VLSI Architectures for Programmable Sorting of Analog Quantities with Multiple-Chip Support

Fabio Ancona, Giorgio Oddone, Stefano Rovetta, Gianni Uneddu, and Rodolfo Zunino

DIBE - Department of Biophysical and Electronic Engineering - University of Genoa
Via all'Opera Pia 11a - 16145 Genova - Italy - e-mail: zunino @ dibe.unige.it

Abstract

The paper describes VLSI architectures for sorting analog quantities. The elementary circuit unit yields analog representations of sorted values and digitally encodes the corresponding ranks in the list. The length of the sorted list can be digitally programmed at run time, hence partial sortings are also supported. The modular, mixed analog/digital structure is arranged into elementary cells operating at the local level. This greatly facilitates the layout design and enables multi-chip integration. A suitable coupling of current-mode and voltage-mode signals minimizes the number of transistors.

1. Overview

Sorting is a basic operation in many information and signal processing paradigms. For example, pattern recognition [1], filtering based on order statistics [2], neural networks [3], and evolutionary computation models [4] typically require sorting at crucial steps in the training processes. Recent research on VLSI implementations of such approaches strongly enforces the interest in hardware support of sorting functions. On the other hand, the global information necessarily involved in sorting typically makes this operation difficult to implement in VLSI efficiently; in digital approaches, this is especially due to the relatively large number of interconnecting wires required to propagate information [9,14].

This paper presents a design method for a programmable VLSI realization of sorting functions. The elementary VLSI unit processes a set of N analog input currents, and outputs k sorted voltages proportional to the largest k input values; the circuit can be programmed by digitally setting the desired length, k, of the sorted list. The system completes a sorting cycle in $O(k)$ time. The simple and modular architecture makes the integration of several chips easy and effective.

Most sorting circuits available in the literature process digital quantities [5]. On the one hand, digital circuits are simpler to design and can be used in a variety of situations. On the other hand, however, the analog modules that we have exploited are very fast, whereas the size and time complexity of the whole system compare favourably with those of other approaches, with the added benefit of a fully modular structure. Finally, one of the main drawbacks of analog signal processors, namely, the lack of flexibility, has been overcome by introducing a programmable structure.

2. Principle of operation

The principle of operation is straightforward: sorting results from iteratively 1) detecting and consequently 2) inhibiting the highest value in the list. Thus sorting proceeds sequentially; the inhibiting circuitry is digital, hence a simple counter is needed to control the process.
Maximum detection - Detecting the highest value represents a crucial step in the overall system. To single out the largest input current, the well-known Winner-Take-All (WTA) circuit described in [6] is used. The basic WTA circuit is arranged into elementary cells operating at the local level and interacting through a single wire (Fig.1). As a result of a competitive process, the whole
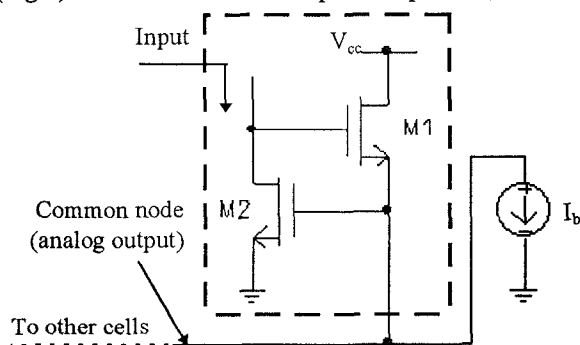

