An ART1 Microchip and Its Use in Multi-ART1 Systems

Teresa Serrano-Gotarredona and Bernabé Linares-Barranco

***Abstract*—Recently, a real-time clustering microchip neural engine based on the ART1 architecture has been reported. Such chip is able to cluster 100-b patterns into up to 18 categories at a speed of 1.8 μs per pattern. However, that chip rendered an extremely high silicon area consumption of 1 cm^2, and consequently an extremely low yield of 6%. Redundant circuit techniques can be introduced to improve yield performance at the cost of further increasing chip size. In this paper we present an improved ART1 chip prototype based on a different approach to implement the most area consuming circuit elements of the first prototype: an array of several thousand current sources which have to match within a precision of around 1%. Such achievement was possible after a careful transistor mismatch characterization of the fabrication process (ES2-1.0 μm CMOS). A new prototype chip has been fabricated which can cluster 50-b input patterns into up to ten categories. The chip has 15 times less area, shows a yield performance of 98%, and presents the same precision and speed than the previous prototype. Due to its higher robustness multichip systems are easily assembled. As a demonstration we show results of a two-chip ART1 system, and of an ARTMAP system made of two ART1 chips and an extra interfacing chip.**

***Index Terms*—Adaptive resonance theory, analog circuits, analog conputers, analog integrated circuits, analog processing circuits, analog systems, ART neural networks, circuits, clustering methods, CMOS integrated circuits, CMOS memory integrated circuits, integrated circuit design, large-scale integration, learning systems, neural-network hardware, nonlinear circuits, real-time systems.**

I. Introduction

SINCE the invention of the ART1 architecture in 1987 [2] many high-level neural processing systems have been developed [3] which are based on the ART1 or more evolved but similar architectures [4]–[8]. These high-level neural systems have internal complex structures, but many times they are based on a small number of ART-like building blocks. When these high-level neural systems have to be used in real-world applications, portable equipments, robots, industrial control applications, etc., it is not always possible to rely on software programs running on expensive workstations. In such cases it is mandatory to build a piece of hardware that realizes physically the neural processing system. The availability of ART-like modular chips would significantly boost the proliferation of ART-based neural hardware systems. Due to the inherent internal hierarchy of ART-based neural systems their hardware realization would be significantly simplified if robust and low-cost ART-like chip modules would be readily available.

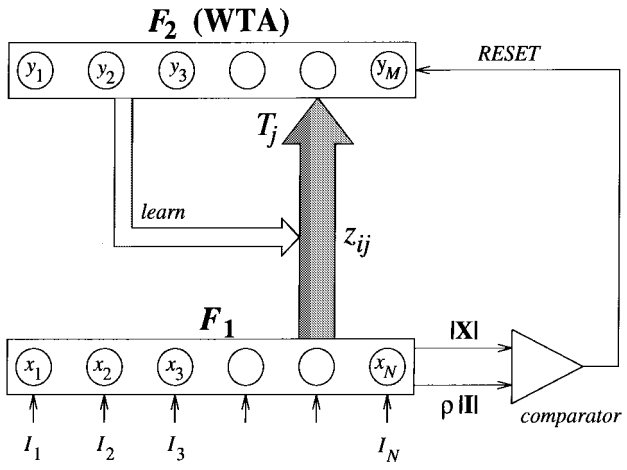

Fig. 1. (a) ART1 architecture diagram. (b) Algorithmic operation description of VLSI-friendly fast-learning ART1 system.

Although some preliminary work was done to build ART-based hardware prototypes [9], [10], it is not until recently that a fully functional reasonable size real-time clustering microchip neural engine based on the ART1 architecture has been reported [1]. It is based on a slightly modified version of the ART1 algorithm which was shown to preserve all its

Manuscript received June 25, 1996; revised April 2, 1997.

The authors are with National Microelectronics Center (CNM), 41012 Sevilla, Spain.

Publisher Item Identifier S 1045-9227(97)05251-X.

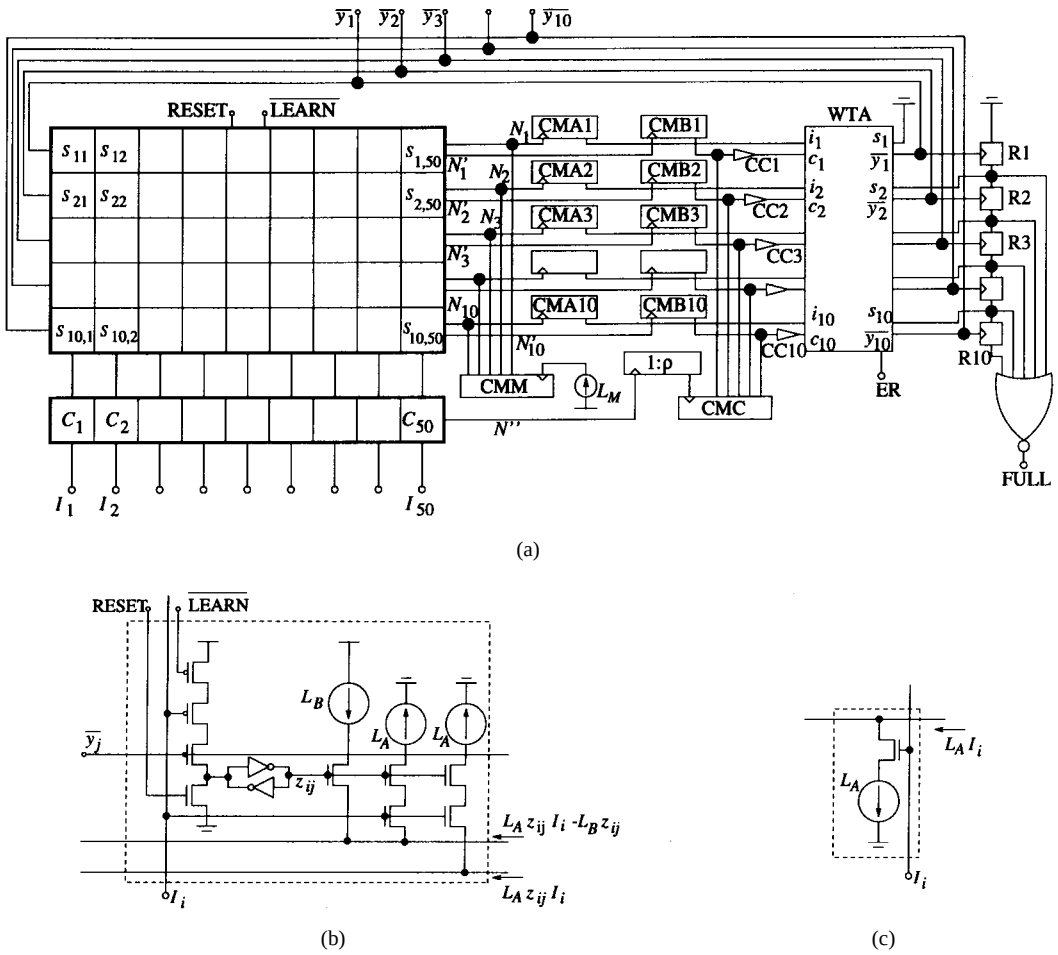

Fig. 2. (a) Circuit diagram of current-mode ART1 chip. (b) Detail of synapse S_{ij}. (c) Detail of controlled current source C_i.

original computational properties [11], 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 μs. 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 (1 cm^2), and consequently its yield[1] 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 [1].

