A 0.5μm CMOS Random Access Analog Memory Chip for TeraOPS Speed Multimedia Video Processing

*Ricardo Carmona[1], Servando Espejo[1], Rafael Domínguez-Castro[1], Ángel Rodríguez-Vázquez[1], Tamás Roska[2], Tibor Kozek[3], Leon O. Chua[3]

[1] Instituto de Microelectrónica de Sevilla-CNM-CSIC-Universidad de Sevilla.
Edificio CICA, Avda. Reina Mercedes s/n, 41012-Sevilla, Spain.
Ph. No.: 34+ 954 239923, Fax: 34+ 954 231832 E-mail: rcarmona@imse.cnm.es
[2] MTA-SZTAKI, Analogic & Neural Computing Laboratory,
Computer and Automation Institute of the Hungarian Academy of Science,
Budapest, H-1111, Hungary.
[3] Electronics Research Laboratory, University of California, Berkeley
258M Cory Hall, Berkeley, CA 94720, USA.

Submitted for revision to the IEEE Transactions on Multimedia
September 7, 1998

ABSTRACT

Data compressing and coding and communications in object oriented multimedia applications like telepresence, computer-aided medical diagnosis or telesurgery require an enormous computing power – in the order of Trillion Operations per Second (TeraOPS). Compared with conventional digital technology, Cellular Neural/Nonlinear Network (CNN) based computing is capable of realizing these TeraOPS-range image processing tasks in a cost-effective implementation. To exploit the computing power of the CNN Universal Machine (CNN-UM), the CNN Chipset architecture has been developed – a mixed-signal hardware platform for CNN-based image processing. One of the non-standard components of the chipset is the cache memory of the analog array processor, the Analog Random Access Memory (ARAM). This paper reports an ARAM chip that has been designed and fabricated in a 0.5μm CMOS technology. This chip consists of a fully addressable array of 32×256 analog memory registers and has a packing density of 637 analog-memory-cells/mm^2. Random and non-destructive access of the memory contents is available. Bottom-plate sampling techniques have been employed to eliminate harmonic distortion introduced by signal-dependent feedthrough. Signal coupling and interaction have been minimized by proper layout measures, including the use of protection rings and separated power supplies for the analog and the digital circuitry. The prototype features an equivalent resolution of up to 7 bits – measured by comparing the reconstructed waveform with the original input signal. Measured access times for writing /reading to/from the memory registers are 200ns and 800ns, respectively. I/O rates via the 16-line wide I/O bus exceed 10Msamples/s. Storage time at room temperature is in the 80 to 100ms range, without accuracy loss.

EDICS: 2-CIRC, 2-EXTN

Front page footnotes[1] [2]

1. This work is supported by the JSEP Grant No. FDF49620-97-1-0220-03/98 and by the ONR Grant No. N00014-98-1-0052
2. Research of the authors from IMSE-CNM (CSIC) has been supported by the spanish CICYT (Project TIC96-1392-C0202 SIVA) and the EU (Project ESPRIT IV 27077-DICTAM).

I. INTRODUCTION

*C*ellular *N*eural *N*etwork**s** (CNNs) are analog nonlinear dynamic processor arrays in which direct interconnections among the basic processing units are restricted to a finite local neighborhood [1]. Their potential for image processing applications was advanced shortly after their invention [2] and is based on the fact that many image processing tasks can be realized by means of weighted local interactions between neighbouring pixels [1][3]. Because of their inherently parallel processing architecture, CNNs achieve a high computation speed in the realization of these tasks. Besides, their uniformity and local connectivity make them especially suited for VLSI implementation [4][5][6][7][8].

The CNN paradigm provides the framework for the definition of an algorithmically programmable analog array computer with supercomputer power on a chip: the *CNN* *U*niversal *M*achine (CNN-UM) [9]. Its dual-computing property enables the realization of highly complex image processing tasks by means of an on-chip analogic – analog and logic – stored program, and renders it a highly competitive alternative to the conventional digital approach to parallel image processing [3]. For example, almost 10^4 Pentium® are required for the TeraFLOPS array computer shipped by Intel® in 1997 [10]. Whenever accuracy in the computation is not a critical issue, as it actually happens in early-vision tasks [11], CNN-UM analogic chips are advantageous in terms of power consumption and computation speed as compared to these digital counterparts [12].

The working CNN-UM chips reported to date, with up to 20×22 [5], 16×16 [6] and 48×48 [7] cells, respectively, contain a much smaller number of pixels than practical image sizes. For instance, conventional television applications require 644×483 pixels per frame – not including the necessary scanning overhead involved in any display system [13]. Although larger chips will be available in the near future – 64×64 [8] – processing of practical size images requires the adoption of system-level solutions to overcome technology limitations on the number of parallel processing cells [14]. Particularly, multiplexing the CNN-UM processors, i.e. making them operate onto a fraction of the complete input image at a time, appears sometimes the only way of operation.

One possible strategy is using space-multiplexed, or multichip, CNN hardware [15]. In a multichip CNN, large arrays are built by interconnecting chips with a smaller number of cells. Each module operates simultaneously onto a fraction of the input image which is, in this way, processed in parallel. One drawback of this approach are the random fluctuations of the process

parameters among the different processors. This may cause incorrect or inaccurate operation and, thus, requires the incorporation of different correction strategies; for instance, using tuning to correct parameter deviations during the generation of the analog weights [16]. However, the major drawback of multichip CNNs is the very large number of chip modules and, specially, off-chip interconnections needed. For instance, around 8E3 chips and 4.1E5 connections are required to process a 644×483 pixels video frame using the 6×6 CNN module reported in [17]. And around 75 chips and 3.8E4 connections are needed using the 64×64 last generation processor reported in [8].

A different approach to using small size CNN chips for large images is time-multiplexing. By taking advantage of the computing power of the CNN-UM, a single chip can be used to process a complete video frame by operating on a fraction of the image at a time. A frame rate of 40Hz – adequate for high quality video applications [13] – represents a data flow of 12.3E6 pixels per second. Real-time processing of such rate demands 81ns processing time per pixel. Thus, by allowing for a 2-pixel wide overlap between image subsets in each scan direction – required for correct processing of the border pixels [18] –, a 32×32 CNN chip should be capable to process each subimage in about 73.8μs; and 320μs for a 64×64 chip. Because the time constant of CNN-UM chips is in the range of 1μs [4][8] we can conclude that the time-multiplexed approach is feasible and, hence, constitutes a more cost-effective solution than the multichip one.

The time-multiplexed approach requires the definition and development of an appropriate hardware platform for the CNN processor: the CNN chipset [19]. It is designed to support high speed data transmission and interfacing of the analogic processor to the sensory devices and the digital host circuitry. The *Analog RAM* (ARAM) is one of the non-standard parts of this chipset. It is a high-speed short-term memory buffer that operates as the cache memory [20] of the CNN processor. A straightforward realization of the required functionality would be the use of a conventional digital RAM interfaced with A/D and D/A converters. However, the resulting I/O rates between the memory and the processor would render this solution impractical. In order to realize a direct data interchange between the memory and the processor, avoiding data conversion, the implementation of a truly analog RAM chip is proposed. For full compatibility with the digital host environment and reduced fabrication cost, this ARAM should be designed using standard CMOS.

The problem of on-chip analog signal storage has been faced by different authors in connection to quite diverse applications. Particularly, CMOS realizations of scanning delay-lines

for video processing are presented in [21] and a high-speed SC sampling circuit is reported in [22] to capture analog waveforms from an array of sensory devices. However, no random access or non-destructive reading of the memory contents can be done. An ARAM for early vision applications was reported in [23]. However, its accuracy relays on mismatch compensation and no switching error reduction strategies are adopted. In this paper an improved version of a well-known *S*ample-and-*H*old (S/H) circuit is proposed to implement a fully addressable analog memory chip. It is realized in a $0.5\mu m$ CMOS single-poly triple-metal technology and allows non-destructive reading and random access to 32×256 memory locations with a cell density of 637 cells/mm^2. It features around 7 bits equivalent resolution with writing/reading access times of 200ns/800ns, respectively, and storage time at room temperature in the 80 to 100ms range. Besides, its power consumption is of only 73mW from a 3.3V power supply – achieved through multiplexing of the active S/H circuitry.

In the next section, a brief review of video signal processing with CNNs is given together with the specifications of the ARAM in the CNN chipset. The, Sect. III reports the details of the ARAM prototype chip architecture and circuit design. Test results are displayed and discussed in Sect. IV. And finally, a summary of concluding remarks is given.

II. VIDEO SIGNAL PROCESSING WITH CNNs

A. CNN based image processing and ARAM chip specifications

In the CNN Universal Machine – which has been demonstrated to be universal in the Turing sense [24] – programmable nonlinear analog dynamics are combined with programmable logic operations and analog and logic distributed memories. Complex image processing tasks are described by an analogic program [25], consisting of a sequence of analog and logic operations. This analogic program has to be compiled into a platform-dependent machine code to be executed by a particular hardware implementation. Fig. 1 depicts a diagram of the CNN-UM and its principal building blocks: the basic processing units (cells), and the *G*lobal *A*nalogic *P*rogramming *U*nit (GAPU). The GAPU stores the analogic program and controls its execution. For this purpose, it is divided into two main functional blocks. First, the storage unit consisting of the *A*nalog *P*rogram *R*egister (APR), the *L*ogic *P*rogram *R*egister (LPR) and the *S*witch *C*onfiguration *R*egister (SCR). They contain the machine code instructions for the analog and logic operations and the switch configuration, respectively. Second, the *G*lobal *A*nalogic *C*ontrol *U*nit (GACU) that decodes these instructions into a microcode that is transmitted to the cells. Inside

the basic cell, three parts can be distinguished which are responsible for signal processing, storage and control of the operation – Fig. 1. For the implementation of the programmable analog dynamics, the CNN core contains the integrator and the limiter blocks. Synaptic operators can be considered as a part of the analog processing unit. A *L*ocal *L*ogic *U*nit (LLU) realizes programmable logic operations between stored binary magnitudes. Short-term storage of intermediate signals is realized by *L*ocal *A*nalog and *L*ogic *M*emories (LAMs and LLMs). Signal transference and operation control is performed by the *L*ocal *C*ommunication and *C*ontrol *U*nit (LCCU). And, finally, data exchange between the cell array and the external circuitry is realized via the *L*ocal *A*nalog *O*utput *U*nit (LAOU).

