A Bio-Inspired Two-Layer Mixed-Signal Flexible Programmable Chip for Early Vision

Ricardo Carmona Galán, *Member, IEEE*, Francisco Jiménez-Garrido, Rafael Domínguez-Castro, Servando Espejo, *Member, IEEE*, Tamás Roska, *Fellow, IEEE*, Csaba Rekeczky, István Petrás, and Ángel Rodríguez-Vázquez, *Fellow, IEEE*

***Abstract*—A bio-inspired model for an analog programmable array processor (APAP), based on studies on the vertebrate retina, has permitted the realization of complex programmable spatio-temporal dynamics in VLSI. This model mimics the way in which images are processed in the visual pathway, what renders a feasible alternative for the implementation of early vision tasks in standard technologies. A prototype chip has been designed and fabricated in 0.5 μm CMOS. It renders a computing power per silicon area and power consumption that is amongst the highest reported for a single chip. The details of the bio-inspired network model, the analog building block design challenges and trade-offs and some functional tests results are presented in this paper.**

***Index Terms*—Cellular neural networks, machine vision, neural networks hardware, visual systems.**

I. Introduction

THE RETINA is found to be responsible for a rather involved treatment of visual information at early stages in the process of vision [1]–[3]. Through the close interaction of sensory and processing structures, complex spatio-temporal processes are realized in the retina which reduces the enormous amount of information associated to the visual flow into a data set of manageable size. Although retinas are not yet fully understood, and defines a challenging basic research area, the construction of vision processing devices with retina-like features shows large potential to overcome the limitations of conventional vision technologies. In that sense, during the last few years, several neuromorphic [4] vision chips have been developed and reported in literature. Some of these works are listed and examined in [5] and [6].

Recently, the behavior of the more external strata of the multi-layered structure of vertebrate retina has been successfully modeled by using the *C*ellular *N*eural *N*etwork (CNN) framework [7]. Such model has been based on studies and observations about the mammalian retina which have been recently published in Nature [3]. In this model, interactions between cells in the retinal fabric are realized on a *local* basis; each cell interacts with its *nearest neighbors*. Also, every cell belonging to the same layer has the same interconnection pattern. For each retinal layer, the same set of interconnection weights is applied to each and everyone of its cells; i.e., layers are *spatially-invariant*. In addition to this, the signals supporting intra- and inter-layer interactions are continuous in magnitude and time.

The phenomena observed in [3] are modeled in [7] by two coupled sets of two-dimensional (2-D) nonlinear differential equations. Because of the local interactions and the spatial-invariance, the behavior of such a model is fully described by some 25 parameters. This set of controlling parameters include interaction strengths, time constants and bias terms. By properly setting their values complex, interacting waves are generated which emulates the phenomena observed in the mammalian retina. This paper presents a fully-programmable *mixed-signal*[1] implementation of the model in [7] on a silicon chip. The chip, fabricated in a standard 0.5 μm CMOS technology, have a core composed of 32×32[2] elementary processors to implement the behavioral model, and embeds, in addition to this core circuitry, a set of circuit structures needed to render it a complete retina-like visual *microprocessor* system on a chip, namely, the following:

- establishing boundary conditions for the network dynamics;
- storing intermediate images, through 2-D short-term analog and digital memory banks;
- content-controlled and programmable intra-cell dataflow;
- global control and timing;
- addressing and buffering of the core cells;
- input–output;
- storing user-selectable analog and digital programming parameter configurations (for the coding of interaction weights and the setting of reconfiguration conditions);
- storing user-selectable instructions (programs) to control the sequence of operations of the processing core;
- controlling and timing the sequence of operations for the whole chip.

Manuscript received September 15, 2002. This work was supported in part by ONR Project N-000140210884, CE Project IST-1999-19007 (DICTAM) and the Spanish MCyT Project TIC1999-0826.

R. Carmona Galán, F. Jiménez-Garrido, R. Domínguez-Castro, S. Espejo, and A. Rodríguez-Vázquez are with the Instituto de Microelectrónica de Sevilla-CNM-CSIC, Campus de la Universidad de Sevilla, Sevilla 41012, Spain (e-mail: rcarmona@imse.cnm.es).

T. Roska, I. Petrás, and C. Rekeczky are with the Analogic and Neural Computing Laboratory, Computer and Automation Institute of Hungarian Academy of Science, Budapest H-1111, Hungary.

Digital Object Identifier 10.1109/TNN.2003.816377

[1]The circuit embodies both analog and digital circuitry. Analog techniques are used to implement the core behavioral model, including the dynamic operation and the interactions among cells, while digital techniques are employed to control the operation of the chip and to make the control interfacing to the external world.

[2]Each elementary processor is double, due to the layered structure of the chip. Hence, it is more correct saying that the new chip includes $2 \times (32 \times 32)$ horizontally and vertically interacting processors.

The integrated system belongs to the category of the so-called *S*ingle-*I*nstruction *M*ultiple-*D*ata processors [8], although works directly on analog signal representations. It reports significant advantages in terms of area and power efficiency. For instance, leaving aside the resources needed to obtain the digital image representations used by the chip in [8], it features 4 $\mathrm{MOPS/mm^2}$ (OPS: *OP*erations per *S*econd) and $1\mathrm{MOPS/mW}$; while the chip in this paper features 6 $\mathrm{GOPS/mm^2}$ and 1.56 GOPS/mW. In addition to that, and to the best of our knowledge, no other SIMD microprocessor-on-a-chip with retina-like behavior has been reported to date, although a number of remarkable, pioneering vision chips have been successfully implemented as for instance those listed in [6].

This paper is organized as follows. Section II is dedicated to the bio-inspired network models, the foundations of the mathematical network model in a sketch of the biological retina. Section III describes the architecture of the APAP chip and its main components. Section IV explains how the analog building blocks of the basic processing units have been designed. Section V reviews the peripheral circuitry design. Experimental results obtained from testing a prototype chip are shown in Section VI. Finally, Section VII displays some conclusions.

II. Bio-Inspired Network Model

A. Sketch of the Vertebrate Retina

Due to the vast amount of information contained in the visual stimuli, nature has developed a specialized part of the nervous system to handle it: the retina. On one side, the neuronal impulses conveying information along the nerves do not support such a large data rate. On the other side, because of the high correlation found between the elements of the image —most of the energy of the signal, in images displaying natural scenes, is concentrated in the lower spatial and temporal frequencies—, not every bit of information has to be passed to the brain to accomplish vision. Therefore, the retina, brought to the sensory periphery instead of being integrated in the central nervous system, processes the visual information at the focal plane, realizing what is called early vision. By performing so, the data flow to the visual cortex is greatly reduced, thus solving the problem of intelligent processing of visual information in a tight time frame.

The vertebrate retina has the structure displayed in Fig. 1 [9]. A first layer of photodetectors at the outermost layer of the retina, the cone cells—a different type of cell, the rods, are specialized in sensing in very dim light conditions and saturate very easily, captures light and converts it to activation signals. Bipolar cells carry these signals across the retina layers to the ganglion cells that interface the retina with the optical nerve, in a trip of several micrometers [3]. The ganglion cells convert the continuous activation signals, proper of the retina, to spike-coded signals that can be transmitted over longer distances by the nervous system. On the way to the ganglion cells, the information carried by bipolar cells is affected by the operation of the horizontal and amacrine cells. They form layers in which activation signals are weighted and promediated in order

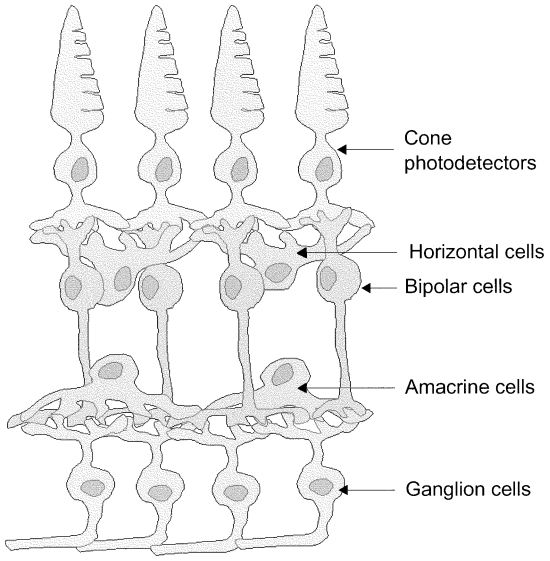


Fig. 1. Schematic diagram of the vertebrate retina [9] showing the layer of photosensors at the top and the ganglion cells connecting to the optic nerve.

to, first, bias photodetectors and, second, to account for inhibition on the vertical pathway. The four main transformations that take place in this structure are: the photoreceptor gain control, the gain control of the bipolar cells, the generation of transient activity and the transmission of transient inhibition. Briefly, captured stimulus are promediated and the high-gain characteristics of the cones and the bipolar cells are shifted to adapt to the particular light conditions. These operations have a local scope and depend on the recent history of the cells. Once adaptation is achieved, patterns of activity are formed dynamically by the presence or absence of visual stimuli. Also inhibition is generated and transmitted laterally through the layers of horizontal and amacrine cells. As a result of these transformations, the patterns of activity reach the layer of ganglion cells. At this point, the patterns are converted into pulse-coded signals that are sent to the brain to be interpreted. In a sense, the layered structure of the retina translates the visual stimuli into a compressed language that can be understood by the brain in recreating vision.

B. CNN Analogy of the Inner and Outer Plexiform Layers

There are, in this description, some interesting aspects of the retinal layers that markedly resemble the characteristics of a CNN: the 2-D aggregation of continuous signals, the local connectivity between elementary nonlinear processors, the analog weighted interactions between them. Also, the complete signal pathway in the retina have the topology of a 3-D, or more properly, two-and-a-half dimensional pile of 2-D layers connected vertically network. Motivated by these coincidences, and based on physiological and pharmacological studies [2], a CNN model has been developed that approximates the observed behavior of the vertebrate retina [10].

The outer plexiform layer of the retina, OPL, is responsible for the image capture. It has been characterized by experimental measurements [11], leading to a model with three different layers of cells. The first one, the photosensing layer, consists in an aggregation of cone cells. It is assumed here that the retina is adapted to lighting conditions and so the rods are saturated

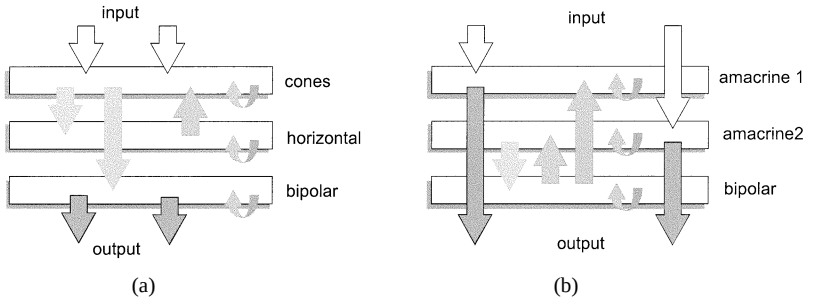


Fig. 2. Conceptual diagram of the (a) OPL of the retina and (b) the wide-field activity in the IPL.

and remain silent. In addition to the layer containing the cones, there is a second layer composed of horizontal cells and a third one composed of bipolar cells. Each of these layers has the structure of a 2-D CNN itself. Each of them has its own interaction patterns (CNN templates) and its particular time constant. Cell dynamics are sustained by a first or a second order core. The structure of the OPL is depicted in Fig. 2(a), where interactions between layers of cells are represented by arrows. The input signal is captured by the cones and feedforward to the layers of horizontal and bipolar cells. From the experiments it has been concluded that no feedforward connection exists between the horizontal cells and the layer of the bipolar cells. No feedback has been observed neither from the output of the bipolar cells to the previous layers. It has been deduced that the feedback connection of the horizontal cells to the layer of cones acts as a modulator of the feedforward functions rather than affecting directly to the cones state. This feature, that is not implemented in this chip, is realized in [12].

Regarding the inner plexiform layer, IPL, it is responsible for the generation of the retinal output. A simplified model of the IPL is described in [11]. It has three layers of cells and supports the so called wide field activity, observed in certain amacrine cells. Wide field activity consists in the integration of the action potentials along a widely extended area previous to the ganglion cells. Based on the experimental records, the model consists in two layers of wide field amacrine cells excited by the input signal, which in this occasion is the output of the bipolar cells, and a third layer that controls the dynamic of the previous layers by means of feedback signals. As before, the three layers are supposed to be 2-D CNNs with their own internal coupling and their own time constant [see Fig. 2(b)].

Because of the relative simplicity of these models, a programmable CNN chip has been proposed [12]. The programmable array processor of the chip consists in two coupled CNN layers, and a third layer, of a much faster dynamics ($\tau_3 \ll \tau_1, \tau_2$) that supports analog arithmetic [see Fig. 3]. Each elementary processor contains the nodes for both CNN layers. The third layer is inherently implemented by these analog cores, with the help of the local facilities for analog signal storage. The evolution of the coupled CNN nodes of a specific cell $C(i,j)$ is described by these coupled differential equations as shown in (1) at the bottom of the page where the nonlinear losses term and the output function in each layer are those of the FSR CNN model [14]:

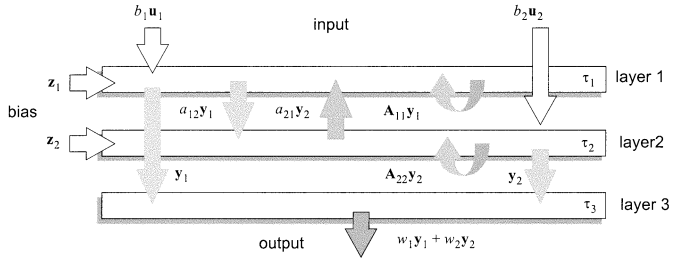


Fig. 3. Conceptual diagram of the second-order three-layer CNN.

$$g(x_{n,ij}) = \lim_{m\to\infty} \begin{cases} m(x_{n,ij}-1)+1 & \text{if } x_{n,ij} > 1 \\ x_{n,ij} & \text{if } |x_{n,ij}| \le 1 \\ m(x_{n,ij}+1)-1 & \text{if } x_{n,ij} < -1. \end{cases} \quad (2)$$

and

$$y_{n,ij} = f(x_{n,ij}) = \frac{1}{2}(|x_{n,ij}+1| - |x_{n,ij}-1|) \quad (3)$$

Fig. 4 depicts the block diagram of the vertically coupled CNN nodes. Synaptic connections between cells are linear. Each CNN layer incorporates feedback connections, by means

$$\begin{aligned} \tau_1 \frac{dx_{1,ij}}{dt} =& -g[x_{1,ij}] \\ &+ \sum_{k=-r_1}^{r_1} \sum_{l=-r_1}^{r_1} a_{11,kl} y_{1,(i+k)(j+1)} + b_{11,00} u_{1,ij} + a_{12} y_{2,ij} + z_{1,ij} \\ \tau_2 \frac{dx_{2,ij}}{dt} =& -g[x_{2,ij}] \\ &+ \sum_{k=-r_2}^{r_2} \sum_{l=-r_2}^{r_2} a_{22,kl} y_{2,(i+k)(j+1)} + b_{22,00} u_{2,ij} + a_{21} y_{1,ij} + z_{2,ij} \end{aligned} \quad (1)$$

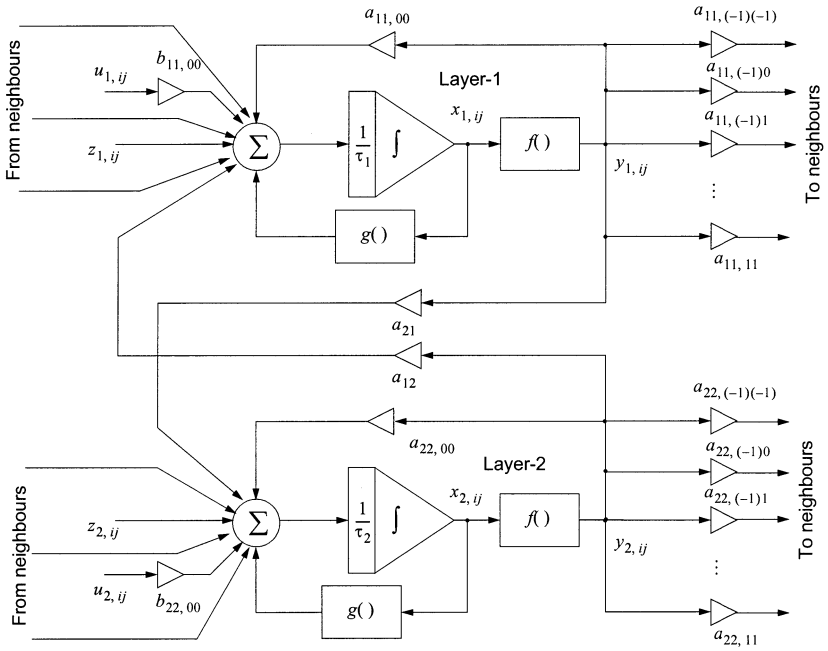


Fig. 4. Block diagram of the two coupled CNN layer nodes.

of which the output of each cell contributes to the state of its neighbor, weighted by the elements $\{a_{nn,kl}\}$; a feedforward connection, weighted by $b_{nn,00}$, that regulates the contribution of the cell's input; a bias term $z_{n,ij}$, that can be different for each cell; and finally coupling connections between both layers, weighted by a_{21} and a_{12}. Each layer has its own time-constant τ_n. Programming different dynamics in this CNN model is possible by adjusting the template elements and the time-constants of the layers. The total number of synapses to be implemented on each cell is 22, plus the 2 bias maps multipliers, which will be treated as a second input image for each layer.