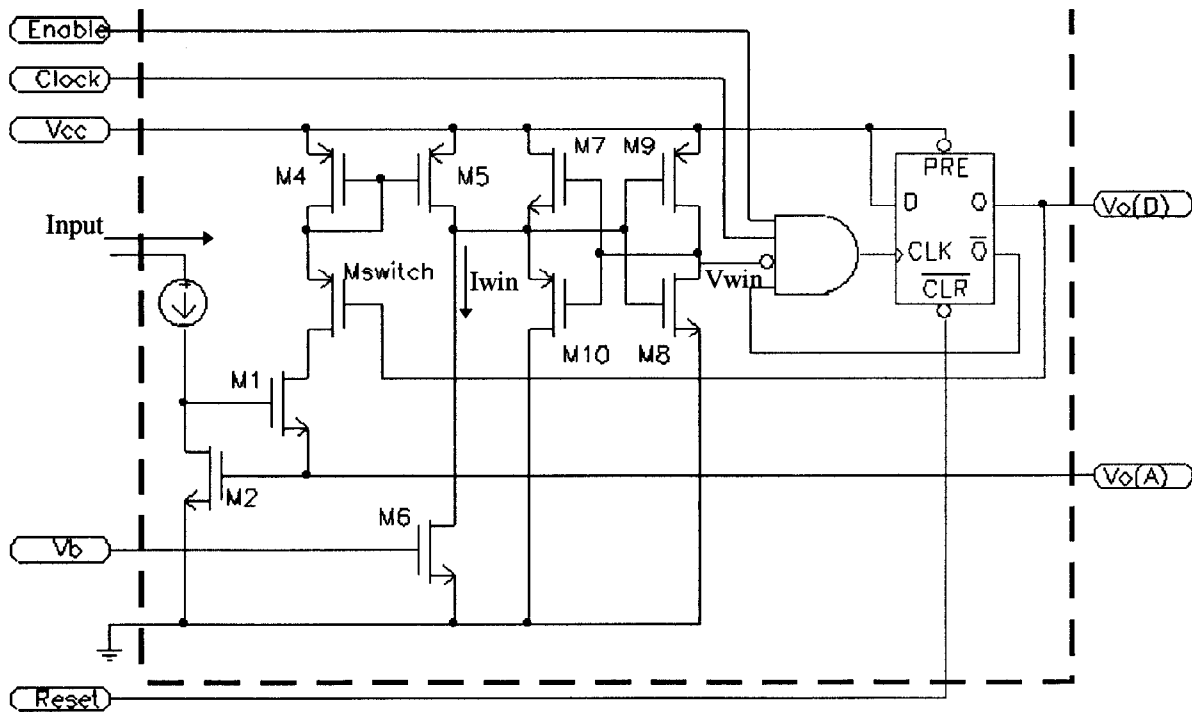Fig.1 - WTA cell [5] for max detection

Fig.2 - Elementary cell of the sorting circuit

sink current I_b flows through the winning cell, whereas all other cells remain switched off.

Linear output - An enhancement of Lazzaro's original circuit makes it possible to linearize the output characteristic and to attain an analog, voltage-mode representation of the winning value from the potential at the common node [7].

Sequential sorting - The mechanism for progressively inhibiting the largest values exploits the peculiar current distribution of the WTA circuit. Each cell can locally detect whether it is currently the winner by checking the drain current in transistor M1. Such a current is then converted into a digital signal and used to drive a switch that disconnects the entire cell from the rest of the circuit. This virtually removes the largest value from the list; in the next iteration, the cell driven by the second largest input current will result as the winner, and so on. A synchronization clock supplies delays for current redistribution.

In summary, potential readings at the output analog node will show up in a decreasing ordered sequence. The process can be stopped by simply counting the number of extracted values; the total timing for this sequential sorting will be proportional to the number of elements extracted.

3. The Sorting Chip (SCHIP)

3.1. Cell-level functioning

This section describes the functioning of the sorting circuitry in detail. The overall schema is arranged into cells operating at the local level. Figure 2 presents the structure of the elementary cell that processes each sorted quantity. Input values are represented in current-mode. Potential Vb forces M6 to pull a drain current equal to the WTA sink current, Ib. For the analysis of the elementary single-chip circuit, the Enable signal will be assumed to be constantly low, hence its effect can be disregarded and wil be reconsidered when desribing the Multi-Chip functioning. The D-flip-flop is triggered by the falling edge of the clock; at initialization, the Reset signal forces Vo(D) to low, hence M_{switch} is on.

The transistor couple (M1,M2) reflects Lazzaro's classical schema; node Vo(A) is the cell-connecting node broadcasting the winning potential. As a result, disregarding additional circuitry, the set of cells keeps supporting the WTA behaviour. Thanks to a suitable polarization technique of the input currents [7], potential Vo(A) represents the largest input value linearly. The transistor set M4, M5, M6 operates as a current com-
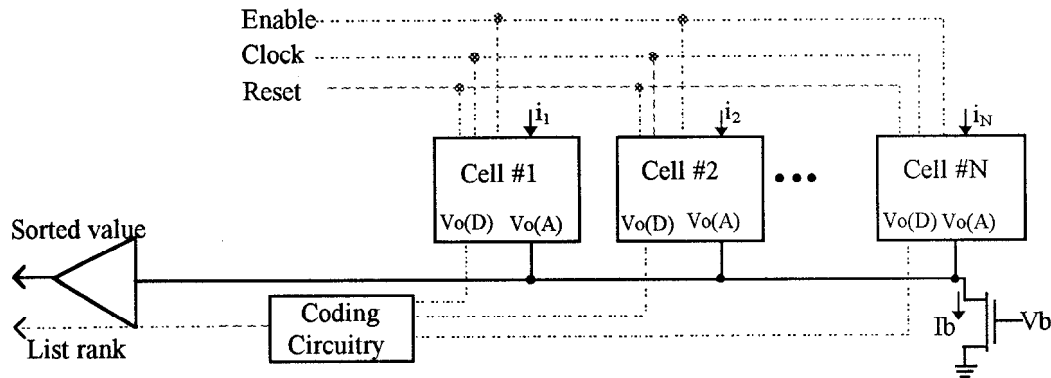
95

