HIGH-SPEED HIGH-DENSITY CMOS CNNs

by

J. M. Cruz and L. O. Chua

Memorandum No. UCB/ERL M91/28

19 April 1991

HIGH-SPEED HIGH-DENSITY CMOS CNNs

by

J. M. Cruz and L. O. Chua

Memorandum No. UCB/ERL M91/28

19 April 1991

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

HIGH-SPEED HIGH-DENSITY CMOS CNNs

by

J. M. Cruz and L. O. Chua

Memorandum No. UCB/ERL M91/28

19 April 1991

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

High-Speed High-Density CMOS CNNs *

J. M. Cruz and L. O. Chua†

Abstract — **A CMOS-compatible Cellular Neural Network circuit architecture is presented. It is based exclusively on the use of operational transconductance amplifiers and capacitors. The operation is in continuous-time. CMOS implementations of this circuit architecture are extremely well suited for processing applications requiring both large array size and high speed. A systematic design approach for those circuits is described, and the design, fabrication and testing of two chip prototypes is presented. These prototypes implement the connected component detector described in [3], and have been fabricated using a 2μm CMOS technology. These 2,000-transistor chips are the first successfully tested hardware implementation of a CNN. Each cell contains only two OTAs and a very small capacitor. The density is 32 cells per square milimiter of silicon, and the time constant of the processing is of the order of 10^{-7}s. Experimental results of static and dynamic tests are given, including complete image processing examples.**

1 Introduction

Cellular Neural Network [1] is a novel analog network architecture having many potential applications, especially in image processing [2]–[4]. It consists of a two dimensional array of locally interconnected *analog* processors.

For image processing applications, each pixel of the image to be processed is usually associated with one cell. The processing is therefore fully parallel, and in principle should be adequate for high-speed real time computation. Furthermore, as opposed to other neural network topologies, the interconnections between the processing elements

*This work is supported in part by a Fulbright/MEC Fellowship, the Semiconductor Research Corporation, and the Office of Naval Research under Grant N00014-89-J-1402

†The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA.

are restricted to be only local. This represents a clear advantage over other proposed parallel neural network architectures from the point of view of VLSI implementation[4]-[7].

We present here a circuit architecture for the implementation of these networks in a standard CMOS technology. Our goal has been to develop compact and reliable silicon designs, that will allow for cost-effective monolithic integration of large arrays of processors. Another consideration has been to maximize the processing speed at an effective power cost.

The circuit architecture presented in this paper fulfill these goals. It is suitable for realizing high density CNN implementations on a CMOS technology that are reliable and have an high speed/power ratio. It is based on the exclusive use of two elements, Operational Transconductance Amplifiers (OTAs) and capacitors. They can be considered, therefore, as a type of circuits OTA-C. The processing is time-continuous. Taking advantage of the non-linear behavior of the CMOS OTA block to implement the PWL sigmoid nonlinearity in each cell makes extremely compact designs possible.

Here we report on the design, fabrication and testing of two integrated prototypes intended for Connected Component Detection [2]. This is an application useful in handwritten character recognition and in many other important feature extraction processes. This is just one example of a *global* image procesing task that can be realized by a CNN with a simple *local* interconnection topology pattern.

The structure of this paper is as follows. First, Section 2 gives ideal circuit model of single layer CNNs. In Section 3 we present a CNN circuit architecture suitable for CMOS implementation. A model of the CMOS OTAs to be used is given, and the influence of some non-linearities is discussed. Section 4 presents a systematic procedure for the design of CMOS CNNs, and describes the complete design and fabrication of two prototypes for connected component detection in a 2μm technology. Part 5 shows the experimental results of static and dynamic testing of those chips. Finally in part 6 we give some final conclusions.

2 Ideal Models of a CNN

A Cellular Neural Network consists of a two dimensional array of locally interconnected cells. For each cell C_{ij} define *input*, *state* and *output* variables, denoted respectively by u_{ij}, x_{ij} and y_{ij}. The nature of the processing of the network is determined by a set of

parameters. If for simplicity we consider the case of space invariant templates, these parameters are:

- The neighborhood size, r, which determines the maximum distance between two cells which are interconnected. For convenience, we define $\aleph_r = \{k, l | |k|, |l| \leq r\}$.
- The normalized feedback template, composed of a matrix of parameters denoted by $a_{k,l}$, $(k,l) \in \aleph_r$
- The normalized control template, composed of a matrix of parameters denoted by $b_{k,l}$, $(k,l) \in \aleph_r$
- The offset. Denoted by x_o.

The dynamics of each cell is given by the equation

$$\begin{aligned} R_x C_x \frac{dx_{ij}}{dt} &= -(x_{ij} - x_o) + \sum_{k,l \in \aleph_r} a_{k,l} y_{i+k,j+l} + \sum_{k,l \in \aleph_r} b_{k,l} u_{i+k,j+l} \\ y_{i,j} &= f_1(x_{ij}) \end{aligned} \tag{1}$$

where $f_1(\cdot)$ is picewise-linear fuction defined by

$$f_1(x) = \frac{1}{2}(|x+1| - |x-1|) \tag{2}$$

Figure 1 shows an ideal circuit model of a Cellular Neural Network cell. The *input*, *state* and *output* variables are represented by voltages at cell nodes. The offset, x_o, is implemented by a voltage source; the feedback and control operators are implemented with linear VCCSs with transconductance values given, respectively, by $a_{k,l}G_x$ and $b_{k,l}G_x$, where the parameters $a_{k,l}$ and $b_{k,l}$ are the dimensionles template coefficients, and G_x is the conductance of the linear state resistor.

In this ideal model the values of the state resistor and capactitor, R_x and C_x, only affect the time constant of the circuit.

This same circuit was presented in [1], but using a Norton equivalent of the DC source, and using non-normalized template parameters given by $A_{k,l} = a_{k,l}G_x$ and $B_{k,l} = a_{k,l}G_x$.

2.1 An ideal circuit model for CMOS implementation

Lets consider an equivalent circuit with the same topology but in which the PWL non-linearity is incorporated into the *feedback* VCCSs. Each of these current sources will now be controlled by the *state* variable, their current being given by $a_{k,l}G_x f_1(x_{i+k,j+l})$. The *control* VCCSs remain unchanged. Only two variables of each cell, those associated to the input node and the state node, are now significant for the processing.

The complete network composed of those cell circuits can be redrawen associating to each cell the current sources controlled by its *own* input and state variables. Figure 1b shows a schematic of one of those new cells. The cell $C_{i,j}$ drives a current $i_{o_{k,l}}$ into the state node of the cell $C_{i+k,j+l}$. That current is the sum of the contribution of the corresponding feedback and control currents:

$$i_{o_{k,l}} = i_{a_{k,l}} + i_{b_{k,l}} = a^*_{k,l}G_x f_1(x_{i,j}) + b^*_{k,l}G_x u_{i,j} \tag{3}$$

The new weight factors $a^*_{k,l}$ an $b^*_{k,l}$ are given by

$$\begin{aligned} a^*_{k,l} &= a_{-k,-l} \\ b^*_{k,l} &= b_{-k,-l} \end{aligned} \tag{4}$$

Observe that now we require only one connection originating from cell $C_{i,j}$ to each other cell in its r-neighbourhood.

Figure 1b also includes a VCCS for generating the output variable. However, as the output variables do not have to be generated in real-time during the processing, in a VLSI implementation that VCCS could be realised outside the cell, for example, in the read out circuitry.

In the implementation of the entire network, it is neccesary to compensate for the effects of the limited size of the array. This can be done by modifying the offset values of the cells near the edges (distance smaller than r). Another possibility is to use some extra cells outside the main array and setting their state and input nodes to permanent voltages. Although this option requires more area in a VLSI implementation, it actually may be preferable to the first due to symmetry and testability purposes.

The voltage ranges are limited in any integrated circuit implementation. Therefore, it is neccesary to scale the two voltage variables of a cell: the *input* variable and the *state* variable.

Define v_x and v_u as the scaled state and input voltages. They are given by

$$\begin{aligned} v_x &= n_x x \\ v_u &= n_u u \end{aligned} \tag{5}$$

where n_x and n_u are the normalizing factors, and x and u are the normalized variables used until now. The same ideal circuit of Figure 1b can be used, but with a voltage source of value v_{x_o} and VCCSs whose i/o function is given by $i = a^*_{k,l} f_a(v_x)$ and $i = b^*_{k,l} f_b(v_u)$, when

$$\begin{aligned} v_{x_o} &= n_x x_o \\ f_a(v_x) &= G_x n_x f_1(\frac{v_x}{n_x}) \\ f_b(v_u) &= G_x \frac{n_x}{n_u} v_u \end{aligned} \tag{6}$$

3 A CMOS-Compatible circuit architecture

In this section we present a CMOS-compatible circuit architecture for the efficient implementation of CNNs. The circuit architecture is based entirely on the use of capacitors and Operational Transconductance Amplifiers (OTAs). Both elements can be integrated efficiently in CMOS technology.

We first will show an OTA-based architecture that is equivalent to the ideal circuit models presented above if a first-order model of the OTA block is used. Then we will consider a more sophisticated model of the OTA that includes the characteristics that will affect the performance of the network. This model is the basis of set of design rules for the efficient design of these structures.

3.1 An OTA-based cell circuit

In a first order model an OTA block acts as a current source controlled by a differential voltage,

$$i = g(v) \tag{7}$$

If we are able to design OTAs with intrisic characteristics g similar to the linear and piece-wise-linear functions given in the previous section, then an extremely compact circuit architecture can be used to implement each CNN cell.

Figure 2 shows a diagram of the circuit architecture for a cell of a single layer CNN. Only three OTAs and 2 capacitors are required. The corresponding pixels of the input and initial image are stored in C_u and C_x, respectively. C_x also is the output load of the OTAs. For simplicity, the number of connections shown in the figure corresponds to a neighbourhood of size $r = 1$ in only one dimension.

OTA R is connected to emulate the state resistor. It has a transconductance characteristic given by

$$i = g_R(v) \tag{8}$$

This OTA should be designed to have a linear input-output characteristic over the whole dynamic range of the state voltage v_x. Its transconductance in that linear region is G_x.

OTA A is used to implement all the current sources controlled by the state voltage. It has multiple outputs, one for each element of the feedback template. The currents driven by them are given by

$$i_{k,l} = a^*_{k,l} g_A(v) \tag{9}$$

The function g_A should ideally be identical to the piecewise-linear fuction f_a given in (6). Its slope in the central linear region is denoted by G_x.

OTA B is used to implement all the current sources controlled by the input voltage. It has also multiple outputs, one for each element of the control template. The currents driven by them are given by

$$i_{k,l} = b^*_{k,l} g_B(v) \tag{10}$$

This OTA should have linear input–output characteristic over the whole dynamic range of the input voltage v_u.

The speed of the processing is determined by the state time constant $\tau_x = R_x C_x$. R_x is the equivalent resistance of the OTA R configuration. Its value can be externally controlled through the OTA control bias. Three switches are added to each cell for loading the input and initial image, starting the processing asynchronous operation, and reading the results. They are denoted as SW_{write}, SW_{start} and SW_{read}.

This OTA-based circuit model is equivalent to the ideal CNN circuit model when considering an OTA model with characteristics identical to the functions f_a and f_b. Of course, in any real CMOS implementation of these amplifiers many second-order effects and non-idealities are present. In the following paragraphs we discuss the effect of the

main characteristics and non-idealities of CMOS OTA structures on the behavior of the network. These considerations will later be taken into account in the design procedure described in Section 4.

3.2 CMOS OTAs: Structures and Models

We will consider the two possible OTA structures shown in Figure 3a. Each is composed of a differential transconductance stage and of current mirrors and current sources. There many other possible CMOS OTA structures; however these simple ones are especially suited for our requirements of low area and power. There are also several possible device-level structures for implementing the transconductance stage, the current sources and the current mirrors. We have chosen to use the structures shown in Figure 3b

In a CMOS OTA, the relationship between the output current and the input differential voltage is determined mainly by the transconductance stage. For most transconductance stages, including the differential pair shown in Figure 3b, this function has a sigmoid characteristic with output saturation for $v > |\sigma|$. For a simple differential pair an approximate analytical expression for the transconductance is given by

$$g(v) = \begin{cases} I & v \geq \sigma \\ \beta v \sqrt{\frac{2I_B}{\beta} - v^2} & -\sigma < v < \sigma \\ -I & v \leq -\sigma \end{cases} \quad (11)$$

where I is the tail bias current, β is the transconductance parameter of the differential pair transistors, and σ is given by

$$\sigma = \sqrt{\frac{I}{\beta}} \quad (12)$$

Later we will show how by, selecting adquately the value of σ, this function can implement, approximately, both the piecewise-linear behavior required for OTA A and the linear behavior required in OTA R and OTA B.

If we consider real CMOS current mirrors and current sources, then the i-v relationship given above remains valid only for a range of values of the input and output voltages. That range is given by an upper bounds on the voltages at the input and at the output, which we denote respectively by $v_{i_{max}}$ and $v_{o_{max}}$, and lower bounds on the input voltage, the common mode input voltage and the output voltage, which we denote by $v_{i_{min}}$,$v_{i_{CMmin}}$ and $v_{o_{min}}$, respectively.

The output node of the OTA has an associated equivalent output resistance that we denote by R_o. Its value can be very large.