In order to exploit the computing power of this architecture, The CNN chipset of Fig. 2 has been developed to interface the CNN-UM processor to the sensors and the digital environment. Data transmission is supported by three different buses. A high-speed analog bus connects the processor, the ARAM and the video signal sources. The width of this analog bus is determined by the I/O bus of the CNN-UM chip, otherwise it will limit the total throughput of the

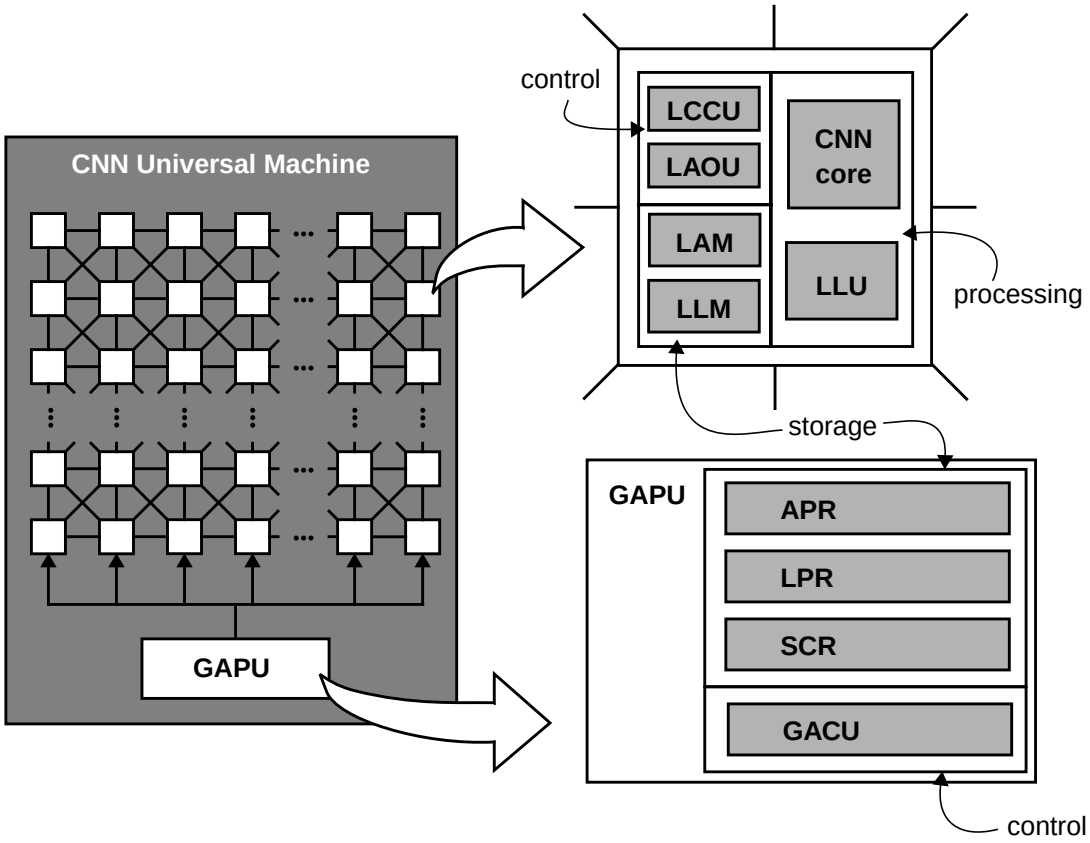

Fig. 1: CNN Universal Machine architecture, basic processing cell and global analog-and-logic programing unit.

system. Digital data are transmitted via the digital bus, which is interfaced to the analog bus through A/D and D/A converters. In addition, there is a digital instruction bus. The required storage capacities and local throughput values have to be evaluated to determine the specifications for the non-standard parts, i. e. the CNN-UM and the ARAM.

Assume an input image composed of $M_i \times N_i$-pixels (Fig. 2). It has to be decomposed into $M_a \times N_a$-pixel subsets that are temporarily stored one-by-one in the analog RAM chip for their processing. However, pixels in the border of this $M_a \times N_a$ window will not be properly processed unless a certain overlap between the image fractions is allowed. Therefore, m_o and n_o pixel overlaps in the vertical and the horizontal direction, respectively, are considered. Taking this into account, a straightforward calculation shows that,

$$k = \frac{(M_i - m_o) \times (N_i - n_o)}{(M_a - m_o) \times (N_a - n_o)} \tag{1}$$

subimages are needed to cover the whole image. Each of these subimages has to be captured, processed and downloaded, thus resulting into the following total processing time T_i for the $M_i \times N_i$ input frame,

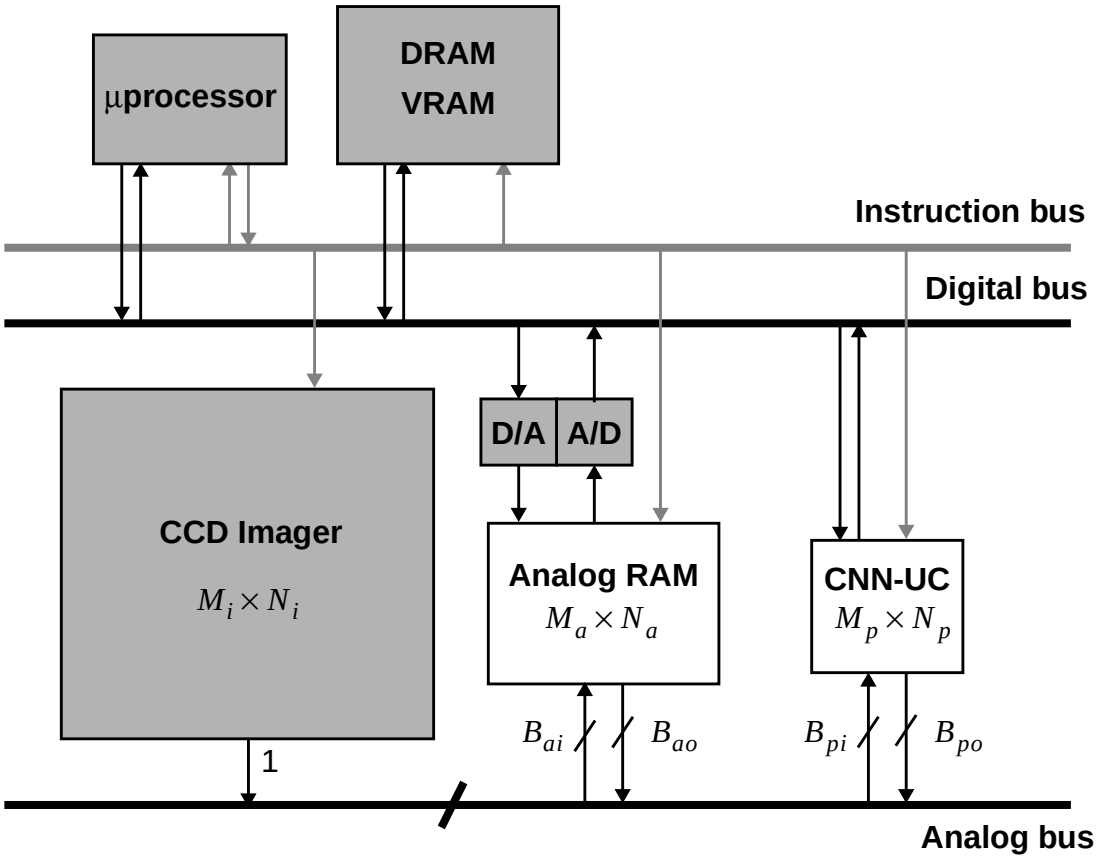

Fig. 2: Diagram of the CNN chipset architecture.

$$T_i = \frac{(M_i - m_o) \times (N_i - n_o)}{(M_a - m_o) \times (N_a - n_o)} \cdot (T_{ai} + T_{ap} + T_{ao}) \tag{2}$$

where T_{ai}, T_{ao} and T_{ap} are the times required to acquire, download and process each subimage, respectively. For the former two times, and assuming that B_{ai} and B_{ao} are the widths of the input and output buses of the ARAM, the following is obtained,

$$\begin{aligned} T_{ai} &= \frac{M_a \times N_a}{B_{ai}} \cdot \tau_{ai} \\ T_{ao} &= \frac{M_a \times N_a}{B_{ao}} \cdot \tau_{ao} \end{aligned} \tag{3}$$

where τ_{ai} and τ_{ao} are the times required for writing and reading, respectively, an analog register of the ARAM chip.

With regards to the processing time T_{ap} in (2) we have to take into account that, in the more general case, the processor size is smaller than the ARAM size. Hence, the necessity arises for another multiplexation. Assume the size of the processor is $M_p \times N_p$ and that each analogic program contain n_i data acquisition steps, n_{ap} analog processing steps, n_{lp} logic processing operations, and n_d data downloads. Thus, the time needed to perform the analogic algorithm on each $M_a \times N_a$ subset is given by,

$$T_{ap} = \frac{(M_a - m_o) \times (N_a - n_o)}{(M_p - m_o) \times (N_p - n_o)} \cdot T_{pp} \tag{4}$$

where,

$$T_{pp} = n_i T_{pi} + n_{ap} T_{pap} + n_{lp} T_{plp} + n_d T_{po}, \tag{5}$$

and T_{pap} and T_{plp} are the times required for the analog and the digital circuitry of the CNN-UM to settle and complete the logic operation, respectively. These parameters are part of the timing specs of the CNN-UM chip. T_{pi} and T_{po} in the expression above represents I/O times which are given by,

$$\begin{aligned} T_{pi} &= \frac{M_p \times N_p}{B_{pi}} \cdot \tau_{pi} \\ T_{po} &= \frac{M_p \times N_p}{B_{po}} \cdot \tau_{po} \end{aligned} \tag{6}$$

where B_{pi} and B_{po} are the widths of the input and output buses of the CNN-UM, respectively, and τ_{pi} and τ_{po} are the times required for updating and downloading analog data from one cell

of the CNN array – also defined as temporal specs of the processing chip.

Assume a frame rate of N_f frames per second. The following must be accomplished in order to process the whole input image ($M_i \times N_i$) in real-time:

$$T_i \leq \frac{1}{N_f} \tag{7}$$

Thus, from the mathematics above, the following design equation can be obtained,

$$\frac{1}{N_f} \geq \frac{M_a N_a (M_i - m_o)(N_i - n_o)}{(M_a - m_o)(N_a - n_o)} \left(\frac{\tau_{ai}}{B_{ai}} + \frac{\tau_{ao}}{B_{ao}} \right) + \frac{(M_i - m_o)(N_i - n_o)}{(M_p - m_o)(N_p - n_o)} T_{pp} \tag{8}$$

We find convenient to illustrate this design equation using typical values. For instance, consider a frame rate of 40 frames per second, an input image of 512×512 pixels, an analog RAM buffer of 32×256 registers and a CNN array of 32×32 cells. Consider as well a 2-pixel wide overlap in both, vertical and horizontal, scan directions and 16-line wide I/O buses. Then, for a typical I/O time of 500ns per memory cell, the CNN-UM chip should be capable to complete the analogic algorithm over each $M_p \times N_p$ subimage in less than 26μs – well within the specs of CMOS CNN-UM chips [4][8]. The larger the CNN processor size the faster the system is. Besides, pipelined architectures and some interleaving of the memory blocks can be used for a more relaxed constrain on the processing time.

Let us now derive the specifications for the ARAM block. It must exhibit the following features for proper usage within the CNN chipset architecture,

- *Non-volatility.* The analog information contained in the memory registers should be maintained for a sufficiently long time. In this case, and because of the high-speed of the computation, a storage time of 100-200ms should be enough. Being a cache memory, power-off non-volatility is not necessary.
- *Resolution*. Accuracy levels for a wide range of early-vision tasks are in the 0.8-1.5% range. It represents an equivalent resolution of 6-7 bits. Cooperative phenomena derived from the parallel processing nature of CNNs, like hyperacuity [26], allow for a moderate resolution requirement.
- *Random access*. Some analogic algorithms designed for the CNN Universal Machine [27] require repeated reading and writing to a specific location of the memory. Thus, random access to any memory register should be provided.
- *Non-destructive reading*. For the same reason, reading any memory location should not affect the contents, because access to then might be required several times in an ana-

logic program.

- *High-speed*. Narrow access times to the memory allow a faster operation. Although difficult to achieve, access times smaller than 100ns will be required to realize complex image processing tasks in real-time.
- *Input/Output*. On the one hand, a serial analog input channel is needed to interface the image acquisition devices – CCD imager, composite-video signal source, ... On the other, the communication with the CNN-UM processor is accelerated by the use of parallel analog channels of width B_{pi} and B_{po} – see Fig. 2.

Obviously, the memory cell should be the smallest possible to allow obtaining the larger possible memory arrays without important yield problems. Besides, compatibility with digital CMOS voltage levels is implicitly assumed for integration with a digital environment at the system level via the instruction and digital data buses.