III. APAP CHIP ARCHITECTURE

A. Analog Programmable Array Processor Chip

The proposed chip consists in a mixed-signal parallel processing array of 32×32 identical cells [see Fig. 5]. It is surrounded by a ring of circuits implementing the boundary conditions for the CNN dynamics. The peripheral circuitry, required for the proper operation of the central array processor consist in the timing and control unit, the program memory and the I/O interface.

The timing and control unit is composed by a micro-instruction decoder, generating the appropriate signals to configure the network, and an internal clock/counter with a set of finite state machines that generate the internal signals that enable program memory accesses and other data transfers. The operation control unit constitutes the interface between the program memory and the processing array. The program memory is composed, on one side, of 16 blocks of SRAM of 64 bytes of capacity dedicated to the analog weights, and four blocks of 128 bytes each for the logic program, including bits for the network configuration and control signals for the I/O interface. Digital signals buffering can be considered part of the operation control unit. In addition, the analog instructions and reference signals, codified in one section of the program memory, need to be transmitted to every cell in the network in the form of analog voltages. Thus, a bank of D/A converters interfaces these memory blocks with the processing array. Distributing analog references across large distances within a chip is not a trivial task. Apart from the problems derived from electromagnetic interference, voltage drops in long metal lines carrying currents can be quite noticeable. Signal buffering and low-resistance paths must be provided to avoid this, especially in the case of weights, that enter the synapses through a low impedance node.

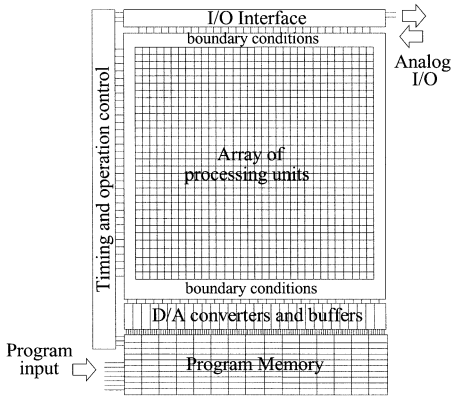


Fig. 5. Floorplan of the prototype chip.

Finally, the image I/O interface consists in a serializing-deserializing analog multiplexor. It accommodates the serial analog I/O channel to the 32 I/O lines corresponding to the 32 columns of the array by means of a battery of blocks. The corresponding row and column address decoders, controlled by the timing unit, are part of this block.

B. Basic Cell Structure

The basic cell of the CNN-based array processor has a similar architecture to that of the CNN universal machine cells

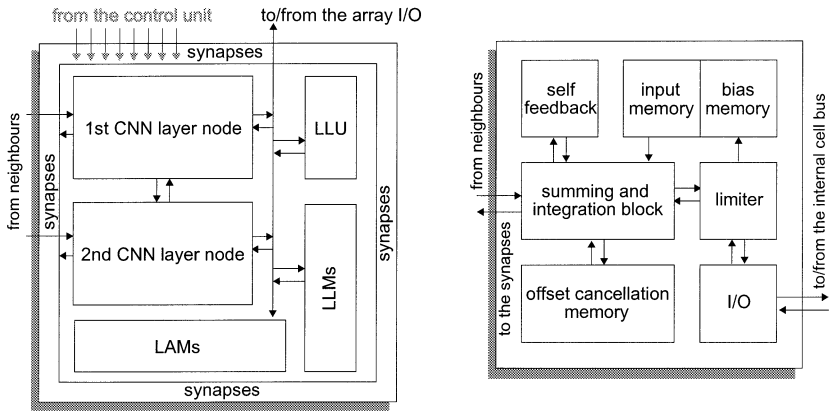


Fig. 6. Conceptual diagram of the (a) basic cell and the (b) internal structure of each CNN layer node.

[15]. However, in this occasion, the prototype includes two different continuous-time CNN layers. Therefore, as depicted in Fig. 6(a), together with the local analog and logic memories (4 LAMs and 4 LLMs), for to the storage of intermediate results, the local logic unit (LLU), responsible for pixel-level logic operations, two different analog CNN core blocks are found, each one belonging to one of the two different CNN layers implemented. The synaptic connections between processing elements of the same layer are built around the cell core, as shown, while interlayer coupling, kept within the pixel scope in this model, is placed inside the cell (represented by arrows between the processing layers in the diagram). All the blocks in the cell communicate via an intracell data bus, which is multiplexed to the array I/O interface. Control and cell configuration bits are passed directly from the control unit.

The internal structure of each of the CNN cores of the cell is depicted in the diagram of Fig. 6(b). Each core receives contributions from the rest of the processing nodes in the neighborhood which are summed and integrated in the state capacitor. The two layers differ in that the first layer has a scalable time constant, controlled by the appropriate binary code, while the second layer has a fixed time constant. The evolution of the state variable is also driven by self- feedback and by the feedforward action of the stored input and bias patterns. There is a voltage limiter which helps to implement the limitation on the state variable of the FSR CNN model. This state variable is transmitted in voltage form to the synaptic blocks, in the periphery of the cell, where weighted contributions to the neighbors' are generated. There is also a current memory that will be employed for cancellation of the offset of the synaptic blocks. Initialization of the state, input and/or bias voltages is done through a mesh of multiplexing analog switches that connect to the cell's internal data bus.

Running complex spatio-temporal dynamics in this network requires following several initialization and calibration steps. First of all, acquisition of the input image and auxiliary masks and/or patterns. For this purpose, the array I/O interface is directed to specific LAM locations in a row-by-row basis. After that, the analog instruction, i.e., the set of synaptic weights required for a specific operation, is selected and transmitted to all the cells in the array. Then, the offset of the critical OPAMPs is quenched in a calibration step. After that, the time-invariant offsets of the synaptic blocks are computed and stored in the current memories. Now the network is almost ready to operate. Then, the state capacitors and the feedforward synapses are initialized by means of the appropriate switch configuration, and the network evolution is run by closing the feedback loop in each processing element. Before stopping the network evolution, the final state is stored in a LAM register for further operation.

IV. The Basic Processing Unit

A. Single-transistor Synapse

One of most important blocks in the cell is the synaptic block. The synapse is, *simply*, a four- quadrant analog multiplier. Their inputs are the cell state, V_x, or input, V_u, variables and the corresponding weight signal, V_w, while the output is the cell's contribution to a specific neighboring cell. The multiplier is required to have voltage inputs and current output. On one side, both the cell state and the weight signal, the multiplier inputs, must be distributed over different points in the circuit. The cell state must drive every synapse in the local scope and the weight signal must be transmitted to every cell in the array. If these signals are represented by voltages, they can be easily conveyed to any high-impedance node by a simple wire. On the other side, because the contributions of all the neighbors are summed at the input of the processing core, this summation can be readily achieved by wiring all these contributions concurrently to a low-impedance node if they are in current format. In addition to this, in this particular application, there is no need to have a strictly linear relation between the weight signal, V_w, and the output current, I_o. Moreover, one thing that is common in this type of processing is that the weight signal does not change during the evolution of the network. It means that any deviation depending on V_w is not a gain error, but an offset error, i.e., an error which can be cancelled by autozeroing in a preprocessing calibration step.

Different CMOS compatible circuits can be employed to realize the multipliers. For instance, synapses can be implemented by MOS transistors in weak inversion [16], exploiting the exponential law that governs this regime of operation to achieve multiplication. There are multipliers based on MOS transistor in strong inversion, operating in the saturation region, where their large-signal characteristic exhibits a quadratic law, which is the

principle behind the well- known Gilbert cell [17]. Direct multiplication can also be achieved by a MOS transistor operating in the ohmic region. Its low-frequency large-signal characteristic is given into first-order approach by (if n-type)

$$I_{DS} = \beta_n \left[V_{GS} - V_T(V_{SB}) - \frac{V_{DS}}{2} \right] V_{DS} \tag{4}$$

where $\beta_n = \mu_o C_{ox}{}'(W/L)$. A multiplication can be realized with this device as long as $V_{DS} \ll 2[V_{GS} - V_T(V_{SB})]$ holds [18]. This alternative has several advantages [19]: it requires a reduced amount of area, because four-quadrant behavior is achieved with one single transistor. Second, it has a better relation between bias power and signal power, thus leading to higher accuracy at lower power consumption, while in the saturation region the information is carried by a small fraction of the actual currents flowing through the devices. Third, the use of the ohmic region shows better mismatch figures than any other region [20].

The one-transistor synapse works as follows. Consider a p-type MOS transistor operating in ohmic region [see Fig. 7]. The transistor is selected type p because the more resistive p-type channel allows smaller currents, and so power consumption, for the same transistor lengths. Or, equivalently, for the same current levels, the required p-channel MOS is shorter than its n-type counterpart. The source-to-drain current of a PMOS transistor in the ohmic region is given by [21]

$$I_o = -\beta_p (V_A - V_L) V_G - \beta_p (V_A - V_L) \left(\left| V_{T_p} \right| - \frac{V_A + V_L}{2} \right) \tag{5}$$

where the threshold adopts one of these two analogue forms:

$$V_{T_p} = \begin{cases} -\left|V_{T0_p}\right| - \gamma \left(\sqrt{\phi_B + V_{\mathrm{DD}} - V_A} - \sqrt{\phi_B} \right) & \text{if } V_A \geq V_L \\ -\left|V_{T0_p}\right| - \gamma \left(\sqrt{\phi_B + V_{\mathrm{DD}} - V_L} - \sqrt{\phi_B} \right) & \text{if } V_A \leq V_L \end{cases} \tag{6}$$

V_L must be kept fixed in order to use V_A and V_G as single-ended input voltages, and to sense I_o as the output of the synapse. For this purpose, we can employ a current conveyor [22] at the current input node of each cell. The current conveyor permits current sensing while maintaining a virtual reference at node Ⓛ. All the synapses contributing to the same cell can be connected to the same virtual reference. The only objection being that the impedance seen at this node must be well below the parallel of the output impedances of all the synaptic blocks.

Back to (5), notice that the second term in the right side of the equation does not depend on V_G, therefore node Ⓖ is a strong candidate to hold the cell state variable voltage. But V_G must be always positive for the MOS transistor to operate above threshold, thus let V_G be composed of a reference voltage V_{x_0}, sufficiently high, and a superposed cell state signal V_x

$$V_G \equiv V_X = V_{x_0} + V_x. \tag{7}$$

And, in order to achieve four-quadrant multiplication, V_A must be permitted to go up and below V_L. Let us select V_L as the reference for the weight signal, V_{w_0}, being

$$V_A \equiv V_W = V_{w_0} + V_w. \tag{8}$$

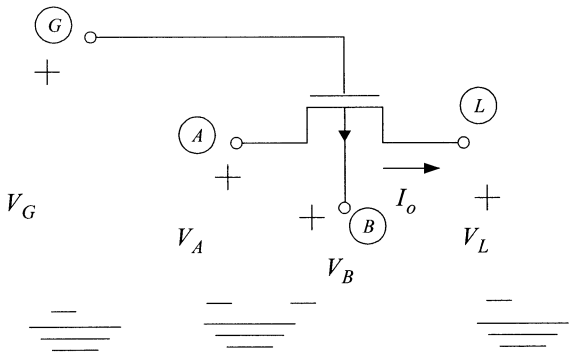


Fig. 7. Multiplier using one single MOS transistor in ohmic region.

Then (5) can be rewritten as

$$I_o = -\beta_p V_w V_x - \beta_p V_w \left(V_{x_0} + \left| V_{T_p} \right| - V_{w_0} - \frac{V_w}{2} \right) \tag{9}$$

which is a four-quadrant multiplier with an offset term that is time-invariant—at least during the evolution of the network—and not depending on the cell state. Therefore, we have arrive to a four-quadrant multiplier with single-ended voltage inputs and a current output, with a offset that can be eliminated by a calibration step, with the help of a current memory

$$I_o = -\beta_p V_w V_x + I_{\mathrm{offset}}(V_w). \tag{10}$$

The limitations found to this behavior are, in the first order, the upper and lower boundaries of the ohmic region in strong inversion [21]. From them, it can be concluded that

$$V_{x_0} \leq V_{w_0} - V_{w_{\max}} - V_{x_{\max}} - \left| V_{T_p} (V_{\mathrm{DD}} - V_{w_0}) \right|. \tag{11}$$

Another restriction is found in the degradation of the mobility. The transversal electric field (normal to the surface of the channel) pushes the carriers toward the semiconductor surface where they suffer scattering, which renders a reduction in the speed of the carriers, thus degrading the mobility. This transversal electric filed depends on the gate voltage, thus the first summand in (10) will no longer be linear with V_x. Using a widely accepted model for this effect in a MOS transistor [21], we arrive to

$$V_{w_0} + V_{w_{\max}} + V_{x_{\max}} - \left| V_{T_p} (V_{\mathrm{DD}} - V_{w_0} - V_{w_{\max}}) \right| \leq V_{GE_{\max}} + V_{x_0} \tag{12}$$

where $V_{GE_{\max}} = (V_{SG} - |V_{TP}(V_{SB})|)\,|_{\max}$ is a maximum effective gate voltage, beyond which the distortion introduced by mobility degradation exceeds the linearity requirements. Combining these two equations:

$$V_{w_{\max}} + V_{x_{\max}} \leq \frac{1}{2} V_{\mathrm{GE}} - \frac{1}{2} \left[\left| V_{T_p} (V_{\mathrm{DD}} - V_{w_0}) \right| - \left| V_{T_p} (V_{\mathrm{DD}} - V_{w_0} - V_{w_{\max}}) \right| \right]. \tag{13}$$

For moderate linearity requirements, in a typical CMOS technology, the right hand side of (13) becomes approximately equal to 1 V. If V_x and V_w are assigned the same voltage ranges, ± 400 mV around their reference values, then

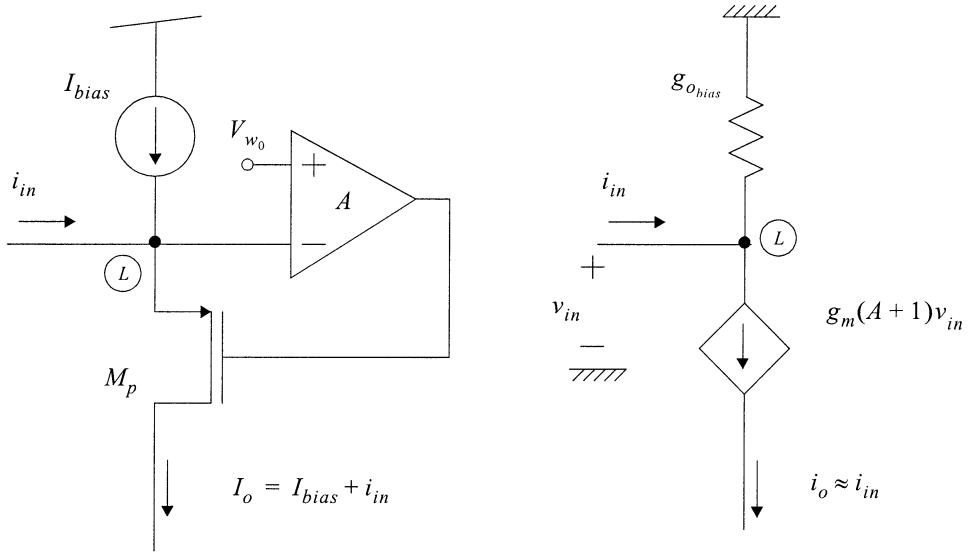


Fig. 8. Current conveyor realization and small-signal equivalent.