OTA CMOS structures, as can be observed in Figure 2, do not contain any internal high impedance node. The dominant pole determining the bandwidth of the OTA is determined by

$$p_o = \frac{g_{out}}{C_{out}} \tag{13}$$

where C_{out} is the total capacitance associated to the output node. It is composed of the sum of parasitic capacitance associated with the OTA output terminal, C_o, and the load capacitance, C_L. And g_{out} is the equivalent conductance from the output node to ground. It is determined by the OTA intrinsic output resistance, R_o and by the resistive load $g_{out} = 1/R_o + 1/R_L$.

The intrinsic reactive behavior of the OTA can be modelled for our applications by an internal LHP pole, p_1. That pole is caused by the internal node connected at the input of a current mirror, and is located at a higher frequecy than p_o. In the frequency domain, below the bandwidth limit its effect is a phase error in the frequency response. In the time domain, its effect can be approximated by a time delay [9]-[10]

$$\tau_d = 1/p_1 \tag{14}$$

3.3 Influence of the OTA nonidealities in the CNN Cell

First we will analyze the effect of the nonidealities of the DC characteristics of the OTAs on the operation of CNNs.

Figure 4 shows the typical form of those characteristics, normalized to have the same slope in the origin. From that figure it is immediate to determine the value of the state normalizing factor, n_x.

$$n_x = g_R^{-1}(I_A^{sat}) \approx R_x I_A^{sat} \tag{15}$$

The condition $a_{0,0} > 1$ given in [1] for assuring that the network settle to a stable equilibrium point in which the output saturats translate in this case to the condition $\sigma_A < g_R^{-1}(a_{0,0} I_A^{sat})$ where σ_A is the voltage value at which OTA A saturates. The actual form of the sigmoid function between the saturation limits does not affect to the value of the the equlilibrum points if the derivative of that fuction is positive. The non-linearity of the OTA R introduces an error in the voltage value of the stable points.

In the case of a simple differential pair the relative error is approximately

$$\varepsilon_r = \left(\frac{v}{2\sigma_R}\right)^2 \tag{16}$$

The input voltages of the three OTAs must be maintained inside the bounds $v_{i_{max}}$ and $v_{i_{min}}$ or $v_{i_{CMmin}}$ limits. Exceeding these lmits will drastically diminish the linearity of OTAs A and B and will introduce a non-zero slope in the saturation levels of OTA A. This may cause instability in some cases[12]. The design must be also done to maintain the output voltage of the OTAs inside the range $(v_{o_{min}}, v_{o_{max}})$. Not doing so would approximatelly equivalent to limiting the state voltage to an artificially smaller range. This can make the array to settle to different stable points. Therefore the following inequalities must be imposed

$$\begin{aligned} v_x &< \min(v_{o_{maxA}}, v_{o_{maxB}}, v_{o_{maxR}}, v_{i_{max_A}}, v_{i_{maxR}}) \\ v_x &> \max(v_{o_{minA}}, v_{o_{minB}}, v_{i_{min_A}}, 2(v_{i_{CMminR}} - v_{x_o})) \end{aligned} \tag{17}$$

Regarding the dynamic behavior, we must analyze how different sigmoid functions affect the behavior of the entire network. We have made simulations using the differential pair sigmoid function for CNNs with the connected component detector template [3], and we have obtained the same final stable points as in the ideal case, with very similar dynamical evolution of the state voltage. Further simulation with other templates will indicate if the collective properties of these networks are to some degree robust to variations in the actual characteristic of the sigmoid function, as has been observed in other neural network architectures [8].

The speed of the proccesing will be given by determing $\tau_x = C_x^{total} R_x$. The capacitance C_x^{total} is given by a double-poly capacitor C_x^{poly} and the parasitic capacitances associated to the nodes x and x'

$$C_x^{total} = C_x^{poly} + C_x^{par} + C_{x'}^{par} \tag{18}$$

where

$$\begin{aligned} C_x^{par} &= C_{i+A} \\ C_{x'}^{par} &= C_{i-R} + C_{oA} + C_{oB} + C_{oR} \end{aligned} \tag{19}$$

The parasitic capacitances in parallel with C_x will in general have a beneficial effect, as they are contributing to the stability of the network. However, some of them may cause some variation in the initial state voltage value of the cell. This can be seen if we study the effect of changing the position of the start switch. When this switch is OFF and the initial state voltage value, $v_x(0)$, is stored in node x, the voltage at node x' is given by

$$v_{x'_{i,j}}(0) = v_{x_o} + \sum_{k,l \in \aleph_r} a_{k,l} f_A(v_{x_{i+k,j+l}}(0)) + \sum_{k,l \in \aleph_r} b_{k,l} f_B(v_{u_{i+k,j+l}}) \tag{20}$$

When the start swith is turned on, at the beginning of the processing, the voltages at both nodes x and x' evolute to the value $v^{eff}_{x_{i,j}}$.

$$v^{eff}_{x_{i,j}} = v_{x_{i,j}}(0) + \frac{C_{x'}}{C_x + C_{x'}} \left(-v_{x_{i,j}}(0) + v_{x_o} + \sum_{k,l \in \aleph_r} a_{k,l} f_A(v_{x_{i+k,j+l}}) + \sum_{k,l \in \aleph_r} b_{k,l} f_B(v_{u_{i+k,j+l}}) \right) \tag{21}$$

This happen with a time costant given by $R^{on}_{start-sw} C_{x'}$ that is much smaller than the time constant of the processing, τ_x. $R^{on}_{start-sw}$ is the equivalent resistance of the *start* switch in the ON position. The maximum tolerated difference between $v_{x_{i,j}}(0)$ and v^{eff}_x sets the minimum value for C_x.

Finally, the stability of the network depends upon the relative position of the two critical frequencies p_o and p_1. This can impose a minimum requirement on the value of the total state capacitance C^{total}_x. However, in the next section it will be shown how it is possible to make designs with high values of p_1 that give stable operation of the network with very small state capacitors.

4 Design of CNNs for Connected Component Detection

In this section we present the complete design of a CNN and its fabrication in a CMOS technology.

4.1 Connected Component Detection

The CNN design presented here is designed for detecting the number of connected components in each row of the image. This is just one example of global processing that can be performed by a CNN with a very simple local interconnection pattern. This

processing application is realized using the Connected Component Detector Cloning Template reported in [3], and is given by

$$a = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad b = 0 \quad x_o = 0 \tag{22}$$

The image to be processed is input as the initial state voltage values in the cells of the network. It is assumed that the input image is binary valued. An image pixel is represented as $x(0) = 1$ and a background pixel as $x(0) = -1$ in the corresponding cell. After the processing the number of cells with positive outputs in each row will be the number of the original connected components in that row. This positive outputs alternate with negative outputs starting from the rightmost cell. Figure 12 shows an example.

This simple cloning template ilustrates very well one of the outstanding features of Cellular Neural Networks; namely, the propagation of information over time through the network. This property allows global computation with a very simple interconnection topology. We will demostrate that this propagation can be so fast in some physical implementations of CNNs that the total processing time of the entire array is significantly shorter than other approaches, such as conventional digital signal processing or fully connected neural networks.

4.2 Circuit structure of a cell

As no control template is needed in this application, only the state capacitor C_x and two OTAs, one for the feedback template and the other for the resistor, are required. The resulting circuit architecture is shown in Figure 5. Our design is based on the structure shown in the Figure 6a. Each cell requires only one capacitor, two differential pairs, two current mirrors and some current sources. Three switches have been included to input the initial state voltage value, to output the state voltage and to start the transient continuous-time operation. The *start* control line is global to all the cells of the array. The control lines $write_i$ and $read_i$ are common for all the cells of row i, and the lines in_j and out_j are common to all the cells of column j.

For the above application we have used CMOS OTA structures with simple input differential pairs and cascade output branches. The device-level schematic of the entire

cell is shown in Figure 6b. The nodes labeled *ol* and *or* are connected to the state nodes of the left-side cell and the right-side cell respectively. To reduce area, only a two-transistor upper branch is used for each of the outputs. Two of these stages (one from the left-side cell and one from the right-side cell) are connected to node x. The DC current through these branches flows through the current source connected to the state node. OTA R has an additional output branch shown at the right of Figure 6b which provides an output current for the non-destructive sensing of the state voltage value. The four biasing current sources of the cell are controllable by five global lines. In this way, the power consumption, the speed and other parameters can be tuned.

Table I gives a summary of the main parameters of the ORBIT 2μm CMOS technology in which the designs have been integrated.

Table I. Technological Data

Parameter	N-channel	P-channel	Unit
V_{th}	1.0	0.8	V
κ	23.6	11.6	$\mu A/V^2$
γ	1.06	0.45	$\sqrt{V}$
ΔL	0.54	0.42	μm
ΔW	0.07	0.17	μm

4.3 Transistor Sizing

For dimensioning all the transistors of the circuit shown above we have developed a design procedure that is described below.

We start by considering the following DC transconductance functions for OTA A and OTA R:

$$g_A = \frac{I_A}{2} + \begin{cases} I_A/2 & v \geq \sigma_A \\ \frac{\beta_A}{2} v \sqrt{\frac{2I_A}{\beta_A} - v^2} & -\sigma_A < v < \sigma_A \\ -I_A/2 & v \leq -\sigma_A \end{cases} \quad (23)$$

$$g_R = \begin{cases} I_R & v \geq \sigma_R \\ \beta_R v \sqrt{\frac{2I_R}{\beta_R} - v^2} & -\sigma_R < v < \sigma_R \\ -I_R & v \leq -\sigma_R \end{cases} \quad (24)$$

In the ideal model the slope of the function of f_a in its linear region is equal to the conductance of the state resistor. In this OTA circuit we require that the derivatives

of g_A and g_R at the origin, represented by g_{mA} and g_{mR}, are equal

$$g_{mA} = g_{mR} \tag{25}$$

The design procedure is summarized as follows:

- An estimation of the maximum voltage range variation in the state node, $v_{x_{max}}$, is made based on the OTA structures to be used, the technology parameters, and the suply voltage values.

- The dynamic range of the normalized state variable, x_{max} is calculated from

$$x_{max} = |x_o| + \sum_{k,l \in \aleph} (|a_{k,l}| + |b_{k,l}|) \tag{26}$$

  This is a slightly stricter limit than the one given in [1].

- The value of the state normalizing factor is calculated.

$$n_x < \frac{v x_{max}}{x_{max}} \tag{27}$$

  Due to noise considerations, this value should be chosen close to the maximum.

- The value of σ_A is calculated from n_x,

$$\sigma_A = \sqrt{2} n_x \tag{28}$$

  This equation is valid for simple differential pairs given by equation (23). For linearized transconductance structures the multiplicative factor is closer to 1.

  It must be verified that $a_{0,0} f_R^{-1}(I_A^{sat}) > \sigma_A$ to guarantee that the cell settles to a stable equlibrium point. In this case, that inequality reduces approximately to $a_{0,0} > \sqrt{2}$.

- The value of σ_R is calculated from the maximum relative variation, ε_r of the DC characteristic of OTA R from a linear one, over the state variable dynamic range.

$$\sigma_R = \frac{x_{max} n_x}{2\sqrt{\varepsilon_r}} \tag{29}$$

- The geometry and current ratios can then be calculated

$$\begin{aligned} \frac{\left(\frac{W}{L}\right)_A}{\left(\frac{W}{L}\right)_R} &= 2\frac{\sigma_R}{\sigma_A} \\ \frac{I_A}{I_R} &= 2\frac{\sigma_A}{\sigma_R} \end{aligned} \tag{30}$$

- The differential pair of OTA A has a smaller bias current, but a bigger W/L transistor ratios. The minimum allowable value of I_A is given by the condition that the OTA A transistors do not enter the weak inversion region. A value close to it should be selected to minimize power.

- The differential pair of the OTA R has a greater bias current, but a smaller W/L transistor ratios. The minimum allowable value of $(W/L)_R$ is given by the condition that the transistors of its bias current source do not enter their triode region.

- The current-source transistors are sized. They are designed to have a $(V_{GS} - V_{th})$ of more than 0.4 volt for robustness against variation of the threshold voltage of the MOS transistors. Also, dimensions of at least 4 μm have been considered for robustness against variation of ΔW and ΔL. They must also be designed so that the following inequality is satisfied

$$-v_{x_{max}} > \max(v_{i_{min_A}}, 2(v_{i_{CMminR}} - v_{x_o})) \tag{31}$$

  If this inequality is impossible to satisfy then the specification of $v_{x_{max}}$ in step 1 must be relaxed.

- The current mirrors are sized. They are also designed to have dimensions of at least 4 μ m and a $(V_{GS} - V_{th})$ of more than 0.4 volt at least. A higher value of $(V_{GS} - V_{th})$ will be used for increasing the value of the intrinsic pole p_1. The current mirror gain in OTA R is 1. In OTA A, the gains are equal to the coefficients $a_{k,l}$ of the normalized feedback templates. These values are typically also on the order of unity, which is very convenient for maximizing the value of the intrinsic pole in the input node of the current mirror. For robustness against the uncertainity in the values of ΔW and ΔL, we use transistor duplication rather than different W/L ratios for the branches associated with the different

weights. The current mirrors must also be designed so that the following equation is satisfied

$$v_{x_{max}} < \min(v_{i_{max_A}}, v_{i_{max_R}}, v_{o_{max}}) \tag{32}$$

If this is impossible the specification of $v_{x_{max}}$ in step 1 must be relaxed.

