An Ultra-Low-Power Voltage-Mode Asynchronous WTA-LTA Circuit

Jorge Fernández-Berni, Ricardo Carmona-Galán, Ángel Rodríquez-Vázquez
Institute of Microelectronics of Seville (IMSE-CNM)
Consejo Superior de Investigaciones Científicas y Universidad de Sevilla
C/ Américo Vespucio s/n, 41092, Seville, Spain
Contact email: berni@imse-cnm.csic.es

***Abstract*—This paper presents an asynchronous mixed-signal WTA-LTA circuit conceived to carry out local minimum-maximum indexing in massively parallel image processing arrays. The hardware is focused on energy-efficient operation. We describe a realization for the standard CMOS base process of a commercial 3-D TSV stack featuring a power consumption of only 20pW per elementary cell at 30fps. The proposed block is also capable of resolving small voltage differences without requiring any external reference. This leads to a hit percentage greater than 90% even when taking into account global process variations and mismatch conditions.**

I. Introduction

Winner-Take-All (WTA) and Loser-Take-All (LTA) circuits constitute basic building blocks for the hardware implementation of theoretical frameworks as dissimilar as neural networks [1], image analysis [2], [3] or signal rectification [4]. These blocks evaluate the highest — WTA — and lowest — LTA — values among a set of inputs. In particular, with regard to image analysis, the WTA and LTA blocks enable the extraction of the local maxima and minima, respectively, across an image. This extraction is suitable for massively parallel operation since the output corresponding to each pixel is independent from the outputs associated with the rest of pixels. A usual approach to address such parallel operation is SIMD-based focal-plane sensing-processing [5]. In this case, a minimum-maximum circuit per pixel is required in order to concurrently compare each pixel value with its neighborhood. The resulting fine-grained lattice demands an area- and power-efficient block realization in order to make up high-resolution processing arrays. Indeed, energy efficiency becomes crucial in 3-D IC technologies [6]. 3-D sensing-processing stacks comprise a top sensor layer followed by vertically interconnected layers exclusively dedicated to processing. This structure enables to improve the spatial resolution when compared to planar implementations by increasing the number of elementary processing blocks. But this is obviously achieved at the cost of increasing the power consumption too. The design of ultra-low-power building blocks is therefore mandatory in order to exploit the additional computational capabilities provided by vertical integration without shooting up the energy required.

There are two main approaches when it comes to implementing WTA-LTA circuits. The first approach is based on the principle of the current conveyor, with certain variations depending on the particular realization considered [7]–[9]. The operation is supported by a set of identical cells, driven by the corresponding voltage or current input signals, which are interconnected through a common source current. As a result of the concurrent competition among the cells, only one of them, the winner (WTA) or the loser (LTA), will remain active. The second approach does not carry out a concurrent interaction among all the inputs involved. Instead, a binary tree topology is deployed [10], [11]. The input signals are grouped by pairs. For each pair, there will be only one signal that will compete at the next layer of the tree. Eventually, the winner or loser among all the inputs initially considered will be obtained. There are also intermediate solutions between current conveyor and binary tree approaches [12]. To the best of our knowledge, all the WTA-LTA blocks reported so far require DC currents of at least several microamps to perform local extrema extraction from nine inputs — this number corresponds to a 3×3 neighborhood in pixel-level image analysis. While such static consumption is affordable for a reduced number of elementary processing cells, high-resolution arrays demand an extremely high power consumption. For example, a VGA-resolution processing grid means $\sim$307k blocks. Considering a DC current of only 1μA per block and a 1.5V standard CMOS process, the total power consumption associated exclusively with the static operation would amount to about 0.46W.