$V_{x_{\max}} = V_{w_{\max}} = 400\,\mathrm{mV}$. With this, back to (11), substituting the values of $V_{x_{\max}}$, $V_{w_{\max}}$ and $|V_{T_p}| \approx 0.8\,\mathrm{V}$

$$V_{x_0} \leq V_{w_0} - 1.6\,\mathrm{V}. \tag{14}$$

Thus, V_{w_0} must be high enough to leave room for V_{x_0}, but not too large because the weight signal will progress up to $V_{w_{\max}}$ above V_{w_0}. In addition, we have to provide range for the current conveyor circuitry to maintain a virtual reference precisely at V_{w_0}, and for the circuits generating the weight voltages, which will have a limited output swing. If we select $V_{w_0} = 2.55\,\mathrm{V}$, then there are 0.75 V above V_{w_0} before hitting the power rail at 3.3 V, what means one $|V_{T_p}|$, approximately. With this value, V_{x_0} results in 0.95 V. Finally, once the voltage ranges are fixed, a maximum current per synapse is selected for meeting power requirements, in our case it will be 1.4 μA. With these values, the synapse is dimensioned. In our chip, it will be 2 $\mu\mathrm{m}$ wide and 2.59 $\mu\mathrm{m}$ long.

B. Current Conveyor

The current conveyor, required for creating a virtual reference node at which the synapses outputs can be sensed, is implemented by the circuit of Fig. 8. Any difference between the voltage at node Ⓛ and the reference V_{w_0} is amplified and the negative feedback corrects the deviation.The input impedance of this block is very low, what means that changes in the small-signal input current Δi_{in} does not disturb appreciably the virtual reference at node Ⓛ, this is $\Delta\nu_{in} \approx 0$. The bias current is required to ensure that node Ⓛ is always the source of transistor M_p. At the same time, this circuit permits the injection of a nearly exact copy of the input current at the state node, whose voltage range differs from that of the weight signals. The only drawback of using this circuit is that a voltage offset, V_{OS}, at the input of the differential amplifier—which can be implemented with a simple OTA as it drives a very high impedance node, the gate of M_p—results in an error of the same amount in the reference voltage implemented at node Ⓛ. Since the main contribution to the offset is random, this error will be distributed all along the array resulting in mismatched synaptic blocks that can degrade performance, e.g., anisotropic evolution of the network yielded by a symmetrical propagation template. As we are impelled to use small-size devices, in order to achieve the highest cell-packing density possible, the random offset can be quite large. In order to avoid this, an offset calibration mechanism has been implemented at the critical OTAs [see Fig. 9]. The input referred offset voltage, V_{OS}, has been taken out of the OTA block symbol. Without the offset cancellation circuit (the shadowed area), at low frequencies, and considering a negligible output conductance, the output of the OTA is

$$I_o = g_m\left(\nu_d + V_{OS}\right). \tag{15}$$

Considering the error cancellation mechanism, when ϕ_{cal} is ON, then the inputs are shorted, $\nu_d = 0$, and M_{mem} is connected as a diode, its source-to-drain is in steady state

$$I_{\mathrm{mem}} = I_B - g_m V_{OS}. \tag{16}$$

After some time, ϕ_{cal} is turned off and, except from a remnant switching error, the current I_{mem} is memorized by means of the voltage stored in C_{mem}. Thus, the total current injected into the load is free of any offset:

$$I_L = I_o + I_{\mathrm{mem}} - I_B = g_m\nu_d. \tag{17}$$

C. Current Memory

As it has been mentioned, the offset term of the synapse current must be removed for the output current to precisely represent the result of a four-quadrant multiplication. For this purpose, before the CNN operation, but right after the new weights has been uploaded, all the synapses are reset to $V_X = V_{x_0}$. Then the resulting current, which is the sum of the offset currents of all the synapses concurrently connected to the same node, is memorized. This value will be subtracted on-line from the input current during the network evolution, resulting in a one-step cancellation of the errors of all the synapses. The validity of this method relies in the accuracy of the current memory. For instance, in this chip, the sum of all the contributions will range

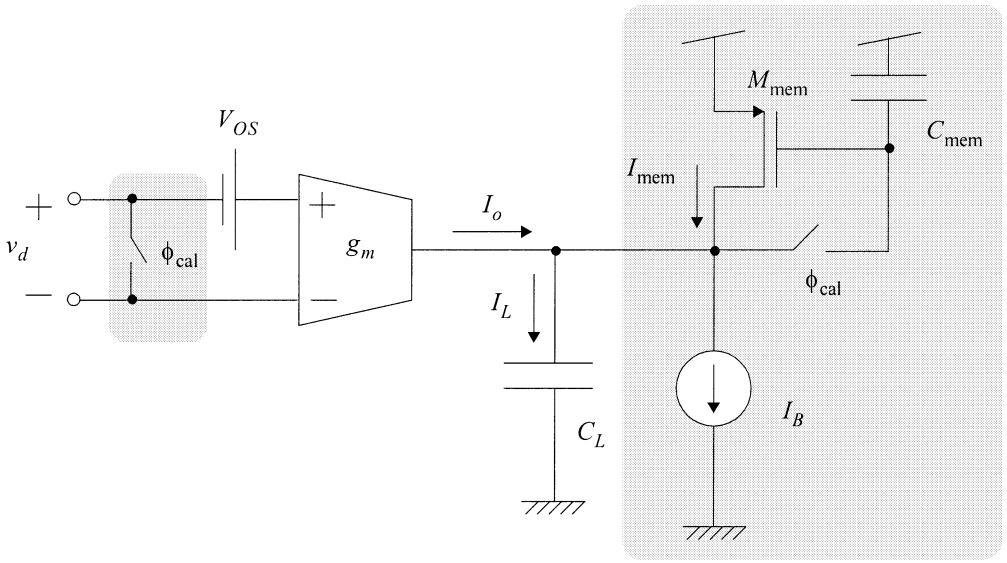


Fig. 9. Offset calibration mechanism for the critical OTAs.

from 18 μA to 46 μA. On the other side, the maximum current signal of the synapse is:

$$I_{\max} = \beta_p V_{x_{\max}} V_{w_{\max}} \approx 0.5\ \mu\text{A} \tag{18}$$

what means a total current range of 1 μA. If an equivalent resolution of 8 bits is intended, then, $(1/2)\text{LSB} = 2$ nA. In these conditions, our current memory must be able to distinguish 2 nA out of the 46 μA. This represents an equivalent resolution of 14.5 bits. In order to achieve such accuracy levels, a so-called S^3I current memory will be employed [23]. As depicted in Fig. 10, it is composed by three stages, each one containing a switch, a capacitor and a transistor. At the beginning, while ϕ_1, ϕ_2 and ϕ_3 are ON, the current I_{in} is divided into I_1, I_2 and I_3, and

$$V_1 = V_2 = V_3 = V_m = V_{T0_n} + \sqrt{\frac{I_{\text{in}}}{\beta_1 + \beta_2 + \beta_3}}. \tag{19}$$

Switches controlled by ϕ_1, ϕ_2 and ϕ_3 are successively turned off. Each time that one of these switches turns off, the voltage stored in its associated capacitor changes, e.g., V_1 changes from V_m to $V_m + \Delta V1$, because of charge injection. The other transistors have to accommodate to absorb the error, as the sum of currents is still forced to be I_{in}, and thus V_2 and V_3 change to

$$V_m{}' = V_{T0_n} + \sqrt{\frac{I_{\text{in}}}{\beta_1 + \beta_2 + \beta_3} + \frac{g_{m_1}\Delta V_1}{\beta_2 + \beta_3}} \tag{20}$$

when ϕ_1 turns off. Correspondingly, V_3 changes to

$$V_m{}'' = V_{T0_n} + \sqrt{\frac{I_{\text{in}}}{\beta_1 + \beta_2 + \beta_3} + \frac{g_{m_1}\Delta V_1}{\beta_2 + \beta_3} + \frac{g_{m_2}\Delta V_2}{\beta_3}} \tag{21}$$

when ϕ_2 falls. Finally ϕ_3 is turned off, and V_3 ends in $V_m{}'' + \Delta V_3$. The final current, I_{out}, is

$$I_{\text{out}} = \beta_1 (V_m - V_{T0_n})^2 - g_{m_1}\Delta V_1 + \beta_2 (V_m' - V_{T0_n})^2 - g_{m_2}\Delta V_2 + \beta_3 (V_m'' - V_{T0_n})^2 - g_{m_3}\Delta V_3 \tag{22}$$

and substituting here the values of V_m, $V_m{}'$ and $V_m{}''$, we find that

$$I_{\text{out}} = I_{\text{in}} - g_{m_3}\Delta V_3 \tag{23}$$

the only error left is that corresponding to the last stage. The former stages do not contribute to the error in the memorized current. If the S^3I block is designed so as to store the most significant bits in the first capacitor, and the less significant bits in the last one, then the error in the memorized current can be made quite small. Consider that the total resolution of the current memory is N. Let us assume that M_1 is conducting the most $N/3$ significant bits of the current I_{in}, then M_2 conducts the next $N/3$ and M_3 conducts the rest, thus, for the last stage an effective resolution can be defined

$$I_3 = \left(\frac{I_{in}}{2^N}\right) \sum_{k=2N/3+1}^{N} 2^{N-k} = \left(\frac{I_{\text{in}}}{2^N}\right) 2^{N_{\text{eff}}}. \tag{24}$$

If the error in the memorized current has to be kept below 0.5 LSB, and $g_{m_3} = 2\sqrt{\beta_3 I_3}$ then

$$\Delta V_3 \leq \sqrt{\frac{I_3}{\beta_3}} \cdot 2^{-(N_{eff}/2+N/2+2)}. \tag{25}$$

And this is the design equation that relates the geometric aspect of transistor M_3, through β_3, with the magnitude of the storage capacitor, via ΔV_3. Once we have β_3, β_1 and β_2 can be easily derived

$$\beta_1 = \left(\frac{\beta_3}{2^{N_{\text{eff}}}}\right) \sum_{k=1}^{N/3} 2^{N-k}$$
$$\beta_2 = \left(\frac{\beta_3}{2^{N_{\text{eff}}}}\right) \sum_{k=N/3+1}^{2N/3} 2^{N-k}. \tag{26}$$

One might think that adding more stages to the current memory will endlessly increase accuracy. However, there is one factor that has not been addressed yet. As the order of the memory increase, the tinier the currents that have to be sensed by the last stages. There comes a point in which the

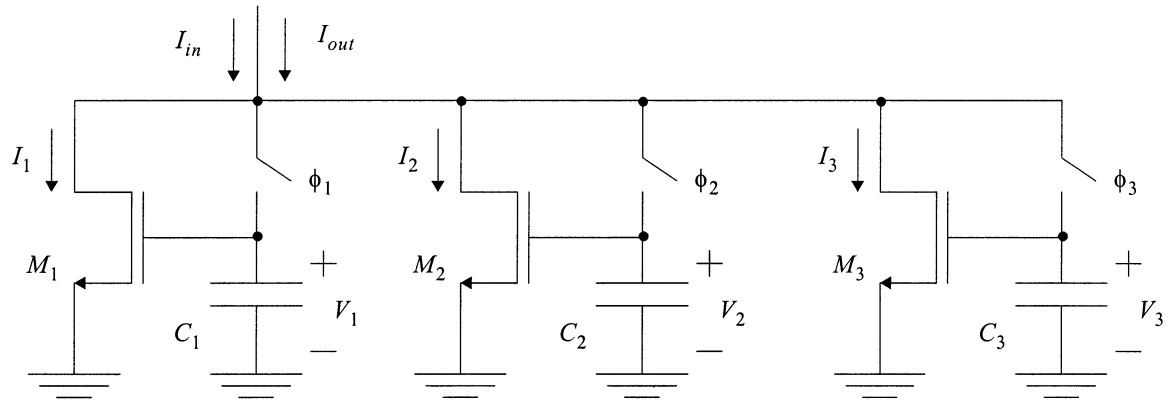


Fig. 10. S^3I current memory schematics an timing.

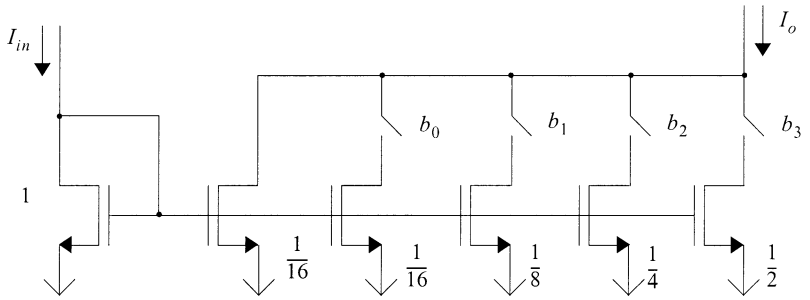


Fig. 11. Binary programmable current mirror (4b).

leakages from the capacitors of the first stages are of the size of the current to be memorized by the last stages, thus making it impossible to reach a steady state current that corrects from the previous errors. This problem worsens as temperature rises. For instance, at 70 °C leakages can introduce changes in the memorized current in the order of 0.2 $nA/\mu s$. If the dynamics of the current memory require several μs to settle—because of the use of large capacitors and the tiny currents involved—the memorized current will display an error that is quite above the initial estimation.

D. Time Constant Scaling Block

The time constant of the CNN layer is defined as $\tau = C_c/G_c$, the ratio between the state capacitor, and the transconductance G_c obtained by multiplying the current factor of the synapse, $\beta_p = 3.13\ \mu A/V^2$, times the weight signal voltage V_w. This time constant depends on the specific set of templates being implemented in the CNN. The state capacitor is composed by the gate capacitances of the 11 synapses driven by the cell's state. As $C_{ox}{}' = 3.45\ fF/\mu m^2$ in this technology, this makes a total of 1.97 pF. In the most favorable case, when every neighbor, even the cell itself, is contributing the maximum amount of current to the cell state, a parallel stack of 18 synapses, a transconductance of 22.5 $\mu A/V$ is found. This represents a minimum CNN time constant of 87.4 ns.

Scaling the time constant of one of the CNN layers involves either modifying the value of the state capacitor or of the synapses transconductance. For the first alternative, we will need to implement a regulable capacitor. If a continuously regulable capacitor is pretended, it does not seem to be easy to realize. If a capacitor with a discrete set of capacitances is adequate, an area of 16 times $2 \times 25.9 = 51.8\ \mu m^2$ will be required to implement a 1:16 time constant ratio.

The second alternative, scaling the transconductances of every synapse contributing to the cell, can be achieved with a current mirror. Scaling up/down the sum of currents entering the cell is equivalent to scaling up/down the transconductances of the synapses, and thus, to scaling down/up the time constant of the CNN core. A circuit for continuously adjusting the gain of a mirror can be designed based on the active-input regulated-Cascode current mirror [24]. The major disadvantage of using this circuit is its strong dependence on the power rail voltage. As we will see later, the power rail voltage can deviate further more than 1% is a densely packed 32×32 -cell parallel array processor chip. This will cause a large mismatch in the time-constants of the different cells in the layer. An alternative to this is a binary programmable current mirror (Fig. 11). The input current, I_{in}, must be always positive, in the sense indicated in the figure, and the output current is given by:

$$l_o = (1 + b_0 + 2b_1 + 4b_2 + 8b_3)\frac{I_{\text{in}}}{16} \tag{27}$$

where b_0, b_1, b_2 and b_3 are the decimal values of the control bits. In this occasion, 4 bits will be more than enough to program the required relations between τ_1 and τ_2. The mismatch between the time constants of the different cells is now fairly attenuated by design.