Table II gives a summary of the values of all the design parameters that have been calculated manually following the above procedure. SPICE simulations are required for obtaining the final transistor dimensions. They are given in Table III. The circuit operates on a 6V power supply voltage and in the nominal operating point the currents biasing the OTA A and OTA R are given by $I_A = 7\mu A$ and $I_A = 17.5\mu A$, respectively. The power consumption of each cell is 0.37mW.

Table II. Design Values

Line	Value	Unit
v_{max}	900	mV
x_{max}	4	V
n_x	0.225	-
σ_A	320	mV
ε_r	0.08	-
σ_R	1.6	V
$\frac{(\frac{W}{L})_A}{(\frac{W}{L})_R}$	10	-
$\frac{I_A}{I_R}$	0.4	-

Table III. Mask Device Dimensions

Device	W (μm)	L (μm)
$T_{1A}, T_{1B}, T_{2A}, T_{2B}$	8	4
$T_{3A}, T_{4A}, T_{5A}, T_{6A}, T_{7A}, T_{8A}, T_{8B}, T_{9A}, T_{9B}, T_{12A}, T_{13A}, T_{14A}, T_{14B}$	4	4
$T_{3B}, T_{4B}, T_{5B}, T_{6B}, T_{7B}, T_{8C}, T_{8D}, T_{9C}, T_{9D}, T_{12B}, T_{13B}, T_{14C}, T_{14D}$	4	2
T_{10}, T_{11}	4	10
T_{15A}, T_{16A}	9	4
T_{15B}, T_{16B}	9	2
T_{startN}, T_{startP}	8	2
$T_{writeN}, T_{writeP}, T_{readN}, T_{readP}$	4	2

Two values for the state capacitance has been considered. The protoype CNN1_C1 contains a 1.1pF capacitor in each cell, while a second prototype, CNN_C6 contains a 5.6pF capacitor in each cell.

4.4 Layout of a cell

The layout has been done using techniques for reducing the mismatch. We have used an input a common-centroid differential pairs, all the transistors of the cell are oriented in the same direction. Each cell of both prototypes occupies an area of only 31,000μm^2 (180 x 177μm^2), corresponding to a density of 32 cells/sqr mm. Figure 7b shows the layout of a cell of the prototype CNN_C1.

4.5 Structure of the chip

Fig. 8 shows the structure of the complete designed chip. It contains a 6×6 array of cells, two 6×1 linear arrays for setting the edge conditions and for testing purposes, and the circuitry for initializing the network and to output the data. The analog input data and output data is multiplexed row by row. In the read out operation, the output currents provided by the selected cells are converted to voltages, using an OTA-based resistor structure similar to those used in the cells. The resulting characteristic of the voltage obtained in the sensing column j versus the voltage in node x', is almost linear, with a slope of –1, over the whole range of state voltage values. These column output voltages are buffered using high-frequency linear opamps.

A die micrograph of one of the fabricated prototypes, CNN_C1, is shown in Figure 9. It contains about 2,000 transistors. CNN_C6 have the same structure.

The propotypes have been encapsulated in a dual 40-pin structure. Figure 10a shows a photograph of one of the sample chips. Figure 10b shows its pin-out diagram.

5 Testing

The fabricated units of both prototypes, CNN_C1 and CNN_C6, worked correctly on the first fabrication run. Experimental tests have shown correct functional behavior, with both static and dynamic characteristics confirming device-level simulations.

The chips can operate over a range of different biasing conditions. In addition to the power supply lines, the chip contains 5 additional lines for voltage and current biasing. These allow us to change the operating point and to adjust the power consumption, the speed of processing, the scale factor in the state node, and a multiplicative factor to the feedback template element values.

The results given here were measured at the chip's nominal operating point. Table IV shows the value of all its external bias voltages and currents.

Table IV. Nominal Operating Point

Line	Value	Unit
V_{DD}	2.5	V
V_{SS}	-3.5	V
V_{GND}	0.0	V
V_c	0.0	V
I_{bias_A}	14.4	μA
I_{bias_T}	14.4	μA
I_{bias_R}	14.4	μA
I_{bias_X}	14.4	μA

In this Section the results of the static tests will be presented first. Then we will present the results of a complete dynamic testing of the entire array. We give the measured voltages of all the state nodes of the network before and after processing, as well as the time history of one row of state voltages *during* high-speed processing.

5.1 Static Tests

We make use of different test structures that have been added to the chip for extracting the DC characteristic of the chip. It is possible to measure the DC characteristic of OTA A and OTA R in any of the 12 edge cells. The cell to be tested is selected using the digital control lines. The input is done throught v_r or v_l. Figure 11 gives the DC transfer characteristic of both OTAs of one of those cells. Note that the characteristic of the OTA A is a sigmoid, and that the slope in the saturation mode is zero. Table V gives the experimental values of the main parameters used to model the DC characteristics. The symbol δ stands for offset. Measurements of the DC characteristics of all those cells in different samples have been made to study the performance variance due to the technological process.

Table V. DC Measurements

Parameter	Value	Unit
I_A	6.5	μA
I_R	17.5	μA
σ_A	0.4	V
δ_A	~ 0	V
I_A^{sat+}	0.0	μA
I_A^{sat-}	-6.3	μA
g_{mA}	11.5	$\mu A/V$
σ_R	2.0	V
δ_R	-0.1	V
I_R^{sat+}	18.0	μA
I_R^{sat-}	15.0	μA
g_{mR}	12	$\mu A/V$

5.2 Dynamic test

In this part we present an example of a dynamic test involving the processing of a complete two dimensional image.

We load a two bi-dimansional image into the analog memory of the chip row by row. Then the image is proccessed in parallel, and the results are read out. We have built a test board for realizing the multiplexed loading row by row of any initial image,

and for reading out.

5.2.1 State variable values before and after the proccessing: Simulation

The black pixels of the input image correspond to an initial voltage v_x of $+1n_x$ volts and the white pixels to $-1n_x$ volts. The initial state normalized matrix is

$$X(0) = \begin{bmatrix} +1 & -1 & -1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 & -1 & -1 \\ +1 & +1 & -1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 & -1 & -1 \end{bmatrix} \tag{33}$$

The state values of all the external edge cells, not shown in the figure, have been set to $-1n_x$.

The normalized value of nodes x' after the image loading and before the processing are given ideally by:

$$X'(0) = \begin{bmatrix} +2 & 0 & -4 & +2 & 0 & -2 \\ +2 & 0 & -4 & +2 & 0 & -2 \\ 0 & +4 & -2 & +2 & 0 & -2 \\ 0 & +2 & +2 & +4 & 0 & -2 \\ +2 & -2 & 0 & +4 & 0 & -2 \\ +2 & 0 & -4 & +2 & 0 & -2 \end{bmatrix} \tag{34}$$

The final normalized state values after the proccessing should be

$$X^{steady} = \begin{bmatrix} -2 & -2 & -4 & +2 & -2 & +2 \\ -2 & -2 & -4 & +2 & -2 & +2 \\ -2 & -2 & -4 & +2 & -2 & +2 \\ -2 & -2 & -2 & -2 & -4 & +2 \\ -2 & -2 & -4 & +2 & -2 & +2 \\ -2 & -2 & -4 & +2 & -2 & +2 \end{bmatrix} \tag{35}$$

For the connected component detector template with this edge conditions set to -1, the only possible steady state values of the network are: $2n_x$, $-2n_x$ and $-4n_x$.

The sign of the state variable determines the output image. A black pixel corresponds a positive state, while a white pixel to a negative state. That image is shown at the bottom of Figure 12.

5.2.2 State variable values before and after the proccessing: Experimental

We now show the results that have been obtained experimentally with the CNN_C1 prototype.

During the writing operation the line *write* is enabled and the line *start* disabled. The input image X is introduced through 6 column input lines. A row is loaded with each clock pulse. At the same time the initial voltages in the nodes $x_{i,j}$ are loaded, the voltages at nodes $x'_{i,j}$ are sensed. After that, *write* is disabled and *start* enabled. One clock period is spent to allow the array to settle. Then the values of nodes $x'_{i,j}$ which corresponds to the final steady values are read out again with each high clock pulse.

Figure 13 shows the experimental scope traces of the $v_{x'}$ voltages for each of the 6 columns. These traces are presented in the middle of the photograhs. The scale is 500 mV/Div. The six data values sampled at the left handside of the waveform are the $x'(0)$ values af all the cells of the column, and the six data values sampled at the left of the waveform are the v_x^{steady} values. The same photographs also include the clock signal (top) and the *write* signal (bottom). The *start* signal, not shown in the photographs, is enabled just after *write* is disabled. At the top of each photograph is a drawing in which the expected normalized value of x' is shown.

The steady voltage values are

$$\begin{aligned} 450 \pm 50mV \quad & for \quad 2n_x \\ -350 \pm 50mV \quad & for \quad -2n_x \\ -750 \pm 50mV \quad & for \quad -4n_x \end{aligned} \tag{36}$$

From this data it can be deduced that a value for the normalizind state factor is approximately $200mV$. There is also a positive systematic offset of $+50mV$ and a maximum random variation of $\pm 50mV$. The systematic offset is caused mainly in the on-chip sensing circuitry. The real systematic offset is smaller. The random variation is due mainly to local mismatches. All these variations, much smaller then n_x, do not have any effect in the output binary image.

5.2.3 State variable values during proccessing

In this section we present the experimentally measured real-time evolution of the state voltages during processing. Figure 14 presents the dynamic evolution of the state voltage for all cells of the 3^{rd} row using the test image presented before.

Figure 14a shows the pixel values of that row of the input image. The pixel represented at the top of the figure is associated with the left hand side cell of the row.

Figure 14b shows the expected final binary output. The numerical values give, in normalized units, the expected final steady voltages of the corresponding state nodes.

Figure 14c shows the transient obtained by simulation using the ideal circuit model of a CNN given in [1], and with a time constant of $0.43\mu s$, that is the R_xC_x value for the prototype CNN_C6 in the nominal operating point. The vertical axis represents the state voltage in normalized units, from –5 to +5. The horizontal axis is time, $1\mu s$/div.

Figure 14d shows the experimental transient waveforms obtained in the prototype CNN_C6. The vertical axis again represents the state voltage. The scale is 2 normalized units/div, i.e. 0.4V/div. The horizontal axis is time, $1\mu s$/div. The offset of 50 mV has been externally compensated. Observe that all the state nodes reach their proper final voltage values in less than $5\mu s$. Moreover, the transient responses are almost identical to the ideal case.

Figure 14e shows the equivalent experimental results in the prototype CNN_C1. The horizontal scale is also $1\mu s$/div. The response is qualitatively the same as for the CNN_C6, except that the state voltages reach their steady value in less than $1.5\mu s$. The obtained increase in speed is in accordance with simulations in which we take into account the parasitic capacitances in parallel with the state node and with the column testing line.

6 Conclusions

Two CNN working chips have been presented. They are the first working hardware implementation of a CNN. They are based on a proposed non-linear OTA-C architecture. Each cell is composed of only two simple CMOS OTAs and a small capacitor. They operate in continous-time at extremelly high speed, with a time constants in the order of 10^{-7}. The total settling times of the complete test arrays are on the order of microseconds. For the reported chips the area of each cell is $177 \times 180\mu m^2$, including the local and global interconnection wiring. With this architecture it is feasible to

realize monolithic VLSI implementations of CNNs containing several thousands cells.

Acknowledgment

The authors would like to thank Professors Paul Gray and Robert Brodersen for providing the resources (MOSIS) necessary in this work, and to Bertram Shi for useful discussions.

References

[1] Leon O. Chua and Lin Yang. Cellular Neural Networks: Theory. *IEEE Transactions on Circuits and Systems*, 32, October 1988.

[2] Leon O. Chua and Lin Yang. Cellular Neural Networks: Applications. *IEEE Transactions on Circuits and Systems*, 32, October 1988.

[3] T. Matsumoto, Leon O. Chua and H. Suzuki. CNN Cloning Template: Connected Component Detector. *IEEE Transactions on Circuits and Systems*, 37, pp 663-935 May 1990.

[4] 1990 IEEE International Workshop on Cellular Neural Networks and their Applications. *Proceedings*, December 1990, Budapest, Hungary. IEEE Catalog No. 90TH0312-9.

[5] L. Yang, L. O. Chua and K. Krieg. VLSI Implementation of Cellular Neural Networks. *Procc. IEEE ISCAS90*, pp 958-961, 1990.

[6] J. M. Cruz and L. O. Chua. A CNN Chip for Connected Component Detection. *Express Letter. To appear in IEEE Transactions on Circuits and Systems.*

[7] J. J. Hopfield. Artificial Neural Networks. *IEEE Circuits and Devices Magazine*, 4, No. 5, Sept 1988.

[8] J. J. Hopfield. Neurons with Graded Response have Collective Computation Properties Like Those of Two State Neurons. *Procc. of the National Academy of Sciences*, 81, pp 3088-3092 May 1984.

[9] A. Rodriguez-Vazquez, B. Linares-Barranco, J. L. Huertas and E. Sanchez-Sinencio. On the Design of Voltage-Controlled Sinusoidal Oscillators Using OTA's. *IEEE Transactions on Circuits and Systems*, 37, No. 2, February 1990.

