An ART1 Microchip and its use in Multi-ART1 Systems

Teresa Serrano-Gotarredona and Bernabé Linares-Barranco

National Microelectronics Center (CNM), Ed. CICA, Av. Reina Mercedes s/n, 41012 Sevilla,

SPAIN. Phone: (34) 5-4239923, FAX: (34) 5-4239923, E-mail: bernabe@cnm.us.es

Abstract

Recently, a real-time clustering microchip based on the ART1 [1] algorithm has been reported [2]. This chip was able to classify 100-bit input patterns into up to 18 categories. However, its high area comsumption ($1cm^2$) caused a very poor yield (6%). In this paper, an improved prototype is presented. In this chip, a different approach has been used to implement the most area consuming elements. The new chip can cope with 50-bit input patterns and classify them into up to 10 categories. Its area is 15 times less than that of the first prototype and it exhibits a yield performance of 98%. Due to its higher robustness, multichip systems are easily assembled.

I. Introduction

Recently, a real-time clustering microchip neural engine based on the ART1 architecture has been reported [2]. It is based on a slightly modified version of the ART1 algorithm which was shown to preserve all its original computational properties [3], but has a more VLSI-friendly algorithmic structure. The reported ART1 chip was able to cluster binary input patterns of up to 100 pixels into up to 18 different categories. The chip was able to classify an input pattern and learn its relevant characteristics by updating its internal knowledge, all in less than $1.8\mu s$. The chip internal circuit architecture also allowed modular expansion of the clustering system. Assembling an $N \times M$ array of these chips would result in ART1 systems able to cluster $N \times 100$ pixel input patterns into up to $M \times 18$ categories. Unfortunately, the resulting area consumption (and cost) of the chip was extremely high ($1cm^2$), and consequently its yield performance was extremely low (6%). Nevertheless, due to the fault-tolerant nature of the algorithm, most of the faulty chips still were able to perform satisfactorily [2].

In this paper, a new ART1 chip is presented which solves the yield problem by reducing chip area. After careful MOS transistor electrical parameter mismatch characterization of the technological process to be used, it was possible to identify the maximum chip area for which the parameter variations would remain within the necessary limits to preserve the required system operation precision. It was found that for the ES2-1.0μm CMOS process, for transistors of size $10\mu m \times 10\mu m$, spread over an area of the order of $2.5mm \times 2.5mm$, and for current levels around $10\mu A$, the transistor current standard deviation is around $\sigma(I) \approx 1\%$. Taking this into account, we designed and fabricated an ART1 chip capable of clustering 50-bit input patterns into up to 10 categories, with a yield performance of 98%, and whose area is 15 times less than that of the first prototype. The chip showed a very robust behavior that allowed us to implement some multi-chip ART1 systems.

II. VLSI-Friendly Algorithm and Its Hardware Implementation

An ART1 system is a neural associative memory capable of generating in an unsupervised way stable recognition codes in response to arbitrary many and complex binary input patterns. An ART1 architecture consists of two layers. The F_1 or input layer has N nodes each of them receiving a component of the input vector $\mathbf{I} = (I_1, \ldots, I_N)$. Each of the M nodes in the F_2 or category layer represents a cluster of input patterns or learned category. Both layers are fully connected by a matrix of binary weights z_{ij}. The weight vector that connects to the j^{th} F_2 node $\mathbf{z}_j = (z_{1j}, \ldots, z_{Nj})$ characterizes the learned F_2 category j. Fig. 1 depicts the operation sequence of the VLSI-Friendly ART1 algorithm:

1. All the binary weights z_{ij} are set to '1'.
2. An input pattern $\mathbf{I}$ is applied to the system.
3. A "*choice function*" T_j is computed for each category j. This function $T_j = L_A|\mathbf{I} \cap \mathbf{z}_j| - L_B|\mathbf{z}_j| + L_M$ is a measurement or distance of the similarity between the input pattern $\mathbf{I}$ and the learned vector $\mathbf{z}_j$ corresponding to category j.
4. The category J whose T_J is maximum is selected. The corresponding output y_J is set to '1' while all others are set to $y_{j \neq J} = 0$.