A straightforward solution to the yield problem is to include extra redundant circuitry in the chip together with some self-testing subsystems that would identify and disconnect faulty subcells. This method is used intensively in large-area high-density commercial DRAM chips. However, this redundancy-based yield enhancement technique increases silicon area, requires more processing circuitry and increases design effort and cost [13]. In this paper a new ART1 chip is presented which solves the yield problem using a different approach: area reduction. After careful MOS transistor electrical parameter mismatch characterization of the technological process to be used, we were able to identify the maximum chip area for which the parameter variations would remain within the necessary limits to preserve the required system operation precision. We concluded that for the ES2-1.0 μm CMOS process, for transistors of size $W = L = 10\,\mu$m spread over a die area of the order of 2.5 $\times$ 2.5 mm, and for current levels around 10 μA, the standard deviation of transistor current mismatch is of the order of $\sigma(I) \approx 1\%$. Taking this into account we were able to design and fabricate an ART1 chip capable of clustering 50-b input patterns into up to ten 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 which enabled us to implement some multichip ART1 systems. As an illustration we will show results of a two-chip ART1 system and of a three-chip ARTMAP system.

This paper is structured as follows. In the next section the VLSI-friendly ART1 algorithm employed is reviewed as well as the circuit design that maps it into hardware. In Section III we show why the first prototype has a very high area consumption, how we performed a careful technology current mismatch characterization, and how we modified the circuit to drastically reduce its area, while maintaining system precision and speed

[1] Percentage of fault-free chips over total number of fabricated chips.

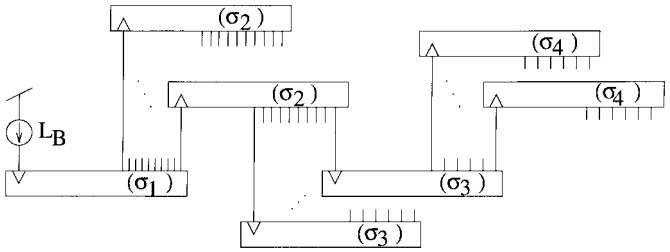

Fig. 3. Tree-like current-mirror structure for generating a large number of matched current sources.

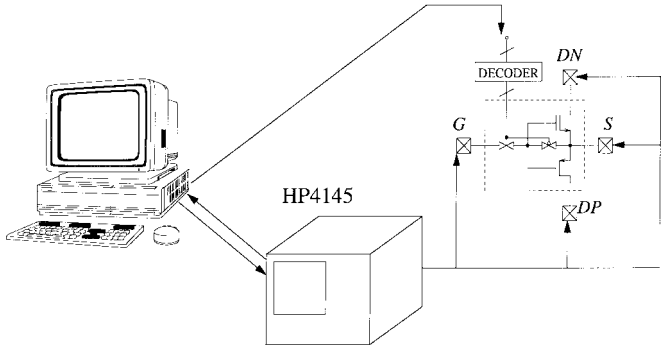

Fig. 4. Simplified diagram of mismatch characterization chip and experimental setup.

performance. In Section IV we will provide measured experimental results of the new ART1 chip and of a two-chip ART1 system. Section V describes how to assemble an ARTMAP system and provides measured experimental results as well. Finally, we conclude in Section VI.

II. VLSI-Friendly ART1 Algorithm and Circuit Implementation

An ART1 system is a self-organizing neural associative memory capable of generating in an unsupervised way stable recognition codes in response to a series of arbitrarily many, arbitrarily ordered, and arbitrarily complex binary input patterns. As shown in Fig. 1(a) the ART1 architecture consists of two layers. The bottom layer F_1 has N nodes each of which receives the ith binary pixel I_i of the external input pattern $\mathbf{I} \equiv (I_i, \cdots, I_N)$. The top layer F_2 has M nodes, each of which represents a learned category or cluster of input patterns $y_j (j = 1, \cdots, M)$. Each F_1 layer node i connects to all F_2 layer nodes through binary weights z_{ij} which can be either "0" or "1." Each F_2 layer category j is characterized by the set of weights $\mathbf{z}_j = (z_{1j}, \cdots, z_{Nj})$ that connects to it. Every time an input pattern $\mathbf{I}$ is presented to the input layer F_1 an internal search process starts which, when finished, results in activating a single F_2 layer category. This category is the one that best represents the input pattern according to the value of a vigilance parameter ρ which can be tuned within the interval [0,1]. For small ρ values many patterns will be clustered into the same category, while for high ρ values only very similar patterns will be considered to belong to the same category. In the original ART1 paper by Carpenter and Grossberg [2] the operation of the system

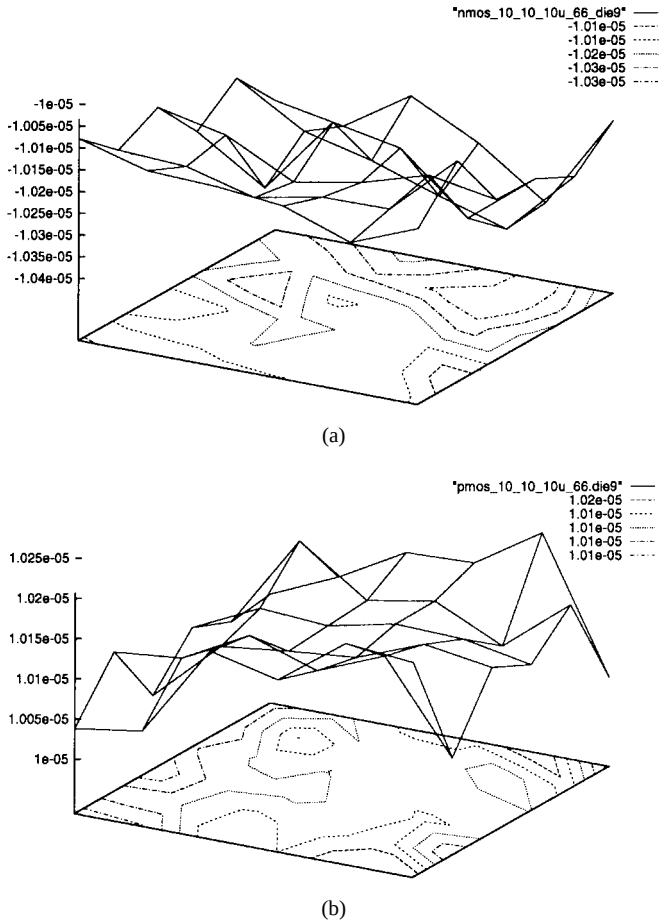

Fig. 5. Measured current for an array of MOS transistors with the same V_{GS} and V_{DS} voltages (for a nominal current of 10 μA), spread over a die area of 2.5 × 2.5 mm. (a) Array of NMOS transistors and (b) array of PMOS transistors.

was described by sets of nonlinear differential equations. It was also mentioned that the operation of the system could be described by an algorithmic flow diagram which basically describes the steady state of the differential equations. This algorithmic description was named as the *fast-Learning* mode of operation. Fig. 1(b) shows a modified version of the original *fast-Learning* ART1 operation which has a higher potential for VLSI circuit implementations. It has been shown that this algorithm preserves all the original computational properties of an ART1 system [11]. The operations to be performed are the following.

1) Reset all binary weights $z_{ij} = 1$.
2) Read a binary input vector $\mathbf{I} \equiv (I_1, \cdots, I_N)$.
3) Compute a set of analog "*choice functions*" or distances

$$T_j = L_A \sum_{i=1}^{N} I_i z_{ij} - L_B \sum_{i=1}^{N} z_{ij} + L_M \qquad (1)$$

or in vector notation[2]

$$T_j = L_A |\mathbf{I} \cap \mathbf{z}_j| - L_B |\mathbf{z}_j| + L_M. \qquad (2)$$

[2]Given a vector $\mathbf{a} \equiv (a_1, \cdots, a_N)$, the notation $|\mathbf{a}|$ represents its ℓ_1 norm $|\mathbf{a}| = \sum_{i=1}^{N} |a_i|$, and the intersection operator between two vectors represents the component-wise logical AND operation.