A new problem arises related with the placement of the scaling block in the signal path. There are several alternatives. First, the scaling block, the binary weighted current mirror, can be placed after the offset cancellation memory, like in

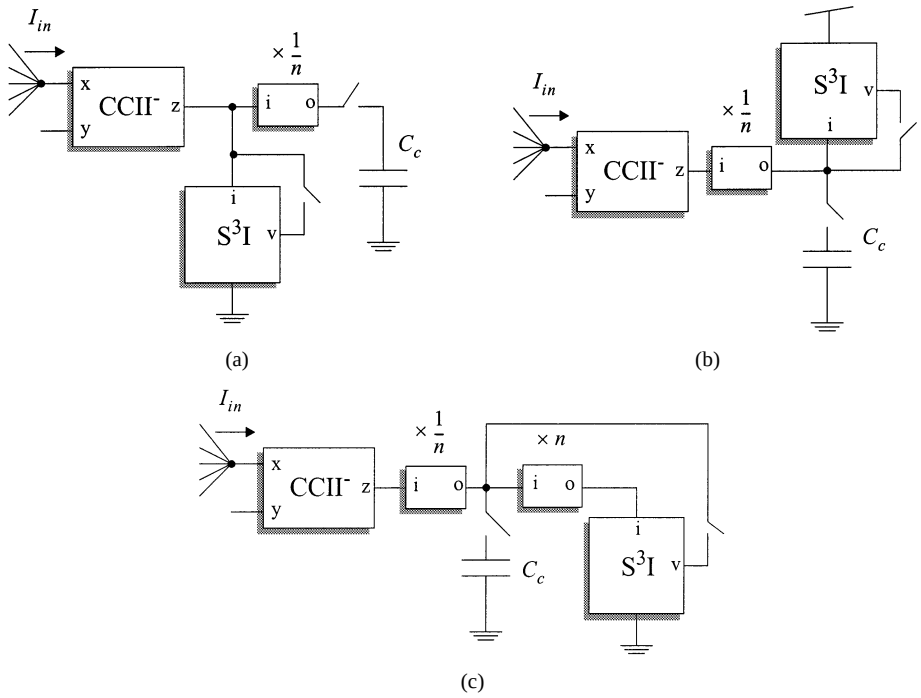


Fig. 12. Alternatives for the placement of the scaling block.

Fig. 12(a). The problem is that any offset introduced by the scaling block is incorporated to the signal path without possible cancellation. The second alternative [see Fig. 12(b)] is to place the scaling block before the offset cancellation memory. It means that the S^3I memory will have to operate over a wider range of currents, and thus complicating its design and surely degrading its performance. Our choice, depicted in Fig. 12(c) has been to place the scaling block in the memorization loop. The current memory will operate on the unscaled version of the input current, and any offsets associated with the scaling blocks will be sensed and memorized to be cancelled on-line during the network evolution.

The resulting CNN core is shown in Fig. 13 [25]. In this picture, the voltage reference generated with the current conveyor, the current mirrors and the S^3I memory can be easily identified. The inverter, A_i, driving the gates of the transistors of the current memory is required for stability. Without it, the output node, Ⓐ, will diverge from the equilibrium. The operation of this circuit is as follows. Before running the CNN dynamics, the current offsets of all the synapses are injected to the virtual reference at node Ⓛ. This current is scaled down to one n-th of its value by means of the adjustable current mirror formed by M_{n1} and M_{n2}. The arrow over M_{n2} stands for the binary programmability of this device. The value of n is

$$n = 1 + b_0 + 2b_1 + 4b_2 + 8b_3. \tag{28}$$

Then, if all the transistors of the S^3I memory are conducting, this is ϕ_1, ϕ_2 and ϕ_3 are ON, then the negative feedback loop makes M_{p_2} to conduct the same current as M_{n2}. M_{p2} is also adjustable so as to make M_{p1} and the current memory to work with the same current ranges than the input stage. The rest of the operation has been already described. The current memory stores successively the remaining most significant bits of the input current, plus the errors accumulated. When it is done, the CNN loop can be closed and the output current I_o represent the scaled sum of the contributions, with the state-independent errors substracted.

The critical aspects of this circuit are related with the feedback loop formed by M_{p1}, M_{p2}, M_{n2}, the inverting amplifier A_i and the transistors M_m, when sensing the offset current. During this process the output current I_o is zero because the current path to the state capacitor is open. Once the input current has been established, V_{nn} can be considered a bias voltage. First of all, it must be taken into account that during the three different phases in which the loop is closed (ϕ_1, ϕ_2 and ϕ_3 ON, ϕ_1 OFF and ϕ_2 and ϕ_3 ON, and, finally, ϕ_1 and ϕ_2 OFF and ϕ_3 ON) the values of g_{m_m} and C_m change, so the stability conditions must hold for any possible set of values. Considering the small-signal equivalent circuit for this loop, a three-pole system is found [see Fig. 14], with pole frequencies: $p_1 = -g_{o_2}/C_o$, $p_2 = -1/C_m R_{oi}$ and $p_3 = -g_{m_{p1}}/C_{pp}$. The nearest pole, that is at node V_o, will be employed to compensate the loop for stability. As g_{m_m} and C_m decrease for the latest phases of the current memorization, the loop will be more stable because this causes the loop dc gain, T_0, to decrease and p_2 to grow, breaking away from p_1 and thus increasing the phase margin. Therefore, the worst situation will occur when ϕ_1, ϕ_2 and ϕ_3 are ON, and thus the circuit is designed to be stable in these conditions. It is also important that A_i is kept reasonably low, otherwise it will displace the unity-gain frequency, ω_u, toward the value of the inversion ω_{180}. This means a loss of phase margin, and can compromise the loop stability.

As we commented before, leakage currents can degrade the S^3I memory operation especially as the operation temperature rises. Although the negative feedback moves the circuit toward the correction of the errors, it maybe too slow to settle at a value before leakages modify the position of the equilibrium point.

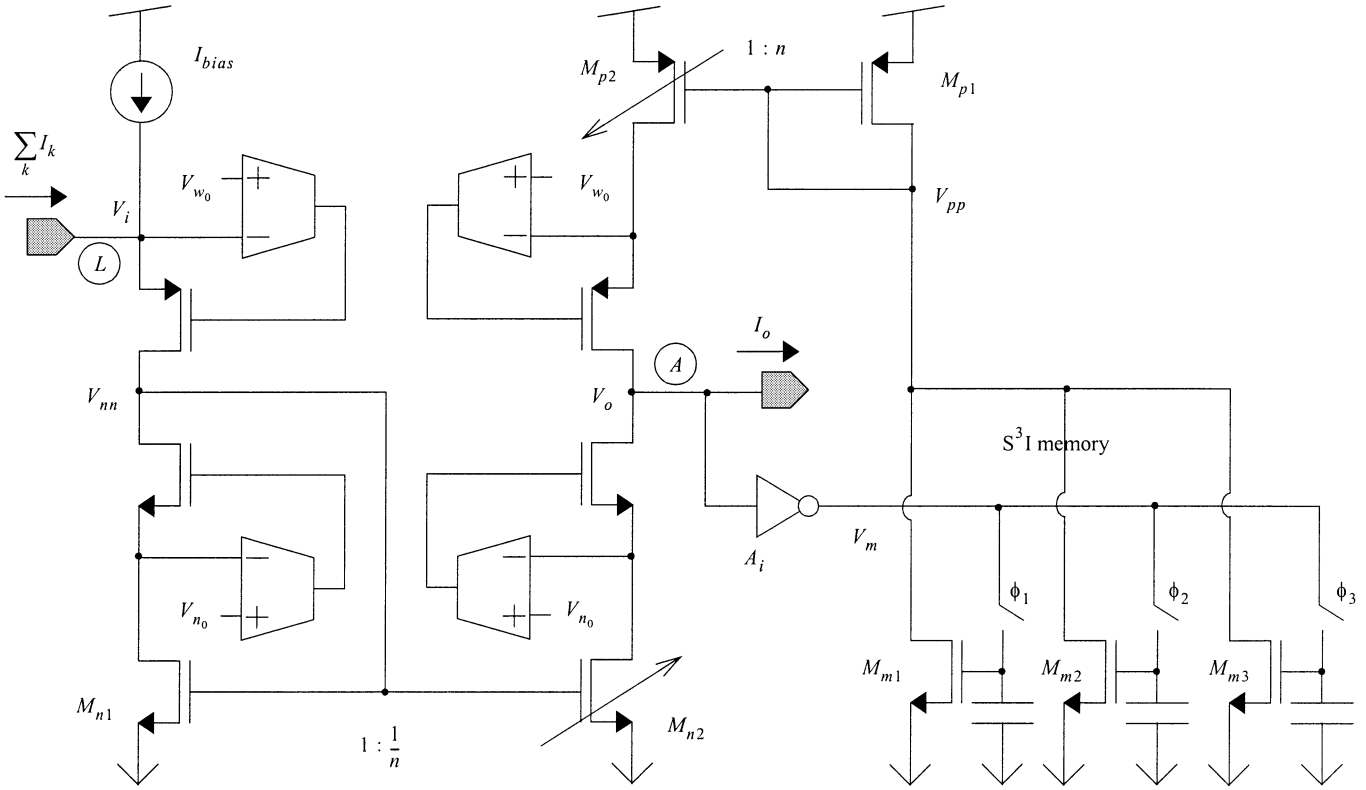


Fig. 13. Input block with current scaling.

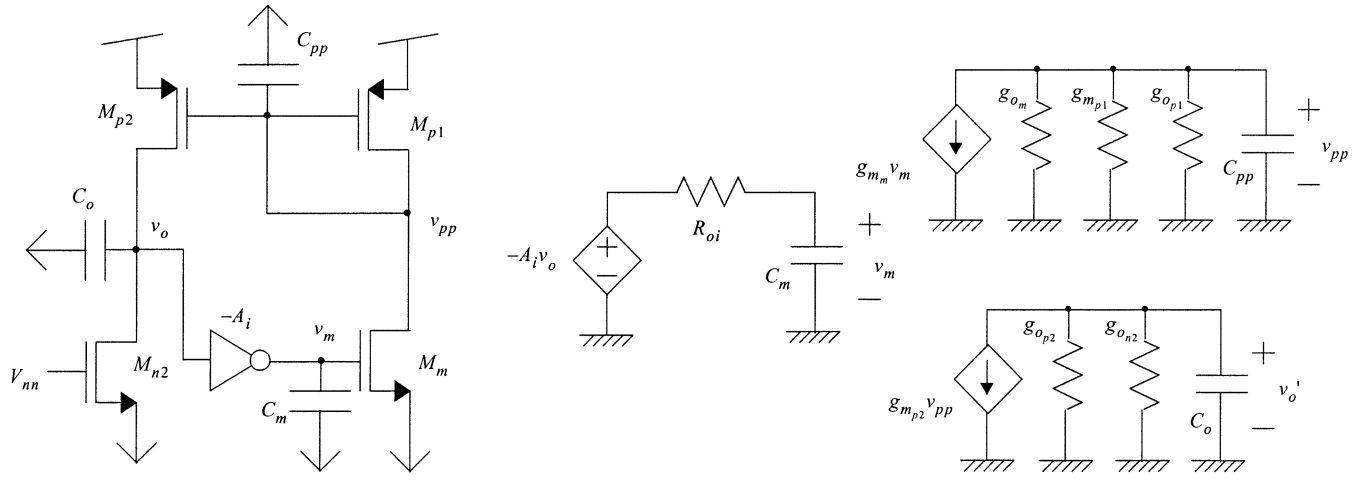


Fig. 14. Simplified schematics of the feedback loop and its small signal equivalent.

Therefore, compensation must be kept under a limit to avoid slowing down the loop dynamics in excess.

V. Peripheral Circuitry

A. Analog Weight Signals Distribution

Conveying analog voltages from the boundaries of the array to the inner cells is not a trivial task. Especially if the metal lines supporting the signals have to carry large currents, and the width of these lines must be reduced because of cell area compromises. These resistive lines carrying some current cause voltage signals to drop. In the case of the weight signals—that have to be transmitted to every synapse in the network, entering through a low-impedance node and, thus, dragging a quite perceptible amount of current—this voltage drop can seriously compromise the appointed resolution. Also for the power supply lines, that carry an important amount of current, voltage drops are a serious problem, as they can cause misfunction of the inner cell's circuitry. In consequence, it is important to develop a reliable model for this phenomenon. It is possible to find a closed expression to compute the maximum error in the propagation of

the reference voltage $V_{\rm REF}$ through a metal line, the one-dimensional model, as a function of the resistance between cells, R, and the current demanded by each cell, I_o [25]:

$$|\Delta V_{\max}| = \left|V_{\rm REF} - V_{(N+1)/2}\right| = \frac{(N-1)^2}{8} I_o R. \quad (29)$$

This expression is useful to determine the appropriate width of the power lines laid across the array. For instance, in the case of this prototype chip ($N = 32$) each cell demands 300 μA under normal operation conditions, let us preview for a much higher peak consumption, say 800 μA. If the maximum error allowed in the power voltage, nominally 3.3 V, will be 50 mV. and the power supply distribution coincides with that of the presented model—the voltage sources are tied to the ends of each row of the array, no vertical connection between horizontal power lines is considered, each segment of the lines will have as much as 520 mΩ

$$R = R_{\square}\left(\frac{L}{W}\right) \leq 520\ \mathrm{m}\Omega \quad (30)$$

where $R_{\square}$ is the sheet resistance of the metal, and L and W are the length and width of the metal track. For the uppermost metal layer in the CMOS process employed, the most conductive of the three, $R_{\square}$ is of 35 m$\Omega/\square$ at room temperature, but can go up to 80 m$\Omega/\square$ at 100 C. If the length of the cells is 190 μm, making a conservative estimation, employing the higher value for $R_{\square}$, it is found that the minimum width needed to distribute the power supply voltage is approximately 30 μm. A similar approach is employed to derive the width of the metal lines carrying the weight signals. Now, currents are much lower, $I_o = 2.0$ μA. But the maximum error permitted is as low as 1.6 mV, this is 0.5 LSB for an equivalent resolution of 8b with a total signal range of 800 mV. Tracing the weight lines with the same metal employed before, it is found that $R \leq 6.66\ \Omega$. The minimum width required to maintain the accuracy is approximately 2.3 μm.

In the chip, the power supply grid is 2-D. There are vertical metal lines connecting the nodes of the network too. Although we have not found a closed form for the maximum error in a 2-D grid, a good estimation can be made making some assumptions.

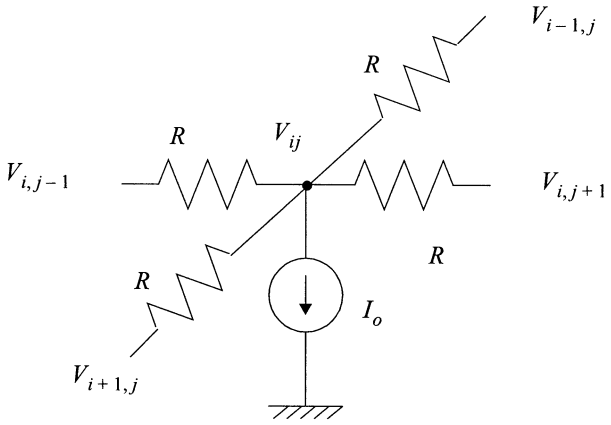


Fig. 15. A 2-D model for the voltage decay toward the center of the network.

First, suppose that the network is a square, $M = N$, and that the power supply is directly connected to every cell in the border:

$$V_{ij} = V_{\rm DD} \quad \forall i, j \in \{1, N\}. \quad (31)$$

The maximum drop will be observed at the centre of the array. An equation can be written for the inner nodes [see Fig. 15], assuming equal resistances of the horizontal and vertical lines:

$$4V_{ij} - V_{i-1,j} - V_{i+1,j} - V_{i,j-1} - V_{i,j+1} = -I_o R \quad \text{if} \quad 1 < i, j < N. \quad (32)$$

These equations constitute a system of N^2 linear equation on N^2 variables, whose matricial form can be automatically computed as shown in (33) at the bottom of the page.