[10] E. Sanchez-Sinencio, J. Ramirez-Angulo, B. Linares-Barranco and A. Rodriguez-Vazquez. Operational Transconductance Amplifier-Based Nonlinear Function Syntheses. *IEEE Journal of Solid-State Circuits*, vol. 24, pp 1576-1586, December 1989.

[11] K. Halonen and J. Vaananen The non-idealities of the IC-realization and the stability of CNN-networks. *Procc of the 1990 IEEE International Workshop on Cellular Neural Networks and their Applications*, pp 226-234, Budapest, Hungary, December 1990.

Figure Captions

Figure 1: Ideal Models of a Cellular Neural Network.

Figure 2: OTA-C cell architecture.

Figure 3:
(a) Two OTA topologies.
(b) Device-level schematic.

Figure 4: Generic DC characteristics of OTA A and OTA R.

Figure 5: OTA-C cell architecture for Connected Component Detection.

Figure 6:
(a) Cell schematic.
(b) Device-level cell schematic.

Figure 7: Core cell layout.

Figure 8: Chip structure.

Figure 9: Chip micrograph.

Figure 10:
(a) Photograph of encapsulated chip.
(b) Pin-out diagram.

Figure 11: Static test.
(a) DC transfer characteristic of OTA G (top trace).
(b) DC transfer characteristic of OTA A (bottom trace).

Figure 12: Dynamic test. An Image processing example.
Figure 13: Experimental results of the dynamic test.

(a) Column 1.

(b) Column 2.

(c) Column 3.

(d) Column 4.

(e) Column 5.

(f) Column 6.

Figure 14: Dynamic test during the processing interval.

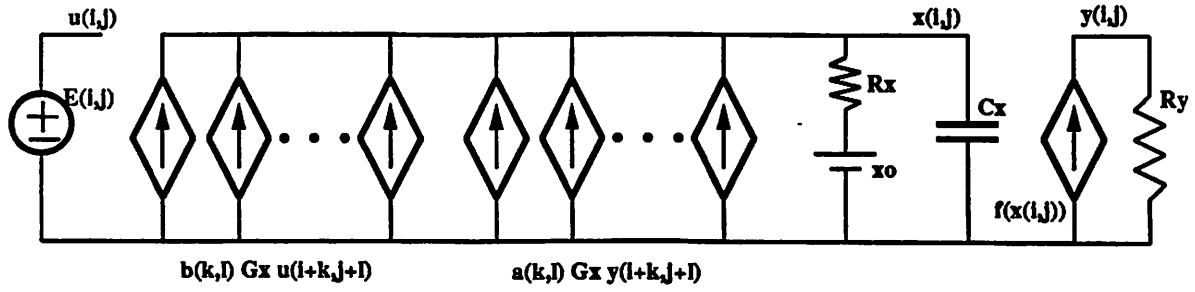
(a) Third row of the input image (top box itself corresponds to leftmost circuit cell).

(b) Final output and steady state values.

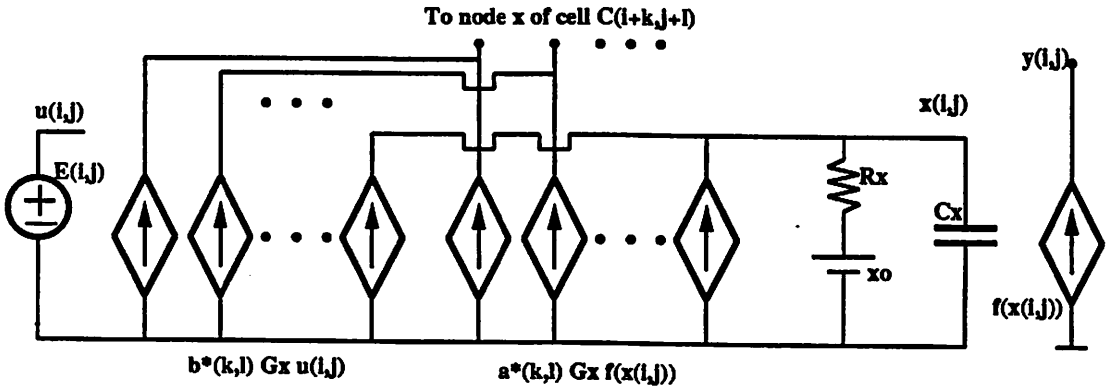
(c) Transient evolution of the state voltages using ideal models.

(d) Experimental transient waveforms of the state voltages in the prototype CNN_C6. ($1\mu s$/div)

(e) Experimental transient waveforms of the state voltages in the prototype CNN_C1. ($1\mu s$/div)


u(i,j)
E(i,j)
x(i,j)
y(i,j)
Rx
Cx
xo
Ry
f(x(i,j))
b(k,l) Gx u(i+k,j+l)
a(k,l) Gx y(i+k,j+l)

(a)


To node x of cell C(i+k,j+l)
u(i,j)
E(i,j)
y(i,j)
x(i,j)
Rx
Cx
xo
f(x(i,j))
b*(k,l) Gx u(i,j)
a*(k,l) Gx f(x(i,j))

(b)

Figure 1

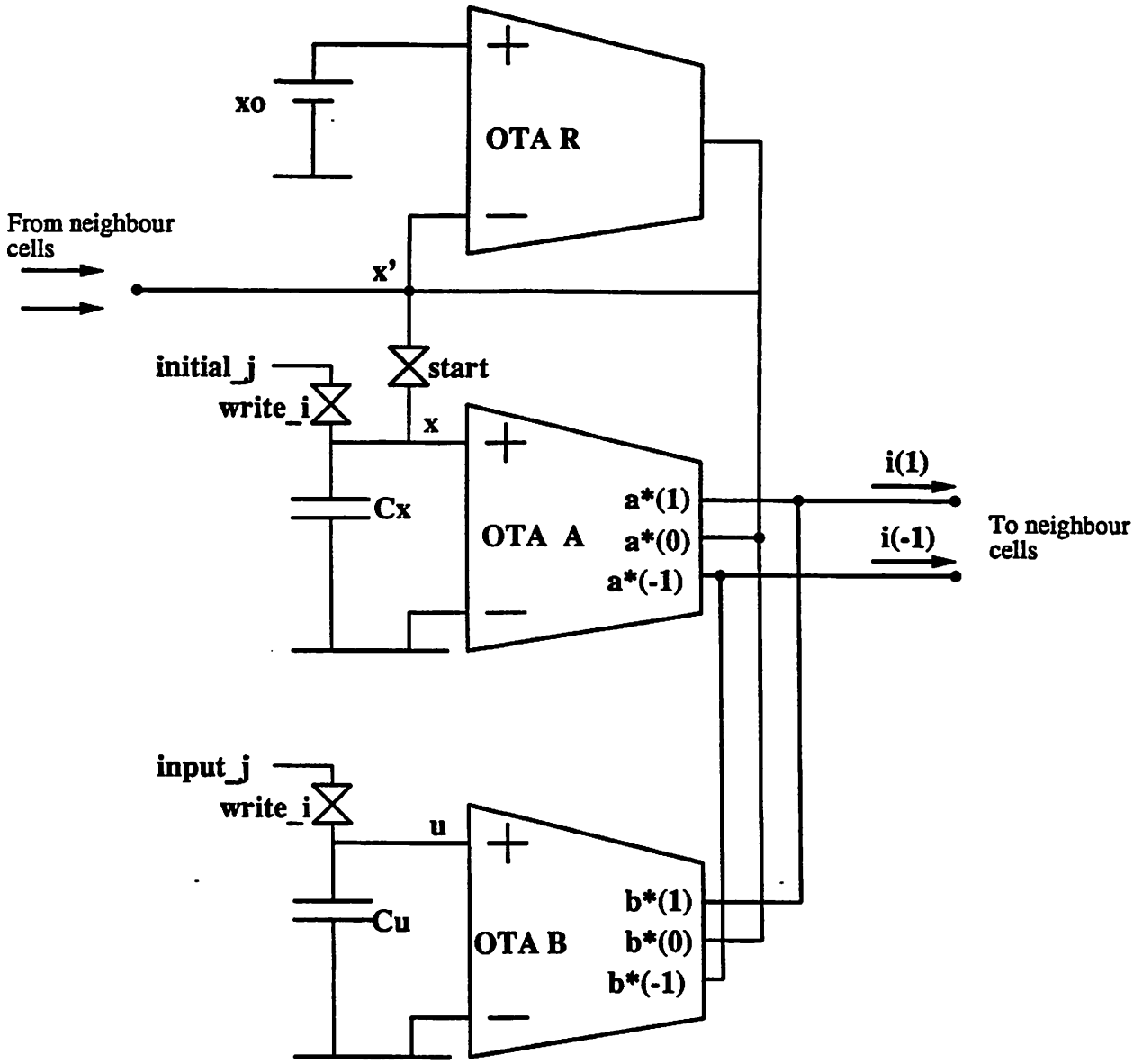
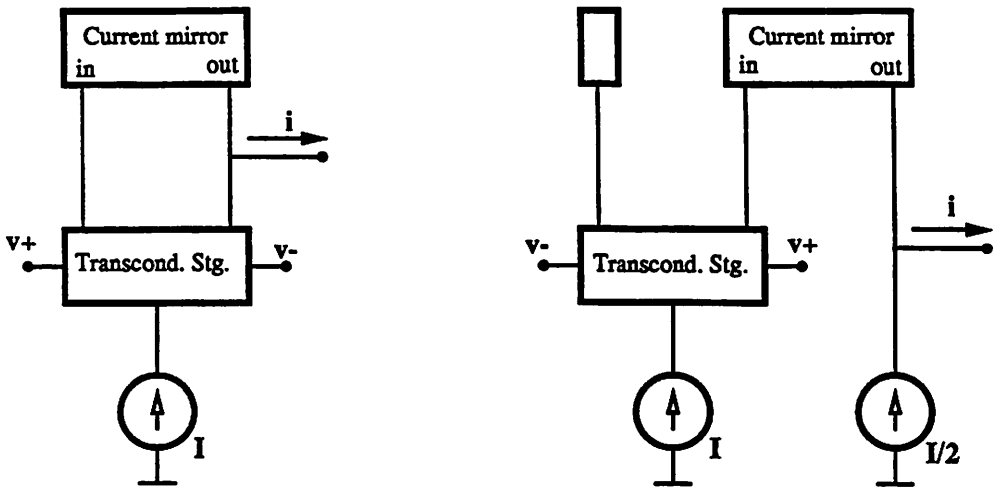
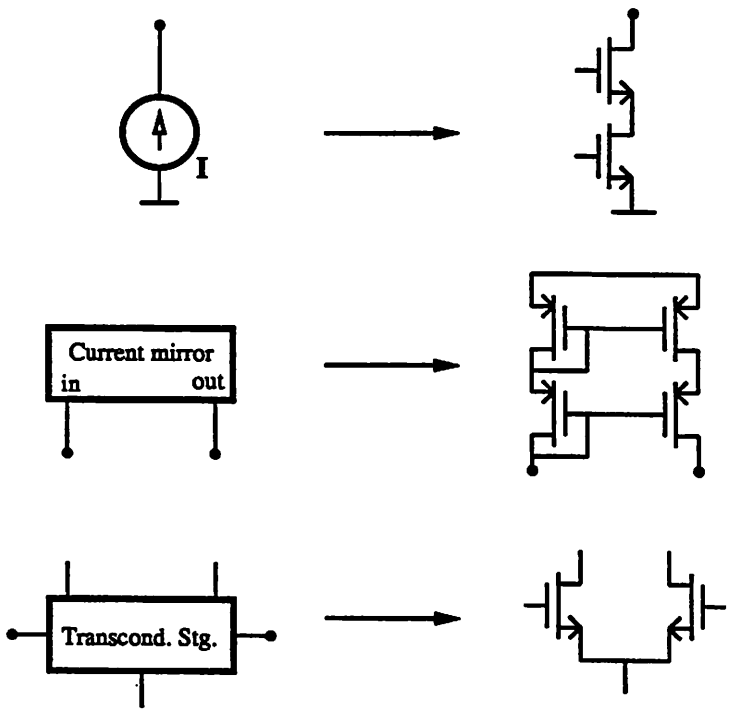

xo
OTA R
From neighbour cells
x'
initial_j
write_i
start
x
Cx
OTA A
a*(1)
a*(0)
a*(-1)
i(1)
i(-1)
To neighbour cells
input_j
write_i
u
Cu
OTA B
b*(1)
b*(0)
b*(-1)

Figure 2


Current mirror
in
out
i
v+
Transcond. Stg.
v-
I
Current mirror
in
out
v-
Transcond. Stg.
v+
i
I
I/2

(a)


I
Current mirror
in
out
Transcond. Stg.