Fig.3 - The chip-level sorting architecture (SCHIP) - Dashed lines indicate digital signals

parator, matching the local status of the WTA cell against the expected status of a winning cell. As soon as the cell wins the WTA competition, the drain current in M5 will be equal to Ib, and I_{win} will drop to zero.

The circuit section M7, M8, M9, M10 converts its input current I_{win} to a digital voltage [8], V_{win}, which is low when I_{win} is not null. As a consequence, a high-to-low transition of V_{win} occurs in the winning cell and, *if the Enable signal is high*, forces a change in the associated flip-flop status.

The flip-flop status is exported from the cell as a digital output Vo(D), and informs the external circuitry about which cell has won the competition. The position of the winning input in the list can be easily detected by a counter. Voltage Vo(A) outputs an analog representation of the actual winning input.

The flip-flop output also controls the internal status of the cell. Further changes in the flip-flop status are inhibited by feeding back its inverted output, $\overline{Q}$, (digital feedback). Instead, the direct output line, Q, switches off the pass transistor M_{switch}. This disconnects the WTA cell from the others, hence the associated input will not affect the competition any longer (analog feedback).

The global result of this operation is that, at each iteration, the largest value is first detected and then automatically removed from the list. Sorting will eventually stem from collecting the series of digital outputs and their corresponding analog values. The Enable signal supports multiple-chip integration: it indicates to a chip that the locally winning value is also the highest one among all inputs, and enables the selected chip to remove the winning input; by contrast, the chips which do not receive an Enable signal will be able to represent their winning values in subsequent iterations

The length of the sorted list can be controlled by feeding digital signals to a counter that eventually dis-

ables the entire system and resets all cells. The schema of the overall chip-level sorting architecture (SCHIP) is given in Fig.3, where cell interconnections are shown.

3.2. Chip-level implementation and performance issues

Architecture – The most significant feature of the cell-oriented architecture is its modularity. Other implementation approaches to the sorting problem require a huge amount of unit interconnections [9,14]. In the described architecture, one wire connects all cells, which operate at the local level. This feature is counterbalanced by the system's sequential functioning. The area complexity is simply $O(N)$, as opposed to other analog solutions (such as the Maxnet [9], requiring $O(N^2)$ connections) or digital approaches [5], requiring $O(N \log N)$ gates.

In theory, the modular approach makes the number of processed inputs virtually unbounded; in fact, this number is limited by the discrimination accuracy of the WTA structure, by the chip area, and by parasitic capacitances in long connections.

VLSI implementation – The mixed analog/digital approach enhances the internal structure as follows. Section M1, M2 implements the actual analog WTA processing. This is the system's most sensitive subpart; the accuracy in matching and dimensioning equals that needed for a correct functioning of the elementary WTA subcircuit. Disregarding the tiny pass transistor M_{switch}, section M4, M5, M6 involves a digital, current-mode representation; the wide range of operating conditions for I_{win} greatly enhances implementation robustness. This also holds for the current-to-voltage mapping section M7, M8, M9, M10, whose nonlinear feedback makes up for mismatchings. In practice, with a careful design the NOT gate on this line and the digital feed-back can be

96

removed; this approach, however, might also affect the overall implementation robustness by tightening precision constraints. From a general perspective, mixing current-mode and voltage-mode representations ultimately increases the implementation simplicity.

Timing performances – The settling time of the competitive section is very short, but it is hard to determine analytically since the WTA structure is based on asynchronous positive and negative feedbacks. Experimentally, the settling time depends on many factors: number of inputs, layout (parasitic capacitances), and heavily on the quantity read as output. If we need to read the sequence of ordered levels, Vo(A), it is necessary to wait for the oscillations in the output voltage to be sufficiently small. If we read the ordered input positions, Vo(D), the only quantities involved in the competitive process are currents, which do not oscillate at all.