In this paper, a completely different approach for the VLSI implementation of WTA-LTA blocks is presented. It is based on a voltage-mode asynchronous circuit featuring no external biasing and no static consumption, apart from the unavoidable leakage. The complexity of the hardware is linear, $O(n)$, requiring only four additional transistors for each additional input. As an example, we describe the design realized for the 0.13μm 1.5V standard CMOS base process of the 3-D TSV stack commercialized by Tezzaron Semiconductor. A pessimistic estimation of the dynamic power consumption for such design, assuming a typical frame rate of 30fps, leads to around 20pW per block, what in turn leads to about 6μW for a VGA-resolution processing array. This estimation is extracted from the simulation results described next and from the power consumption figures of the standard logic cells of the technology provided by the manufacturer.

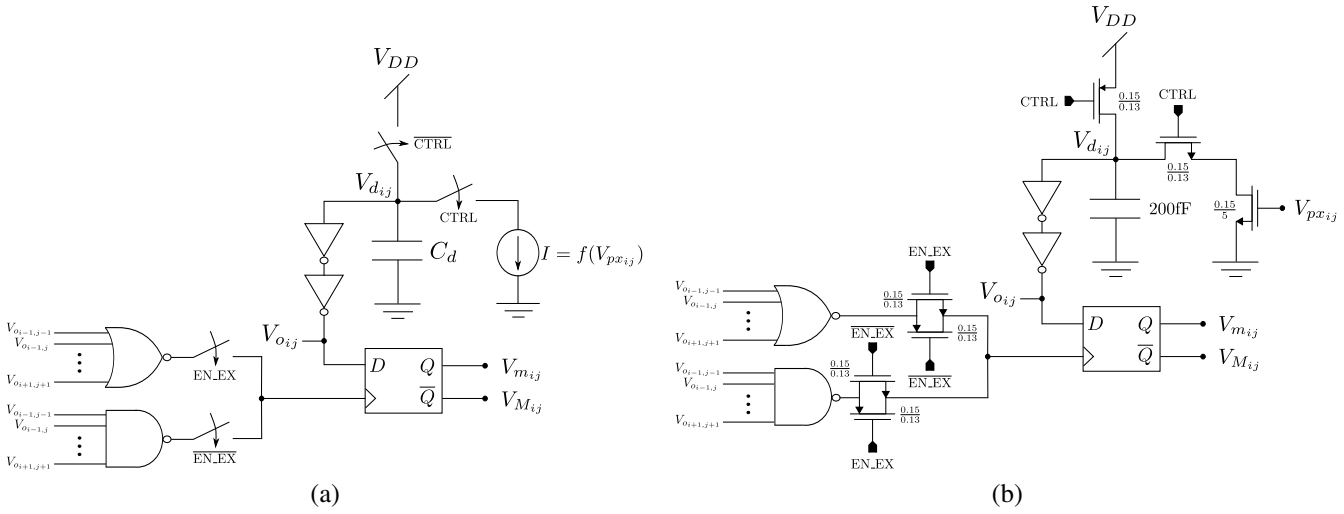

Fig. 1. General schematic of the proposed WTA-LTA block (a) and a particular realization for the Tezzaron/Globalfoundries CMOS process (b).