Solving this system for the central term of the array, a maximum value for the error is found. The voltage drop from $V_{\rm DD}$ in the middle of the network as a function of the number of cells is plotted in Fig. 16. Black stars are the computed minima while the solid line is the best fitting second-order curve. This approximation yields the following relation:

$$|\Delta V_{\max}| \approx \frac{N^2 - 2.0543N + 0.01221}{13.569} I_o R. \quad (34)$$

It means that providing a second mesh of metal connections can reduce the voltage drop in nearly one half. If the horizontal

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 & -1 & 4 & -1 & \dots & 0 & -1 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 & 0 & -1 & 4 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & -1 & 0 & \dots & -1 & 4 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{12} \\ V_{13} \\ \dots \\ V_{1N} \\ V_{21} \\ V_{22} \\ V_{23} \\ \dots \\ V_{31} \\ V_{32} \\ \dots \\ V_{NN} \end{bmatrix} = \begin{bmatrix} V_{\rm DD} \\ V_{\rm DD} \\ V_{\rm DD} \\ .. \\ V_{\rm DD} \\ V_{\rm DD} \\ -I_o R \\ -I_o R \\ .. \\ V_{\rm DD} \\ -I_o R \\ .. \\ V_{\rm DD} \end{bmatrix}. \quad (33)$$

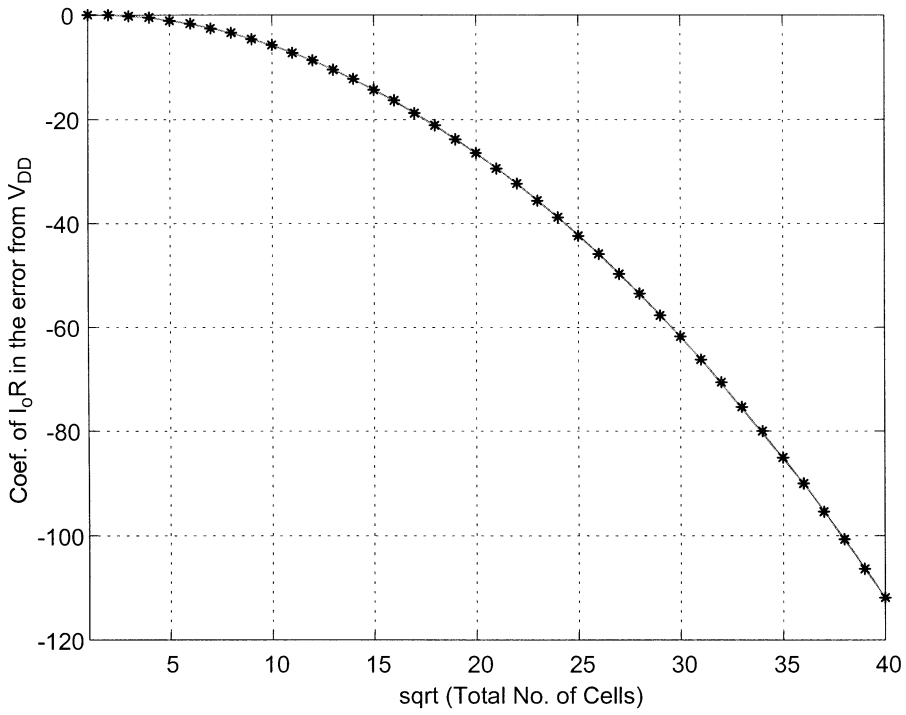


Fig. 16. Voltage drop from $V_{\rm DD}$ experimented by the central cell in the grid.

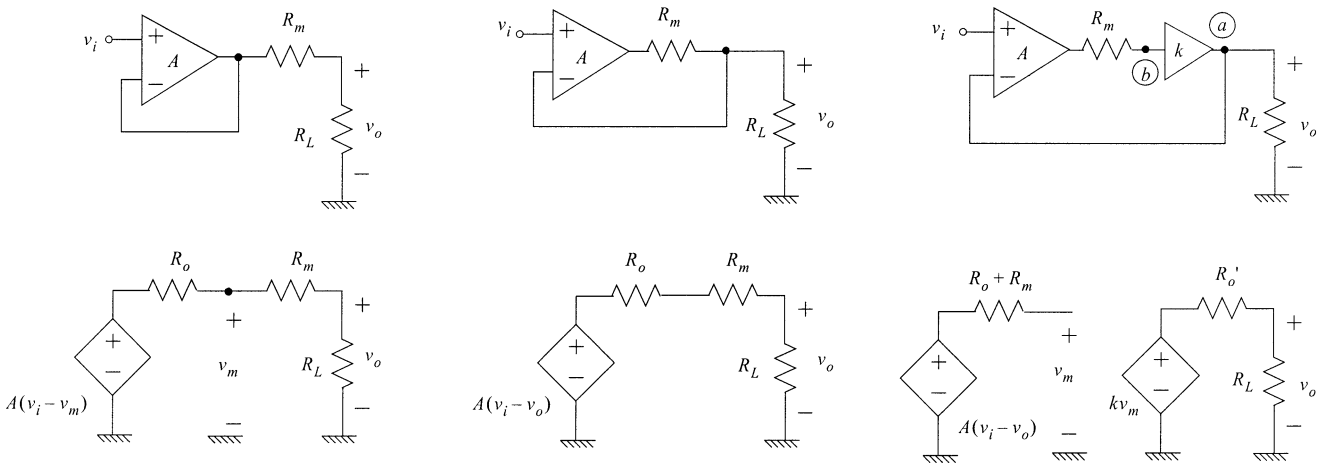


Fig. 17. Weights buffering alternatives and low-frequency small signal models.

and vertical resistances are the same, the error at the centre cell is divided, approximately, by $\sqrt{3}$.

B. D/A Weight Codes Conversion and Buffering

The D/A converters employed in this chip are of an inherently monotonic type [26]. This circuit consists in a long string of equally valued resistors running from the higher to the lower voltage references. This string is tapped at equally spaced points. The access to this points is controlled by a tree of analog switched driven by the outputs of some decoding logic. Monotonicity is assured by construction, as every tap points to a higher voltage that the previous one. Differential nonlinearity in this circuit is introduced by the mismatch between the resistors, therefore, they must be sized to avoid important impairities in the converter steps. Integral nonlinearity in the weights representation is not a problem because it can be corrected by software.

The outputs of the D/A converters need buffering to be transmitted to the array processor. Voltage references driving high-impedance nodes, as the gates of MOS transistors, do not require extra driving other than that afforded by the voltage followers at the converters' output. Weights, on the contrary, have to be transmitted to low-impedance nodes, 1024 sources of MOS transistors in parallel. The loading impedance, R_L, results unbearable for the voltage buffer especially if we take into account the resistance of the long tracks of metal

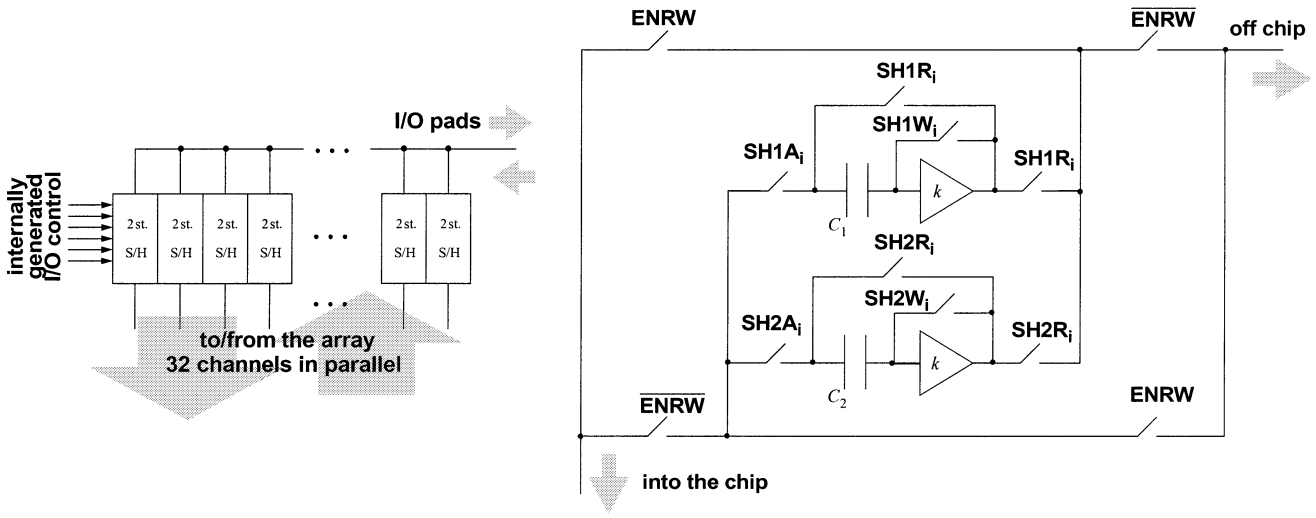


Fig. 18. (a) Serializing-deserializing I/O interface and (b)2-stage circuit for sample and hold.

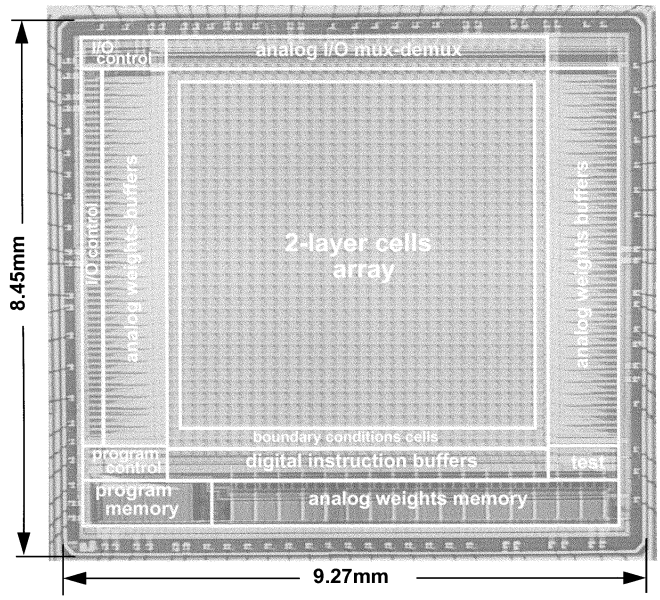


Fig. 19. Microphotograph of the prototype chip.

TABLE I
PROTOTYPE CHIP DATA

Technology	0.5μm CMOS 1-P 3-M
Number of cells	32 × 32 = 1024
Die area	9.27 × 8.45 mm^2
Die area (w/o pads)	8.77 × 7.94 mm^2
Array area	5.98 × 5.83 mm^2
Package	ceramic PGA-100
Power supply voltage	Vdd = 3.3V
Logic "0" / Logic "1"	0V and 3.3V
Accuracy on the weights	8b
Image samples resolution	7-8b
I/O rates	10MS/s
CNN time constant	below 100ns

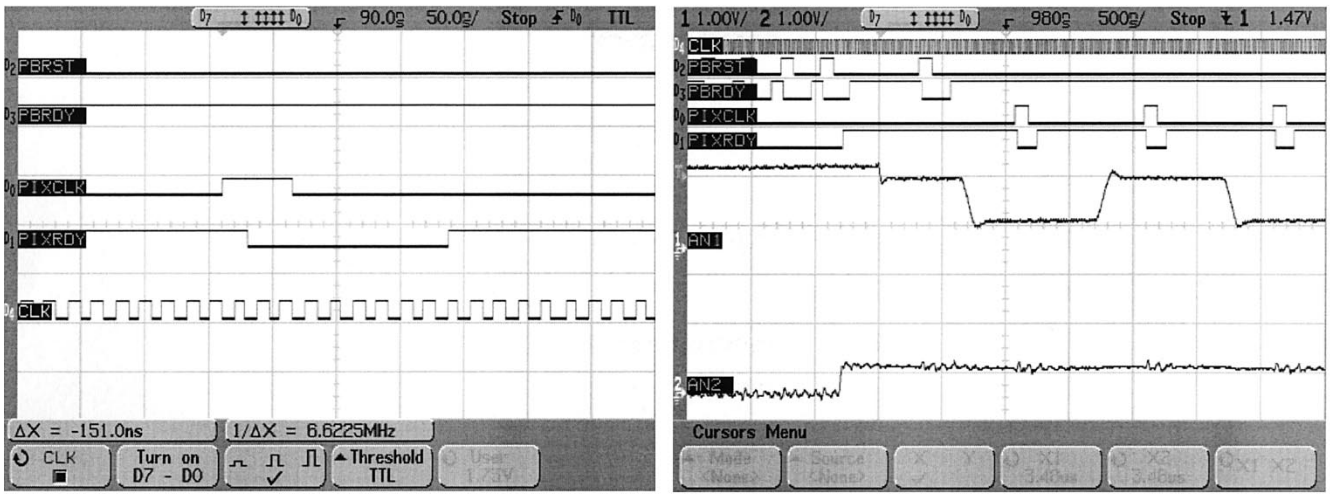
conveying the signal from one end of the die to the other. Basically, the resistance of the metal tracks add up to the buffer output resistance, thus boosting the voltage division. This is the situation in Fig. 17(a). Computing the output impedance of the buffer-plus-metal-tracks circuit, it adds up to

$$Z_o = R_m + \frac{R_o}{1+A} \tag{35}$$

hence, despite the fact that the output resistance of the amplifier is greatly attenuated by feedback, the resistance of the metal tracks, represented here by R_m, makes Z_o to be as large as the actual R_L. The resistance of the metal tracks can be as high as 1 kΩ while the resistance of one of the p-type MOS transistors operating in the ohmic region employed can be about 1 MΩ. 1024 of them in parallel makes an R_L of kΩ, approximately, thus halving the dynamic range of the signals, and therefore losing one bit of resolution in the weights. This error can not be afforded so a different scheme must be employed for the buffers of the weight signals. For instance, in the circuit in Fig. 17(b) is the actual output voltage the one that is fed back to the amplifier, causing the output impedance seen by the load to be

$$Z_o = \frac{R_o + R_m}{1+A} \tag{36}$$

what enhances the performance of the buffer. But still, propagation of signals through metal tracks carrying strong current intensities may end in appreciable disparities between the voltages transmitted to different points along the metal tracks, unless they are made unrealistically wide. In order to avoid this, voltage buffers have been allocated nearer to the cell array [27]. Working in parallel, they can be considered as a voltage amplifier with gain k, a high input impedance and a rather low output



(a) (b)

Fig. 20. (a) Protocol signals handshake and (b) start of image samples acquisition.

impedance [see Fig. 17(c)]. The output impedance at low-frequencies is

$$Z_o = \frac{R_o'}{1+kA} \tag{37}$$

what certainly reduces any possible loading effect. The design challenge is to avoid the loss of phase margin in the feedback loop, being the dominant pole, associated with the output node ⓐ, and the second pole, which is associated with node ⓑ, too close to operate without compensation. Extra capacitance must be added accordingly to node ⓐ to avoid instability, at the expense of a reduction on the circuit speed.

C. I/O Interface

The last major subsystem of the prototype is the I/O interface. This circuit is provided for the acquisition and delivery of image samples, which must be analog voltages ranging from 0.6 V to 1.4 V, nominally. The collection of 32×32 image samples, that will be acquired or delivered on the same batch, are transmitted through a serial channel but passed in parallel to the 32 columns of the array. This is achieved by the circuit structure depicted in Fig. 18(a). It consists in 32 sample and hold circuits, one for each column of the array, connected concurrently to the serial I/O channel. Each S/H circuit, represented by the schematic in Fig. 18(b), consists in 2 S/H stages and several transmission gates. When acquiring an image, $\mathbf{ENRW} = 1$. The serial I/O channel connects the input pad to the input nodes of the sample and hold stages. Then, at 10 MS/s, 32 samples of the signal are stored in the first row of S/H circuits by activating alternatively signals $\mathbf{SH1A_i}$ and $\mathbf{SH1W_i}$. Once the first 32 samples are acquired, the following 32 samples of the input signal are stored in the second row of S/H circuits, by activating $\mathbf{SH2A_i}$ and $\mathbf{SH2W_i}$ for the 32 S/H stages alternatively. At the same time, all the signals $\mathbf{SH2R_i}$ corresponding to the first row of S/H's are activated together, thus uploading the stored samples to the first row of the array. This is realized

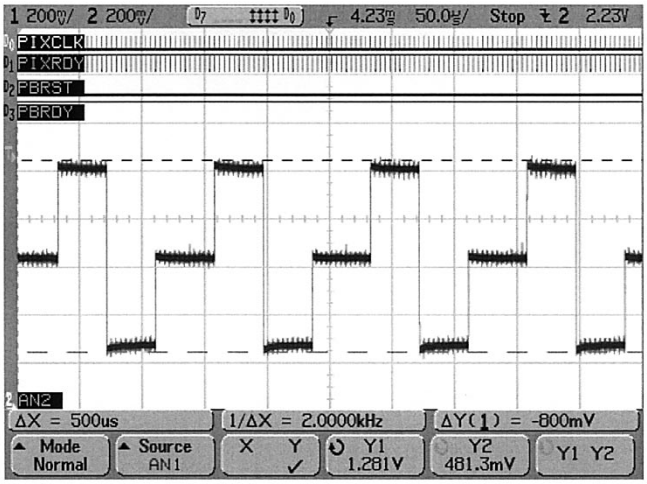


Fig. 21. Read-out of a three stepping stripes image row-by-row.

during $32 \times 100\ \mathrm{ns} = 3.2\ \mu\mathrm{s}$, which is enough time for the driving capabilities of the S/H amplifiers to update the voltage at the prescribed local memory in the cells of the first row of the array. By the time the second row of S/H circuits is done with the acquisition of the second batch of 32 samples of the input, the first row starts acquiring the following 32 samples, while the second S/H row passes the stored sample to the second row of the array of cells. This process continues until the last row of the array is updated with the information from the new image. For the delivery of the output image, the process is inverted, $\mathbf{ENRW} = 0$, then the S/H circuits first read in parallel what is transmitted by the rows of the array, and then deliver the acquired samples of the output one by one, through the I/O channel that now is connected to an output buffer to send the voltages off-chip at 10 MS/s.