(b)

Figure 3

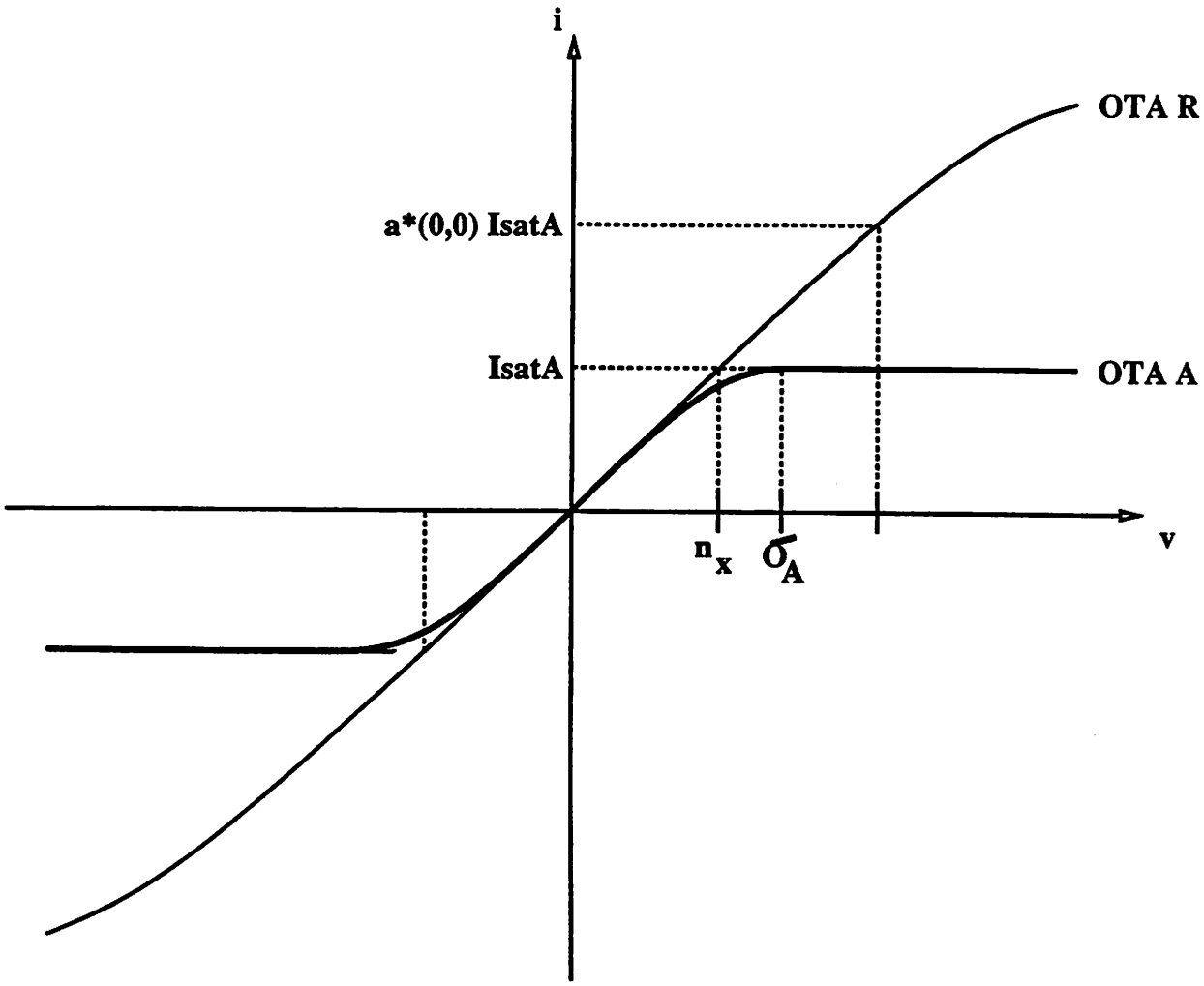


Figure 4

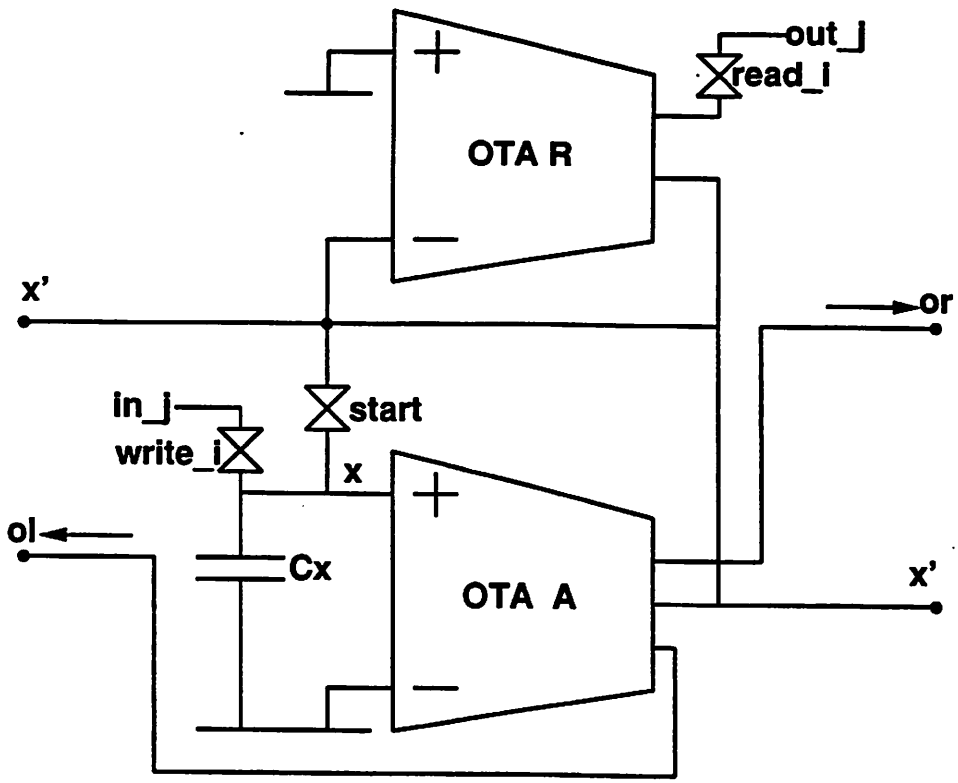
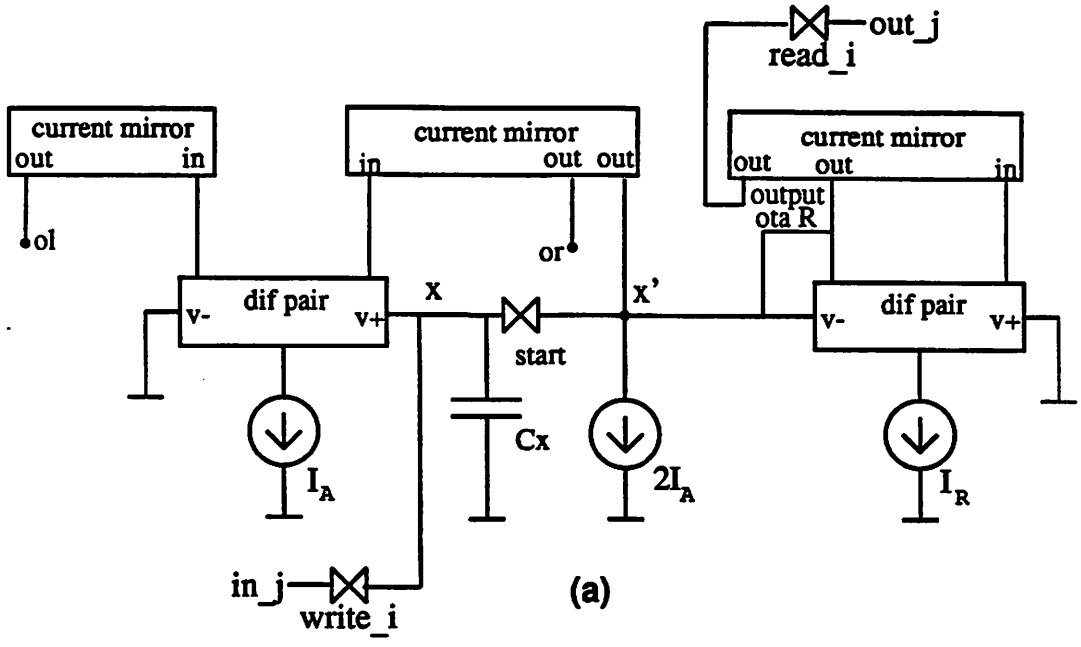
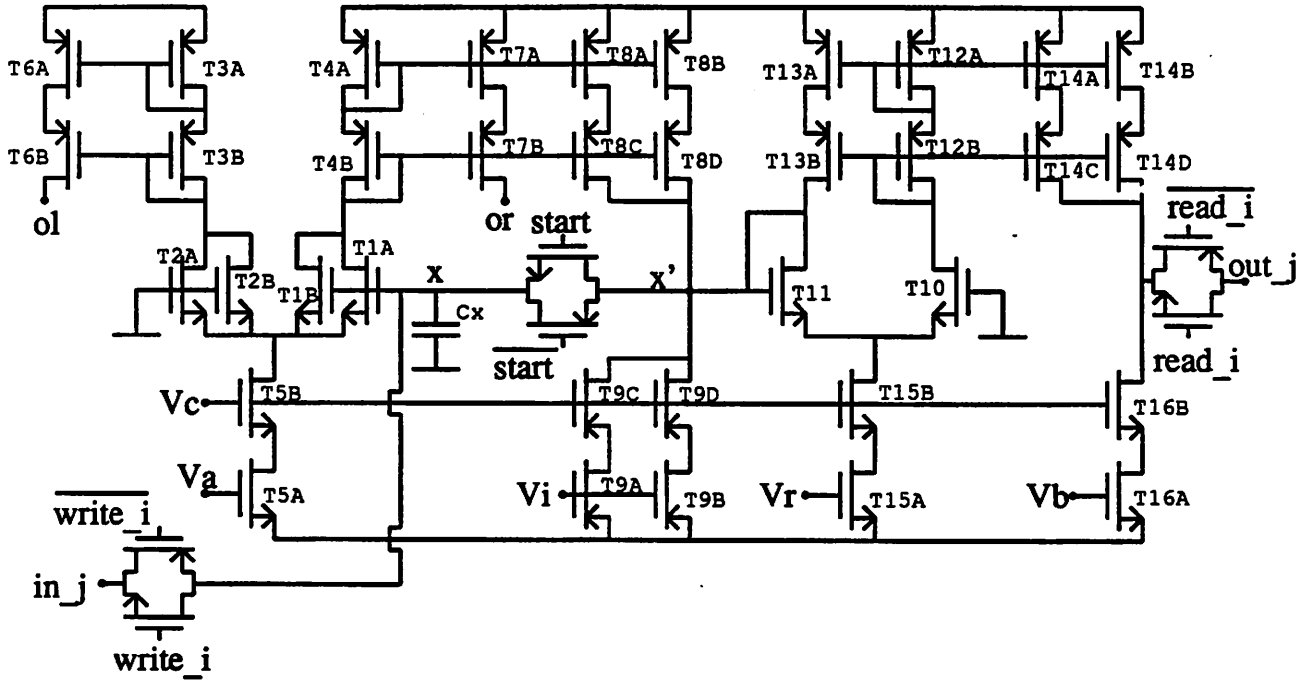


Figure 5


out_j
read_i
current mirror
out in
current mirror
in out out
current mirror
out out in
output
ota R
ol
or
dif pair
v- v+
X
X'
start
dif pair
v- v+
Cx
I_A
$2I_A$
I_R
in_j
write_i
(a)


T6A
T3A
T4A
T7A
T8A
T8B
T13A
T12A
T14A
T14B
T6B
T3B
T4B
T7B
T8C
T8D
T13B
T12B
T14C
T14D
ol
or
start
read_i
out_j
T2A
T2B
T1B
T1A
X
X'
Cx
T11
T10
start
read_i
Vc
T5B
T9C
T9D
T15B
T16B
Va
T5A
Vi
T9A
T9B
Vr
T15A
Vb
T16A
write_i
in_j
write_i

(b)