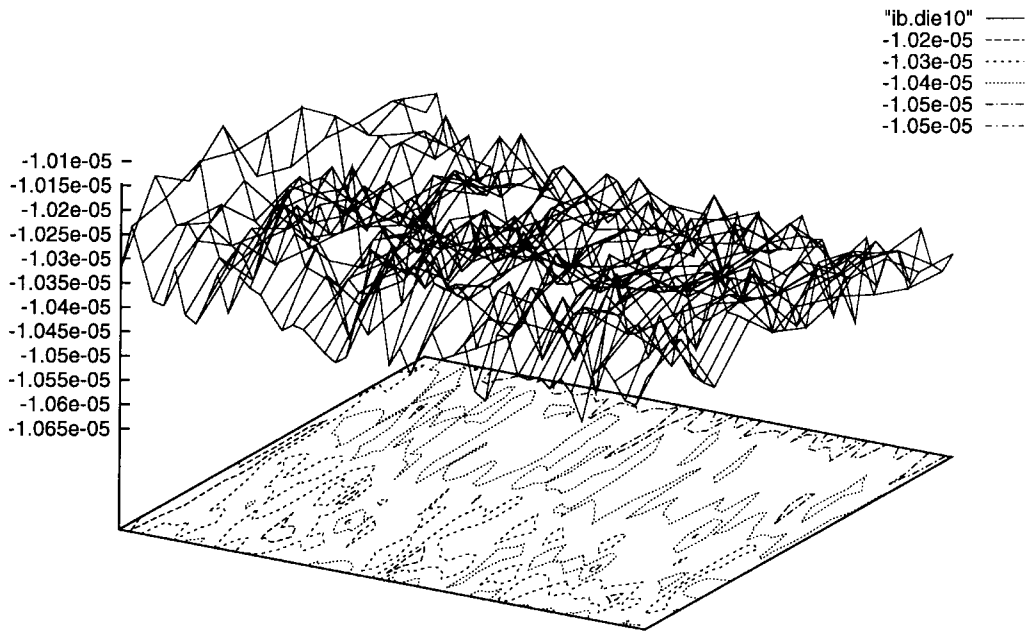

Fig. 6. Measured currents of the L_B array for the new ART1 chip prototype.

4) Select the maximum among all $\{T_j\}$. If T_J is this maximum then the Jth F_2 node is set to $y_J = 1$ while all others are set to $y_{j\neq J} = 0$. Hence, layer F_2 acts as a winner-take-all (WTA).
5) Check the vigilance criterion: if $\rho|\mathbf{I}| > |\mathbf{I} \cap \mathbf{z}_J|$ the criterion is not satisfied. In such case, force $T_J = 0$ and return to Step 4). Otherwise, the criterion is satisfied and the weights $\mathbf{z}_J$ must be updated to incorporate the characteristics of pattern $\mathbf{I}$ into category J

$$z_{Ji}(new) = I_i z_{Ji}(old) \tag{3}$$

or in vector notation

$$\mathbf{z}_J(new) = \mathbf{I} \cap \mathbf{z}_J(old). \tag{4}$$

The way this algorithm can be implemented in a parallel analog current-mode processing circuit is depicted in Fig. 2(a). It consists of a 10×50 array of synapses S_{ij}, a 1×50 array of controlled current sources C_i, two 1×10 arrays of unity-gain current mirrors CMAj, CMBj, a 1×10 array of current comparators CCj, a ten-input WTA circuit, two unity-gain current mirrors CMM and CMC, and an adjustable-gain ($0 \leq \rho \leq 1$) current mirror. Registers R_j and the NOR gate are optional. The circuit diagram of a synapse S_{ij} is shown in Fig. 2(b). It contains three current sources, a latch, and a set of NMOS and PMOS transistors acting as switches. The state of the latch z_{ij} is set to "1" by activating the RESET signal prior to circuit operation, or is set to "0" during circuit operation if $\overline{LEARN} = 0, I_i = 0$, and $\overline{y}_j = 0$ simultaneously. The synapse generates two currents, one of value $L_A z_{ij} I_i$ which is drained from node N'_j in Fig. 2(a), and another of value $L_A z_{ij} I_i - L_B z_{ij}$ drained from node N_j. Nodes N_j and N'_j are shared by all synapses in the same row. Consequently, the total input current to mirror CMAj and injected to input i_j of the WTA is

$$\begin{aligned} T_j &= L_A \sum_{i=1}^{N} z_{ij} I_i - L_B \sum_{i=1}^{N} z_{ij} + L_M \\ &= L_A |\mathbf{I} \cap \mathbf{z}_j| - L_B |\mathbf{z}_j| + L_M. \end{aligned} \tag{5}$$

Note that current L_M is provided by current mirror CMM to all N_j nodes. Similarly, the total input current to each mirror CMBj is

$$L_A \sum_{i=1}^{N} z_{ij} I_i = L_A |\mathbf{I} \cap \mathbf{z}_j|. \tag{6}$$

Fig. 2(c) shows the circuitry for each cell C_i. This cell drains a current $L_A I_i$ from node N''. The total input current for the ρ-gain mirror is thus $L_A|\mathbf{I}|$. This current, amplified by a factor ρ, is replicated by mirror CMC and compared against each current $L_A|\mathbf{I} \cap \mathbf{z}_j|$ at each CCj current comparator. If

$$\rho L_A |\mathbf{I}| > L_A |\mathbf{I} \cap \mathbf{z}_j| \tag{7}$$

comparator CCj deactivates the WTA input current T_j, by making the WTA control input $c_j = 1$. This way, current T_j will not compete in the WTA. Consequently, only the currents T_j that meet the vigilance criterion (7) will compete. The maximum among these T_j currents, let us call it T_J, will make $\overline{y}_J = 0$ while the rest become $\overline{y_{j\neq J}} = 1$. Once a single winner y_J is active the $LEARN$ signal can be activated making

$$z_{Ji}(new) = I_i z_{Ji}(old). \tag{8}$$

An uncommitted F_2 node y_j is one that has not yet been selected as a winner. Such nodes have their initial weight values $z_{1j} = \cdots = z_{Nj} = 1$. Consequently, their corresponding row of synapses will generate the same current

$$T_j|_{\text{uncommitted}} = L_A |\mathbf{I}| - L_B N. \tag{9}$$

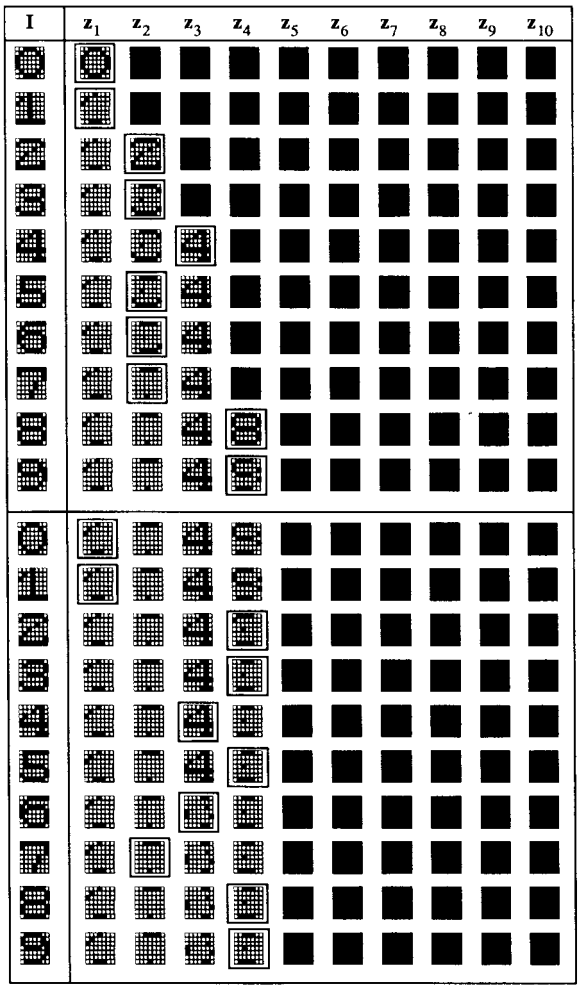

Fig. 7. Training sequence for a one-chip ART1 system with $\rho = 0.3$ and $\alpha = 1.1$.

