Rafał DŁUGOSZ[1,2,3], Tomasz TALAŚKA[2]

Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland (1)
University of Technology and Life Sciences, Faculty of Telecommunication and Electrical Engineering, Bydgoszcz, Poland (2)
Poznań University of Technology, Department of Computer Engineering, Poland (3)

A power-efficient, current-mode, binary-tree min / max circuit for Kohonen self-organizing feature maps and nonlinear filters

Abstract. *A novel current-mode, binary-tree Min / Max circuit for application in analog neural networks and filters has been presented. In the proposed circuit input currents are first converted to step signals with equal amplitudes and different delays that are proportional to the values of these currents. In the second step these delays are compared using a set of time domain comparators in the binary tree structure that determine Min or Max signal. The circuit realized in the CMOS 0.18 µm process offers a precision of 99.5% at data rate of 2.5 MS/s and energy of 0.5 pJ per input.*

Streszczenie. *W pracy zaproponowano nowy, pracujący w trybie prądowym układ Min / Max oparty na strukturze drzewa binarnego, do zastosowań w analogowych sieciach neuronowych oraz filtrach nieliniowych. W układzie tym sygnały prądowe najpierw zamieniane są na sygnały skoku jednostkowego o równych amplitudach i różnych opóźnieniach. Następnie opóźnienia te porównywane są w komparatorach czasu znajdujących się w strukturze drzewa binarnego wskazującej sygnał o minimalnej lub maksymalnej wartości. Układ zaprojektowany w technologii CMOS 0.18 µm charakteryzuje się precyzją działania na poziomie 99.5 %, przy szybkości przetwarzania danych 2.5 MS/s oraz energii 0.5 pJ na każde wejście.* (***Nowy, pracujący w trybie prądowym układ Min / Max oparty na strukturze drzewa binarnego, do zastosowań w analogowych sieciach neuronowych oraz filtrach nieliniowych***)

Keywords: Min / Max operations, analog neural networks, nonlinear filters, CMOS implementation, low energy consumption.
Słowa kluczowe: funckje Min / Max, analogowe sieci neuronowe, filtry nieliniowe, implementacja CMOS, niskie zużycie energii

Introduction

The Max and the Min functions, often called the winner takes all (WTA) and the loser takes all (LTA) operations respectively, are useful in various applications, like artificial neural networks (ANN), signal and image processing and telecommunication systems.

In competitive learning in Self -Organizing Feature Maps (SOFM) proposed by Kohonen the Min function is using [1].

Other possible applications include noise removal, edge detection and object correction in pictures e.g. in pattern analysis [2]. Running Min / Max filters may be combined together to realize more complex tasks, e.g. morphological dilatation and erosion smoothing operations commonly used in image processing [3].

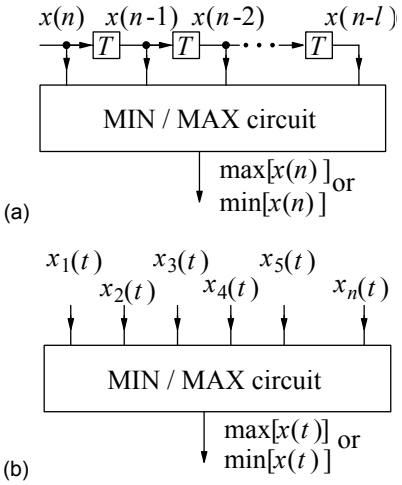

Fig. 1. Two applications of the Min / Max circuit: (a) the nonlinear filter (b) the WTA circuit used neural networks

A large similarity exists between the nonlinear running Min / Max filters and the LTA / WTA circuits, as shown in Figure 1. In both cases the core circuit performs the same task, which relies on searching for either the minimum or the maximum signal among a set of the input signals. The only difference lies in type of the input signals. The Min / Max nonlinear filters, shown in Fig. 1 (a), process a single signal, which is sampled either in time domain (1-D signal) or in the 'pixel domain' in image processing (2-D signal). On the other hand, in case of the LTA / WTA circuits, shown in Fig. 1 (b), all input signals come from independent sources. In both cases the same core circuit can by used.

Numerous realizations of the Min and the Max circuits have been reported in the literature [4-9], but two main groups of these circuits may be clearly distinguished. The first group embraces these solutions, in which all, *M*, input signals are compared simultaneously one with each other in a single step. In most cases they utilize the principle of the current conveyor (CC), in which 1-dimensional source coupled array of MOS transistors (or emitter coupled array in case of bipolar transistors) conveys the common source current (or common emitter current) i.e. the bias current to drain of the transistor with the largest input signal [4, 5], while the other branches remain cut-off in this time. Circuits of this group usually feature a simple structure but their performance is limited in terms of, *M*, since the precision linearly degrades with *M* [4].