B. Video signal interface to the CNN chipset

A standard composite-video signal has a limited bandwidth of 5MHz and must, hence, be sampled at a minimum rate of 10Msamples/s. The maximum time interval between consecutive-samples is hence 100ns. In addition, the composite-video signal carries information on the luminance and chrominance of each pixel, and a synchronization pulse generated by the raster scanning of the object picture. Fig. 3 displays the envelope spectrum of a NTSC coded signal and the waveform of a scan line. Although NTSC is a color encoding standard, it is also commonly used to refer to its associated scanning standard 525/59.94. A simple implementation of a video-signal interface to the CNN chipset is portrayed in Fig. 4. It can be built up by using off-the-shelf components. Here, the incoming video signal (NTSC coded in this case) is fed into a video decoder chip. It is decomposed into its luminance (**Y**) and chrominance (**C**) components plus the recovered timing signals. By now, only the luminance component will be of interest as we are not considering color information processing. After some amplification and level shifting, if required, the ARAM chip take samples of the input via the serial input channel. Control signals and memory address codes are generated by some programmable logic device from the synchronization pulses extracted from the raw input by the NTSC decoder. Time requirements for the ARAM in this video interface can be easily derived. Using a square pixel grid -- equal horizontal and vertical sample pitch, each frame in the 525/59.94 scanning standard is composed of 780×525 pixels, this includes the required blanking intervals. It means that each line of the image, containing 780 pixels, will be transmitted in 64μs approximately. Acquisition of this

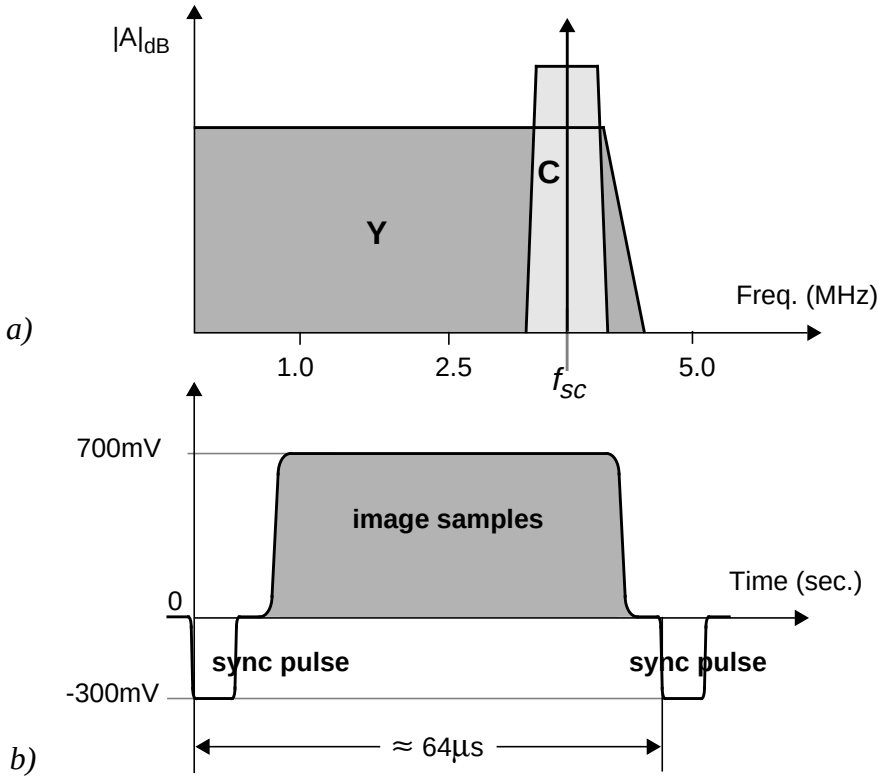

Fig. 3: a) Envelope spectrum of a NTSC signal, showing the luminance (Y) and the chrominance (C) centered around the color subcarrier (f_{sc} = 3.86MHz), and b) waveform of a scan line of the associated 525/59.94 scanning standard.

serial data stream has to be realized at more than 12Msamples/s, a time interval of 82ns between samples. By the use of the ARAM input bus (16-line wide) and an appropriate de-multiplexing of the analog data stream, samples can be taken at 1.31µs. It is interesting to point out that one of the tasks to be performed by the ARAM chip is the re-organization of the information in a shape that can be processed by the CNN chip. Lines of the image are sampled by the ARAM

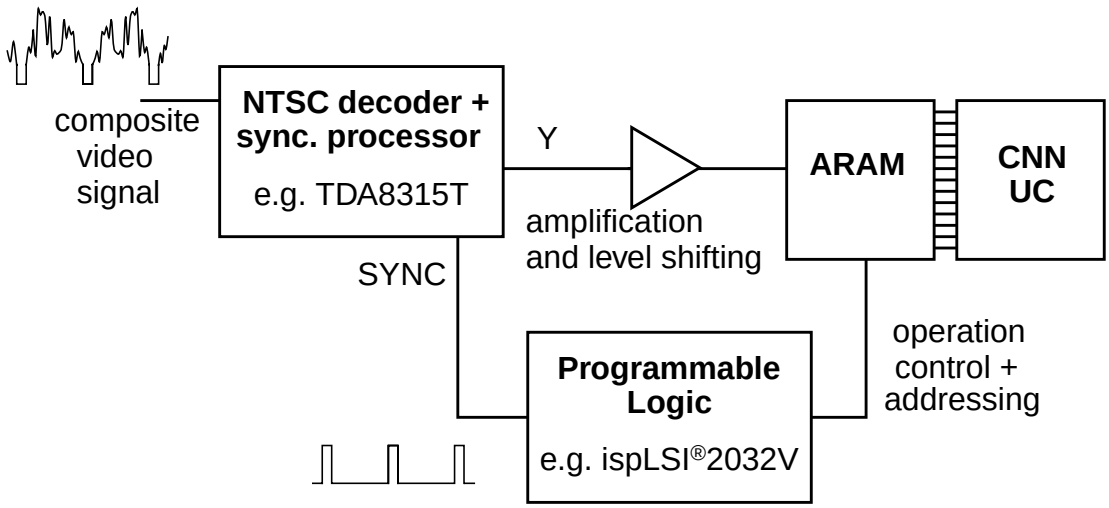

Fig. 4: Composite-video signal interface to the CNN chipset.

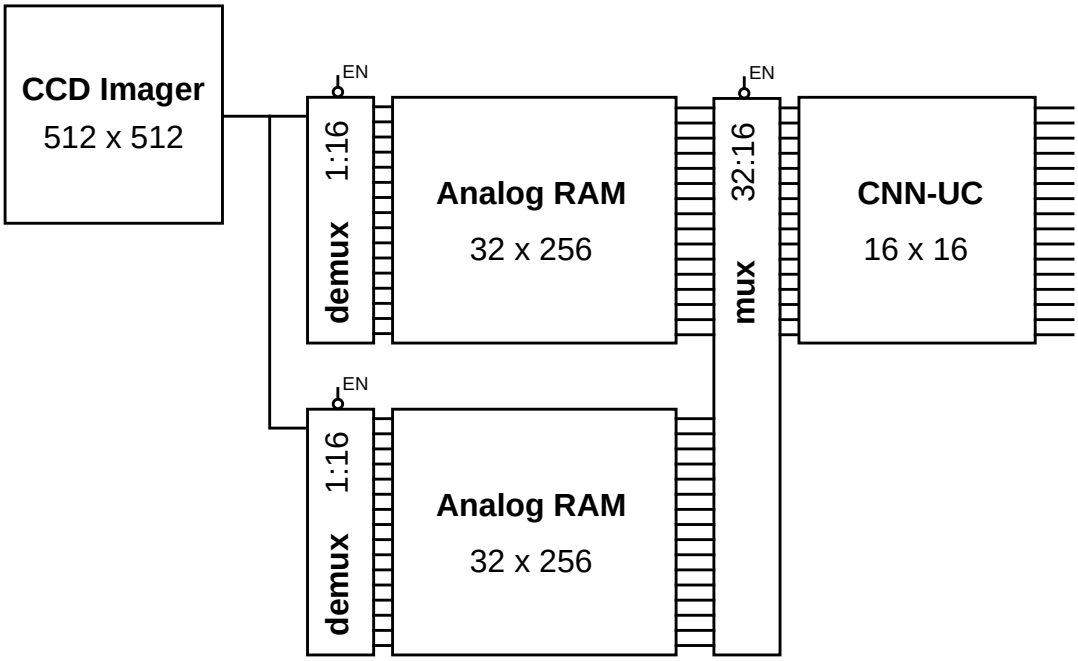

Fig. 5: ARAM chip interleaving for a pipelined architecture of the CNN chipset.

one-by-one but the processor operates on $M_p \times N_p$-pixel pieces of the input. Full addressing of the memory array and random access to its contents make this re-ordering feasible.

A different approach using interleaved memory chips and a pipelined structure is depicted in Fig. 5. Pixel rate of the CCD imager is about 10.49Mpixels/s for a frame rate of 40Hz, what means 95ns per pixel in a serial transmission. Using the 16-line input bus of the ARAM samples can be taken at 1.53μs intervals. It means that we have a 0.78ms time period for updating the contents in the memory chip. By the use of a second analog RAM, the process can be pipelined in the way that the first ARAM is working with the CNN processor while the second is being updated with the next 16 lines of the input image. After that, the role of the memory chips is reversed thus speeding up the system to meet the requirements for real-time image processing.

III. CIRCUIT DESIGN AND PROTOYPE SYSTEM ARCHITECTURE

A. Errors in the sampling process: speed-accuracy trade-off

The non-idealities in the *S*ample-and-*H*old (S/H) process are evaluated by using the circuit of use of Fig. 6. a, composed of a pass transistor and a storage capacitor [28]. Its operation is affected by deterministic and random errors.

Let us consider the deterministic errors first. During the track phase – while the pass transistor is ON – the finite ON resistance of the pass transistor originates a sample acquisition

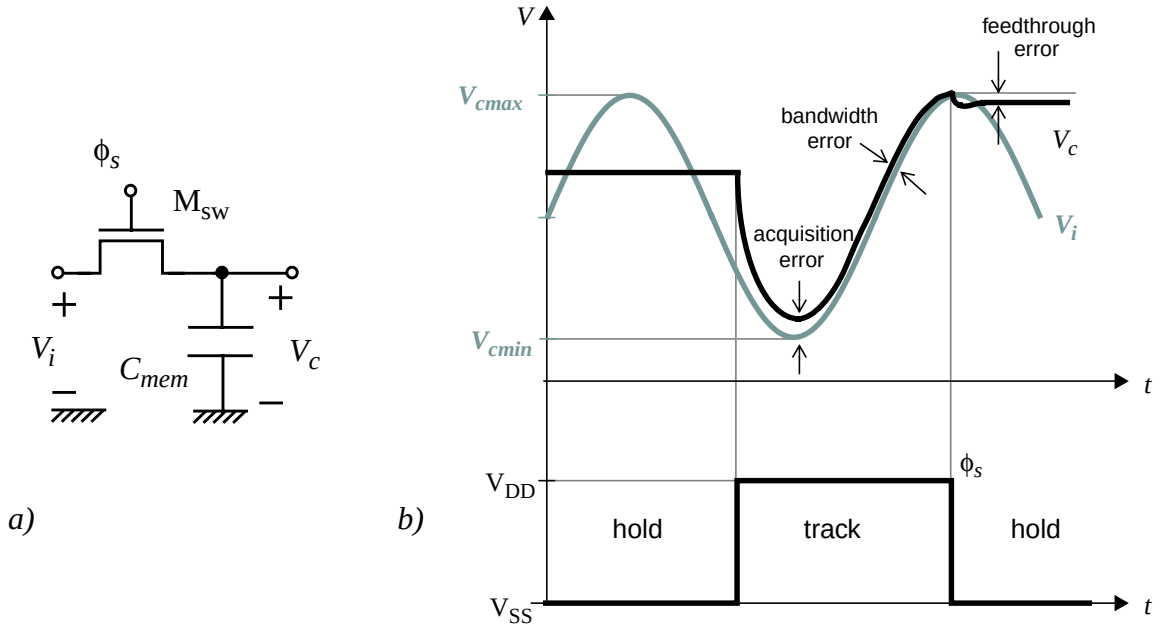

Fig. 6: a) Storage capacitor and pass transistor and b) errors in the sampling process