The function of the shift register R_j is to enable only one uncommitted cell to compete for the winner. Every time an uncommitted cell wins, the shift register content is shifted one position and the next uncommitted F_2 node is enabled for WTA competition. The NOR gate signals that all F_2 nodes are already committed.

III. Yield and Area Optimization by Process Mismatch Characterization

From a system precision point of view it is important to make all L_A and L_B current sources to match within the required precision. When we designed our first ART1 prototype [1] we had no information concerning the long distance matching behavior of large arrays of current sources for the technology we were using. Therefore, we decided to use a mirror tree-like structure to generate all current sources from two external L_A and L_B current references. This approach is shown in Fig. 3. Each multiple-output current mirror had at the most ten outputs. Each current mirror was laid out using common centroid techniques, thus minimizing gradient-induced mismatch at the expense of increasing die area. If each current mirror of stage k introduces a mismatch error characterized by a standard deviation σ_k, the total error of a

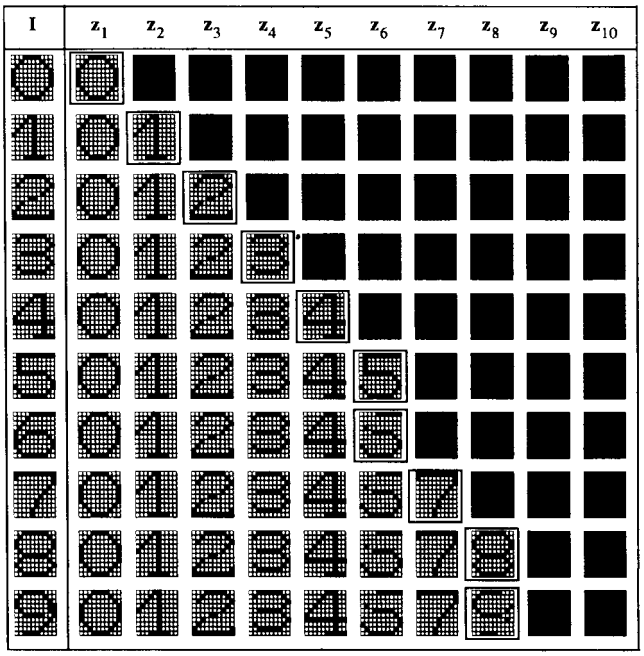

Fig. 8. Training sequence for a two-chip ART1 system with $\rho = 0.5$ and $\alpha = 2$.

q-stages cascade is given by

$$\sigma_{Total}^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_q^2. \tag{10}$$

The last stage q is the most numerous and will occupy most of the area. Pure random mismatch is inversely proportional to transistor area and current level [12]. If we want to keep σ_{Total} around 1%, each stage q must have smaller errors. In our first ART1 prototype chip [1] most of the die area was spent by the qth stage of common centroid low-mismatch multiple-outputs current mirrors. The resulting ART1 chip had a die area of 1 cm^2 while having a 100-node F_1 layer and an 18-node F_2 layer.

The yield performance of a microchip has the following approximate dependence on die area Ω:

$$yield(\%) = 100e^{\rho_D \Omega} \tag{11}$$

where, for this technology the estimated average defect density is $\rho_D \approx 3.2$ cm^{-1}. For $\Omega = 1$ cm^2 yield results to be around 6%.[3] Although most of the faulty chips rendered satisfactory clustering behavior [1] we decided to increase yield by reducing die area. In order to keep the system precision around 1% without using a large-area-consuming tree-like mirror structure a careful long distance mismatch characterization of the technological process to be used was necessary.

A special purpose chip was designed in the ES2-1.0 μm CMOS technology to estimate the matching behavior of large transistor arrays, for different transistor sizes. The chip contains a matrix of cells, each of which has several NMOS and PMOS transistors of several sizes, plus a transistor selection circuitry. Fig. 4 shows schematically the chip together with an experimental set-up to measure all transistors. In the chip all

[3] The chip pad ring area is not included in the yield computation.

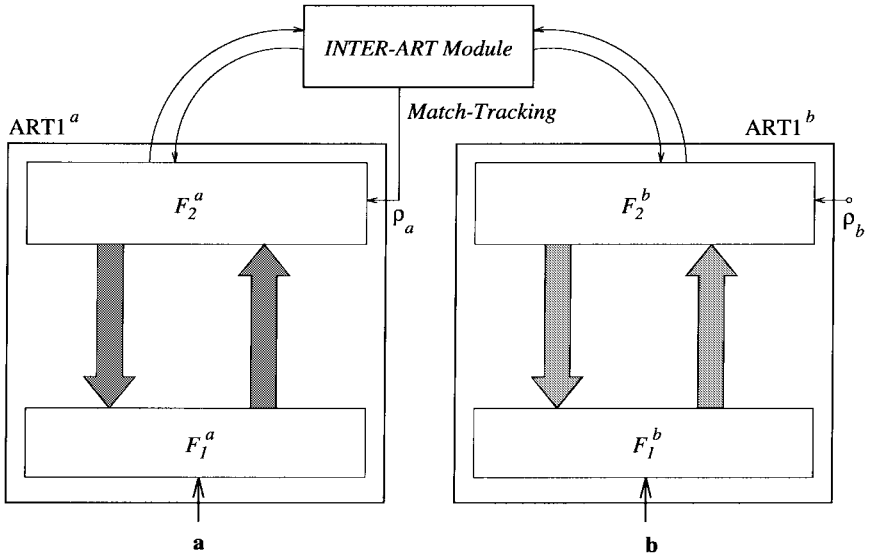

Fig. 9. ARTMAP architecture.

NMOS and PMOS transistors have their sources connected to pin S, all NMOS transistors have their drains connected to pin DN, all PMOS transistors have their drains connected to pin DP, all transistors have their gates short-circuited to their sources, except for one pair of NMOS and PMOS transistors. This pair has their gates connected to the external pin G. A digital bus and internal decoding circuitry selects one pair among all. By connecting a curve tracing instrument (in our case, the HP4145) to pins S, G, and DN the selected NMOS transistor can be accessed and characterized, while by using pins S, G, and DP the selected PMOS transistor can be measured. This technique has been used to characterize the mismatch behavior of several technological processes [14].

For transistors of size 10 μm $\times$ 10 μm spread over a chip area of 2.5 mm $\times$ 2.5 mm, biased by the same gate-to-source V_{GS} and drain-to-source V_{DS} voltages so that their nominal current was around 10 μA, we measured the current spreads depicted in Fig. 5. Fig. 5(a) shows, as a function of transistor position, the current measured for each transistor of an NMOS array. Fig. 5(b) shows the same for a PMOS array. As can be seen, the surfaces present a long distance gradient component and a short distance noise component. Let us call the measured currents surface $I_o(x,y)$. For this surface we can compute the best fit plane $I_o^p(x, y) = Ax + By + C$. Then, for each point (x, y) we can define

$$\Delta I_o(x, y) = I_o(x, y) - I_o^p(x, y). \tag{12}$$

By computing the standard deviation of $\Delta I_o(x, y), \sigma(\Delta I_o)$, we are extracting 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

$$\Delta I_o^p = \max\{I_o^p(x, y)\} - \min\{I_o^p(x, y)\}. \tag{13}$$

On the other hand, for the noise component, 98% of the points remain within the $\pm 3\sigma(\Delta I_o)$ interval. Consequently, let us define the maximum deviation due to the noise component as $6\sigma(\Delta I_o)$. Let us now define