Figure 6

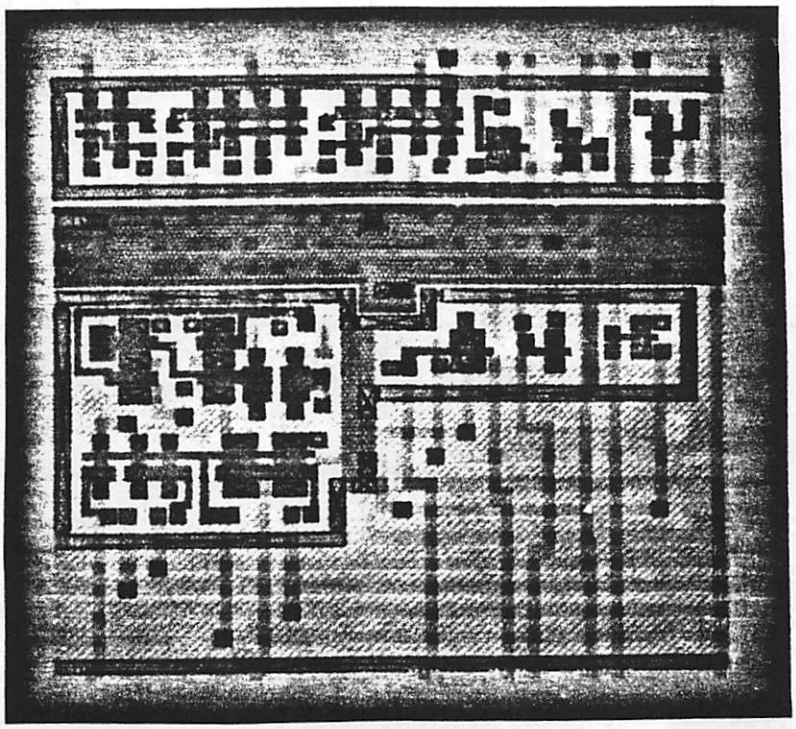


Figure 7

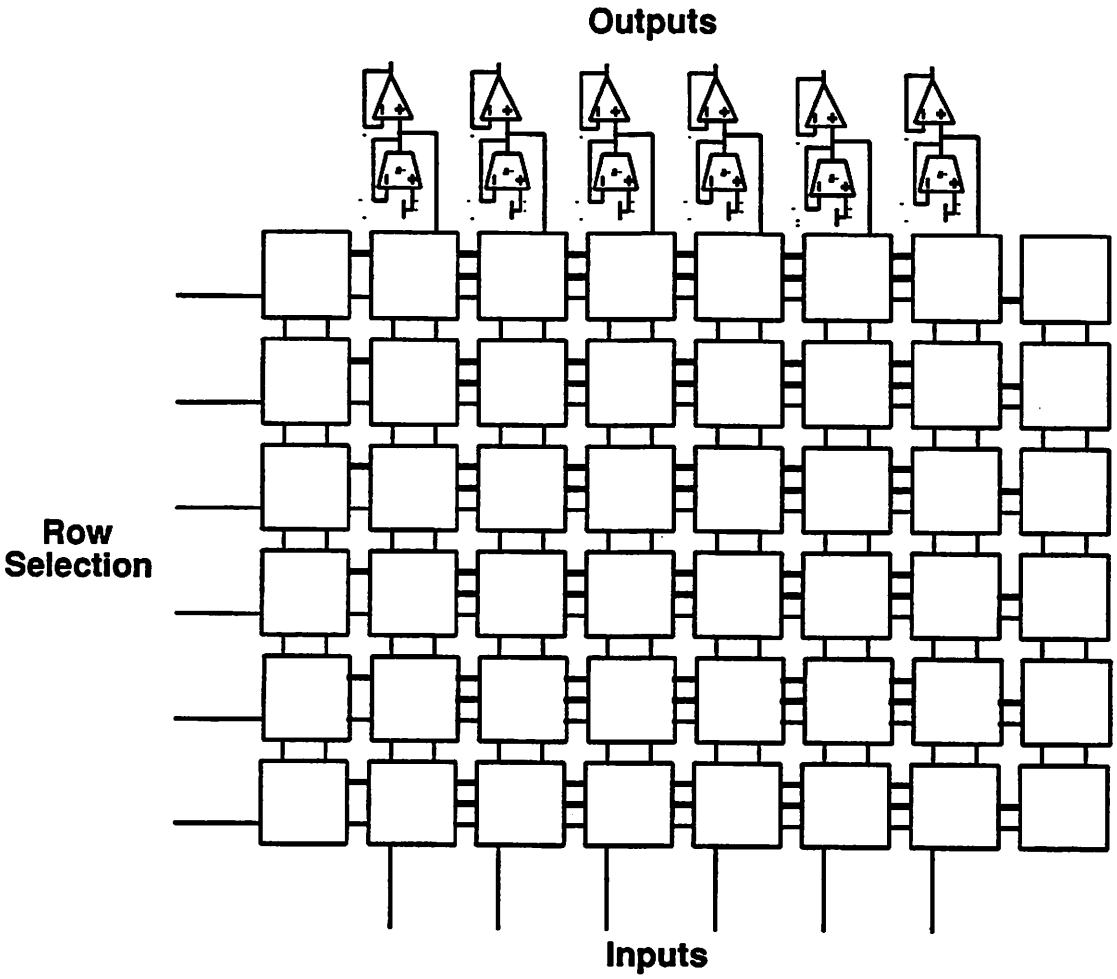


Figure 8

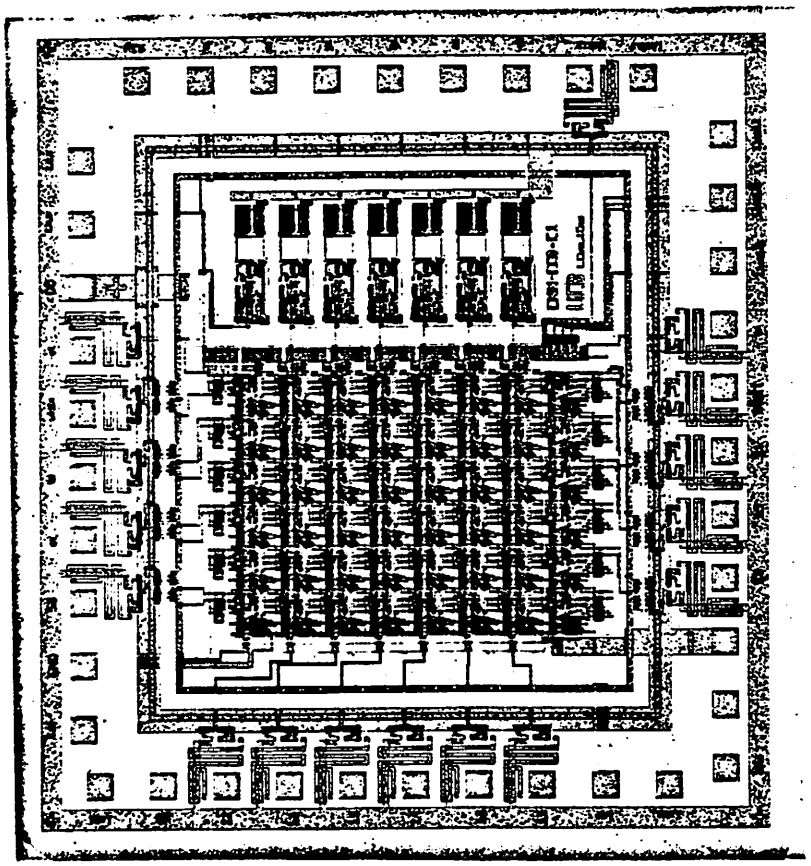
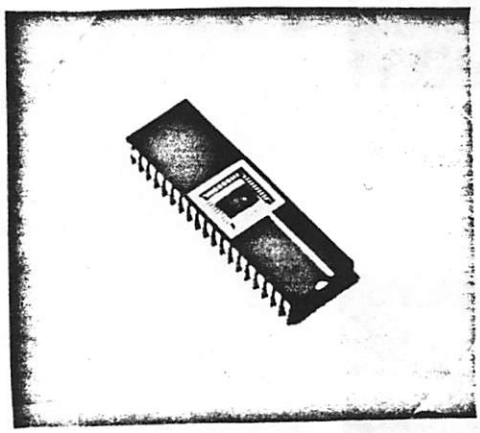
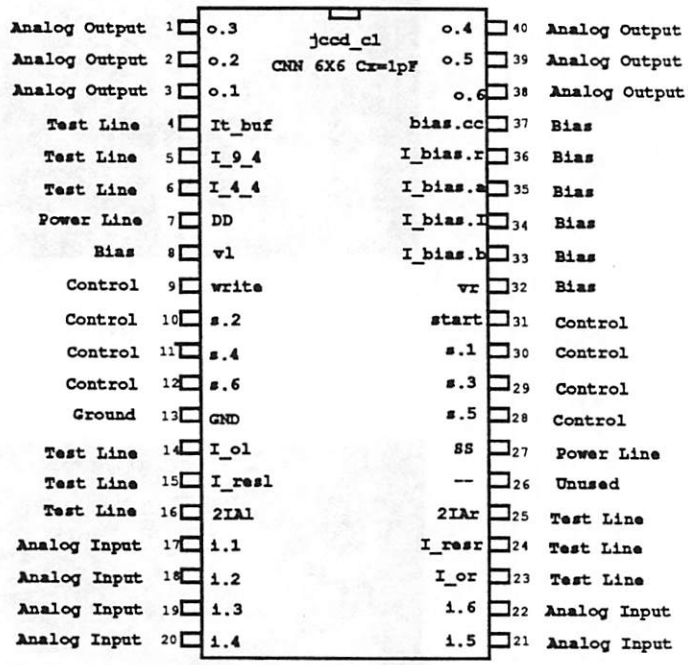


Figure 9

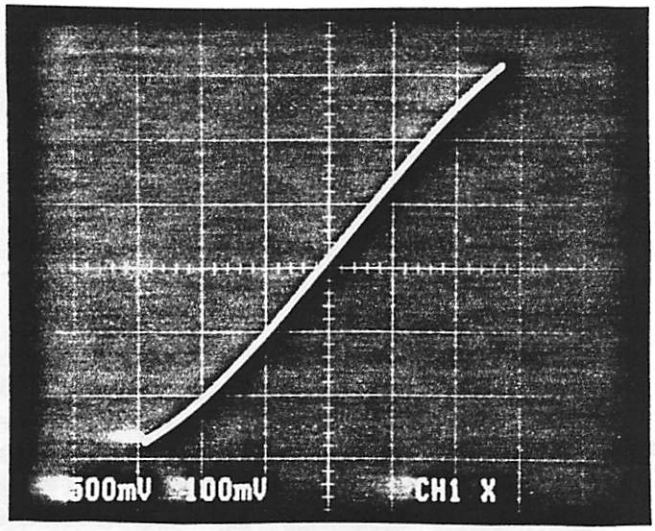


(a)

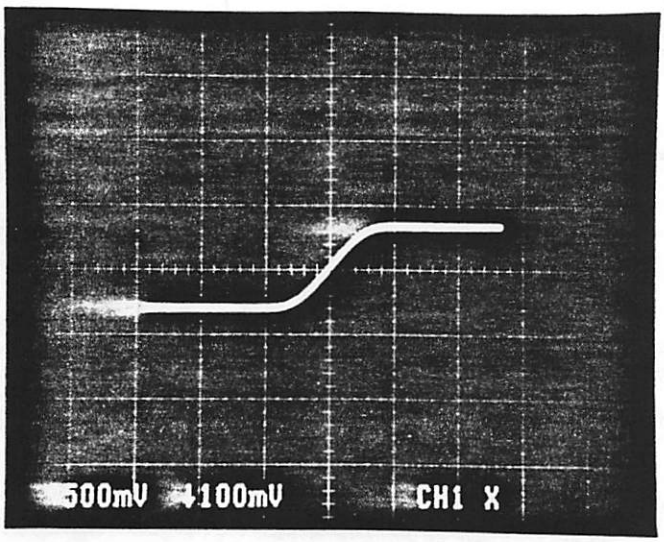

jccd_c1
CNN 6X6 Cx=1pF
Analog Output 1 o.3
Analog Output 2 o.2
Analog Output 3 o.1
Test Line 4 It_buf
Test Line 5 I_9_4
Test Line 6 I_4_4
Power Line 7 DD
Bias 8 vl
Control 9 write
Control 10 s.2
Control 11 s.4
Control 12 s.6
Ground 13 GND
Test Line 14 I_ol
Test Line 15 I_resl
Test Line 16 2IAl
Analog Input 17 i.1
Analog Input 18 i.2
Analog Input 19 i.3
Analog Input 20 i.4
o.4 40 Analog Output
o.5 39 Analog Output
o.6 38 Analog Output
bias.cc 37 Bias
I_bias.r 36 Bias
I_bias.a 35 Bias
I_bias.I 34 Bias
I_bias.b 33 Bias
vr 32 Bias
start 31 Control
s.1 30 Control
s.3 29 Control
s.5 28 Control
ss 27 Power Line
-- 26 Unused
2IAr 25 Test Line
I_resr 24 Test Line
I_or 23 Test Line
i.6 22 Analog Input
i.5 21 Analog Input

(b)

Figure 10



(a)



(b)