delay. Let us call V_{cmin} and V_{cmax} to the minimum and maximum input signal voltages, respectively. The maximum step size occurs when the preceding sampled value is V_{cmin} and the next one is V_{cmax} or vice versa – see Fig. 6. b. In this case, the capacitor voltage is given by,

$$V_c(t) = V_{cmin}\exp\left(-\frac{t}{\tau}\right) + V_{cmax}\left[1 - \exp\left(-\frac{t}{\tau}\right)\right] \tag{9}$$

where $\tau = R_{sw}C_{mem}$. Therefore, the acquisition error decreases exponentially with time:

$$\varepsilon_a(t) = -(V_{cmax} - V_{cmin})\exp\left(-\frac{t}{R_{sw}C_{mem}}\right) \tag{10}$$

establishing the maximum sampling rate. The minimum acquisition time required for this error to be smaller than 1/2 *L*east *S*ignificant *B*it (LSB) – that is less than $(V_{cmax} - V_{cmin})/2^{N+1}$, where N is the number of bits corresponding to the equivalent digital resolution, is given by:

$$\Delta t \geq (N+1)\tau \ln 2 \tag{11}$$

Once the acquisition transient settles, the S/H circuit is in track mode. Now, the voltage at the capacitor attempts to follow the input voltage. The circuit formed by the pass transistor and the storage capacitor C_{mem} acts as a single-pole low-pass filter. Although it does not have an important incidence on the amplitude of the tracked output, the phase shift introduced by this low-pass characteristic can be specially harmful when it is operating onto signals modulated in phase. For a given frequency, this phase shift is calculated as:

$$\phi = -\text{atan}(R_{sw}C_{mem}2\pi f) \tag{12}$$

A further deterministic error arises at the falling edge of the clock due to clock feedthrough. It manisfests as a small discrepancy between the sampled voltage and the actually held magnitude which can be expressed as

$$\varepsilon_f \approx \frac{C_{gds}}{C_{mem}+C_{gds}}[V_{SS}-V_{c0}-V_T(V_{c0}-V_{SS})] \tag{13}$$

where V_{c0} is the undegraded sampled voltage value, C_{gds} is the parasitic overlap capacitor and $V_T(V_{c0}-V_{SS})$ is a nonlinear function which accounts for the substrate effect. The feedthrough error is, hence, signal-dependent, and therefore, it will induce harmonic distortion at the output. The second and third harmonic distortion terms can be calculated as:

$$\text{HD}_2 = \frac{1}{8}A\left|\frac{\dfrac{C_{gds}}{C_{mem}+C_{gds}}\cdot\dfrac{\gamma}{4(\phi_B+V_C\big|_Q-V_{SS})^{\frac{3}{2}}}}{1-\dfrac{C_{gds}}{C_{mem}+C_{gds}}\cdot\left(1+\dfrac{\gamma}{2\sqrt{\phi_B+V_C\big|_Q-V_{SS}}}\right)}\right| \tag{14}$$

and

$$\text{HD}_3 = \frac{1}{96}A^2\left|\frac{\dfrac{C_{gds}}{C_{mem}+C_{gds}}\cdot\dfrac{3\gamma}{8(\phi_B+V_C\big|_Q-V_{SS})^{\frac{5}{2}}}}{1-\dfrac{C_{gds}}{C_{mem}+C_{gds}}\cdot\left(1+\dfrac{\gamma}{2\sqrt{\phi_B+V_C\big|_Q-V_{SS}}}\right)}\right| \tag{15}$$

where $A = V_{cmax}-V_{cmin}$ is the full scale signal voltage.

The equation (11) shows that the acquisition time decreases with the capacitor decreasing and the aspect ratio – W/L – of the pass transistor increasing. Such measures also reduce the phase shift given by (12). However, these measures make the feedthrough error to increase. Also, the use of large switching devices results into heavy clock loads, producing an excessive clock skew and, consequently, serious aperture jitter – a random variation of the delay between the edge of the gate signal and the actual time instant in which the circuit enters in hold mode.

This speed-accuracy trade-off prompts for the calculation of optimum sizes for the sampling capacitor and the access switch. On the one side, the acquisition error is especially noticeable at higher frequencies because of the single-pole low-pass characteristic of the switch-

capacitor circuit. Forcing a desired acquisition time, this error can be expressed as a function of the capacitor size and the access transistor width, given that $R_{sw} = f(W)$. On the other side, while sampling low frequency signals, the tracking period is large enough to allow a proper settling of the S/H circuit dynamics. Therefore, the switching error is the major source of inaccuracy. Being $C_{gds} = W \cdot \mathrm{CGSO}$, where CGSO is the vendor provided SPICE parameter for this process, feedthrough error can also be expressed as a function of the capacitor and switch sizes. Now, combining (10) and (13), and assuming a comparable influence of these two effects being the optimal solution, the capacitor size for each value of the pass transistor width can be obtained by solving this equation:

$$\varepsilon_a(C_{mem},W) = \varepsilon_f(C_{mem},W) \tag{16}$$

what has been done applying a graphical method in Fig. 7.

There is also a random contribution to the sampling error. When the pass transistor is ON, it can be considered as a resistance that introduces a white noise with gaussian amplitude and distribution of a thermal origin. Its noise power density being $\overline{v_i^2} = 4kTR_{sw}\Delta f$. In addition, the single-pole low-pass filter formed by the access switch and the sampling capacitor limits the bandwidth of the noise to an equivalent noise bandwidth $(4R_{sw}C_{mem})^{-1}$. This results in a total noise power at the output, the storage node in this case,

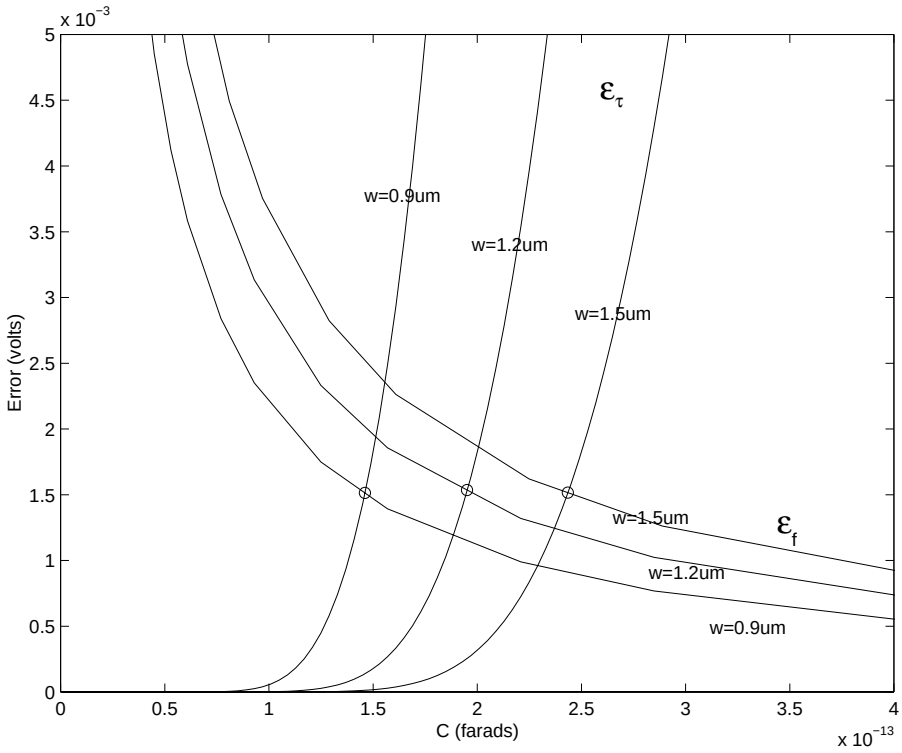

Fig. 7: Speed-accuracy trade-off.

$$\overline{v_c^2} = \int_0^{\infty} |A_v(j2\pi f)|^2 \cdot \frac{\overline{v_i^2}}{\Delta f} df = \frac{kT}{C_{mem}} \tag{17}$$

which is independent of the transistor size and can be reduced by using, once again, a larger sampling capacitor. Notice that $\sqrt{kT/C_{mem}}$ constitutes a limit to the dynamic range of the system. For a S/H circuit that operates over the whole rail-to-rail scale (3.3V power supply voltage) with a 0.1pF sampling capacitor, the maximum achievable dynamic range will be approximately 84dB. This is not a crucial issue in our case because of the relatively low system resolution requirements, but it can restrain the use of analog signal processing in some other applications.

B. S/H stage design

The ARAM chip includes 32 identical S/H lines whose schematic is depicted in Fig. 8. It is based on the S/H circuit reported in [29] and employs bottom-sampling of the analog signal to realize an offset-free and non-destructive recovery of the sampled data with reduced harmonic distortion. Assume first that the opamp has infinite DC gain and that clock feedthrough and the parasitic capacitor C_p are negligible. The difference between the input voltage and the opamp offset voltage V_{os} is stored at C_k during phase ϕ_1 yielding $V_{c_k} = V_i - V_{os}$. Then, during the next phase ϕ_2, the positive capacitor electrode is switched to the output node giving $V_o = V_{c_k} + V_{os}$ and, hence, $V_o = V_i$, with no trace of the opamp offset voltage.

Consider now that feedthrough is not negligible. Because this S/H circuit employs bottom-plate sampling, the harmonic distortion introduced in the sampling process due to feedthrough can be eliminated. Here, an extra switch is employed to isolate the bottom-plate of the capacitor

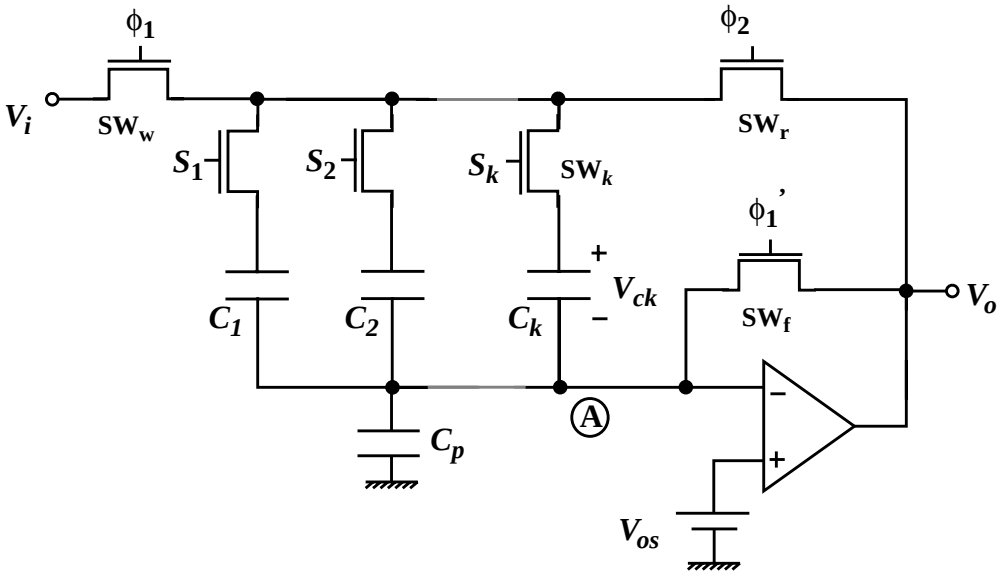

Fig. 8: S/H line schematic including parasitic capacitance and opamp offset.

at the end of the sampling phase. It is controlled by the signal ϕ_1^* that falls slightly before ϕ_1. In this way, the feedthrough error is introduced via the bottom-plate of C_{mem}, which is maintained at a constant voltage V_{REF} by the opamp. Now, ε_f is independent of the input and, therefore, its derivatives with respect to V_i are equal to zero. Consequently, no harmonic distortion due to clock feedthrough will be present at the output. The stored voltage is only affected by an additional voltage offset. A small pedestal error of magnitude

$$\varepsilon_f = -\frac{C_{gds}}{C_{mem}+C_{gds}} \cdot [V_{REF} + V_T(V_{REF}-V_{SS}) - V_{SS} \tag{18}$$

If the finite DC gain and the parasitic capacitor are accounted for, the output voltage is an attenuated copy of the input and an offset term appears,

$$V_o \approx \left[1+\frac{1}{A_0}\left(1+\frac{C_p}{C_k}\right)\right]^{-1} V_i + \frac{1}{A_0}\left(1+\frac{C_p}{C_k}\right)V_{os} \tag{19}$$

Fig. 10 shows the opamp schematics, which has been realized through a folded cascode architecture to better fit the 3.3V power supply voltage. For 7 bits equivalent resolution of the S/H crcuit, and assuming that a 16mV error is allowed for each sample, the opamp output swing has to be larger than 2V. Other opamp specifications are: GBW of 20MHz – required to follow the input during the tracking phase; and Slew-Rate (SR) of 8V/μs – required to sample 4MHz band limited signals with up to 2V amplitude (peak-to-peak).