It must be regarded that most of the signals controlling the I/O processes are generated on- chip, hence, for the user, external control is reduced to follow a simple protocol. First of all,

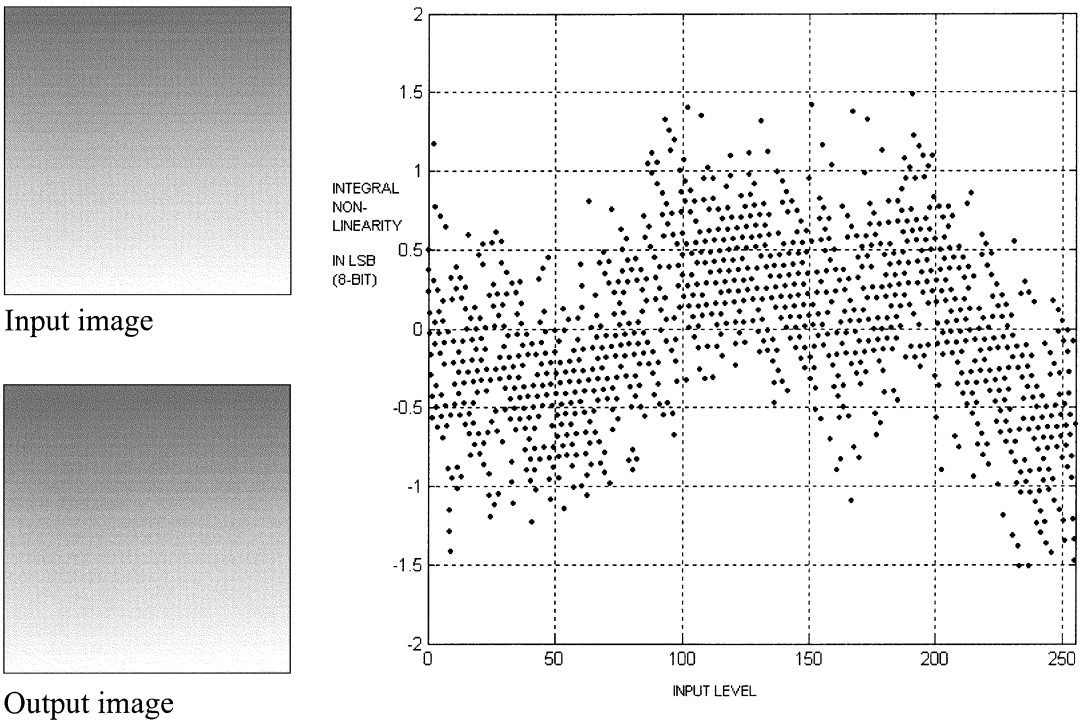


Fig. 22. Integral nonlinearity of the I/O map.

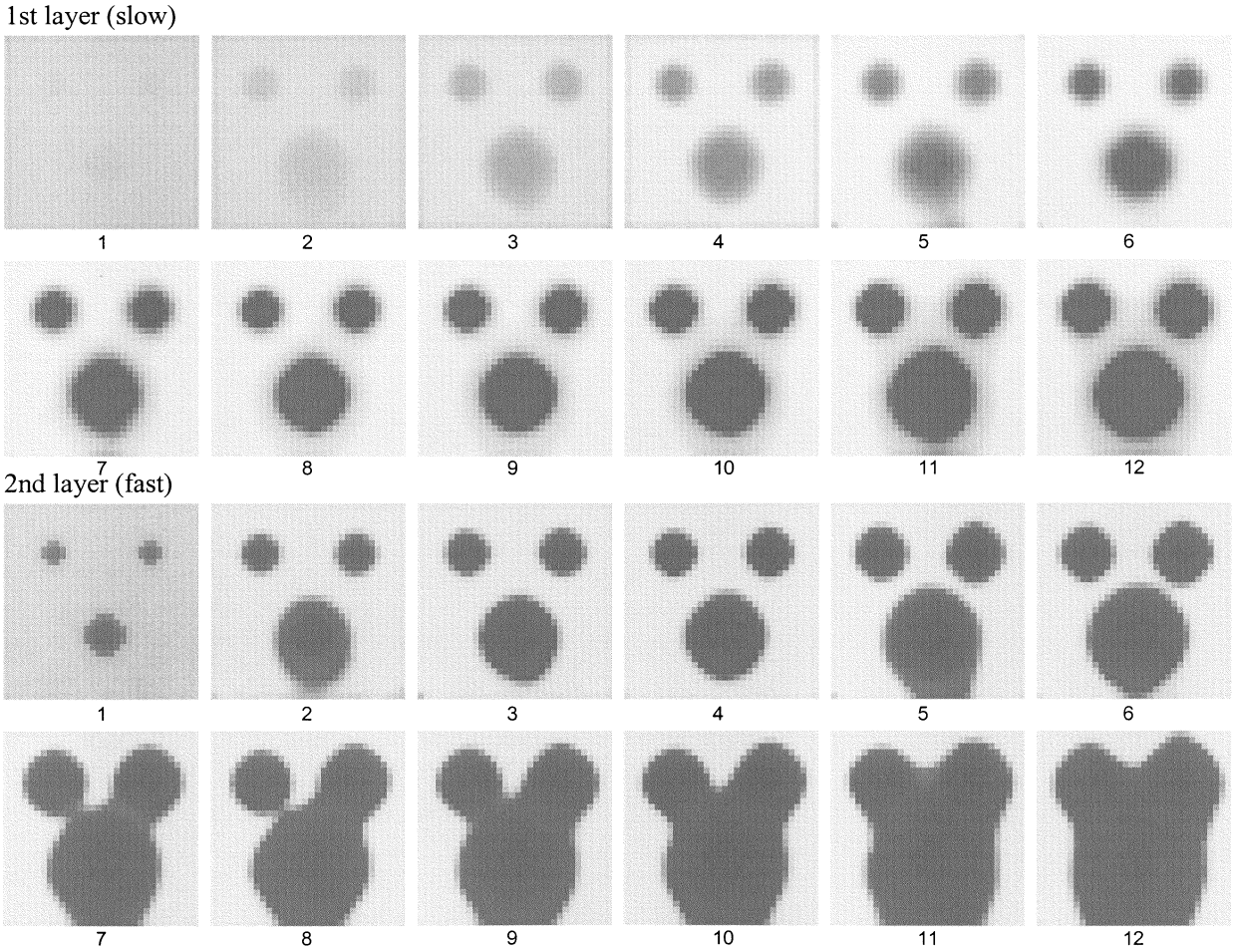


Fig. 23. Triggered waves across the fastest and slowest layers.

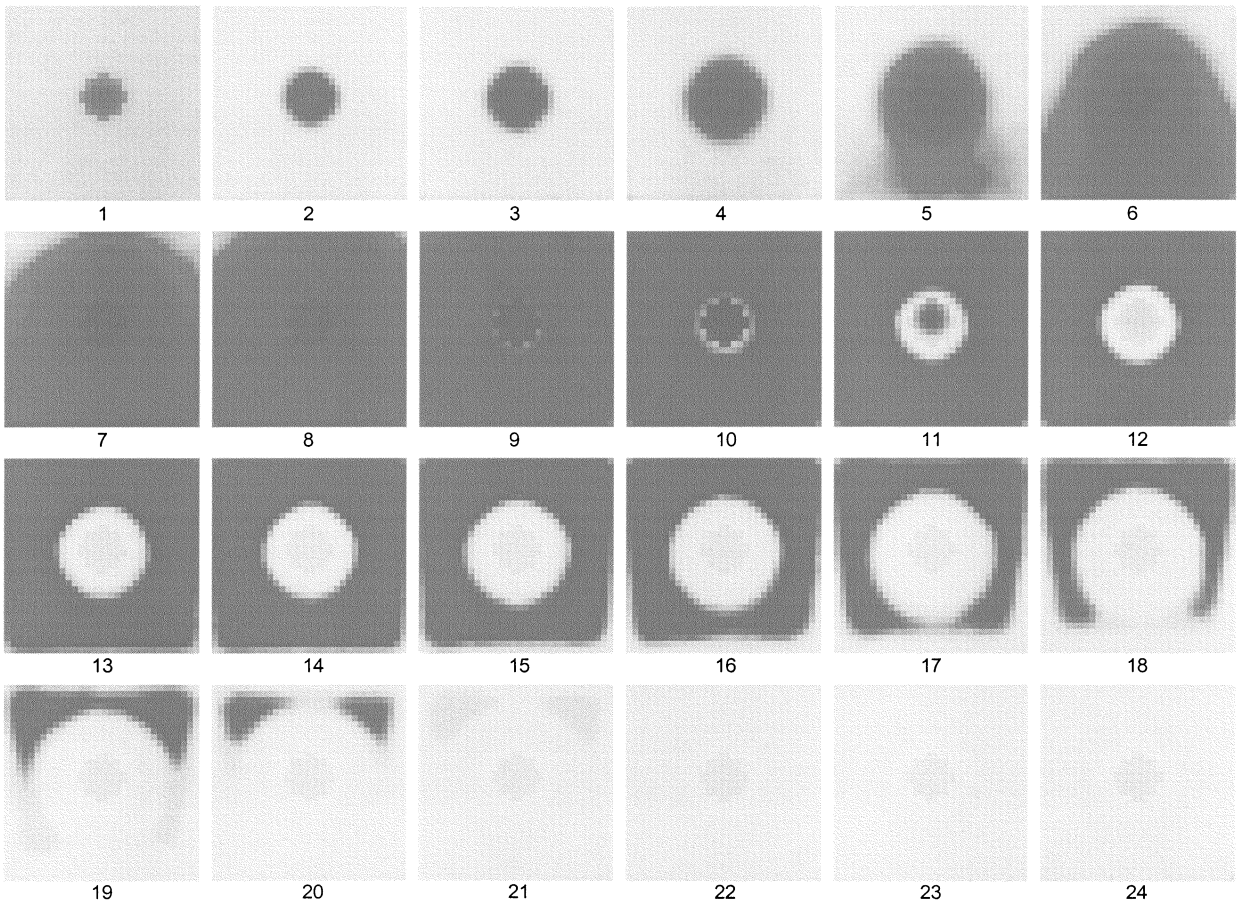


Fig. 24. Wide field erasure effect (only the fastest layer shown).

there is an internal clock, made up from a ring of inverters, with frequency controlled by an analog voltage, constituting a nonharmonic VCO. This is the subpixel clock, generating signal **SubPixCLK**. From the program memory, bits 54th and 55th of the logic program instruction indicate the initiation of an I/O process (the 54th bit corresponds to signal **ENIO**) and specify whether a 32×32 -pixel image is to be acquired or delivered by the chip (what is done by signal **ENRW**, 55th bit of the logic program instruction). Then, when an instruction containing $\mathbf{ENIO} = 1$ is selected, the I/O interface waits for a rising edge of an external signal named **PixCLK**, referred as the pixel clock. Once it occurs, an internal process controlled by **SubPixCLK** generates the appropriate internal signals ($\mathbf{SH1A_i}$, $\mathbf{SH1W_i}$, $\mathbf{SH1R_i}$, $\mathbf{SH2A_i}$, $\mathbf{SH2W_i}$ and $\mathbf{SH2R_i}$, and the appropriate row selection) to acquire or deliver one voltage sample. When this is achieved, the I/O interface generates a pulse named **PixReady**, that must be sensed by the user, that means that the system is ready to receive, or send, the following pixel sample. The critical steps in the design of this scheme are two.

- Overlapping in the S/H selection signals must be avoided. Guard times must be provided that do not rely on the internal delays that are not controlled precisely.
- Control of the local memories access must be passed to the I/O interface while acquiring or delivering an image.

Therefore, with the activation of **ENIO** and the arrival of the first **PixCLK** rising edge, a shift- register clocked by **SubPixCLK** passes a pulse from the leftmost stage, say **SR0**, to the rightmost, **SR5**, permitting the generation of synchronized edges that will constitute the rising and trailing edges of the S/H control signals. These pulses can be combined to obtain the desired control signals. The internal generation and separation of the signal edges prevent uncontrolled delays to alter the precedence between signals, almost independently of the clock frequency at which these circuits are operated. At the end of the count realized by the shift-register, a **PixReady** pulse is generated, notifying the user that the next voltage sample can be recorded in the following capacitor of the S/H battery. The combinational and sequential logic circuits employed to implement the I/O interface has been designed full-custom, in order to allow as much integration with the processing array as possible. Using a library of cells would not have permitted the intricate routing employed to tailor the I/O control.

VI. Experimental Results

A. Prototype Chip Data

A prototype chip has been designed and fabricated in a standard 0.5 μm CMOS technology with single-poly and triple-metal layers. Fig. 19 displays a microphotograph of the chip. It

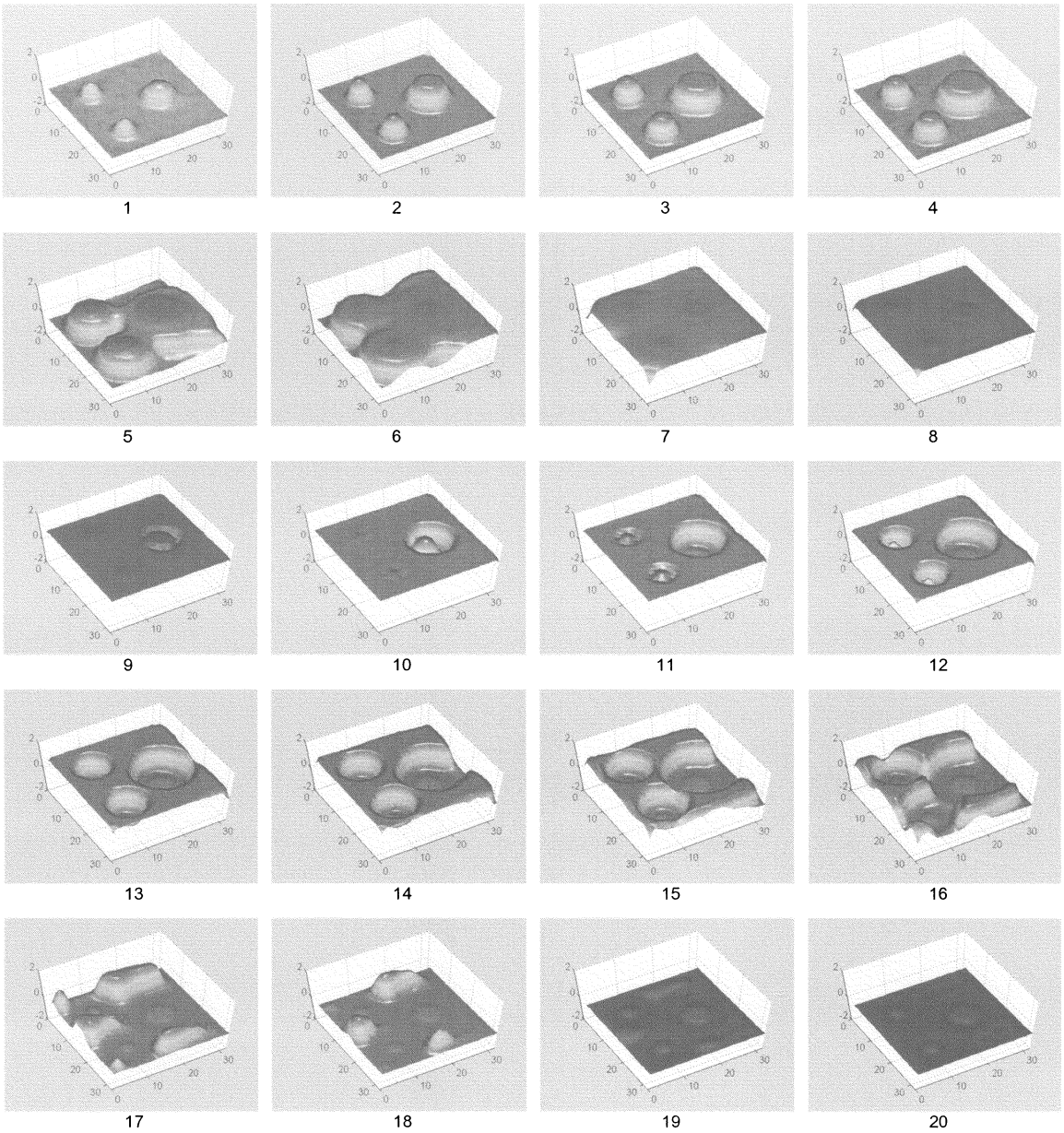


Fig. 25. Wide field erasure effect, represented in 3-D.

