceedings of the IEEE International Symposium on Circuits and Systems*, New Orleans, pp. 2417–2420.

TSAY, S. W., and NEWCOMB, R., 1991, VLSI implementation of ART1 memories. *IEEE Transactions on Neural Networks*, **2,** 214.