Let g_{m1} be the small-signal transconductance of the transistors in the input differential-pair of the opamp, and I_B the tail-current. A relation between the transistors aspect ratio and I_B can be derived from the GBW specifications. Because $GBW = g_{m1}/(2\pi C_L)$ and assuming

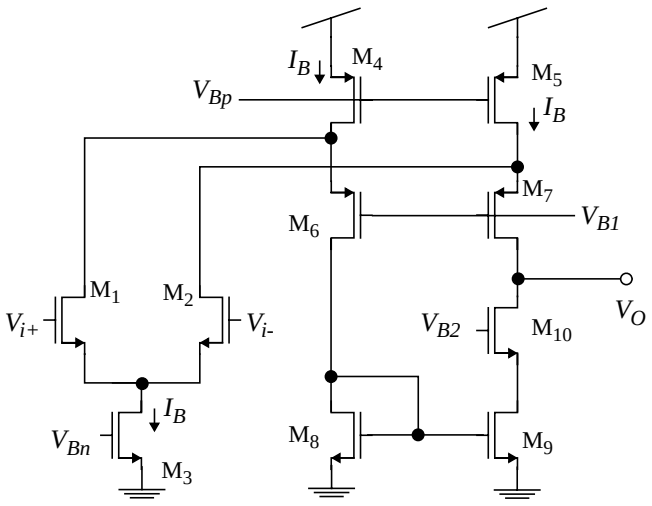

Fig. 9: Opamp schematic.

Table I: Transistor sizes

M_1–M_2	24/1.2
M_3	16/1.2
M_4–M_5	48/2.4
M_6–M_7	48/1.2
M_8–M_9	24/2.4
M_{10}	24/0.6

operation within saturation region in strong inversion, one obtains,

$$\frac{W}{L} = \frac{(2\pi C_L \mathrm{GBW})^2}{2k_n I_B} \tag{20}$$

where k_n is the intrinsic transconductance of the MOS transistor. On the other hand, the necessary tail-current is fixed by the slew-rate,

$$I_B = SR \cdot C_L \tag{21}$$

This current determines the appropriate aspect ratio of the input differential-pair for a constant GBW of 20MHz. The folded-cascode output stage is specified by the DC gain. By providing at least 60dB for the DC gain – $A_0 = g_{m1}R_o$ – the error introduced by the parasitic capacitance is reduced to 0.1%. As g_{m1} is now fixed, the output stage has to be designed so as to achieve the necessary output impedance. Final compromises are resolved by phase margin and matching considerations.

C. Leakage currents and storage time

During the hold period, several leakage currents attempt to discharge the storage capacitor, contributing to degrade the sampled voltage value. In the first place, the reverse-biased junction formed by the n-diffusion area, corresponding to the source terminal of the pass transistor and the substrate pumps out of the upper plate of the capacitor a current that can be approximated by the reverse-biased saturation current of the parasitic diode. Another leakage is due to the subthreshold drain-to-source current of the pass MOS transistor. These effects add up resulting in a total current in the range of the pA. In this occasion, capacitors are implemented by a poly-over-diffusion structure lying on top of a weakly-doped n-well (Fig. 10). Then, the n-well/p-substrate junction is reverse-biased and the current that flows out of the bottom plate of the capacitor correspond to the associated reverse-bias saturation current. Since it is in the fA range, it limits the effect of the upper plate leakage. Stored voltage degradation in time during the hold period is now given by

$$\frac{dV_c}{dt} = -\frac{1}{C_{mem}} \cdot \frac{dq^-}{dt} \approx -\frac{I_{self}}{C_a A} \tag{22}$$

where C_a is the capacitance per unit area of the poly-over-diffusion structure. In these conditions, a self-discharge rate, independent of the capacitor size, is defined:

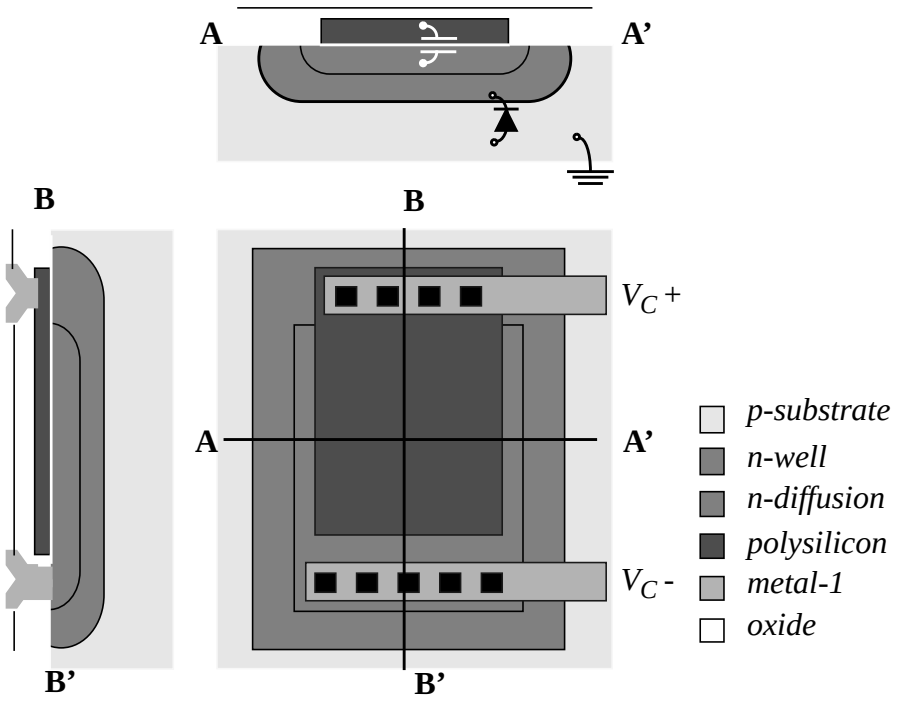

Fig. 10:Polysilicon over n-diffusion capacitor.

$$r_{self} = \frac{q}{C_a}\left(\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right), \tag{23}$$

where q is the charge of an electron, D_p and D_n are the diffusion coefficients for holes and electrons, L_p and L_n their diffusion lenghts and p_{n0} and n_{p0} are the minority-carrier concentrations in each side of the junction. In this technology r_{self} is 50mV/s. Then, the voltage at the capacitor decays linearly in time during the hold period. A maximum storage time can be defined in terms of the accuracy requirements. For an equivalent resolution of N bits and a full scale range of the input signal given by A, the maximum storage time (t_{sto}) is the period in which the difference between V_c and the initially stored voltage does not exceed $A/2^{N+1}$, that is 1/2 LSB. That is

$$t_{sto} = \frac{A}{r_{self} \cdot 2^{N+1}} \tag{24}$$

which is in the 200ms range for a 10mV error. These figures, however, must be understood only as orientative because of the strong sensitivity of the leakage currents to the operating temperature. Also, incidence of light on the circuit surface can seriously degrade the contents of the memory because of the light induced generation of an extra amount of carriers.

D. ARAM chip floorplan

This CMOS ARAM chip is composed of an array of 32×256 analog memory cells. Each one contains a capacitor, a pass transistor and some local logic for address decoding. The system includes as well some digital control circuitry and an I/O interface consisting in an analog MUX/DEMUX and 16 output buffers. Fig. 11 shows a picture of the ARAM chip floorplan. The memory matrix is arranged into 32 S/H lines with 256 capacitors each. Random access to any memory location is available with the help of two binary-to-one-hot address decoders. A code of 5 bits activates one out of the 32 row selection lines, by means of the row address decoder. Similarly, each one of the 256 columns is selected by an 8-bit code. Different access schedules can be implemented by an adequate programming of the address codes. In order to avoid the selection of more than one capacitor per row at a time, what would seriously degrade the operation, a global clock controls the duty cycle of the access signals leaving a tunable guard time interval for address codes to change. Now, with respect to the I/O interface, the 32 data lines of the array are multiplexed either to the 16-line wide I/O bus or the serial I/O channel. A digital control signal sets the serial or parallel I/O mode. Row selection signals are employed to scan the 32 data lines with either the I/O serial channel or the 16-line wide I/O bus. Some test pads have been added to characterize the output buffers for a better analysis of the test results.

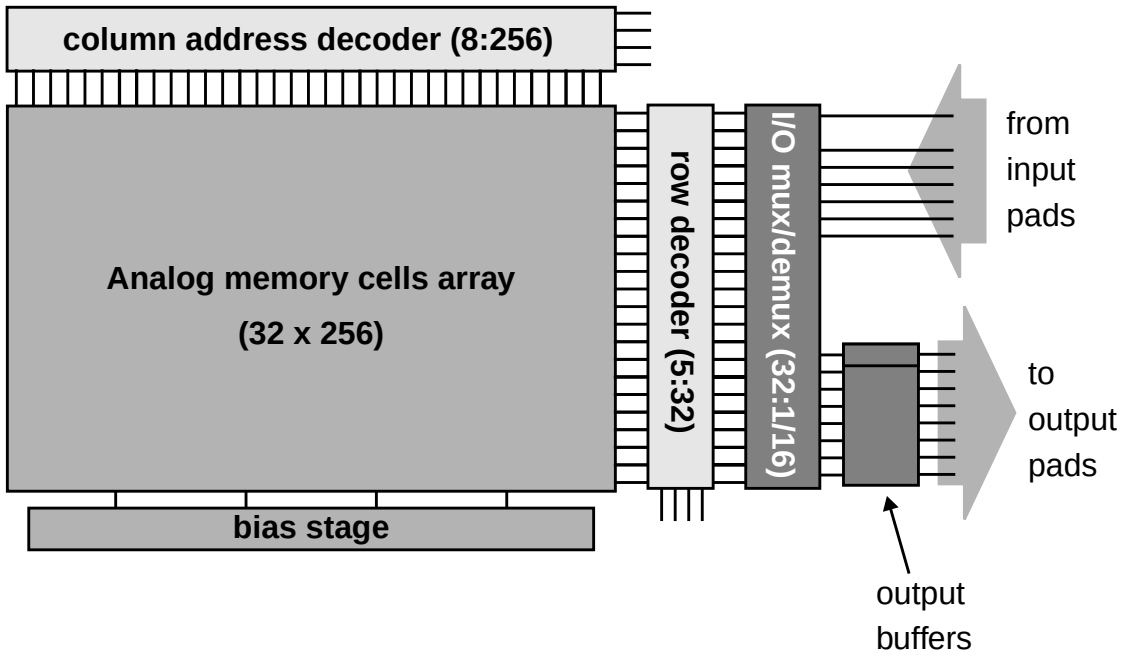

Fig. 11:System architecture of the ARAM chip.