$$r = \frac{6\sigma(\Delta I_o)}{\Delta I_o^p} \tag{14}$$

as the ratio between noise component and gradient component contributions. Table I shows these ratios measured for NMOS and PMOS transistors of size 10 μm $\times$ 10 μm, driving nominal currents of 10 μA and for different chips. Also shown in Table I are the standard deviations of the noise component $\sigma(\Delta I_o)$, the maximum deviation of the gradient component ΔI_o^p, and the total standard deviation of transistor currents $\sigma_T(I_o)$, computed as

$$\sigma_T(I_o) = \sqrt{\overline{I_o^2} - \overline{I_o}^2}. \tag{15}$$

The current mirror tree-like structure of Fig. 3 was intended to suppress the gradient component of a 1 cm^2 chip. The noise component can only be reduced by increasing transistor area [12]. Table I reveals that for die areas of 2.5 mm $\times$ 2.5 mm, transistor sizes of 10 μm $\times$ 10 μm, and nominal currents of 10 μA, the contribution of noise component is equal or higher than the gradient component, while the standard deviation of current mismatch $\sigma_T(I_o)$ is kept below 1%. Consequently, for these dimensions we can avoid the use of high area consuming

TABLE I
CURRENT MISMATCH COMPONENTS FOR TRANSISTOR ARRAYS WITH 10 μA NOMINAL CURRENT, 10μm $\times$ 10 μm TRANSISTOR SIZE, AND 2.5 $\times$ 2.5 mm DIE AREA FOR THE ES2-1.0 μm CMOS PROCESS

	NMOS				PMOS			
chip	$\sigma(\Delta I_o)$ (%)	ΔI_o^p(%)	r	$\sigma_T(I_o)$ (%)	$\sigma(\Delta I_o)$ (%)	ΔI_o^p(%)	r	$\sigma_T(I_o)$ (%)
1	0.57	1.30	2.652	0.67	0.58	1.53	2.278	0.67
2	0.62	1.98	1.874	0.83	0.47	0.74	3.830	0.51
3	0.47	3.09	0.921	0.79	0.48	0.82	3.519	0.51
4	0.52	0.90	3.456	0.56	0.40	2.18	1.100	0.63
5	0.54	1.65	1.959	0.64	0.46	0.60	4.666	0.49
6	0.58	3.01	1.160	0.88	0.45	2.18	1.236	0.72
7	0.65	1.96	1.996	0.82	0.44	0.83	3.171	0.50
8	0.73	2.15	2.027	0.90	0.41	1.28	1.926	0.50

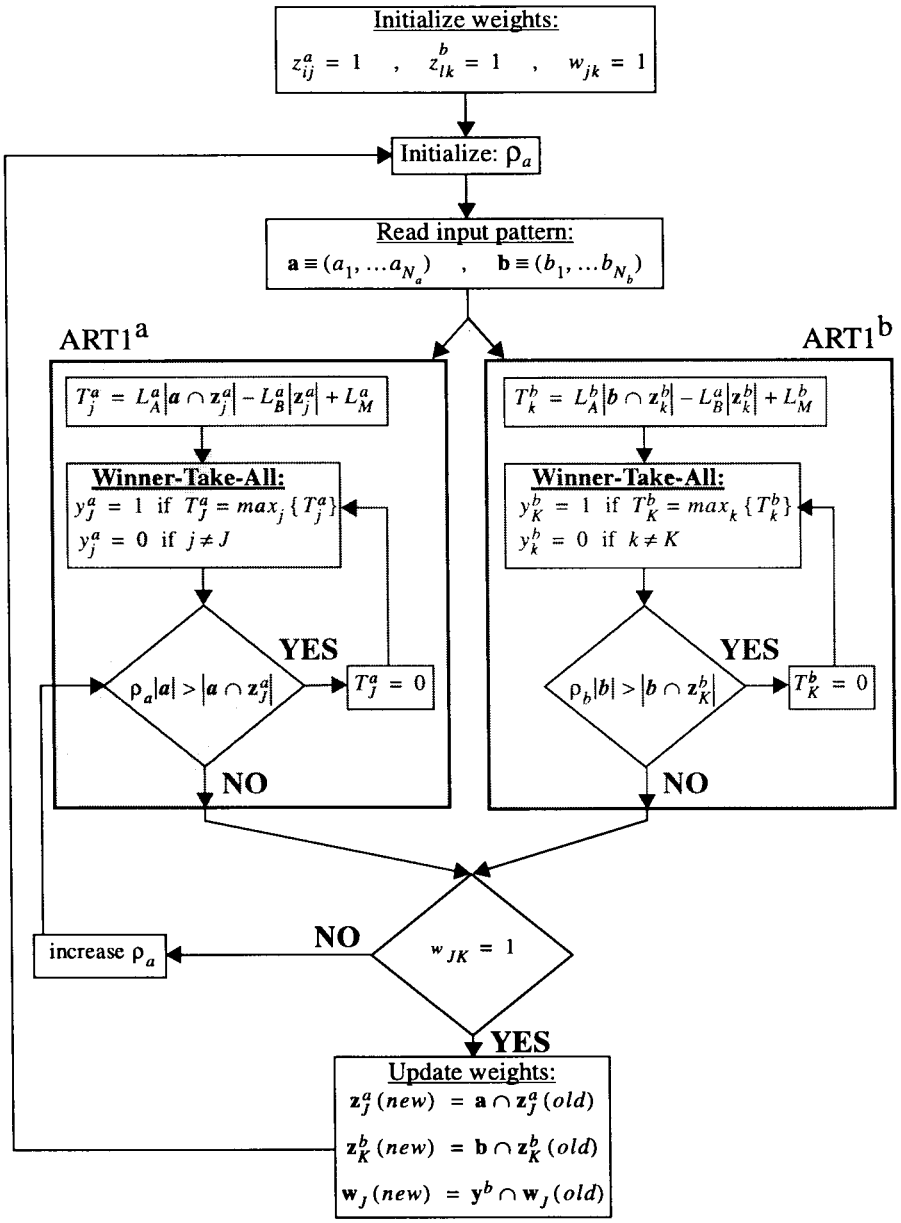

Fig. 10. Flow diagram of ARTMAP training mode operation.

circuit structures (like common centroid mirrors arranged in a tree-like fashion) to eliminate the gradient component, and directly implement a single current mirror with all the outputs needed. This is the approach we used in the present ART1 chip prototype. This chip has a die area of 2.5 mm $\times$ 2.2 mm, and contains an array of 50 $\times$ 10 synapses, each synapse with two L_A and one L_B current sources. The current sources transistors are of size 10 μm $\times$ 10 μm and drive a nominal current of 10 μA. Fig. 6 shows the measured currents of the L_B array. Table II shows the measured values of the mismatch components of the L_A and L_B current sources arrays for all fabricated chips. Note that the total current mismatch standard deviation is less than 1% for all chips.

Due to the much smaller chip area its fabrication cost is much less and its yield performance is significantly higher: 98% by applying (11). In the next sections single chip operation experimental results are described as well as results for systems assembled with several ART1 chips.

IV. Experimental Results of ART1 Systems

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 [1], and consequently their characteristics will not be repeated in this paper. Here we will only provide some illustrative examples on system level behavior.