5. The vigilance criterion is checked for the winning category.

   If $\rho|\mathbf{I}| < |\mathbf{I} \cap \mathbf{z}_J|$ the criterion is not satisfied, T_J is forced to '0' and a new winner is selected.

   If $\rho|\mathbf{I}| \geq |\mathbf{I} \cap \mathbf{z}_J|$ the weights $\mathbf{z}_J$ are updated according to the law

$$\mathbf{z}_J(new) = \mathbf{I} \cap \mathbf{z}_j(old). \quad (1)$$

Fig. 2(a) shows the schematic of the circuit implementing an ART1 network with 50 input nodes and 10 category nodes. The schematic of a synapse s_{ij} is depicted in Fig. 2(b) and Fig. 2(c) shows the schematic of an input cell C_i.

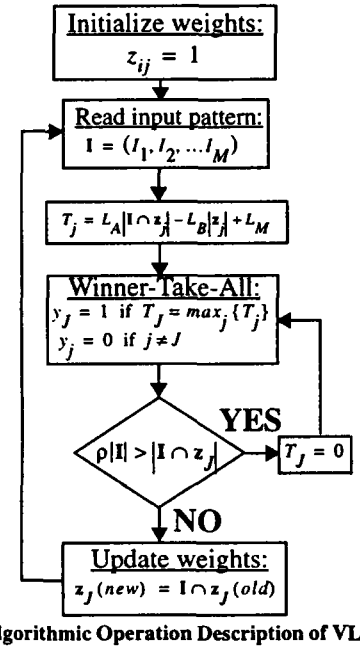

Fig. 1: Algorithmic Operation Description of VLSI-Friendly ART1 System

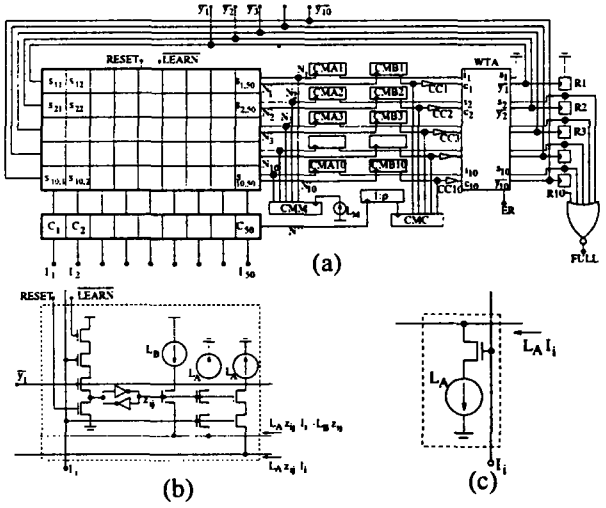

Fig. 2: (a) Schematic of the Circuit that Implements the VLSI-Friendly ART1 Algorithm, (b) Detailed Schematic of a Synapse, and (c) Schematic of an Input Cell

The array of input cells generates a current $L_A\sum I_i = L_A|\mathbf{I}|$, which enters a tunable gain current mirror of gain ρ. This mirror distributes a current $\rho L_A|\mathbf{I}|$ to the input of ten current comparators CC_j.

Each synapse outputs two currents:

- A current $L_A I_i z_{ij}$ which flows to a common node for all the synapses in the $j-th$ row, resulting in a total current $L_A\sum I_i z_{ij} = L_A|\mathbf{I} \cap \mathbf{z}_j|$ that enters in the $j-th$ current comparator CC_j.
- A current $L_A I_i z_{ij} - L_B z_{ij}$, that results in a total current $T_j = L_A\sum I_i z_{ij} - L_B\sum z_{ij} + L_M$ which enters the $j-th$ branch of the WTA.