Guidelines concerning signal interaction prevention in mixed-signal IC's have been followed in the development of the prototype. It is a well-known fact that the integration of a significant amount of digital circuitry along with analog signal processing in the same substrate can potentially degrade system performance. A conservative layout style, with an extensive use of grounded guard rings, reduces signal coupling by opening alternative return paths to the currents induced into the substrate [30]. This is reinforced by the implementation of separated power supply and ground connections for the analog and digital circuitry and guard rings [31]. Digital lines switching at higher rates have been routed over insensitive areas and critical crossings have been shielded with a grounded metal intermediate layer. Also, analog bus lines are made wider and are separated to a larger distance than recommended by technology rules, in order to reduce cross-talk at higher frequencies.

IV. EXPERIMENTAL RESULTS

The first prototype of this ARAM chip has been integrated in the Hewlett-Packard 0.5μm CMOS process offered by the MOSIS service. The 24 available samples of the chip has been tested and proved to be functional. No major discrepancies have been found during the test of the different samples. First of all, a functional characterization test has been developed. Several input sine waves of different frequencies have been sampled at different rates. Fig. 12 shows a plot of the measured root-mean-square error during the reconstruction of the input waveform. It has been computed by taking the square root of the average of the squared difference between the input signal and the recovered waveform over the N samples of the input wave:

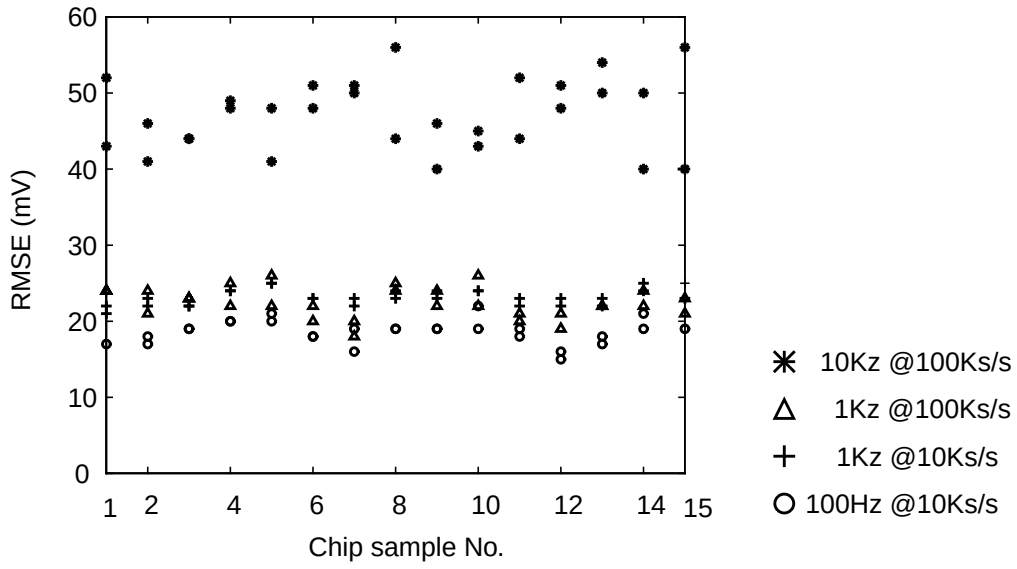

Fig. 12:Measured RMS error in the reconstructed waveform

$$\text{RMSE} = \sqrt{\frac{1}{N} \cdot \sum_{k=1}^{N} (V_{i_k} - V_{o_k})^2} \tag{25}$$

It is important to mention that no correction of the output buffer offset or the feedthrough induced pedestal error has been made. Fig. 13 displays a reconstructed triangular wave sampled at 10KHz and a recovered sine wave sampled at 100KHz. The computed absolute RMSE is in the 13-25mV range, which means a relative error of 0.7-1.4% for a 1.8V output swing.

A revealing picture of the test results is obtained by computing the FFT of the output signal. In this case, a 10KHz sine wave has been sampled at 250Ksamples/s. It has been fed to the ARAM chip through the serial input channel, therefore, 8192 samples of the input waveform

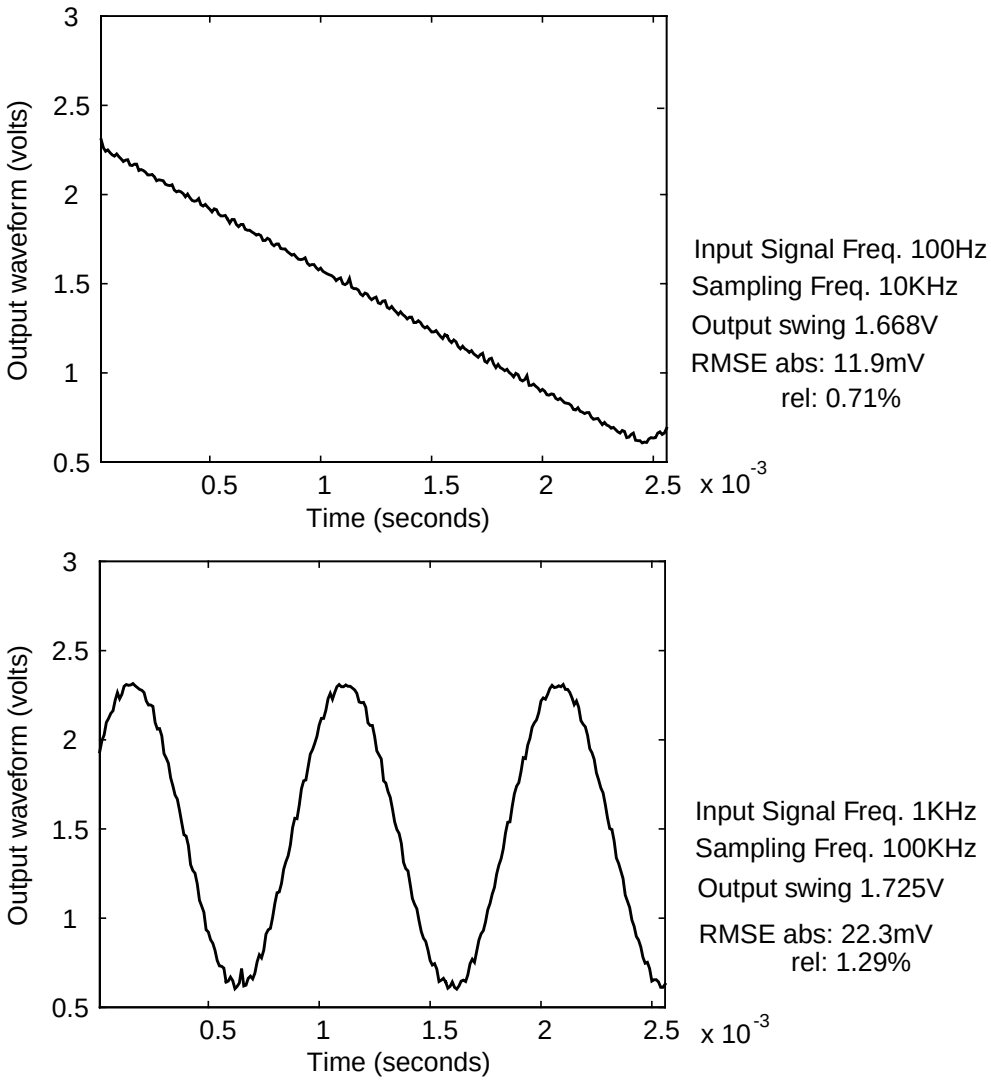

Fig. 13:Recovered triangular and sine waveforms

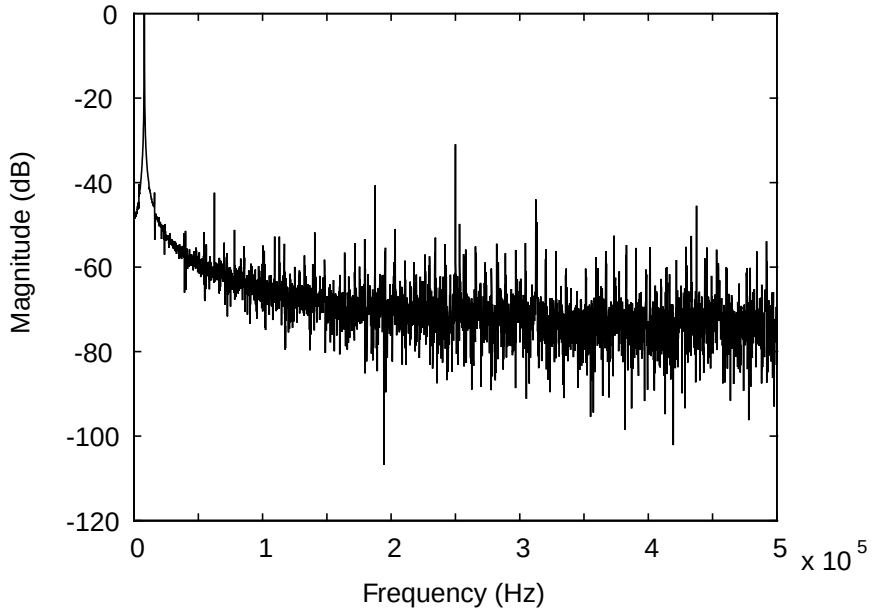

Fig. 14:Spectrum of the output sinewave at 10KHz (no filtering of the readings)

have been taken. Fig. 14 shows the spectrum of the output signal, directly measured from the output of the chip without eliminating irrelevant information or filtering of the digitizer readings. It means that not only the stored voltage samples but also the voltage peaks occurring during address changes are captured. The magnitude of the single-tone at 10KHz is nearly 80dB above the background level. The following peak in magnitude, that takes place at the sampling rate, is approximately 30dB below the sine wave tone. Fig. 15 displays the input and the output signals as 32×256 -pixel images using a linear 256-levels grayscale (8 bits deep). Each pixel in the image represents the voltage at a memory capacitor in the array. The absolute value of the difference between the input and output images is represented in the same grayscale.

Besides, some real images have been loaded to the chip at 200ns per pixel and downloaded at 800ns. Fig. 16 displays the input and output pictures together with a grayscale representation of the absolute difference between them. The first two examples are 512×512 -pixel pictures in a 256-level grayscale. The last one is a 256×256 color picture. They have been processed in 32×128 -pixel pieces because of test equipment requirements. Some spatial noise can be detected in the output picture. It is partly due to image partitioning and, on the other side, due to an improper tracking of the input at the beginning of each pixel group -- vertical lines at the 1^{st}, 129^{th}, 257^{th} and 385^{th} pixels. Because of the clocking scheme adopted to avoid the selection of more than one memory register at a time, the feedback loop of the opamp in the S/H stage is left open for a certain period. Consequently, the voltage of the output node goes up to the power supply voltage or down to the negative rail. In these conditions, the slew-rate of the opamp is insufficient to catch up with the input in the required acquisition time.