II. Elementary WTA-LTA cell

The proposed WTA-LTA circuit is depicted in Fig. 1(a). The operation is governed by a single global signal, 'CTRL', with loose timing requirements. It must be first set lo logic '0' during a certain time interval in order to precharge the capacitor C_d to V_{DD}. Then, it switches to logic '1', and from that instant on, the circuit performs the extraction of the corresponding local extremum asynchronously. Notice that when 'CTRL' switches to logic '1', the discharge of C_d through the pixel-dependent current source is enabled. A key point is that this current source must depend monotonically on the local pixel value, represented by the voltage $V_{px_{ij}}$. Thus, the larger $V_{px_{ij}}$, the larger the current I provided by the source and therefore the faster the discharge of C_d. Consequently, the time instant t_c at which $V_{d_{ij}}$ reaches the input threshold voltage of the first inverter will ultimately depend on the local pixel value. That time instant is known by the neighbors through the switching of the voltage $V_{o_{ij}}$ from logic '1' to logic '0'. Likewise, the cell under consideration receives this temporal information from each of its eight neighbors through the switching of the corresponding voltage $V_{o_{kl}}$. These voltages constitute the inputs for a NOR and a NAND gate. Let us now describe the ideal operation of the circuit. The NOR gate is associated with the extraction of the local minimum. Just before 'CTRL' switches from '0' to '1', its output is '0' since all the inputs have been set to '1'. As the discharge at the neighbor cells evolves, the inputs will be progressively switching to '0'. Eventually, the last one will switch, switching in turn the output of the NOR gate to '1'. At that time instant, let us call it t_m, the input of the positive-edge triggered flip-flop will be latched — we are assuming that 'EN_EX' was previously set to '1'. This input corresponds to $V_{o_{ij}}$. If this voltage still remains '1' at t_m, the discharge of C_d is the slowest when compared to its neighbors, that is, the pixel value $V_{px_{ij}}$ is the minimum. It will be represented by a logic '1' latched at $V_{m_{ij}}$. On the other hand, if $V_{o_{ij}}$ switched to '0' before t_m, the discharge of $V_{d_{ij}}$ is not the slowest. This fact is indicated by a logic '0' latched at $V_{m_{ij}}$. The NAND gate is similarly associated with the extraction of the local maximum. In this case, the switching to '0' of only one of its inputs determines the time instant t_M at which $V_{o_{ij}}$ is latched — 'EN_EX' had to be previously set to '0'. If $V_{o_{ij}}$ remains '1' at t_M, it can be concluded that the discharge of C_d is not the fastest among its neighbors and therefore $V_{px_{ij}}$ is not the maximum. However, if $V_{o_{ij}}$ switched to '0' before t_M, the discharge is then the fastest, being $V_{px_{ij}}$ the maximum. This is represented by a logic '1' at $V_{M_{ij}}$.

The design of the block just described for the Tezzaron/Globalfoundries CMOS process is depicted in Fig. 1(b). We make use of standard logic cells of the technology. The implementation of the current source is carried out by a single nMOS driven by $V_{px_{ij}}$. Notice that its drain voltage is always swept in the same way, no matter the pixel voltage considered. In such conditions, the inherent physics of the transistor makes sure a monotonically increasing dependance of the drain current with $V_{px_{ij}}$. Notice also that, in order to achieve low-power operation, this MOSFET must be adequately sized according to the signal range chosen. Otherwise, low pixel voltages could generate too low drain currents, thereby slowing down the discharge. This in turn could cause high power consumption during the — consequently long — time interval in which $V_{d_{ij}}$ is around the threshold voltage of the first inverter. The signal range for the pixels is [0.75V,1.5V]. The lower limit could be extended, but we constrain ourselves to this range for compatibility with other functional blocks already designed for this technology [13]. For the sake of a better visualization, simulation results for only three cells interconnected are shown in Fig. 2. The representations along the first row constitute an example of extraction of the maximum for pixel values of $V_{px_1} = 0.75V$, $V_{px_2} = 1.125V$ and $V_{px_3} = 1.5V$. It can be first seen how the precharge and subsequent discharge of the capacitors take place, then the signals mutually provided by