Each current comparator receives a total current $L_A|\mathbf{I} \cap \mathbf{z}_j| - \rho L_A|\mathbf{I}|$ and compares it versus '0'. If this current is negative the vigilance criterion is not satisfied and signal c_j is activated preventing current T_j from competing in the WTA.

Once a winning node ($y_J = 1$) is stable, signal "LEARN" is activated and weights $\mathbf{z}_J$ are updated changing its stored value to $I_i z_{ij}$.

III. Yield and Area Optimization

To obtain good system precision it is important to make all L_A and L_B synapse current sources to match within a certain limit. In our first prototype, a tree-like current-mirror structure was implemented to generate all L_A and L_B currents from two external current references. The external current references enter to a multiple-output current mirror which delivers several output currents which enter as inputs to another stage of multiple-output current-mirrors. Each multiple-output current mirror has at the most ten outputs and is laid out using common centroid techniques to reduce the gradient-induced mismatch. After a few stages several thousands of L_A and L_B currents are available which match with a precision better than 1% for currents levels higher than $5\mu A$. However, this structure is very area consuming, which results in a very poor yield. That prototype had a die area of $1cm^2$ while having a 100-node F_1 layer and an 18-node F_2 layer and exhibits a yield performance of 6%.

Our results of mismatch characterization showed that is was possible to eliminate the tree-like current mirror structure while maintaining a current precision better than 1%. A new ART1 prototype was designed with an area 15 times less than that of the first prototype and a 98% yield performance. This prototype chip occupies an area of $2.5mm \times 2.2mm$ having a 50-node F_1 layer and a 10-node F_2 layer.

For the mismatch characterization, a special purpose chip in the ES2-1.0μm technology was designed [4]. The chip contains a matrix of cells, each of them

containing different sized PMOS and NMOS transistors, plus decoding circuitry. A simplified diagram of the chip and the experimental set-up to measure the transistors is depicted in Fig. 3. All NMOS transistors in the chip have their sources connected together to pin S. All NMOS transistors share their drains at pin DN and all PMOS transistors have their drains connected to pin DP. Every transistor in the chip has its gate short-circuited to its source except for one pair of NMOS and PMOS transistors. The selected pair transistors have their gate connected to pin G. A host computer controls the selection decoder and a curve tracer (HP4145). If pin DP is left unconnected and the curve tracer is connected to pins S, DN and G each NMOS transistor can be separately characterized. In a similar way, if pin DN is left unconnected, each PMOS transistor can be measured by connecting the curve tracer to pins S, DP and G.

NMOS and PMOS transistors of size $10\mu m \times 10\mu m$ spread over an area of $2.5mm \times 2.5mm$ were forced to the same V_{GS} and V_{DS} voltages so that their nominal current was around $10\mu A$. The effective measured currents flowing through the transistors are depicted in Fig. 4. Fig. 4(a) shows the currents flowing through the NMOS transistors as a function of the transistor position in the array. Fig. 4(b) depicts the same but for the PMOS transistors. As can be seen, each surface $I_o(x,y)$ has two deviation components: a long-distance gradient component, and a short-distance noise component. For each surface, the plane $I_o^p(x,y) = Ax + By + C$ that best fits the points of the measured surface $I_o(x,y)$ is computed. Afterwards, the standard deviation $\sigma(\Delta I_o)$ of the difference