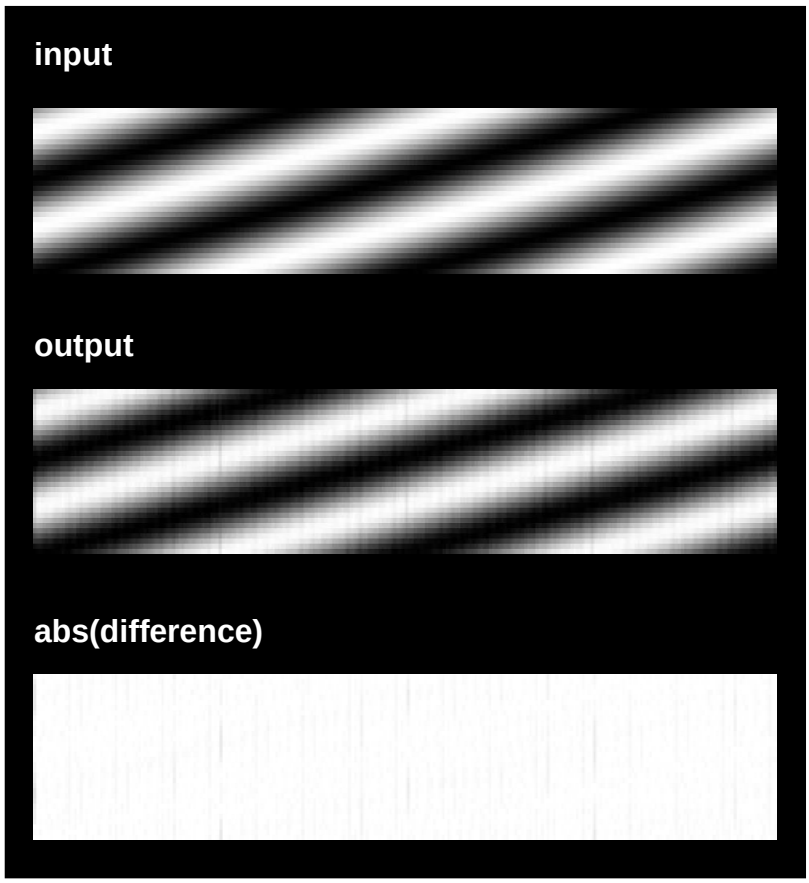

Fig. 15:Input and output images (256 gray levels)

Finally, storage time has been measured for randomly selected cells of the array. Fig. 17 shows the difference between the initially stored voltage and the instant value through time. These data represent 24 cells in the 24 different samples of the chip. Stored voltage degradation exceeds the required accuracy levels after 80-100ms. Recursive reading of the same memory spot does not have a noticeable influence on the stored voltage.

Finally, Fig. 18 shows a photograph of the prototype circuit and Table II provides a survey of data extracted from the tests results.

V. CONCLUSIONS

The only missing part of the CNN chipset architecture has been implemented. A random access analog memory chip has been designed and integrated in a standard 0.5 μm CMOS single-poly triple-metal technology. Measured equivalent resolution is around 7 bits. Storage time is larger than 80ms. DC power dissipation remains 73mW for a 3.3V power supply. Access times of 200ns have been obtained, while reading time is 800ns. Higher sampling and output rates can be achieved using the 16-line wide analog I/O bus. In future generations of the CNN

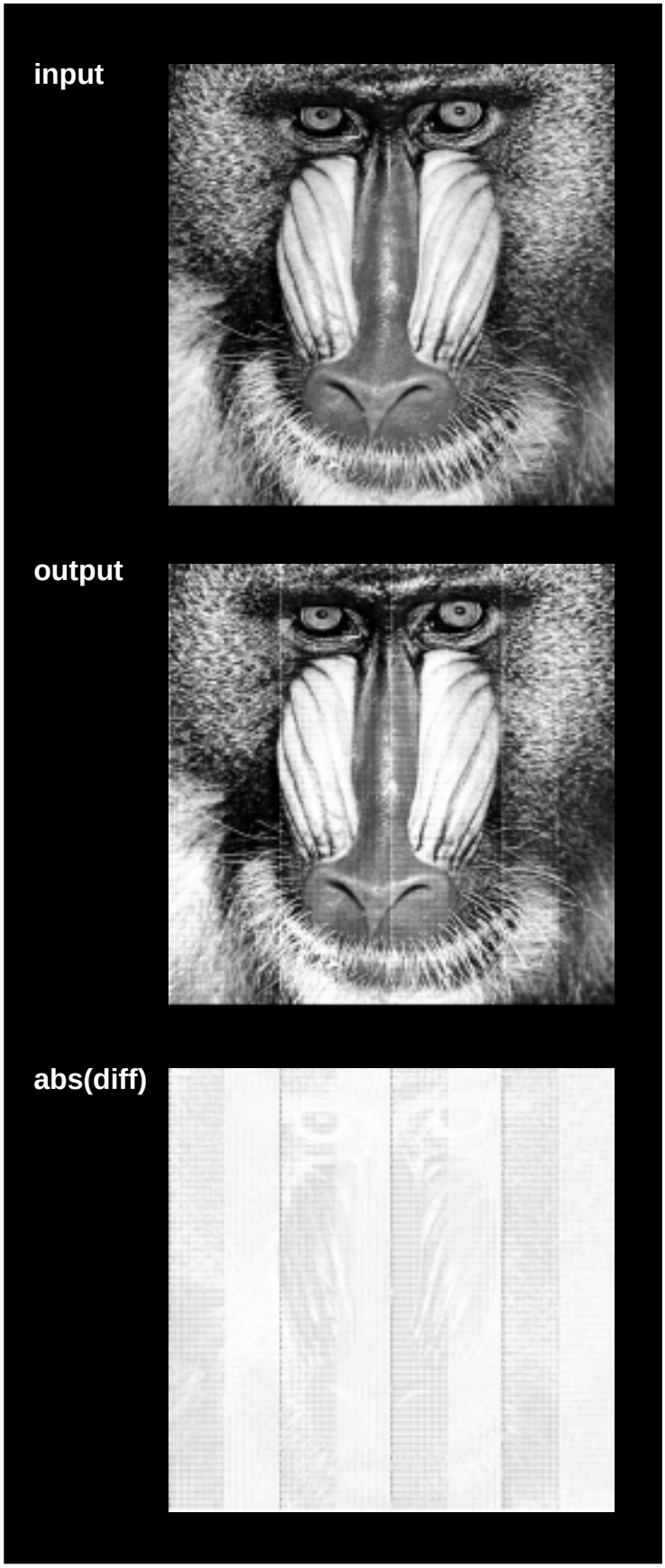

Fig. 16:Test input and output images

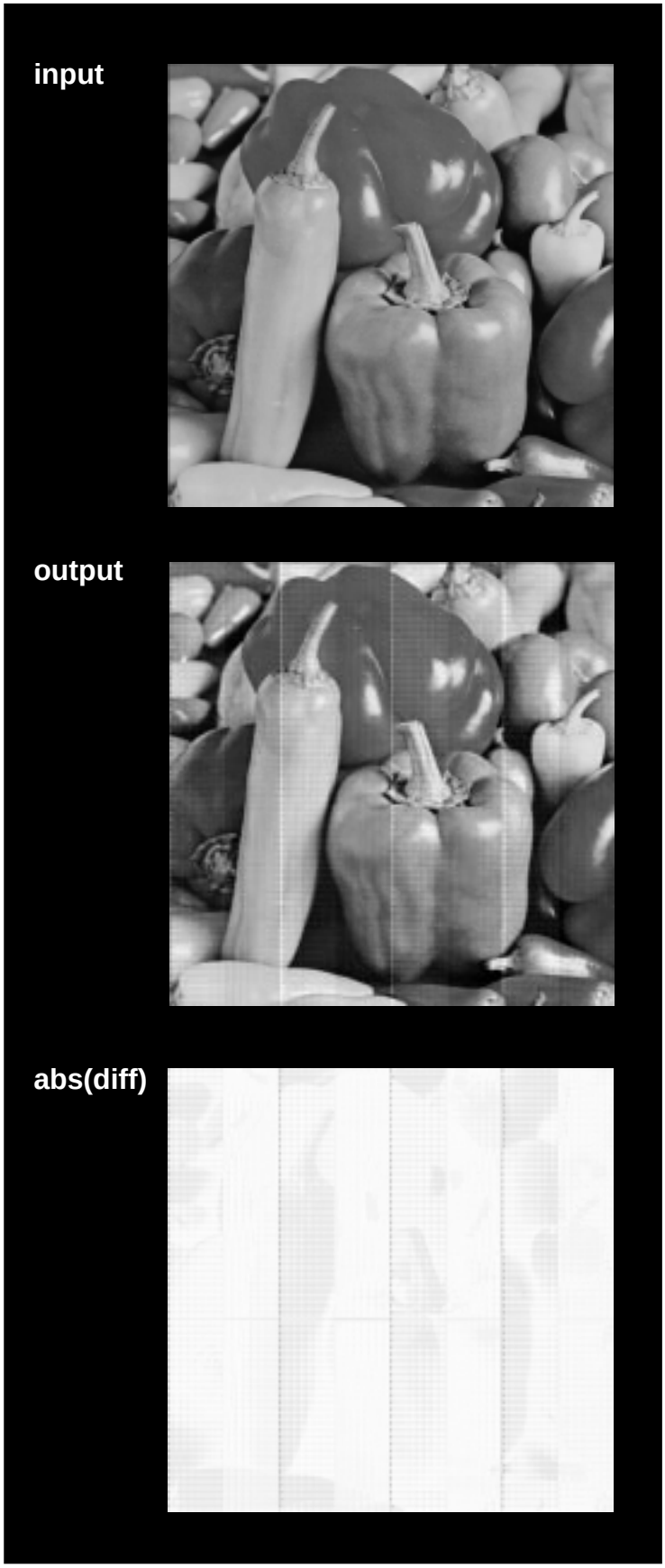

Figure 16: (Continued)

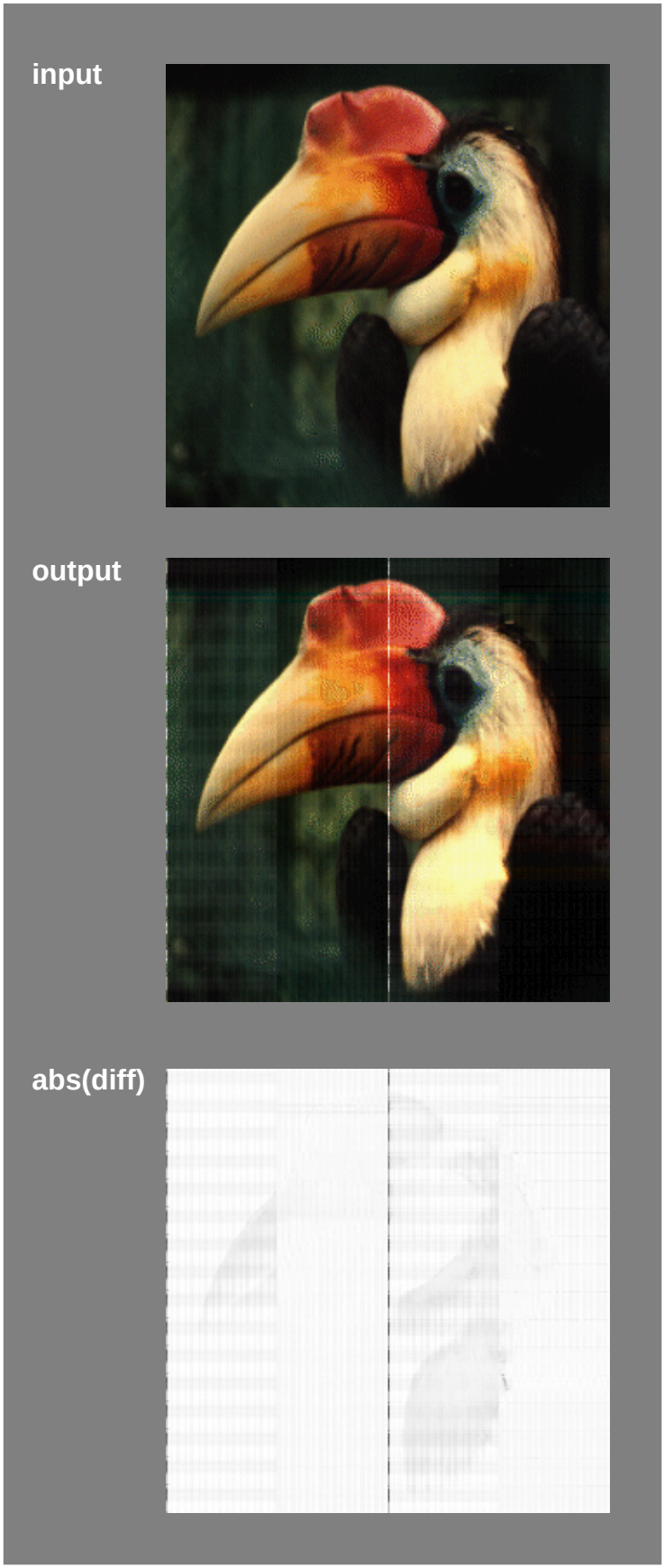

Figure 16: (Continued)