Although the chip is analog in nature, its inputs and outputs are digital. Therefore, it is possible to test its system level behavior using a digital test equipment (in our case, the HP82000). This equipment applies digital input vectors ($\mathbf{I}$),

TABLE II
MEASURED MISMATCH COMPONENTS FOR THE FABRICATED ART1 CHIP PROTOTYPES

	L_A current sources				L_B current sources			
chip	$\sigma(\Delta I_o)$ (%)	ΔI_o^o(%)	r	$\sigma_T(I_o)$ (%)	$\sigma(\Delta I_o)$ (%)	ΔI_o^o(%)	r	$\sigma_T(I_o)$ (%)
1	0.63	1.39	2.694	0.71	0.62	0.62	6.076	0.64
2	0.51	0.68	5.311	0.62	0.59	0.22	16.497	0.60
3	0.68	1.91	2.128	0.81	0.56	3.32	1.015	0.89
4	0.59	0.28	12.607	0.61	0.63	0.90	4.196	0.64
5	0.64	1.27	3.204	0.69	0.65	1.83	2.118	0.76
6	0.65	1.29	3.028	0.71	0.64	1.49	2.565	0.73
7	0.66	0.41	9.535	0.68	0.60	1.58	2.255	0.67
8	0.64	0.92	4.174	0.67	0.62	1.48	2.524	0.71
9	0.79	2.20	2.157	0.91	0.63	0.37	10.080	0.63
10	0.74	0.43	10.368	0.75	0.57	2.16	1.573	0.73

reads digital output vectors ($\mathbf{y}$), and reads the internal weights (z_{ij}) at each processing step. Three external reference currents need to be supplied to the chip: L_A, L_B, and L_M. Current L_M [see Fig. 2(a)] is needed to assure that all currents T_j reaching the WTA are positive. The ART1 system behavior is controlled by two externally adjustable parameters, ρ and α. ρ is the gain of a current mirror and is adjusted through a digital word applied externally [1], while $\alpha = L_A/L_B$ is controlled by appropriately setting currents L_A and L_B.

To test the system behavior it was trained with a set of ten $7 \times 7 = 49$-b input patterns. Each pattern represents each of the ten digits from "0" to "9." The last input pixel was always set to zero and it is not shown in the figures. The classification of the set of input patterns was repeated for different values of the vigilance parameter and several values of parameter $\alpha = L_A/L_B$.

Fig. 7 shows the training sequence for $\rho = 0.3$ and $\alpha = 1.1$. The first column represents the input pattern applied to the system. The remaining ten columns correspond to the weights $\mathbf{z}_j$ stored in each category when the input pattern has been classified and learned. The boxed category is the winning category after the WTA competition. In this case, learning self-stabilizes after two input pattern presentations. That is, no modification of the winning category or the stored weights take place in subsequent presentations of the input pattern sequence. As shown in Fig. 7, the system has clustered all ten input patterns into four categories.

A two-chip ART1 system was assembled. In this case, the input patterns had $10 \times 10 = 100$ binary pixels. Fig. 8 depicts a training sequence performed on this system. The system classifies the ten input patterns into eight categories after a single presentation of the input pattern set. The sequence of Fig. 8 was obtained for a vigilance parameter of $\rho = 0.5$, and $\alpha = 2 (L_A = 10\,\mu\text{A}, L_B = 5\,\mu\text{A})$.

V. ASSEMBLING AN ARTMAP SYSTEM USING ART1 CHIP MODULES

An ARTMAP system [7] consists of two ART1 subsystems connected through an *Inter-ART* module, as depicted in Fig. 9. Let $\mathbf{a} \equiv \lfloor a_l, \cdots, a_{N_a} \rfloor$ be an N_a-dimensional input vector to the first ART1 subsystem ART1^a, and $\mathbf{b} \equiv \lfloor b_l, \cdots, b_{N_b} \rfloor$ an N_b-dimensional one for the second ART1^b subsystem. An

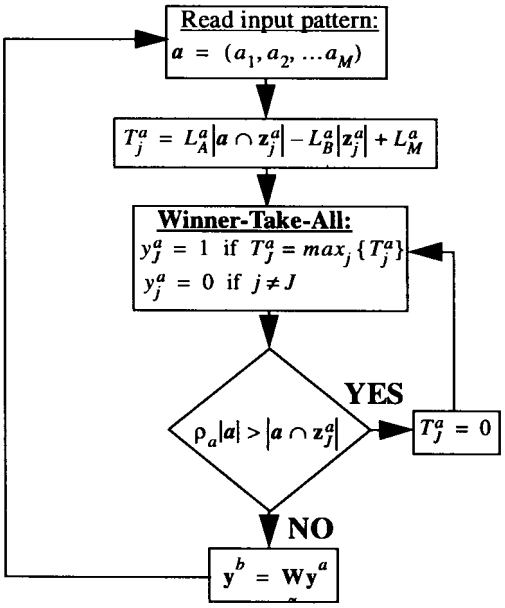

Fig. 11. Flow diagram of the prediction ARTMAP operation.

ARTMAP system is a supervised learning neural network that learns the correspondence between two simultaneous input patterns $\mathbf{a}$ and $\mathbf{b}$. Two modes of operation can be distinguished:

- **Training Mode**, during which pairs of input patterns ($\mathbf{a}$, $\mathbf{b}$) are provided, and the ARTMAP system learns their correspondence.
- **Prediction Mode**, during which only patterns $\mathbf{a}$ are provided to the first ART1^a subsystem, and ARTMAP predicts the corresponding ART1^b cluster.

Fig. 10 illustrates the algorithmic description of ARTMAP operation in training mode [7]. After reading two input vectors $\mathbf{a}$ and $\mathbf{b}$ each ART1 module selects an F_2 winning node (y_J^a for ART1^a and y_K^b for ART1^b) that meet their vigilance criteria. The inter-ART module, which is simply an $M_a \times M_b$ array of binary weights w_{jk} initially set to "1," learns the correspondence between the ART1^a winning category y_J^a and the ART1^b one y_K^b by making

$$w_{Jk} = \begin{cases} 1 & \text{if } k = K \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

or, in vector notation

$$\mathbf{w}_J(new) = \mathbf{y}^b \cap \mathbf{w}_J(old). \tag{17}$$

However, if ART1^a category y_J^a and ART1^b category y_K^b become simultaneously active and the Inter-ART weight w_{JK} has already been set to "0," this means that ART1^a category y_J has already been assigned to a different ART1^b category. In this case ART1^a vigilance parameter ρ_a is increased until y_J^a deactivates and a different ART1^a category is selected.

During the prediction mode of operation subsystem ART1^b does not receive any inputs. Only subsystem ART1^a receives external input patterns and selects a winning category y_J^a. ART1^b outputs, which are the outputs of the complete