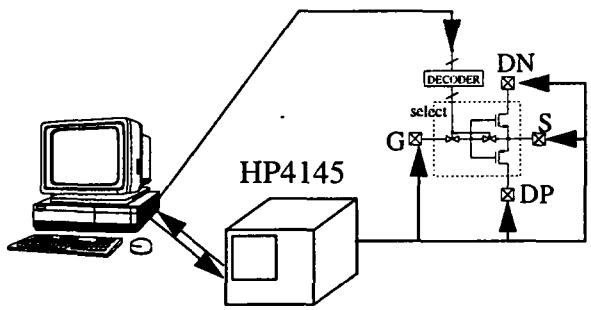

Fig. 3: Experimental Set-Up for Transistor Mismatch Characterization

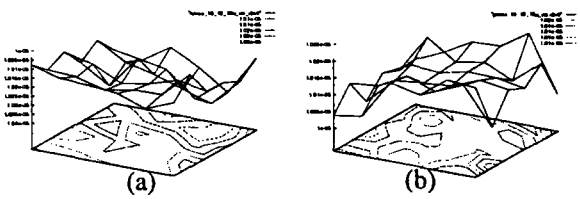

Fig. 4: Measured Currents for an Array of MOS Transistors with the same V_{GS} and V_{DS} values, (a) NMOS Transistors, and (b) PMOS Transistors

$$\Delta I_o(x,y) = I_o(x,y) - I_o^p(x,y) \quad (2)$$

is computed. This deviation is due to the noise component of surface $I_o(x,y)$. The gradient component is defined by plane $I_o^p(x,y)$. The maximum deviation due to the gradient component is given by

$$\Delta I_o^p = max\{I_o^p(x,y)\} - min\{I_o^p(x,y)\}. \quad (3)$$

On the other hand, for the noise component, 98% of the points remain within the $\pm 3\sigma(\Delta I_o)$. Consequently, let us define the ratio between noise component and gradient component contributions as

$$r = \frac{6\sigma(\Delta I_o)}{\Delta I_o^p}. \quad (4)$$

Eight chips could be fully characterized. Each chip contains several arrays of NMOS and PMOS transistors of different sizes spread over an area of $2.5mm \times 2.5mm$. Table I shows the results for NMOS transitors of size $10\mu m \times 10\mu m$ driving a nominal current of $10\mu A$. The table shows the noise error component $\sigma(\Delta I_o)$, the gradient error component ΔI_o^p, the ratio r, and the total error component $\sigma_T(I_o)$ (gradient+noise). Table II contains the same information but for PMOS transistors. As can be seen, for this chip dimensions, this current level and transistor geometries the noise error contribution is of the same order or higher than the gradient error contribution, and the total current error $\sigma_T(I_o)$ is always less than 1%. Consequently, for these conditions it is possible to avoid the use of high area consuming tree-like mirror structures and directly implement a simple current mirror with all the needed outputs. This is the approach used in the present ART1

chip	$\sigma(\Delta I_o)$ (%)	ΔI_o^p (%)	r	$\sigma_T(I_o)$(%)
1	0.57	1.30	2.652	0.67
2	0.62	1.98	1.874	0.83
3	0.47	3.10	0.921	0.79
4	0.52	0.90	3.456	0.56
5	0.54	1.65	1.959	0.64
6	0.58	3.01	1.160	0.88
7	0.65	1.96	1.996	0.82
8	0.73	2.15	2.027	0.90

Table I: Output current error in an NMOS array

chip	$\sigma(\Delta I_o)$ (%)	ΔI_o^p(%)	r	$\sigma_T(I_o)$(%)
1	0.58	1.53	2.278	0.67
2	0.47	0.74	3.830	0.51
3	0.48	0.83	3.519	0.51
4	0.40	2.18	1.100	0.63
5	0.46	0.60	4.666	0.49
6	0.45	2.18	1.236	0.72
7	0.44	0.83	3.171	0.50
8	0.41	1.28	1.926	0.50

Table II: Output current error in a PMOS array

chip prototype. The chip has a die area of $2.5mm \times 2.2mm$, and contains an array of 50×10 synapses. Fig. 5 depicts the measured L_B output currents as a function of the output transistor position in one chip for an input current level of $10\mu A$. Table III contains the deviation components measured in the L_B output currents. The random component is always higher than the gradient component and the total deviation is less than 1%. Similar results are obtained for the two L_A current sources.

IV. Experimental Results

All ten fabricated chip samples were fully operational and for none of them we were able to detect any fault in its subcircuits. All system components could be isolated and independently characterized. The circuit performances of the different subcircuits were similar to those of the first prototype [2].