contains a central array of 32×32 cells of the type formerly described. Surrounding the array, a ring of boundary cells, implementing the contour conditions for the CNN dynamics, is found, together with the necessary buffers to transmit digital instructions and analog references to the array. On the lower part of the chip, the program control and memory blocks can be found. The last major subsystem is the I/O interface including S/H batteries, decoders, counters and different sequential logic. The whole system fits in 9.27×8.45 sq.mm., including the ring of bonding pads. Without pads, the total area is $8.77 \times 7.94\ \mathrm{mm}^2$, this includes the CNN array and the necessary circuit overhead. The array of CNN cells alone occupies $5.98 \times 5.83\ \mathrm{mm}^2$, which is, roughly, the 50% of the total area of the chip. The resulting cell density, excluding circuits outside the array, is $29.24\ \mathrm{cell}/\mathrm{mm}^2$. In order to cautiously handle this data, it is important to notice that the area occupied by the cell array scales linearly with the total number of cells, what is not the case of the overhead circuitry, which tends to be a smaller fraction of the total chip size as long as the number of cells rises. The power consumption of the whole chip has been estimated in 300 mW. Data I/O rates are nominally 10 MS/s. The time constant of the fastest layer (fixed time constant) is designed to be under 100 ns. Table I summarizes some characteristics and measured features of the chip.

The peak computing power of this chip is of 470 GOPS. Here, OPS means analog arithmetic operations per second. In a time constant, 100 ns, each CNN core performs 12 multiplications and 11 additions, then, for each cell, with two cores, we have 46 operations within each cell in 100 ns. Having 1024 processing cells, the chip can reach 470 GOPS when running the network dynamics. If the computing power per unit area —considering the main array alone— and per unit power are calculated we have $6.01\ \mathrm{GOPS}/\mathrm{mm}^2$ and 1.56 GOPS/mW.

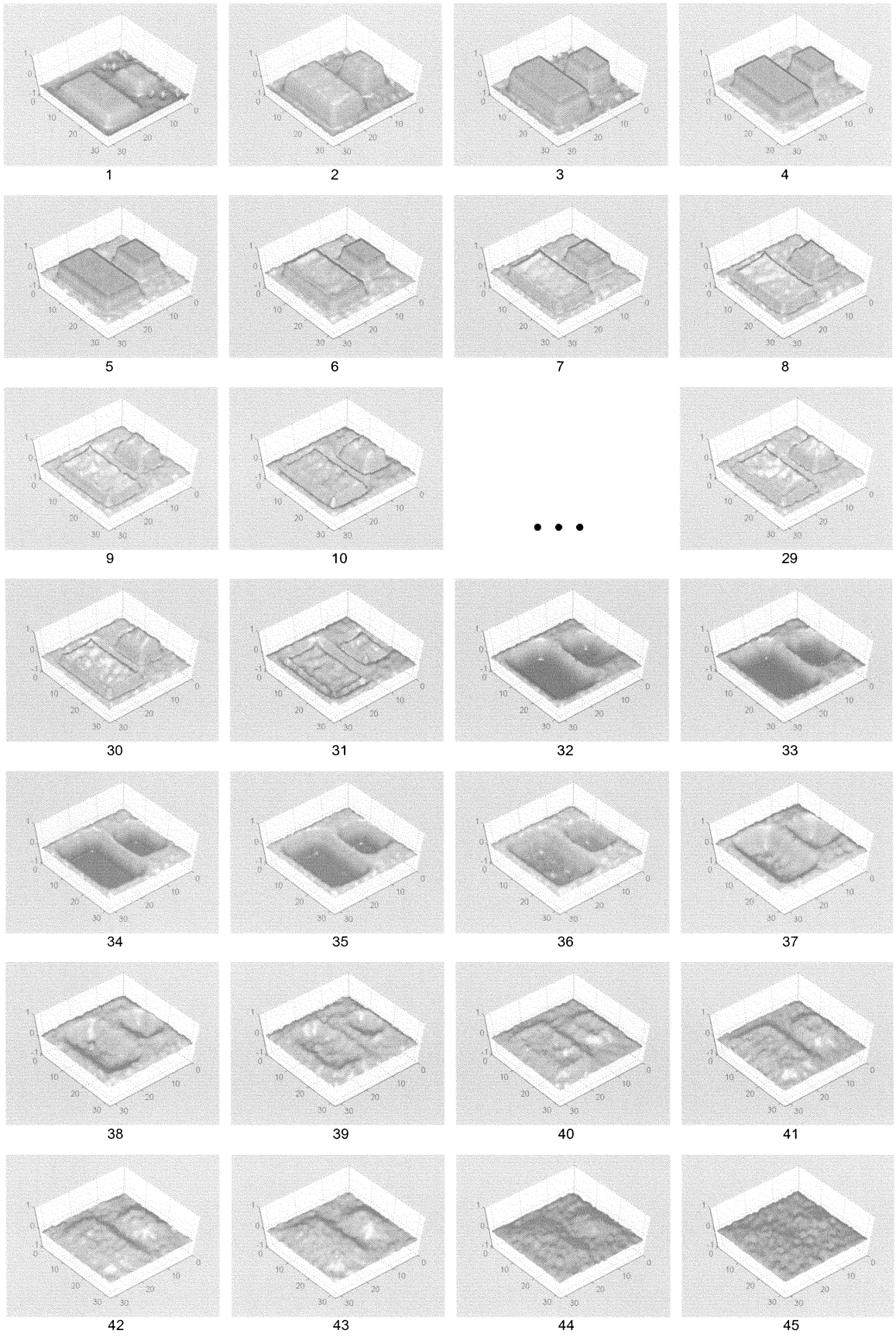


Fig. 26. Spatio–temporal edge detection and deactivation (fast layer).

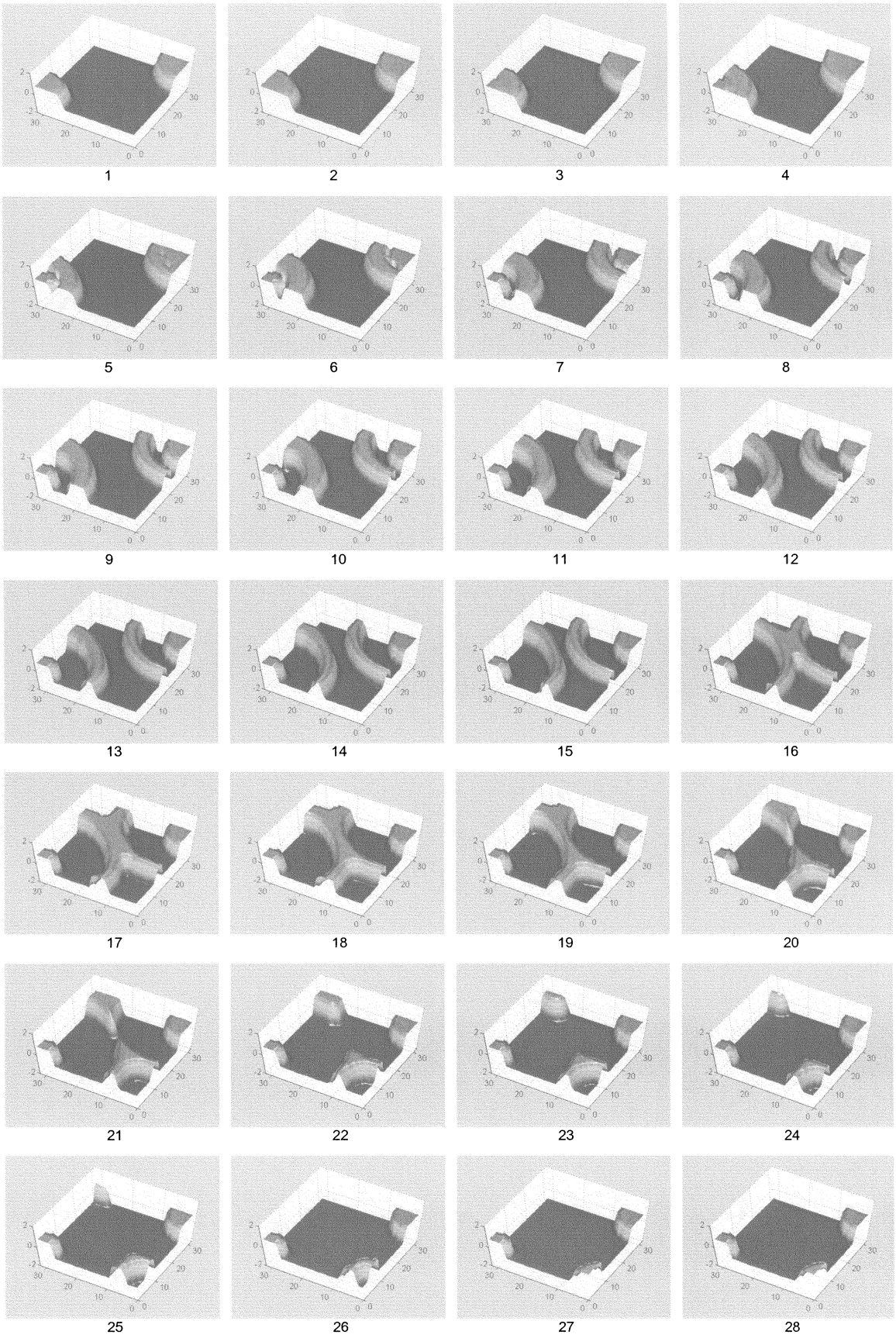


Fig. 27. Traveling-wave generation (fast layer only).

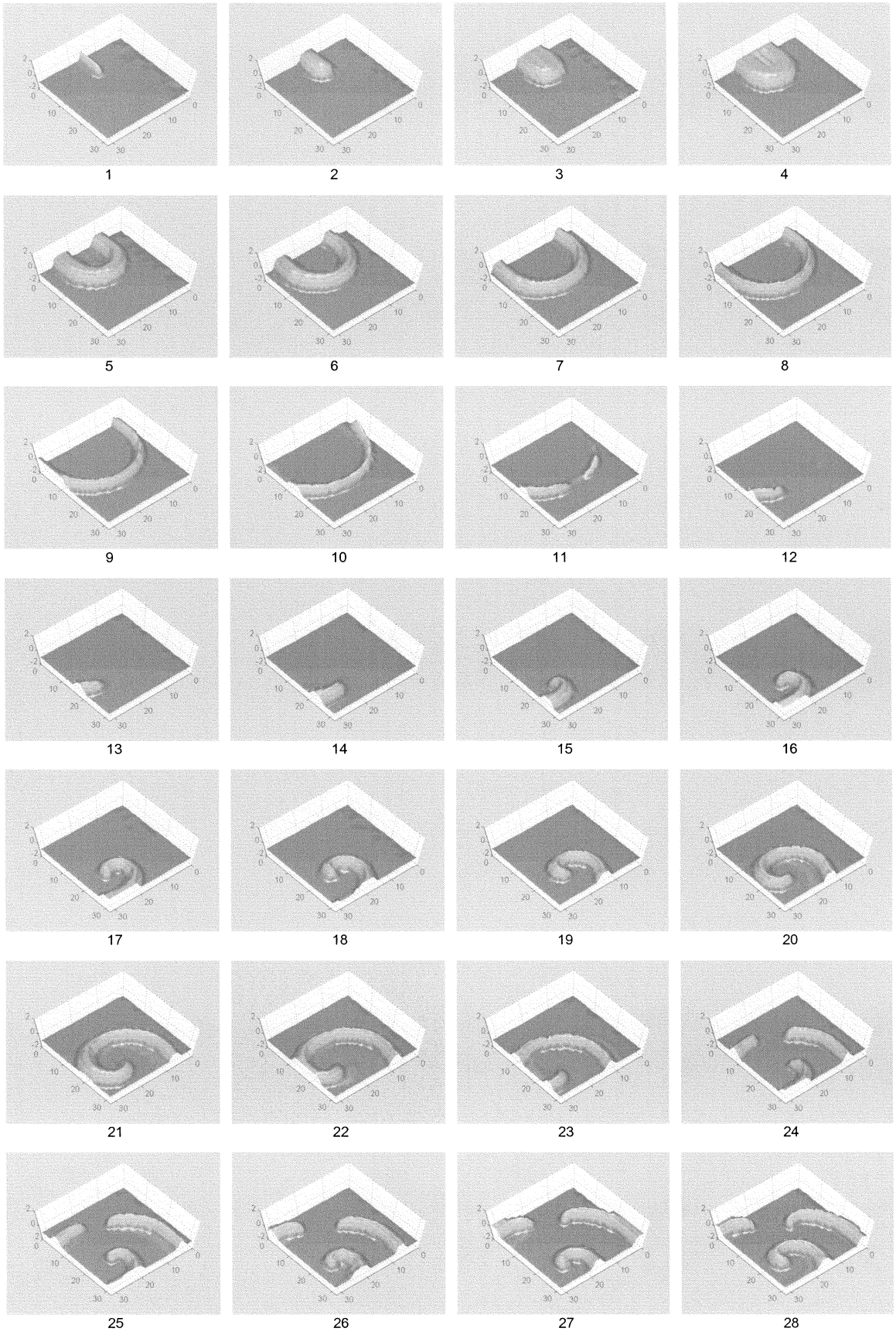


Fig. 28. Spiral wave (only the fastest layer shown).

B. Electrical Test Results

First of all, the I/O interface has been tested. In Fig. 20(a), it can be seen how the chip gives back the **PixReady** pulse, a certain time after the signal **PixCLK** shows a rising edge. Fig. 20(b) displays the start of the image acquisition process. Because of limitations of the test setup, **PixCLK** has a frequency of 2 MHz in this tests, but the chip has been designed to operate with a 10-MHz pixel clock. Fig. 21 shows an image consisting in three vertical stripes of different values (0.6 V, 1.0 V and 1.4 V) being delivered through the I/O channel. In order to test the accuracy of the image acquisition and delivery processes, a ramp of analog values ranging from the minimum to the maximum valid inputs has been transmitted to the chip, stored in the LAM's, and finally recovered. The separation of the recovered samples from their corresponding input in a 256-level representation, is represented in Fig. 22 in LSB as the INL of the input acquisition, storage and recovery. The chip can handle analog data with an equivalent resolution of 7.5 bits.

C. Retinal Behavior Emulation

Image processing algorithms can be programmed on this chip by setting the corresponding switches configuration and by tuning the appropriate interconnection weights — the programming interface is digital while internal coding of the weights is analog. Propagative and wavelike phenomena, similar to those found at the biological retina, can be observed in this chip by just setting the proper coupling between cells in the same or in different layers. For instance, Fig. 23 shows how some spots in the faster layer (second layer) grow until reaching the boundaries of the network, these same spots trigger a slower set of waves in the first layer. These pictures has been generated with the prototype chip by running the network dynamics, from the same initial state, during successively larger periods of time. This permits the reconstruction of the evolution of the state of the cells during the CNN dynamics.