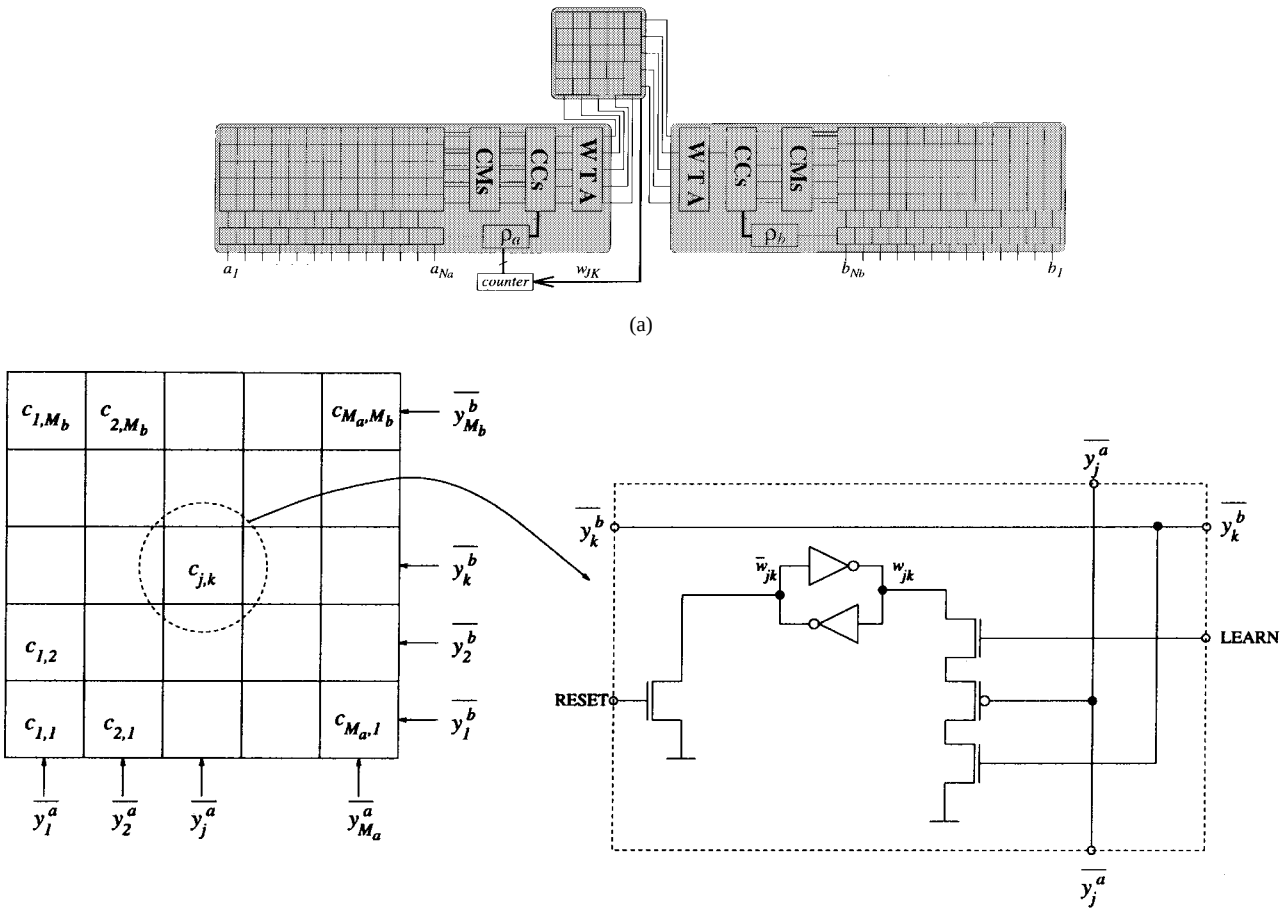

Fig. 12. (a) ARTMAP hardware assembly. (b) Diagram of Inter-ART chip. (c) Detail of Inter-ART chip cell.

ARTMAP system, are activated by the Inter-ART module

$$y_k^b = \sum_{j=1}^{M_a} w_{jk}\, y_j^a = w_{Jk}\, y_J^a, \quad k = 1, \cdots, M_b \tag{18}$$

or equivalently in matrix notation

$$\mathbf{y}^b = \mathbf{W}\mathbf{y}^a \tag{19}$$

where $\mathbf{W}$ is the weight matrix of the Inter-ART module.

According to the way the Inter-ART weights w_{jk} are set, for each ART1^a active category only one ART1^b category will be chosen, but an ART1^b category can be activated by more than one ART1^a cluster.[4] Fig. 11 shows the algorithmic flow diagram of the ARTMAP prediction mode operation.

An ARTMAP hardware system can be assembled using two ART1 chips and an extra chip for the Inter-ART module, as is shown in Fig. 12(a). The Inter-ART chip, shown in Fig. 12(b), is simply an array of cells c_{jk} whose simplified schematic is depicted in Fig. 12(c). Each c_{jk} cell has a latch which is set initially to "1" and changes to "0" if $y_j^a = 1, y_k^b = 0$, and the *LEARN* signal is high. Extra transistors, not shown in Fig. 12(c), are also included to read out the weight values. During training mode the value of weight w_{JK} is used to control a digital counter that increments the value of ρ_a. If w_{JK} the counter will increase its value until the ART1^a winning category changes and w_{JK} becomes "1." At this moment the counter stops and its content represents the appropriate value for ρ_a.

The system level operation of the ARTMAP hardware system has also been tested using the HP82000 digital test equipment. Fig. 13 shows a system training sequence. The first column, named $\mathbf{a}$, represents the input patterns applied to the ART1^a chip. The column named $\mathbf{b}$ represents the input patterns applied to the ART1^b chip. The columns named $\mathbf{z}_j^a$ and $\mathbf{z}_k^b$ represent the stored weights in the ART1^a and ART1^b modules after the classification and learning of each input pattern pair. The boxed categories are the ones that remain active after the search process has finished, and these are the only ones that are updated with learning. Below each ART1^a winning category the final value of the vigilance parameter ρ_a needed in the search process to choose this category is indicated (ρ_a was increased in steps of $\Delta\rho_a = 1/32$). The last column shows the stored weights in the inter-ART module which represent the learned correspondence between the ART1^a and ART1^b categories (index j is coded vertically from top to bottom, while index k is coded horizontally from

[4] This is true unless the input pattern $\mathbf{a}$ activates an uncommitted ART1^a F_2 node (pattern $\mathbf{a}$ is not recognized as belonging to any ART1^a category). In this case, $w_{jk} = 1\,\forall\, k$, and all ART1^b F_2 nodes would be activated, implying that the applied input pattern is not recognized as belonging to any of the learned categories.