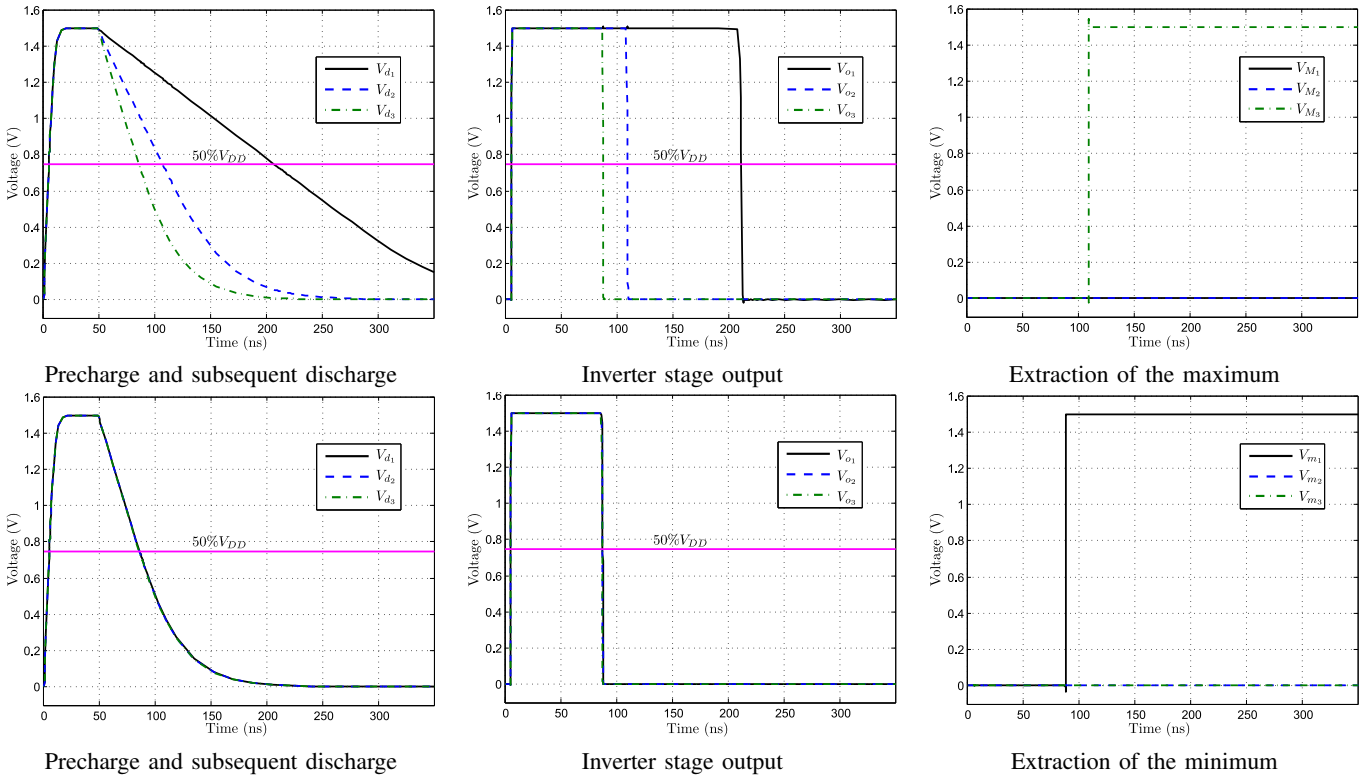

Fig. 2. Simulation results for three interconnected elementary cells and two different sets of input voltages. First row: extraction of the maximum for $V_{px_1} = 0.75V$, $V_{px_2} = 1.125V$, $V_{px_3} = 1.5V$. Second row: extraction of the minimum for $V_{px_1} = 1.496V$, $V_{px_2} = 1.498V$, $V_{px_3} = 1.5V$

the cells are depicted and finally the resulting output voltages are shown. The straight line crossing the first two diagrams corresponds to the input threshold voltage of a standard logic gate. The time required to carry out the operation is about 215ns. The second row is an example of extraction of the minimum for pixels differing only 2mV, $V_{px_1} = 1.496V$, $V_{px_2} = 1.498V$, $V_{px_3} = 1.5V$. This is really a worst-case scenario for the circuit to resolve the extremum. First of all, the differences between the pixel values are really small. Secondly, such differences occur at the highest part of the signal range, what implies that the discharges are faster. This in turn means that the time interval to transmit the local state to the neighbors is shorter in order not to latch wrong information. Pixel values around the lowest part of the signal range are easier to resolve. In such a case, the non-linearity of the transistor acting as the current source makes the discharges of close pixels much more distinguishable when compared to the other extreme of the range. All in all, we have obtained by simulation that, for the nominal case, with typical transistor models, the circuit can resolve differences of 2mV at the highest part of the range and 300μV at the lowest part. Under mismatch conditions, the minimum resolved difference is 9mV due to the mismatched signal paths affecting the timing of the operation. Anyway, notice that these really small differences, below 1.2% of the signal range, will be usually masked by the different noise sources typically existing in sensing-processing arrays, e.g. fixed pattern noise.