Although the chip is analog in nature, its inputs and outputs are digital. Therefore, it is possible to test the system level behavior using a digital test equipment. We used the test equipment HP82000 to fully test the system level operation. The system proved to be very robust and therefore a multichip system was easy to assemble. The operation of two multichip systems was also tested: a two-chip ART1 system and a three chip system forming an ARTMAP architecture.

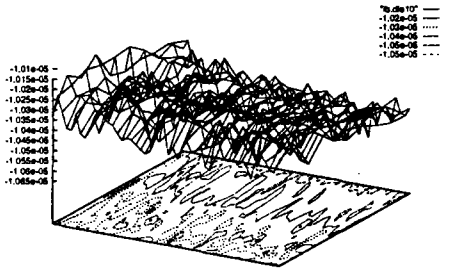

Fig. 5: Measured L_B Current in an ART1 Prototype Chip

chip	$\sigma(\Delta L_B)$ (%)	ΔL_B^p (%)	r_{L_B}	$\sigma_T(L_B)$(%)
1	0.62	0.62	6.076	0.64
2	0.59	0.22	16.497	0.60
3	0.56	3.330	1.015	0.89
4	0.63	0.90	4.196	0.64
5	0.65	1.83	2.118	0.76
6	0.64	1.49	2.565	0.73
7	0.60	1.58	2.255	0.67
8	0.62	1.48	2.524	0.71
9	0.63	0.37	10.080	0.63
10	0.57	2.16	1.573	0.73

Table III: Measured L_B output current error

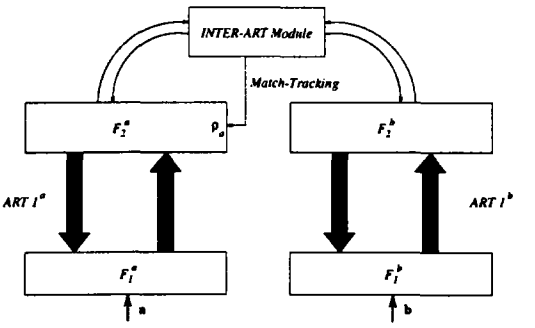

Fig. 6: Diagram of an ARTMAP Architecture

The two-chip ART1 system consists of two horizontally assembled ART1 chips. The resulting system is able to cope with 100-bit input patterns.

The ARTMAP architecture consists of two ART1 subsystems connected through an *Inter-ART* module as depicted in Fig. 6, where **a** is an N_a-dimensional input vector to the first subsystem ART1^a, and **b** an N_b-dimensional input vector for the second ART1^b subsystem. An ARTMAP system is a supervised learning neural network that learns the correspondence between two simultaneous input patterns **a** and **b**. The *Inter-ART* module is simply an $M_a \times M_b$ array of binary weights which learns the correspondence between the ART1^a category which classifies pattern **a** and the ART1^b category which classifies pattern **b**. An ARTMAP hardware system was assembled using two ART1 chips and an extra chip for the *Inter-ARTMAP* module. The system level operation of the ARTMAP hardware system has also been tested using the HP82000 digital test equipment.

V. References

[1] G. A. Carpenter and S. Grossberg, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, pp. 54-115, 1987.

[2] T. Serrano-Gotarredona and B. Linares-Barranco, "A Real-Time Clustering Microchip Neural Engine," *IEEE Transactions on VLSI Systems*, 1996.

[3] T. Serrano-Gotarredona and B. Linares-Barranco, "A Modified ART1 Algorithm more suitable for VLSI Implementations," *Neural Networks*, 1996.

[4] T. Serrano-Gotarredona and B. Linares-Barranco, "Systematic CMOS Mismatch Characterization," *Proceedings of the IEEE Int. Symposium on Circuits and Systems*, 1996.

[5] G. A. Carpenter, S. Grossberg, and J. H. Reynolds, "ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network," *Neural Networks*, vol. 4, pp. 565-588, 1991.