As a reference, we take a realization with $N=64$. With input signals very close to each other (worst case), a typical settling time for Vo(A) is about 300 ns, corresponding to a clock speed of about 3Mhz (3 000 000/k inputs sorted per second). This results from the netlist extracted from the layout simulator, since the plain schematic-based simulation gives much better results (about 10 times faster). Conversely, Vo(D) requires for settling about 10ns + the propagation delay of the flip-flop; this results in a speed of 100 000 000/k inputs sorted per second.

The time complexity is therefore O(k), a function not of the total number of inputs, but only of the programmed number of sorted elements. This is in contrast with other analog approaches, in which the complex dynamics does not ensure a O(1) time for the maximum selection [10]. There are examples of k-WTA circuits with complexity O(N) in area and O(1) in time [11], but they cannot provide the complete sort function.

The voltage-mode analog output is buffered off-chip by a suitably dimensioned op-amp; for multi-chip support, as will be explained in the following, the control lines of such amplifier are ported off-chip, as well, mainly for gain and offset adjustment.

3.3. Multiple-chip integration (MCHIP)

The sorting performance of a single chip is limited by the number of elementary cells that can be included in a die. As a matter of fact, a single wire traverses all cells for maximum-value detection: this makes a multiple-chip structure feasible, whose operation limits are bounded by the current resolution accuracy of a WTA
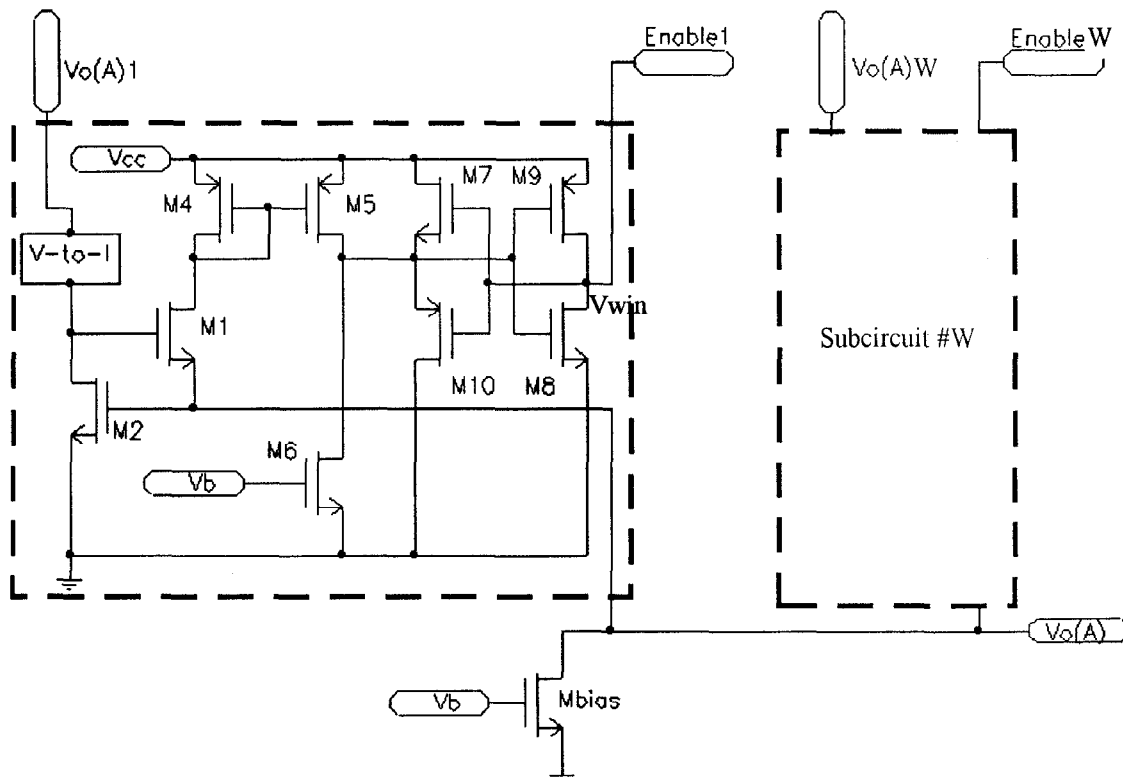