III. Simulation of a 32×32 array

In order to further evaluate the capability of the circuit to extract local maxima and minima, we have built a 32×32 array in HSPICE. A larger array was not possible due to the heavy memory and computational requirements of the simulations. Fortunately, since the binary output images only depend on the immediate neighborhood of each pixel, we could divide a 128×128-px image into 32×32-px subimages. Each subimage was mapped into the array, on which we incorporate additional peripheral cells in order to account for the neighbors of the pixels at the edges. Fig. 3 shows the simulated outcomes for local maxima and minima indexing over the Lena image under typical operation conditions. In order to compare the ideal extractions with the outputs provided by the array, we have analyzed them pixel by pixel. If a pixel of the ideal extraction is different to the corresponding pixel of the simulated extraction, either because a false extremum was triggered or because a extremum was not detected, the error count is increased by one. Otherwise, a new hit is added up. The hit percentages for the simulated extractions are really high: 99.31% for the maxima — 109 errors out of 16384 pixels, 1146 maxima correctly detected out of 1243 — and 97.17% for the minima — 464 errors out of 16384 pixels, 774 minima correctly detected out of 1238. These percentages remain high across the design space according to the simulations performed at the corners of the technology. The worst case is for the 'SS' corner, where the maxima hit percentage is 96.26% — 612

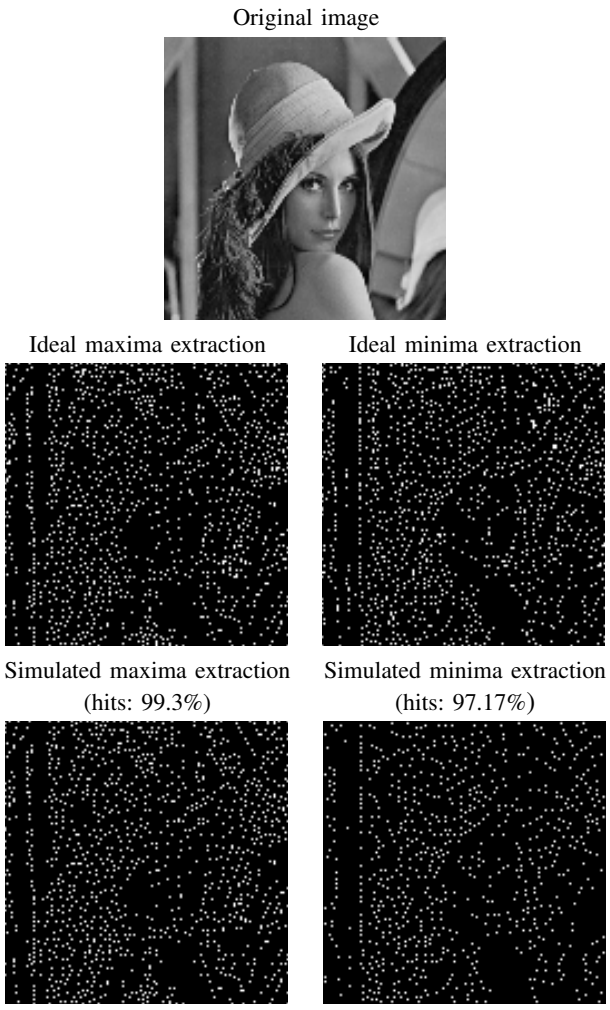

Fig. 3. Simulation results for a 128×128-px image making use of a 32×32-px array under typical conditions.

errors out of 16384 pixels, 903 maxima correctly detected out of 1243. It must be remarked that the ideal extractions consider the possibility of limit cases in which neighbor maxima or minima can arise from neighbor pixels featuring the same highest or lowest value, respectively. The adequate resolution of such cases by any physical realization is extremely difficult due to the unavoidable mismatch. In fact, the main reduction of performance occurs when mismatch variations are taken into account. We have performed 16 Monte-Carlo simulations of the array under typical operation conditions making use of the MOSFET statistical models provided by the manufacturer. These simulations correspond to the 16 subimages which make up the 128×128-px original image. The resulting hit percentage is reduced to 90.81% — 1506 errors out of 16384 pixels, 537 maxima correctly detected out of 1243. A direct way to improve this figure is by increasing the area of the transistor driven by the pixel value, which constitutes the primary source of error. The dimensions will finally depend on the tradeoff between accuracy and area.