The wavefronts generated at the slower layers can be employed to inhibit propagation in the faster layer, thus generating a trailing edge for the waves in the fast layer. This produces the similar results as the wide field erasure effect observed in the IPL of the retina [see Fig. 24]. Fig. 25 displays a 3-D plot of this effect for a different input. Another interesting effect observed in the OPL of the retina [11] is the detection of spatio–temporal edges followed by de-activation of the patterns of activity. This phenomenon has been also programmed in the chip [see Fig. 26].

D. Active Waves Phenomena

By setting the appropriate interconnection weights, active wave phenomena—the propagation of waves in an energetically active medium, can be observed in the chip. For instance, the triggering of a traveling wave [see Fig. 27], or the generation of spiral waves [see Fig. 28].

VII. Conclusion

The proposed approach supposes a promising alternative to conventional digital image processing for applications related with early-vision and low-level focal-plane image processing. Based on a simple but precise model of the real biological system, a feasible efficient implementation of an artificial vision device has been designed. The peak computing power of this chip is 470 GXPS, what outdoes its digital counterparts due to the fully parallel nature of the processing —based on the analogy not on the simulation. In terms of computing power per silicon area and power consumption, this chip features amongst the more powerful devices reported.

In addition to the advantages in terms of performance features highlighted in the previous table, the chip presented in this paper is capable to generate complex spatio-temporal dynamic processes, in a programmable way and storing intermediate processing results.

Acknowledgment

The authors deeply appreciate the many useful and fruitful discussions with G. Liñán related to chip architecture and circuit design, T. Serrano-Gotarredona regarding the implementation of programmable current mirrors and with D. Bálya and P. Földesy related to the experiments.

References

[1] D. H. Hubel, *Eye, Brain and Vision*. New York: W. H. Freeman, 1988.
[2] F. Werblin, "Synaptic connections, receptive fields and patterns of activity in the tiger salamander retina," *Investigative Ophthalmology and Visual Science*, vol. 32, no. 3, pp. 459–483, Mar. 1991.
[3] B. Roska and F. S. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature*, vol. 410, pp. 583–587, Mar. 2001.
[4] M. Carver, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.
[5] C. Koch and H. Li, Eds., *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*. Los Alamitos, CA: IEEE Computer Society Press, 1995.
[6] M. Alireza, *Vision Chips*. Boston, MA: Kluwer Academic Publishers, 1999.
[7] D. Bálya, B. Roska, E. Nemeth, T. Roska, and F. S. Werblin, "A qualitative model framework for spatio-temporal effects in vertebrate retina," *Proc. 2000 IEEE Conf. on Cellular Neural Networks and their Applications*, pp. 165–170, 2000.
[8] J. C. Gealow and C. G. Sodini, "A pixel-parallel image processor using logic pitch – matched to dynamic memory," *IEEE J. Solid-State Circuits*, vol. 34, no. 6, pp. 831–839, June 1999.
[9] F. Werblin, T. Roska, and L. O. Chua, "The analogic cellular neural network as a bionic eye," *Int. J. Circuit Theory and Applications*, vol. 23, no. 6, pp. 541–69, Nov.–Dec. 1995.
[10] A. Jacobs, T. Roska, and F. S. Werblin, "Methods for constructing physiologically motivated neuromorphic models in CNN's," *Int. J. Circuit Theory Appl.*, vol. 24, no. 3, pp. 315–339, May–June 1996.
[11] C. Rekeczky, B. Roska, E. Nemeth, and F. Werblin, "Neuromorphic CNN models for spatio-temporal effects measured in the inner and outer retina of tiger salamander," *Proc. Sixth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 15–20, May 2000.
[12] K. Boahen, "A retinomorphic chip with parallel pathways: Encoding INCREASING, ON, DECREASING, and OFF visual signals," *Analog Integr. Circuits Signal Processing*, vol. 30, no. 2, pp. 121–35, Feb. 2002.
[13] C. Rekeczky, T. Serrano-Gotarredona, T. Roska, and A. Rodríguez-Vázquez, "A stored program 2nd order/3-Layer complex cell CNN-UM," *Proc. Sixth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 219–224, May 2000.

[14] S. Espejo, R. Carmona, R. Domínguez-Castro, and A. Rodríguez-Vázquez, "A VLSI oriented continuous-time CNN model," *Int. J. Circuit Theory Appl.*, vol. 24, no. 3, pp. 341–356, May-June 1996.
[15] T. Roska and L. O. Chua, "The CNN universal machine: An analogic array computer," *IEEE Trans. Circuits Syst.—II*, vol. 40, no. 3, pp. 163–173, Mar. 1993.
[16] S. Sidney, *Analog Integrated Circuits*. Englewood Cliffs, NJ: Prentice-Hall, 1985.
[17] B. Gilbert, "A precise four-quadrant multiplier with subnanosecond response," *IEEE J. Solid-State Circuits*, vol. 3, no. 4, pp. 365–373, Dec. 1968.
[18] Y. P. Tsividis, "Integrated continuous-time filter design —An overview," *IEEE J. Solid-State Circuits*, vol. 29, no. 3, pp. 166–176, Mar. 1994.
[19] R. Domínguez-Castro, A. Rodríguez-Vázquez, S. Espejo, and R. Carmona, "Four-Quadrant one-transistor synapse for high density CNN implementations," *Proc. Fifth IEEE International Workshop on Cellular Neural Networks and Their Applications*, pp. 243–248, Apr. 1998.
[20] A. Rodríguez-Vázquez, E. Roca, M. Delgado-Restituto, S. Espejo, and R. Domínguez-Castro, "MOST-Based design and scaling of synaptic interconnections in VLSI analog array processing chips," *J. VLSI Signal Processing Systems for Signal, Image and Video Technol.*, vol. 23, pp. 239–266, Nov./Dec. 1999.
[21] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. New York: McGraw-Hill, 1987.
[22] K. C. Smith and A. S. Sedra, "The current conveyor —A new circuit building block," *IEEE Proceedings*, vol. 56, pp. 1368–1369, Aug. 1968.
[23] C. Toumazou, J. B. Hughes, and N. C. Battersby, Eds., *Switched-Currents: An Analogue Technique for Digital Technology*. London, U.K.: Peter Peregrinus, 1993.
[24] T. Serrano and B. Linares-Barranco, "The active-input regulated-cascode current mirror," *IEEE Trans. Circuits Syst.—I*, vol. 41, no. 6, pp. 464–467, June 1994.
[25] R. Carmona, "Analysis and design of CNN-based VLSI hardware for real-time image processing," Ph.D., Universidad de Sevilla, 2002.
[26] R. C. Jaeger, "Tutorial: Analog data acquisition technology. Part I–digital-to-analog conversion," *IEEE Micro*, vol. 24, no. 3, pp. 20–37, May 1982.
[27] G. Liñán, S. Espejo, R. Domínguez-Castro, and A. Rodríguez-Vázquez, "ACE4k: An analog I/ O 6464 visual microprocessor chip with 7-bit analog accuracy," *Int. J. Circuit Theory Appl.*, vol. 30, no. 2-3, pp. 89–116, June 2002.

Ricardo Carmona Galán (M'95) received the degrees of *Licenciado* and *Doctor* (Ph.D.) degrees in physics, in the speciality of electronics, both from the University of Seville, Spain, in 1993 and 2002, respectively.

He was a student at the National Center for Microelectronics at Seville, funded by IBERDROLA S. A. He was a Research Assistant at the Electronics Research Laboratory of the Department of Electrical Engineering and Computer Sciences of the University of California, Berkeley, from 1996 to 1998. He is a member of the Department of Analog Design in the Microelectronics Institute of Sevilla (CNM-CSIC). Since October 1999, he has been an Assistant Professor of the Department of Electronics and Electromagnetism at the School of Engineering of the University of Seville. His main areas of interest are linear and nonlinear analog and mixed-signal integrated circuits, in particular, the design and VLSI implementation of cellular neural networks and analog memory devices for real-time image processing and vision chips.

Dr. Carmona Galán has co-received the Best Paper Award of 1999 from the *International Journal of Circuit Theory and Applications*, and the 2002 *Salvà i Campillo Award*, conceded by the Catalonian Association of Telecommunication Engineers.

Francisco Jiménez-Garrido received the B.S. degree in physics in 1998 and the B.S. degree in electronic engineering in 2002 from University of Seville, Spain. Since 1999, he has been with the Department of Analog Circuit Design of the Spanish Microelectronics Center (Institute of Microelectronics of Seville, IMSE). And he is working toward the Ph.D. degree in the Department of Electronics and Electromagnetism of the University of Seville.

He has research interests in linear and nonlinear analog and mixed-signal integrated circuits for image processing and communication devices.

Rafael Domínguez-Castro received the five-year degree in electronic physics in 1987, the M.S. equivalent in microelectronics in 1989 and the *Doctor en Ciencias Fisicas* degree in 1993, from the University of Seville, Spain.

Since 1987, he has been with the Department of Electronics and Electromagnetism at the University of Seville, where he is currently a professor of electronics. He is also a member of the research staff at the Institute of Microelectronics of Seville – Centro Nacional de Microelectrónica (IMSE-CNM-CSIC), where he is a member of a research group on Analog and Mixed-Signal VLSI. His research interests are in the design of embedded analog interfaces for mixed-signal VLSI circuits, design of CMOS imagers and CMOS focal plane array processors and development on CAD for automation of building blocks analog design, specially optimization and automatic sizing of basic building blocks for integrated circuits.

Dr. Domínguez-Castro has co-received the 1995 Guillemin-Cauer award of the IEEE Circuits and Systems Society, and the Best Paper Award of the 1995 European Conference on Circuit Theory and Design.

Servando Espejo (M'96) received the Licenciado en Física degree, the M.S. degree equivalent in microelectronics, and the Doctor en Ciencias Físicas degree from the University of Seville, Spain, in June 1987, July 1989, and March 1994, respectively.

He is currently Profesor Titular of Electronics at the Department of Electronics and Electromagnetism of the University of Seville, and also with the Department of Analog Circuit Design of the Spanish Microelectronics Center. From 1989 to 1991, he was an intern at AT&T Bell Laboratories in Murray Hill, NJ, and an employee of AT&T Microelectronics of Spain. His main areas of interest are linear and nonlinear analog and mixed-signal integrated circuits, including neural networks electronic realizations and theory, vision chips, massively parallel analog array processing systems, chaotic circuits, and communication devices.

Dr. Espejo has co-received the 1995 Guillemin-Cauer award of the IEEE Circuits and Systems Society, and the best paper award of the 1995 European Conference on Circuit Theory and Design.

Tamás Roska (M'87–SM'90–F'93) received the Diploma in electrical engineering from the Technical University of Budapest, Hungary, in 1964 and the Ph.D. and D.Sc. degrees in Hungary in 1973 and 1982, respectively.

Since 1964, he has held various research positions. During 1964–1970, he was with the Measuring Instrument Research Institute, Budapest, between 1970 and 1982 with the Research Institute for Telecommunication, Budapest (serving also as the head of department for Circuits, Systems and Computers) and since 1982, he has been with the Computer and Automation Institute of the Hungarian Academy of Sciences, where for 15 years, he has been the head of the Analogic and Neural Computing Research Laboratory. He has taught several courses at various universities, presently, at the Technical University of Budapest, at the University of California at Berkeley, and very recently at the Pázmány P Catholic University in Budapest. He is teaching courses on "Emergent Computations" and "Cellular Neural Networks." In 1974, and since 1989, he has been a Visiting Scholar at the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, and recently a Visiting Research Professor at the Vision Research Laboratory of the University of California at Berkeley. He is presently a Dean of the Faculty of Information Technology at the Pázmány P. Catholic University, Budapest. His main research areas are cellular neural networks, nonlinear circuit and systems, neural circuits, visual computing and analogic spatial–temporal supercomputing. He has published more than 200 research papers and four books (partly as a coauthor), and held several guest seminars at various universities and research institutions in Europe, the United States, and Japan.

Prof. Roska is a member of several Hungarian and international Scientific Societies. Since 1975, he has been a member of the Technical Committee on Nonlinear Circuits and Systems of the IEEE Circuits and Systems Society. Between 1987–1989, he was the founding Secretary and later he served as Chairman of the Hungary Section of the IEEE. Recently, he has served twice as Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, Guest Co-Editor of special issues on Cellular Neural Networks of the *International Journal of Circuit Theory and Applications* (1992, 1996, 1998, and 2000), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (1993 and 1999), and the *Journal of VLSI Signal Processing Systems* (1999). He is a member of the Editorial Board of the *International Journal of Circuit Theory and Applications*. He is a member of the Technical Committee on Multimedia and the Technical Committee on Neural Networks of the IEEE. He received the IEEE Fellow award for contributions to the qualitative theory of nonlinear circuits and the theory and design of programmable cellular neural networks. In 1993, he was elected to be a member of the Academia Europaea (European Academy of Sciences, London) and the Hungarian Academy of Sciences. For technical innovations he received the D. Gabor Award, for establishing a new curriculum in information technology and for his scientific achievement he was awarded the A. Szentgyörgyi Award and the Széchenyi Award, respectively. In 1994, he became the elected active member of the Academia Scientiarium et Artium Europaea (Salzburg). In 1998, he established and became the first Chair of the Technical Committee on Cellular Neural Networks and Array Computing of the IEEE Circuits and Systems Society. In 2000, he received the IEEE Millennium Medal and the Golden Jubilee Award of the IEEE Circuits and Systems Society.

Csaba Rekeczky received the B.S. degree in electrical engineering and the Ph.D. degree from the Technical University of Budapest, Hungary, in 1993 and 1999, respectively.

He is with the Computer and Automation Institute of the Hungarian Academy of Sciences, working at the Analogic and Neural Computing Research Laboratory. He was a research assistant at Department of Electrical Engineering of the Tokushima University, Japan, from 1994 to 1995. He was a visiting scholar of the Department of Electrical Engineering and Computer Sciences of University of California, Berkeley, from 1997 to 1998. His main areas of interest are cellular neural and nonlinear networks, neuromorphic modeling and image processing with parallel nonlinear array processors.

Dr. Rekeczky received the 1995 award for outstanding Ph.D. students at the Computer and Automation Institute of the Hungarian Academy of Sciences, and the 1993 award of the Hungarian Scientific Society of Measurement and Automation for diploma thesis.

István Petrás, photograph and biography not available at the time of publication.

Angel Rodríguez-Vázquez (M'80–SM'95–F'96) is a Professor of Electronics at the Department of Electronics and Electromagnetism (University of Seville). He is also a member of the research staff of the Institute of Microelectronics of Seville – Centro Nacional de Microelectrónica (IMSE- CNM) – where he is heading a research group on Analog and Mixed-Signal Integrated Circuits. His research interests are in the design of analog front-ends for mixed-signal circuits and systems-on-chip, telecom circuits, CMOS imagers and vision chips, sensory-processing-actuating systems-on-chip and bio-inspired integrated circuits. In these topics, he has published seven books, 36 book chapters in other books, about 100 journal papers, and about 300 conference papers. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I from 1993 to 1995, as Guest Editor of the IEEE TCAS-I special issue on "Low-Voltage and Low-Power Analog and Mixed-Signal Circuits and Systems" (1995), as Guest Editor of the IEEE TCAS-II special issue on "Advances in Nonlinear Electronic Circuits" (1999), as Guest Editor of the IEEE TCAS-I special issue on "Bio-Inspired Processors and Cellular Neural Networks for Vision" (1999), and as chair of the IEEE-CAS Analog Signal Processing Committee (1996). Currently, he is an Associate Editor for IEEE TCAS-II and Guest Editor of the IEEE TCAS-I special issue on "Advances on Analog-to-Digital and Digital-to-Analog Converters". He is also member of the editorial staff of the *International Journal on Circuit Theory and Applications* and the *Analog Integrated Circuits* (New York: Wiley) and *Signal Processing Journal* (New York: Kluwer Academics) He was co-recipient of the 1995 Guillemin-Cauer award of the IEEE Circuits and Systems Society, the Best Paper Award of the 1995 European Conference on Circuit Theory and Design, and the 1999 Best Paper Award of the *International Journal on Circuit Theory and Applications*. In 1992, he also received the Young Scientist Award of the Seville Academy of Science. In 2002, he received el VII Premi Salvá i Campillo al Projecte Més Original. In 1996, he was elected to the degree of Fellow of the IEEE for "contributions to the design and applications of analog/digital nonlinear ICs."