Fig. 4 - The higher-level circuitry (MCHIP) for comparison of multiple chip outputs
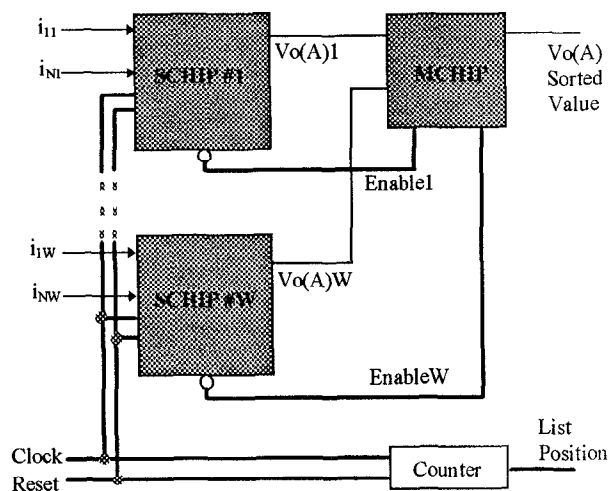
Fig.5 - Simplest interconnection schema for multiple-chip architecture

circuitry. Multiple-chip support simply applies the basic WTA schema (as per Fig.1 and Fig.2) to the output lines of the involved chips; this implements a hierarchical structure, in which the "winning" cell indicates the device showing the largest output value. The digital voltage, V_{win}, generated within each cell can be used as the Enable signal for the associated chip. A cell in such chip-integration module differs from that shown in Fig.2 in that it does not include any digital subcircuitry; the analog feedback line and the M_{switch} transistor are not necessary. In addition, the output voltage from each chip must be converted into a corresponding current to drive the WTA circuit.

Figure 4 displays the overall schema of the circuitry (MCHIP) for a general comparison among W chip outputs Vo(A)1...Vo(A)W, whereas Fig.5 gives a simple representation of the connections within the multiple-chip integrated system. These pictures demonstrate that the chip-integration structure exhibits the same modular nature and low connectivity that characterizes the single-chip sorting architecture.

Multiple chip integration requires tuning facilities to make output voltage levels consistent and insensitive to peculiarities and random fluctuations in the implementation processes. As said above, in each chip the output op-amp driver for analog voltage buffering is completely controllable through external pins, which makes it possible to adjust the amplifier gain and offset. In addition, external (discrete) circuitry can support on-board voltage balancing. The tuning process is easily accomplished 1) by feeding all chips with a common set of values (hence output voltages should ideally be equal), and then 2) by adjusting parameters in such a way that all input lines to

the integration module are equipotential.

4. Implementation Results

The single-chip SCHIP layout has been tested with HSPICE level 13 and 1µm technology. The transistor sizes (W/L) in Fig.2 are as follows: M1,M2: 25/2; Mswitch: 2/2; M4,M5: 5/2; M6:30/2; M7,M9: 6/2; M8,M10: 3/2 – power voltage is 0-5V; Ib=30µA. Figure 6a displays the layout of a pair of WTA cells.

Experimental tests confirmed the validity of the SCHIP circuit design. In Figure 6b, an excerpt from an actual HSPICE printout shows three steps of the system's sorting process. The upper section displays three current inputs with base values of 100, 110, and 120 µA, respectively. In the middle section, the output analog voltage is plotted; the clock is shown at the bottom. At $t=0$, input one prevails and is remapped to the output. The arrows point out two subsequent (clock-triggered) eliminations of the previous winners, hence inputs i2 and i3, respectively, are mapped to the output.

The sorting module is a subpart of a circuit for the VLSI support of vector quantization tools [12]. The whole chip stores 64 codevectors and processes 64-dimensional patterns (analog inputs and codevetors). Each vector unit evaluates the Euclidean distance from the input pattern, and the WTA/sorting module works out the k best-matching codevectors. The application domain is image coding; layout simulations give an effective operating speed of 2MHz, allowing the chip to attain a frame rate of about 240 frames/sec for standard 512x512, 8bpp images. These results outperform digital