Figure 11

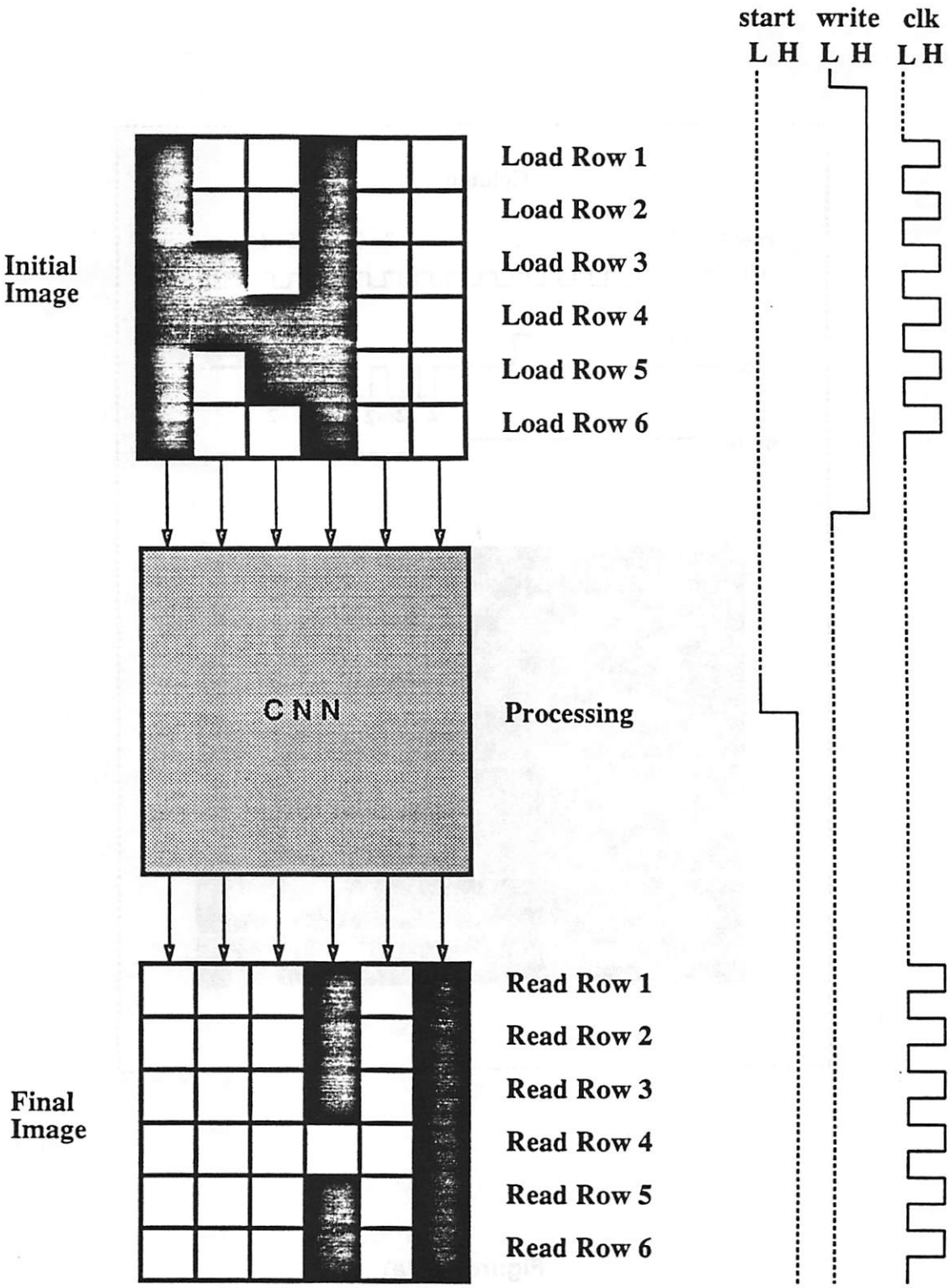

start
write
clk
L H
L H
L H
Initial Image
Load Row 1
Load Row 2
Load Row 3
Load Row 4
Load Row 5
Load Row 6
CNN
Processing
Final Image
Read Row 1
Read Row 2
Read Row 3
Read Row 4
Read Row 5
Read Row 6

Figure 12

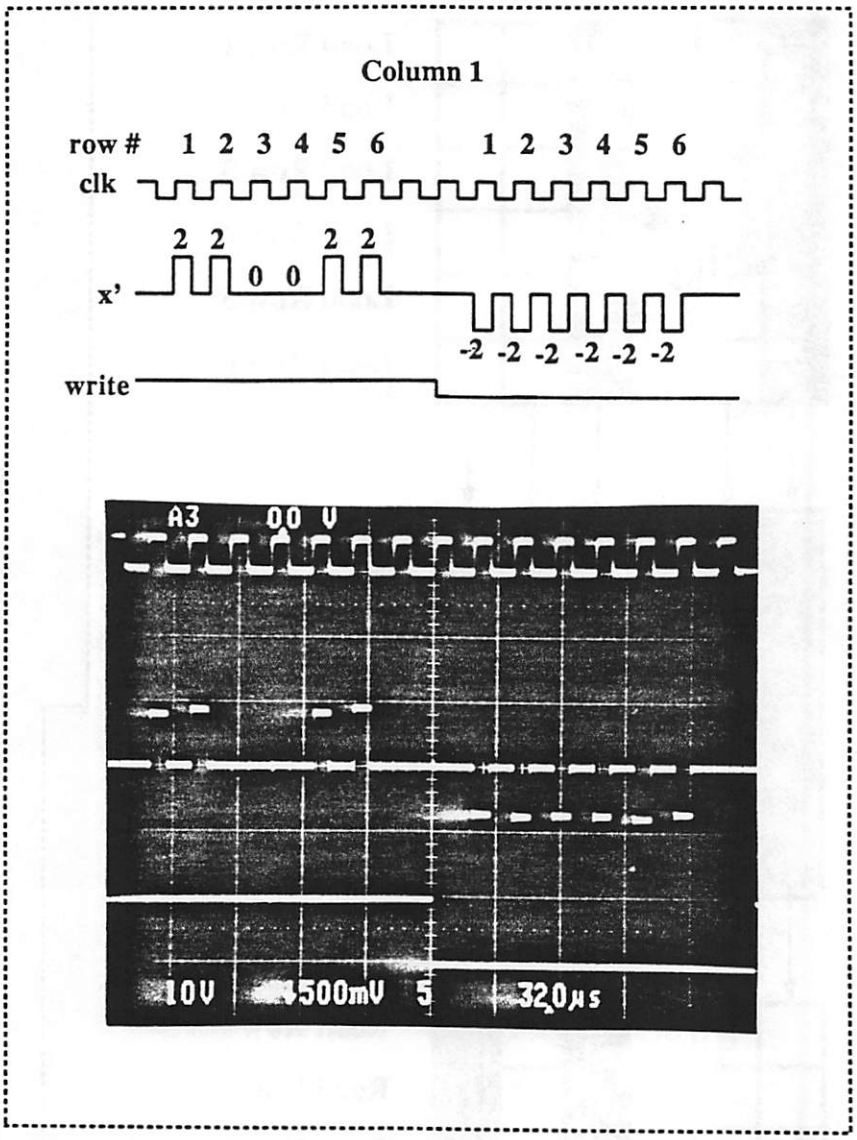

Column 1
row # 1 2 3 4 5 6 1 2 3 4 5 6
clk
2 2 2 2
0 0
x'
-2 -2 -2 -2 -2 -2
write
A3 00 V
10V 500mV 5 320µs

Figure 13 (a)

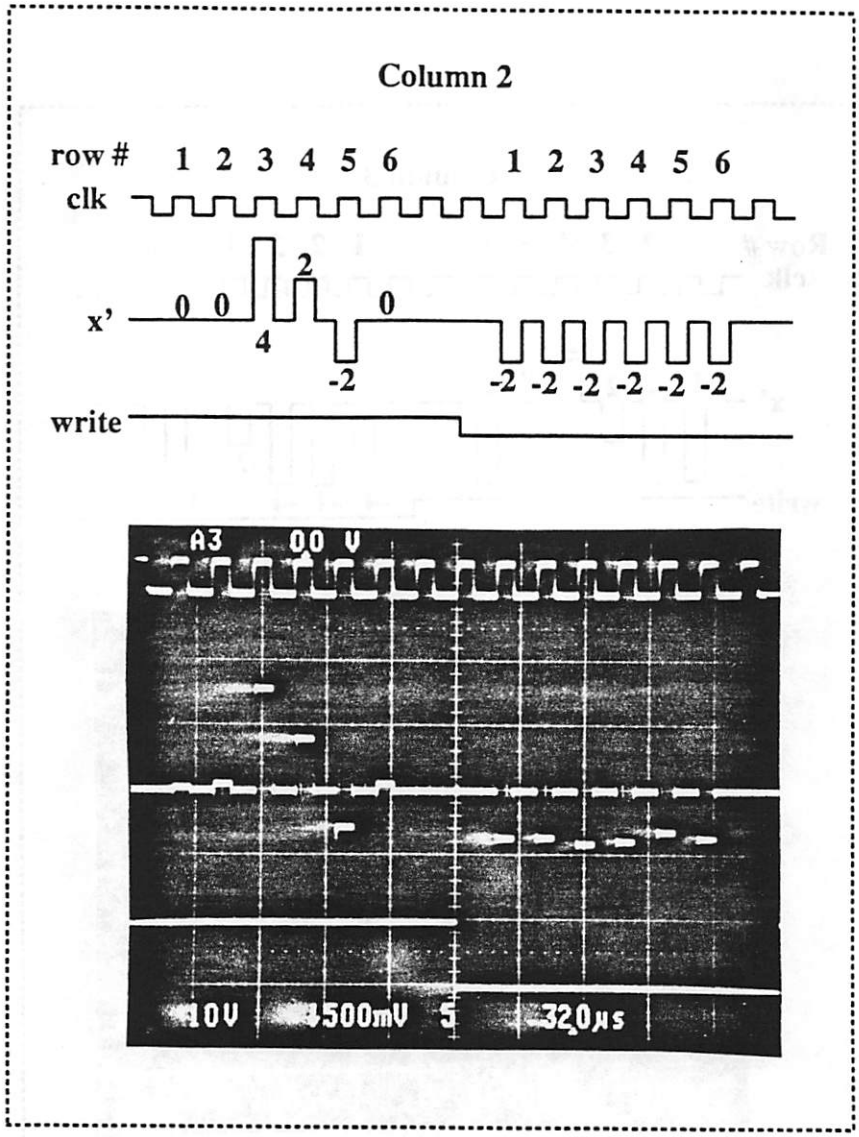

Column 2
row #
1 2 3 4 5 6
1 2 3 4 5 6
clk
x'
0 0
4
2
-2
0
-2 -2 -2 -2 -2 -2
write
A3
00 V
10V
500mV
5
32.0µs

Figure 13 (b)

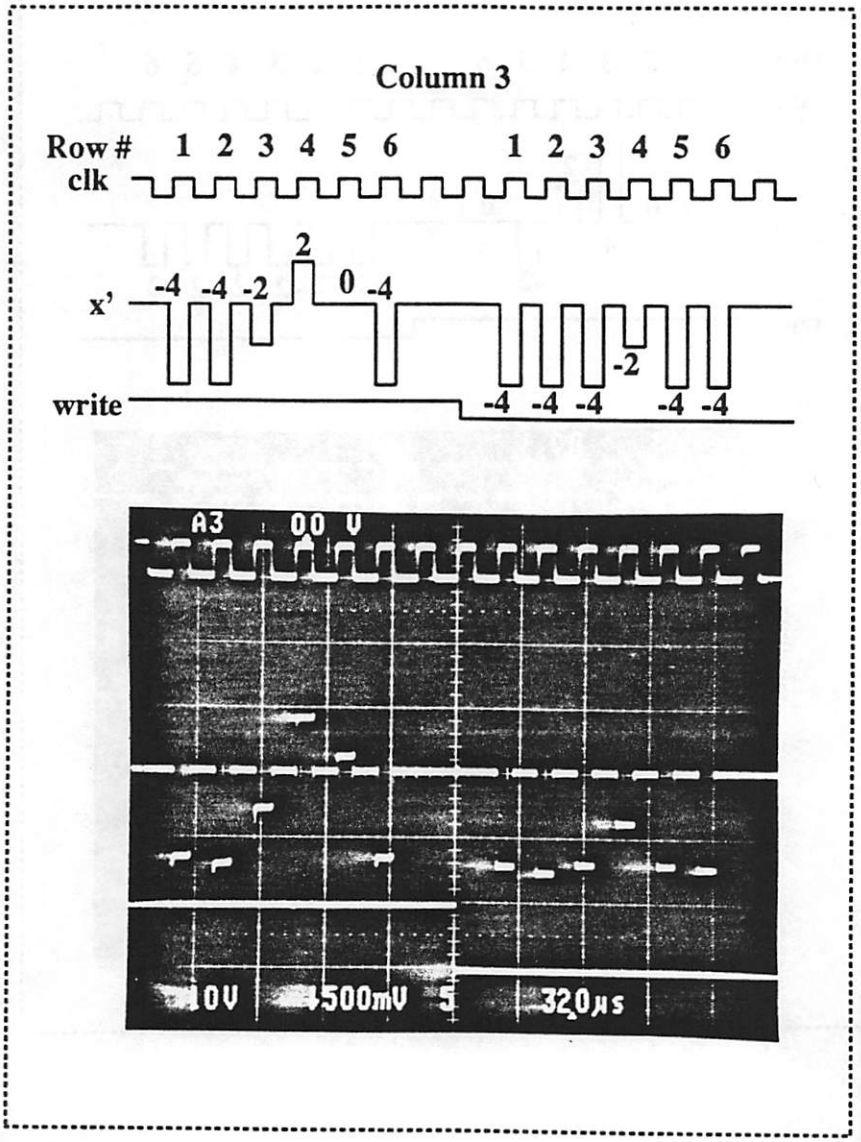

Column 3
Row #
1 2 3 4 5 6
1 2 3 4 5 6
clk
x'
-4 -4 -2 2 0 -4
-2
write
-4 -4 -4 -4 -4
A3
00 V
10V
500mV
5
320µs

Figure 13 (c)