IV. Conclusions

We have described a WTA-LTA block intented for ultra-low-power image processing arrays. In addition to the energy efficiency, the hardware presented stands out for its ability to resolve small voltage differences. Its complexity is linear, requiring only four transistors per additional input. The simulation results have demonstrated a high performance under different operation conditions. Future work will be focused on the design of new blocks which take advantage of the functionality provided by this WTA-LTA block, e.g. Difference of Gaussians within the Single Invariant Feature Transform (SIFT) framework.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments. This work is funded by MEyC (Spain) through projects TEC2012-38921-C02-01, co-funded by the European Regional Development Fund, and IPT-2011-1625-430000, by the Office of Naval Research (USA) through grant N000141110312, and by the Spanish Centre for Industrial Technological Development, co-funded by the European Regional Development Fund, through Project IPC-20111009.

References

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.
[2] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.
[3] P. Rüedi, P. Heim, F. Kaess, E. Grenet, F. Heitger, P. Burgi, S. Gyger, and P. Nussbaum, "A 128×128 pixel 120-dB dynamic-range vision-sensor chip for image contrast and orientation extraction," *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2325–2333, 2003.
[4] J. Koton, A. Lahiri, N. Herencsar, and K. Vrba, "Current-mode precision full-wave rectifier using two WTA cells," in *Int. Conf. on Telecommunications and Signal Processing*, 2011, pp. 324–327.
[5] A. Zarándy, Ed., *Focal-plane Sensor-Processor Chips*. Springer, 2011.
[6] G. Campardo, G. Ripamonti, and R. Micheloni, Eds., *Proceedings of the IEEE*, vol. 97, no. 1, 2009.
[7] J. Ramirez-Angulo, J. Molinar-Solis, S. Gupta, R. Carvajal, and A. Lopez-Martin, "A high-swing, high-speed CMOS WTA using differential flipped voltage followers," *IEEE Trans. Circuits Syst. II*, vol. 54, no. 8, pp. 668–672, 2007.
[8] M. Rahman, K. Baishnab, and F. Talukdar, "A high speed and high resolution VLSI winner-take-all circuit for neural networks and fuzzy systems," in *Int. Symp. on Signals, Circuits and Systems*, 2009.
[9] M. T. Moro-Frias, D.and Sanz-Pascual and C. A. de la Cruz Blas, "A novel current-mode winner-take-all topology," in *European Conf. on Circuit Theory and Design*, 2011, pp. 134–137.
[10] B. Tomatsopoulos and A. Demosthenous, "Low power, low complexity CMOS multiple-input replicating current comparators and WTA/LTA circuits," in *European Conf. on Circuit Theory and Design*, 2005, pp. 241–244.
[11] H. Hung-Yi, T. Kea-Tiong, T. Zen-Huan, and C. Hsin, "A low-power, high-resolution WTA utilizing translinear-loop pre-amplifier," in *Int. Conf. on Neural Networks*, 2010.
[12] R. Dlugosz and T. Talaska, "A low power current-mode binary-tree WTA/LTA circuit for Kohonen neural networks," in *Int. Conf. on Mixed Design of Integrated Circuits and Systems*, 2009, pp. 201–204.
[13] J. Fernández-Berni, R. Carmona-Galán, and A. Rodríguez-Vázquez, "Ultralow-power processing array for image enhancement and edge detection," *IEEE Trans. Circuits Syst. II*, vol. 59, no. 11, 2012.