98

Fig.6 - Implementation and simulation results
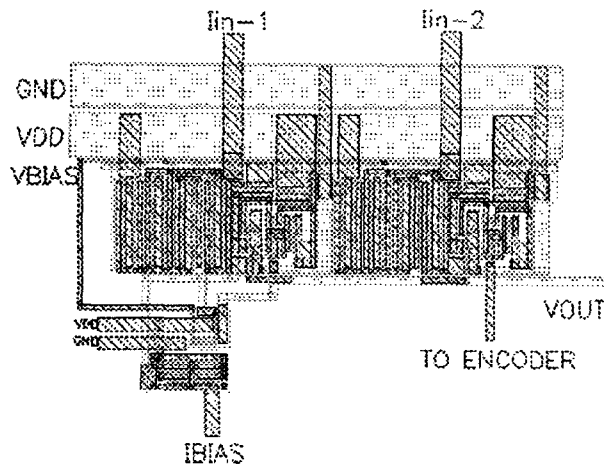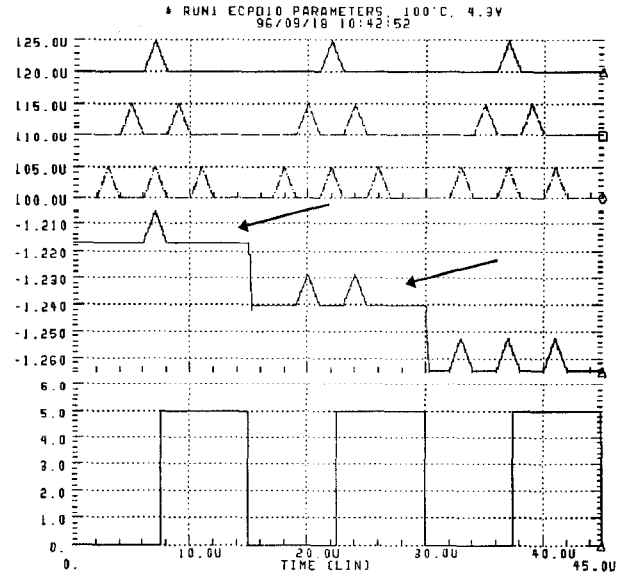
a) WTA elementary cell   b) SPICE trace of sorting-circuit simulation

VLSI approaches to VQ system implementations, which attain maximum frame rates of about 30 frames/sec [13].

Future lines of research in this area will include the chips production for both SCHIP and (later) MCHIP and field testing. Parallel research will concentrate on the development of an integrated environment of VQ-based image coding including additional modules for adaptive coding and quadtree support.

5. References

[1] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd Ed, Academic Pr., 1990.

[2] I. Pitas, "Fast algorithm for running ordering and max/min calculation" *IEEE Trans. Circuits and Sys.*, vol.36, No.6, pp.795-804, June 1989

[3] T. Martinetz, S.G. Berkovich, and K. Schulten, "Neural Gas network for vector quantization and its application to time-series prediction", *IEEE Trans. Neur. Net.*, vol. 4,No. 4, pp. 558-569, 1993.

[4] J.H. Holland, *Adaptation in natural and artificial systems*, Univ. Michigan Pr., Ann Arbor, 1975.

[5] G.M. Blair, "Low cost sorting circuit for VLSI," *IEEE Trans. Circ. Sys. I*, vol. 43, no. 6, Jun. 1996.

[6] J. Lazzaro, R. Ryckebush, M.A. Mahowald, and C. Mead, "Winner take all networks of $O(n)$ complexity," *NIPS 2*, Morgan Kaufmann, pp.703-711, 1989.

[7] G. Oddone, S. Rovetta, G. Uneddu, R. Zunino "WTA circuit with proportional output," *1996 World Congr. Neur. Netw.*, Sept. 1996.

[8] D.A. Freitas and K.W. Current, "CMOS current comparator circuit," *Electronics Letters*, vol. 19, no. 17, Aug. 1983.

[9] R. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol.4, pp. 4-22, 1987.

[10] J.P.F. Sum, P.K.S. Tam, "Note on the Maxnet dynamics" *Neur. Comp.*, vol. 8, pp. 491-499, 1996.

[11] K. Urahama and T. Nagao, "K-Winners-Take-All circuit with O(N) complexity," *IEEE Trans. Neur. Net.*, vol. 6, no. 3, pp. 776-778, May 1995.

[12] F. Ancona, S.Rovetta, and R. Zunino, "Analog implementation of vector quantization," *1996 Int. Workshop on the HDTV*, Oct. 1996.

[13] K. Tsang and B.W.Y. Wei, "A VLSI architecture for a real-time code book generator and encoder of a vector quantizer," *IEEE Transactions on VLSI Systems*, vol. 2, no. 3, pp 360–364, Sept. 1994.

[14] G.V. Russo, M. Russo "A novel class of sorting networks" *IEEE Trans. on Circuits and Sys. I*, vol.43, No.7, pp.544-552, July 1996