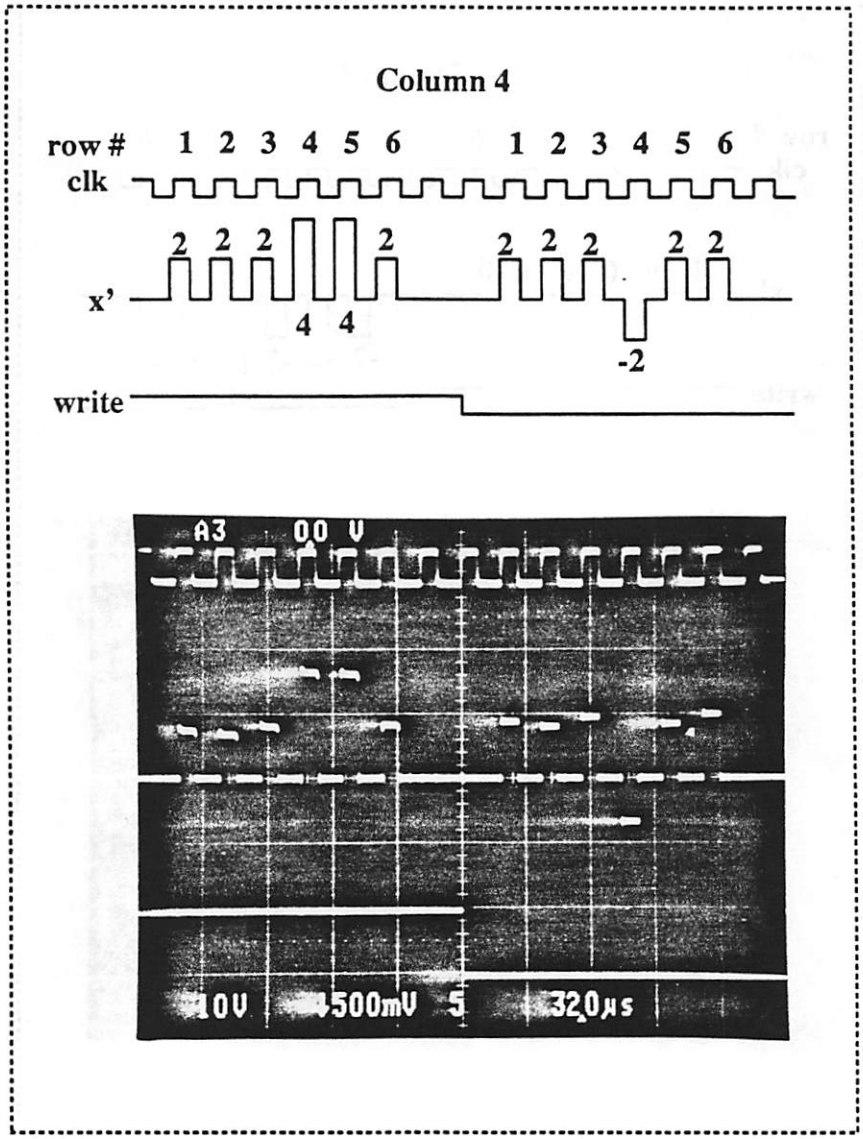

Column 4
row #
1 2 3 4 5 6
1 2 3 4 5 6
clk
x'
2 2 2 2 2 2 2 2 2 2
4 4
-2
write
A3 00 V
10V
500mV
5
320µs

Figure 13 (d)

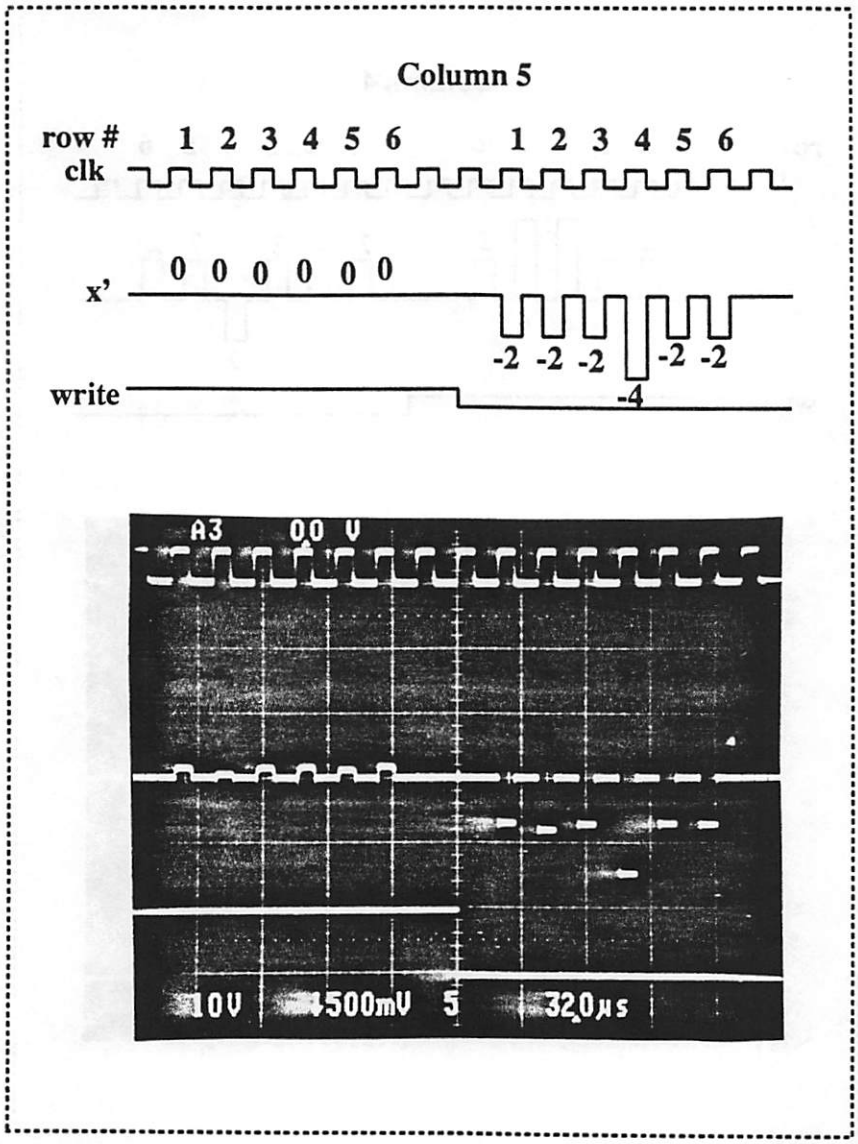

Column 5
row #
1 2 3 4 5 6
1 2 3 4 5 6
clk
x'
0 0 0 0 0 0
-2 -2 -2
-4
-2 -2
write
A3
00 V
10V
500mV
5
32.0µs

Figure 13 (e)

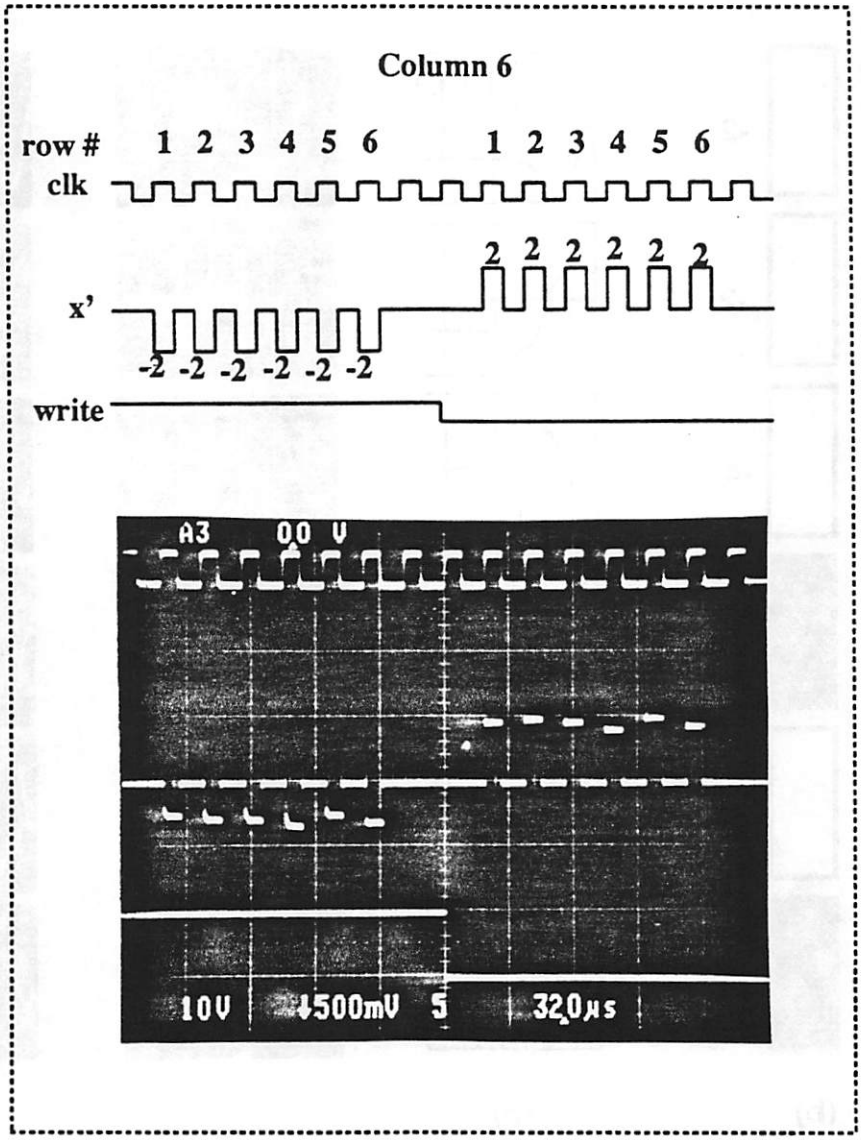

Column 6
row #
1 2 3 4 5 6
1 2 3 4 5 6
clk
2 2 2 2 2 2
x'
-2 -2 -2 -2 -2 -2
write
A3
00 V
10V
500mV
5
320µs

Figure 13 (f)

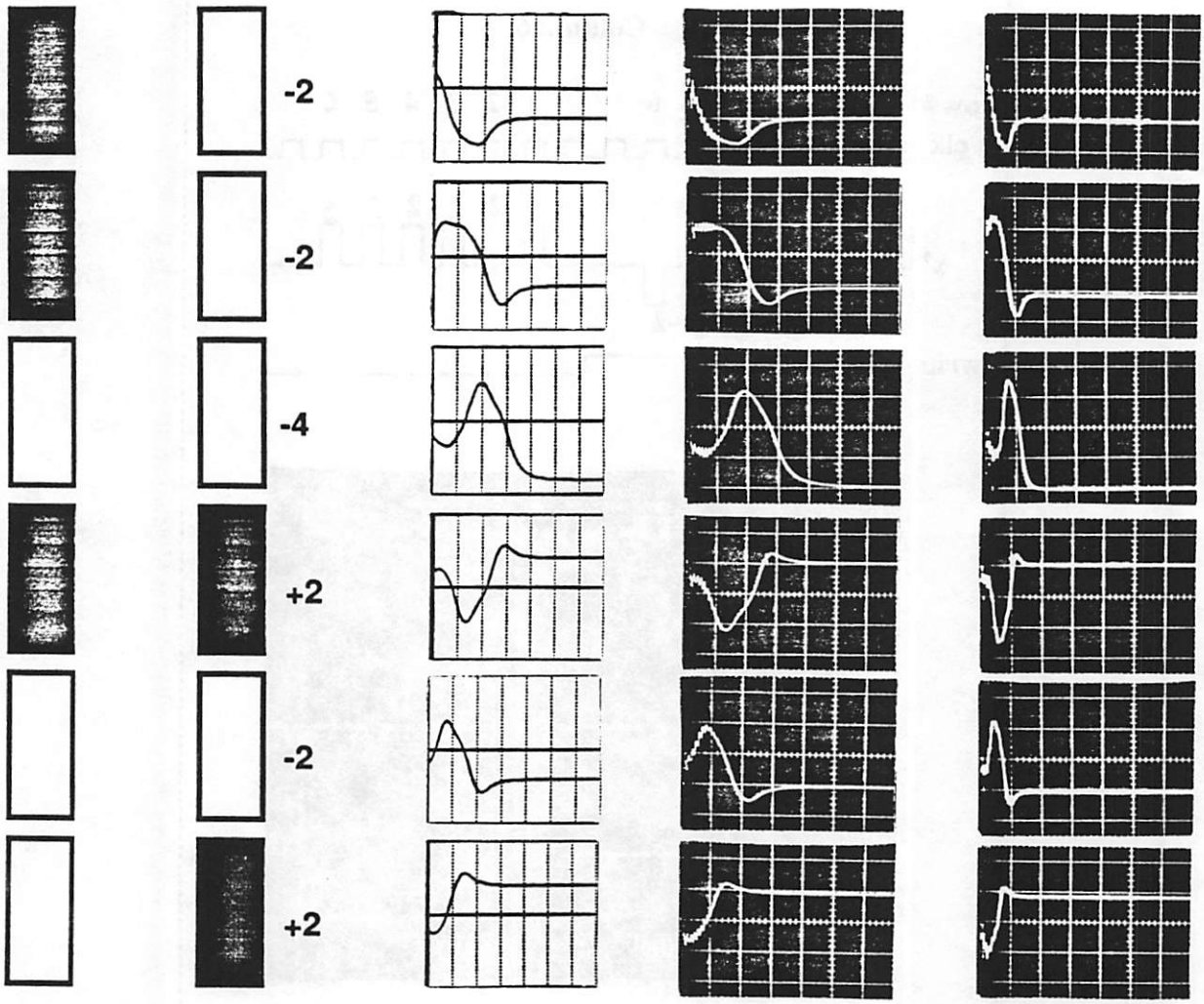


Figure 14