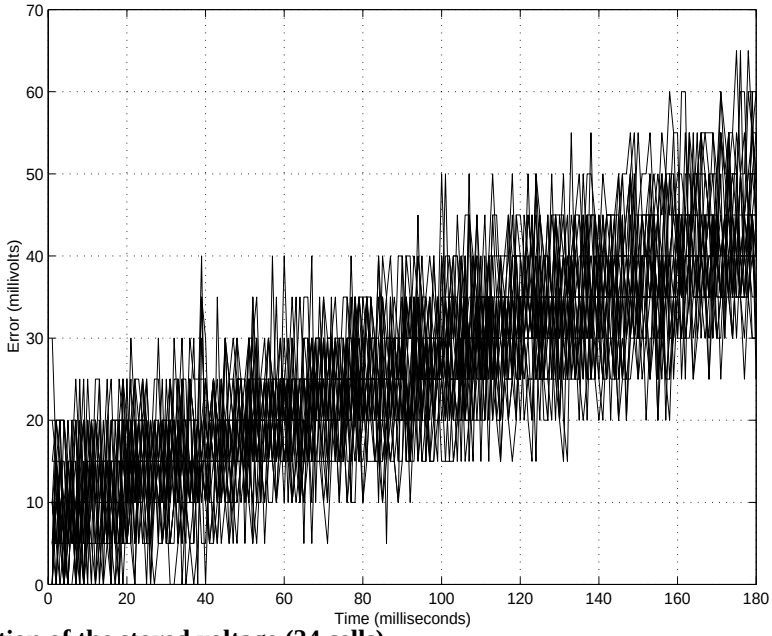

Fig. 17:Degradation of the stored voltage (24 cells)

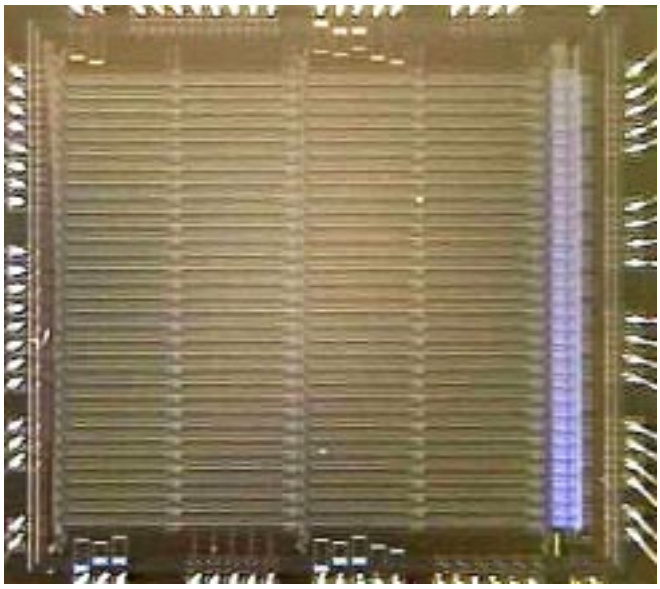

Fig. 18:Die photograph of the ARAM prototype.

Universal Chip, an embedded and distributed version of this analog RAM will be implemented. The reported prototype is now being employed in different experiments related to video signal processing in multimedia applications.

ACKNOWLEDGEMENTS

This work is supported by the JSEP Grant No. FDF49620-97-1-0220-03/98 and by the ONR Grant No. N00014-98-1-0052.

Research of the authors from IMSE-CNM (CSIC) has been supported by the spanish CICYT (Project TIC96-1392-C0202 SIVA) and the EU (Project ESPRIT IV 27077-DICTAM).

The help of Péter Földesy on the experiments with test images is kindly acquainted.

Table II: ARAM Prototype Data

Number of pixels	8192 (32 x 256)
Cell-array area	3.73mm x 3.45mm
Cell-density	637 cells/mm^2
System area (w/o pads)	4.13mm x 3.89mm
Die Area	4.77mm x 4.47mm
Package (used pins)	PGA-84M (81)
Power Supply	3.3V
Power dissipation	72.86mW @ 3.3V
Sampling time	200ns
Reading time	800ns
I/O rates (via 16-lines bus)	10 Msamples /s
Input range	[0.6, 2.4] V
Output swing	[0.6, 2.4] V
Storage time (1.5% error)	80-100ms
Measured resolution	6-7bits (0.7-1.5% error)

REFERENCES

[1] L.O. Chua and T. Roska, "The CNN Paradigm", *IEEE Trans. Circuits and Systems I: Fundamental Theory and Applications*, vol. 40, pp. 147-156, March 1993.

[2] L.O. Chua and L. Yang:, "Cellular Neural Networks: Applications", *IEEE Trans. Circuits and Systems*, Vol. CAS-35, pp 1273-1290, 1988.

[3] Kenneth R. Crounse, *"Image Processing Techniques for Cellular Neural Network Hardware"*. Ph. D. Thesis, University of California, Berkeley, June 1997.

[4] S. Espejo, A. Rodríguez-Vázquez and R. Domínguez-Castro, "Smart-Pixel Cellular Neural Networks in Analog Current-Mode CMOS Technology", *IEEE Journal of Solid-State Circuits*, Vol. 29, pp. 895-905, IEEE August 1994.

[5] R. Domínguez-Castro, S. Espejo, A. Rodríguez-Vázquez, R. Carmona, P. Foldesy, A. Zarándy, P. Szolgay, T. Sziranyi and T. Roska, "A 0.8 μm CMOS Programmable Mixed-Signal Focal-Plane Array Processor with On-Chip Binary Imaging and Instructions Storage", *IEEE J. of Solid State Circuits*, vol. 32, No. 7, pp. 1013-1026, July 1997.

[6] J. M. Cruz and L. O. Chua, "A 16 x 16 Cellular Neural Network Universal Chip: the First Complete Single-Chip Dynamic computer Array with distributed Memory and Grayscale I/O", *Analog Integrated Circuits and Signal Processing*, Vol. 15, No. 3, pp 227-238, March 1998.

[7] A. Paasio, A. Dawidziuk, K. Halonen and V. Porra, "Minimum Size 0.5μm CMOS Programmable 48 x 48 CNN Test Chip", *Proc. of the 1997 European Conference on Circuit Theory and Design*, pp. 154-156, Budapest, Hungary, Septmeber 1997.

[8] S. Espejo, R. Domínguez-Castro, G. Liñán and A. Rodríguez-Vázquez, "64 x 64 CNN Universal Chip with Analog and Digital I/O", *IEEE 1998 Int. Conf. on Electronic Circuits and Systems*, Lisbon, September 1998.

[9] T. Roska and L. O. Chua: "The CNN Universal Machine: An Analogic Array Computer", *IEEE Transactions on circuits and Systems-II: Analog and Digital Signal Processing*, Vol. 40, No. 3, pp. 163-173, March 1993.

[10] Timothy G. Mattson and Greg Henry, "An Overview of the Intel TFLOPS Supercomputer". *Intel Technology Journal*, Issue No. Q1-98, January 1998.

[11] C. Koch and H. Li (Eds.), *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, IEEE Computer Society Press, Los Alamitos, CA, 1995.

[12] L. O. Chua, T. Roska, T. Kozek and A. Zarandy, "CNN Universal Chips crank up the Computing Power". *IEEE Circuits and Devices Magazine*, vol. 12, No. 4, pp. 18-28, July 1996.

[13] Charles A. Poynton, *"A Technical Introduction to Digital Video"*. John Wiley & Sons, New York, 1996.

[14] T. Roska, "Implementation of CNN computing technology". *Proceedings of the International Conference on Artificial Neural Networks*, pp. 1151-1155. Lausanne, Switzerland, October 1997.

[15] M. Salerno, F. Sargeni and V. Bonaiuto, "DPCNN: a modular chip for large CNN arrays". Proc. of the IEEE Symposium on Circuits and Systems, Vol. 1, pp. 417-420, Seattle, WA, USA, April-May 1995.

[16] S. Espejo, R. Dominguez-Castro, R. Carmona and A. Rodriguez-Vazquez, "Hybrid-control of synapse circuits for programmable cellular neural networks". *Proc. of IEEE International Symposium on Circuits and Systems*, Vol. 3, pp. 507-510, Atlanta, GA, USA, 12-15 May 1996.

[17] M. Salerno, F. Sargeni and V. Bonaiuto, "A 720 cells interconnection-oriented system for cellular neural networks". *Proceedings of the IEEE International Symposium on Circuits and Systems*. Vol. 1, pp. 681-684, Hong Kong, June 1997.

[18] J. Pineda de Gyvez, Lei Wang and E. Sanchez-Sinencio, "Large-image CNN hardware processing using a time multiplexing scheme". *Proc. of the Fourth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 405-410, Sevilla, Spain, June 1996.

[19] T. Roska: "CNN Chip Set Architectures and the Visual Mouse". *Proceedings of the 4th IEEE Int. Workshop on Cellular Neural Networks and their Applications*, pp 487-492. Sevilla, Spain, June 1996.

[20] L. Howard Pollard, *Computer Design and Architecture*, Prentice-Hall Inc., Englewood Cliffs, N. J., 1990.

[21] K. A. Nishimura and P. R. Gray, "A Monolithic Analog Video Comb Filter in 1.2μm CMOS". *IEEE Journal of Solid-State Circuits*, Vol. 28, No. 12, pp. 1331-1339, December 1993.

[22] G. M. Haller and B. A. Wooley, "A 700-MHz Switched Capacitor Analog Waveform Sampling Circuit". *IEEE Journal of Solid-State Circuits*, Vol. 29, No. 4, April 1989.

[23] E. Franchi, M. Tartagni, R. Guerrieri, G. Baccarani, "Random Access Analog Memory for Early Vision". *IEEE Journal of Solid-State Circuits*, vol. 27, No. 7, pp. 1105-1109, July 1992.

[24] K. R. Crounse and L. O. Chua, "The CNN Universal Machine is as universal as a Turing Machine". *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Vol. 43, No. 4, pp. 353-355, April 1996.

[25] T. Roska and L. O. Chua and A. Zarandy, "Language, compiler and operating system for the CNN supercomputer". Report UCB/ERL M93/34, University of California, Berkeley, 1993.

[26] F. Werblin, T. Roska and L. O. Chua, "The analogic cellular neural network as a bionic eye". *International Journal of Circuit Theory and Applications*, Vol. 23, No. 6, pp. 541-69, November-December 1995.

[27] T. Roska and L. Kek (Eds.), *"Analogic CNN program library (Version 6.1)"*. Report DNS-6-1995, Analogical and Neural Computing Laboratory, Computer and Automation Institute of the Hungarian Academy of Sciences, Budapest, 1995.

[28] K. A. Nishimura, *"Optimum Partitioning of Analog and Digital Circuitry in Mixed-Signal Circuits for Signal Processing"*. Ph. D. Dissertation, College of Engineering, University of California, Berkeley, July 1993.

[29] Roubik Gregorian and Gabor C. Temes, *"Analog MOS integrated circuits for signal processing"*. John Wiley & Sons, New York, 1994.

[30] T. J. Schmerbeck, "Minimizing mixed-signal coupling and interaction". *Proc. of the 20th European Solid-State Circuits Conference*, pp. 28-37. Ulm, Germany, September 1994.

[31] B. R. Stanisic, R. A. Rutenbar and L. R. Carley, "Addressing Noise Decoupling in Mixed-Signal IC's: Power Distribution Design and Cell Customization". *IEEE Journal of Solid-State Circuits*, vol. 30, No. 3, pp. 321-326, March 1995.