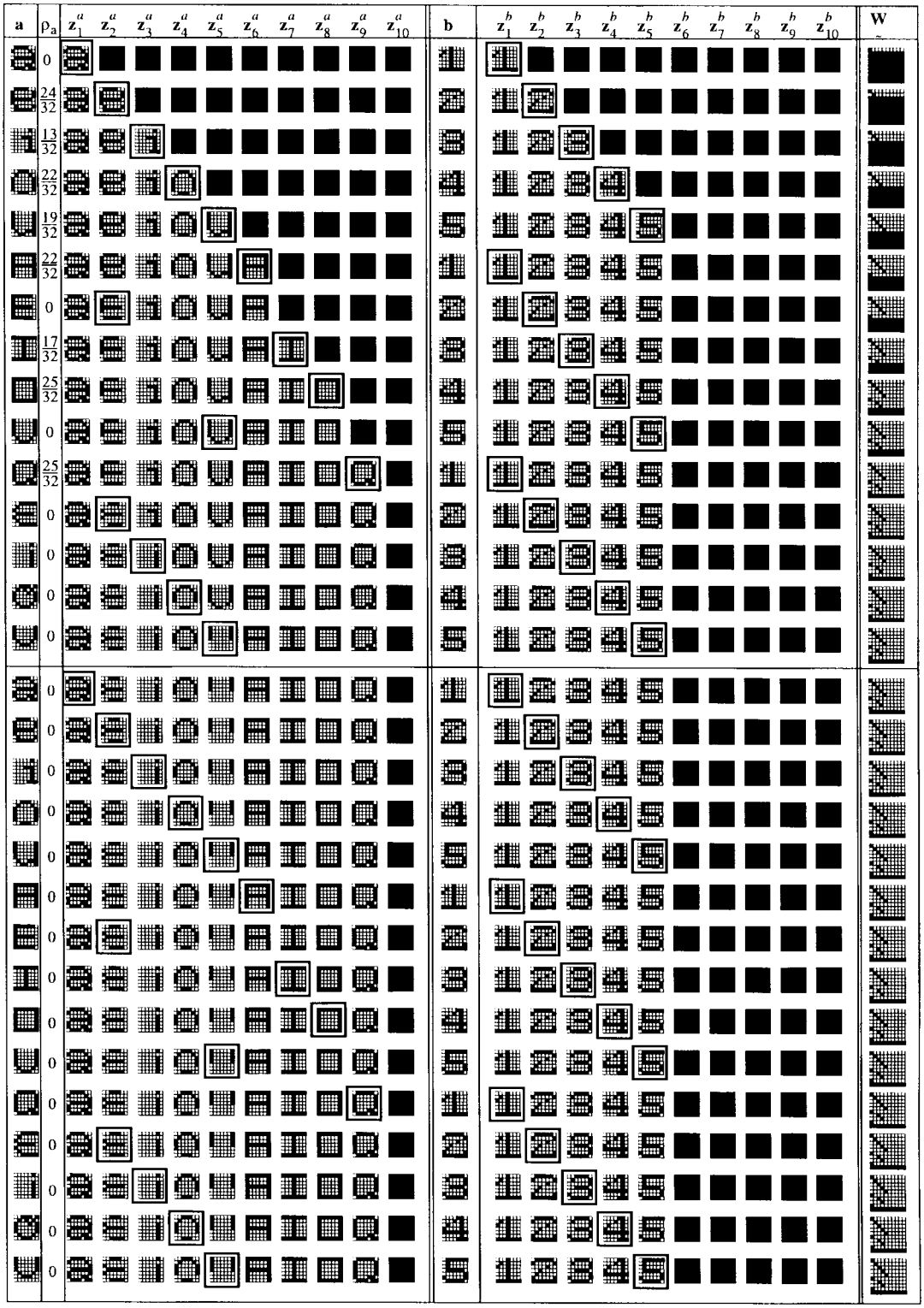

Fig. 13. Complete training sequence of the ARTMAP system for $\rho_a^{inital} = 0$ and $\rho_b = 0.75$.

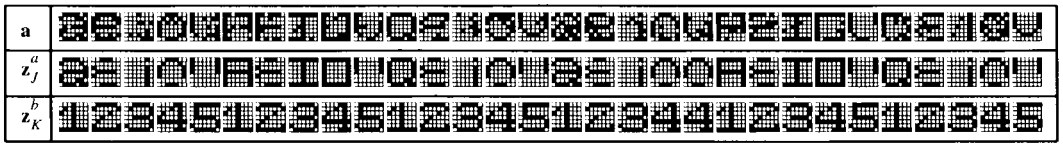

Fig. 14. Recognition sequence performed on the ARTMAP system trained in Fig. 13. Applied input patterns are noisy versions of the training set.

left to right). The vigilance parameter ρ_a was initially set to "0" and the current ratio parameters were ($\alpha^a = \alpha^b = 2, L_A^{a,b} = 10\,\mu$A and $L_B^{a,b} = 5\,\mu$A). For the ART1^b system it was $\rho_b = 0.75$. For this vigilance parameter, the ART1^b chip forms a distinct category for each input pattern.

Fig. 14 shows the results of a prediction sequence. Now instead of showing all stored $\mathbf{z}_j^a$ and $\mathbf{z}_k^b$ templates, only the categories of the chosen F_2 nodes are given. The top row shows the sequence of applied input patterns. The second row shows the ART1^a categories chosen by the chip after each search process. The third row shows the ART1^b categories that the corresponding ART1^a categories have learned to predict through the Inter-ART weights. Note that the applied input patterns are corrupted versions of the ones used during learning.

VI. Conclusions

An improved-yield ART1 chip has been designed, fabricated, and tested. Original prototype yield was 6% for an ART1 chip with 100 F_1 nodes, 18 F_2 nodes, and a die area of 1 cm^2. Present prototype yield is 98% with a die area 15 times less, 50 F_1 nodes, and ten F_2 nodes, while maintaining the same speed and precision. Yield improvement was possible after a careful large CMOS transistor arrays mismatch characterization. This enabled us to identify the maximum chip area for which gradient-induced mismatch is of the same order or less than pure random mismatch, while maintaining the targeted operation precision. Using this information an optimum die area ART1 prototype was designed for which no gradient-induced compensation circuitry is necessary, thus allowing a much more compact design, and consequently with significantly improved yield performance. A two-chip ART1 system and a three-chip ARTMAP system have been assembled and measured experimental results on their system-level behavior are provided.

References

[1] T. Serrano-Gotarredona and B. Linares-Barranco, "A real-time clustering microchip neural engine," *IEEE Trans. VLSI Syst.*, vol. 4, pp. 195–209, June 1996.

[2] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Comput. Vision, Graphics, Image Processing*, vol. 37, pp. 54–115, 1987.

[3] ———, *Pattern Recognition by Self-Organizing Neural Networks.* Cambridge, MA: MIT Press, 1991.

[4] ———, "ART 2: Self-Organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, pp. 4919–4930, Dec. 1, 1987.

[5] ———, "ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures," *Neural Networks*, vol. 3, pp. 129–152, 1990.

[6] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759–771, 1991.

[7] 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.

[8] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural-network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Trans. Neural Networks*, vol. 3, pp. 698–712, Sept. 1992.

[9] S. W. Tay and R. W. Newcomb, "VLSI implementation of ART1 memories," *IEEE Trans. Neural Networks*, vol. 2, pp. 214–221, Mar. 1991.

[10] T. P. Caudell, "A hybrid optoelectronic ART1 neural processor," *Appl. Opt.*, vol. 31, no. 29, pp. 6220–6229, Oct. 1992.

[11] T. Serrano-Gotarredona and B. Linares-Barranco, "A modified ART1 algorithm more suitable for VLSI implementations," *Neural Networks*, vol. 9, no. 6, pp. 1025–1043, 1996.

[12] 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, Oct. 1989.

[13] N. R. Strader and J. C. Harden, "Architectural yield optimization," *Wafer Scale Integration*, E. E. Swartzlander, Jr., Ed. Boston, MA: Kluwer, 1989, pp. 57–118.

[14] T. Serrano-Gotarredona and B. Linares-Barranco, "Systematic CMOS transistor mismatch characterization," in *Proc. 1996 IEEE Int. Symp. Circuits Syst. (ISCAS'96)*, Atlanta, GA, 1996, vol. 4, pp. 113–116.

Teresa Serrano-Gotarredona received the B.S. degree in electronic physics in June 1992 from the University of Seville, Sevilla, Spain. She received the Ph.D. degree in VLSI neural categorizers from the University of Seville in December 1996, after completing all her research at the Department of Analog and Mixed-Signal Circuit Design of the National Microelectronics Center (CNM), Sevilla, Spain. From September 1996 until August 1997, she was enrolled in the M.S. program from the Department of Electrical and Computer Engineering of the Johns Hopkins University, Baltimore, MD, where she is sponsored by a Fulbright Fellowship.

Her research interests include analog circuit design of linear and nonlinear circuits, VLSI implementations of neural computing and sensory systems, and VLSI electrical parameter characterization.

Dr. Serrano-Gotarredona was corecipient of the 1995–1996 IEEE Transactions on VLSI Systems Best Paper Award for the paper "A Real-Time Clustering Microchip Neural Engine."

Bernabé Linares-Barranco received the B.S. degree in electronic physics in June 1986 and the M.S. degree in microelectronics in September 1987, both from the University of Seville, Sevilla, Spain. He received the first Ph.D. degree in high-frequecy OTA-C oscillator design in June 1990 from the University of Seville, Spain, and the second Ph.D. degree in analog neural network design in December 1991 from Texas A&M University, College Station.

From September 1996 to August 1997, he was on sabbatical stay at the Department of Electrical and Computer Engineering of the Johns Hopkins University. Since September 1991, he has been a Senior Researcher with the Analog and Mixed-Signal Circuit Design Department of the National Microelectronics Center, Sevilla, Spain. He has been involved with circuit design for telecommunication circuits, VLSI emulators of biological neurons, neural based pattern recognition systems, hearing aids, precision circuit design for instrumentation equipment, and VLSI elecrical parameters characterization.

Dr. Linares-Barranco was corecipient of the 1995–1996 IEEE Transactions on VLSI Systems Best Paper Award for the paper "A Real-Time Clustering Microchip Neural Engine." Since July 1997 he has been an Associate Editor of the IEEE Transactions on Circuits and Systems Part II.