One of the problems encountered in circuits of this type is the so called corner error, which occurs when two or more input signals have similar values and simultaneously are close to the max or the min input signal (depending on configuration). In this case the bias current is divided into the corresponding transistors. As a result, the output signal does not follow the winning signal but becomes close to an average value of these input signals [5], thus reducing the circuit precision. An additional problem is that CC circuits usually must be supplied with high voltage, although this problem has recently been addressed by several solutions reported for example in [5]. Circuits that belong to this group have limited usage to some applications only, e.g. the nonlinear filters, since they only calculate the winning signal, but are unable to indicate the address of this signal.

The second group of the Min / Max circuits utilizes a conception of the binary tree (BT). In this approach particular input signals are coupled and only one signal from each pair becomes a local winner and takes part in the competition at the next layer in the tree [4, 6-9]. In contrary to circuits that belong to the CC group, the BT solutions are able to calculate the value of the winning signal but also to provide the information about the address of this signal. This feature significantly increases the area of potential applications of these circuit e.g. in artificial neural networks such as Kohonen self-organizing feature maps (SOFM).

The BT solutions usually suffer from limited precision, since the input analog signals are copied many times from one layer to the other in the tree, which is the source of errors that cumulate at the top of the tree. One of the reasons of this error is the offset at the comparator input in particular competing units.

In this paper we propose a new binary-tree solution, in which the relatively high precision is achieved by eliminating copying of analog signals between layers in the tree [10]. This circuit is kind of the intermediate solution between the CC and the BT circuits. The input signals are first converted to corresponding delays, which is performed in a single step under the same conditions, as it is in CC circuits. In the next step the binary tree, in which only digital elements are used, makes an unambiguous selection of one winning signal and determines the address of this signal.

The proposed Min / Max circuit is going to be used in a new prototype of analog Kohonen SOFM, which will be the continuation of the previous authors' work in this topic [11]. The first prototype of this network has been realized in the CMOS 0.18 µm technology and experimentally verified by means of the laboratory tests. In Kohonen neural networks several operations are performed. Just after applying a new input training pattern, the network calculates the Euclidean distance between this pattern that is the vector of the input signals (currents in our approach) and the weights vectors in all neurons. The resultant signals, which are measures of these distances for particular neurons are in the following step provided to the WTA block that determines the address of the neuron with the smaller value i.e. the neuron whose weights are the most similar to the learning pattern. In fact, looking from the formal point of view this block should be called LTA, since it determines the minimum signal, but since this neuron is called the winner, therefore this block is usually called WTA. The winning neuron is in the next step allowed to adapt its weights.

In the previous prototype all required blocks have been designed by the authors from scratch. Although the network behaves properly is still needs the optimization of particular components, which is addressed by this work.

The paper is organized as follows. The conception of the proposed circuit is described in second section. The following section is devoted to performance evaluation of our new circuit, while transistor level verification by means of postlayout simulations is presented in next section. The conclusions are formulated at the end.

The proposed Min / Max bninary tree circuit

The general structure of the proposed BT circuit is shown in Figure 2, for an example case of 8 inputs and 3 layers.

This circuit contains three main blocks. The current-to-time converter (ITC) set up the flag signal F after a delay, which is linearly proportional to the value of the input current. The ICT block features a very simple structure that consists of a single PMOS current-mirror, a capacitor *C*, a switch and two NOT gates. Both gates change logical state once the V_C voltage becomes larger than a given threshold value, which equals about $V_{DD}/2$. The ICT block is the only analog component in the overall circuit. The advantage of this solution is that the conversion of all the input currents is performed under the same conditions. The potential source of limited precision is in this case the threshold voltage, V_{TH}, mismatch between transistors in the current mirrors as well as the insufficient matching between capacitors. The careful layout design helps to improve the matching between capacitors mostly by ensuring similar values of the parasitic capacitances in all the ICT blocks. On the other hand, in the realized circuit the transistors in CMs have sufficiently large sizes (W/L = 20µm / 5µm) to ensure the sufficient matching. In the weak inversion region of operation the resultant error, which is due to the threshold voltage variation, ΔV_{th}, does not exceed 2% [12], while in the strong inversion region the influence of this parameter is much smaller, resulting in the precision which is as high as 99.5% [13]. Such parameters are utterly sufficient in analog Kohonen SOFMs.

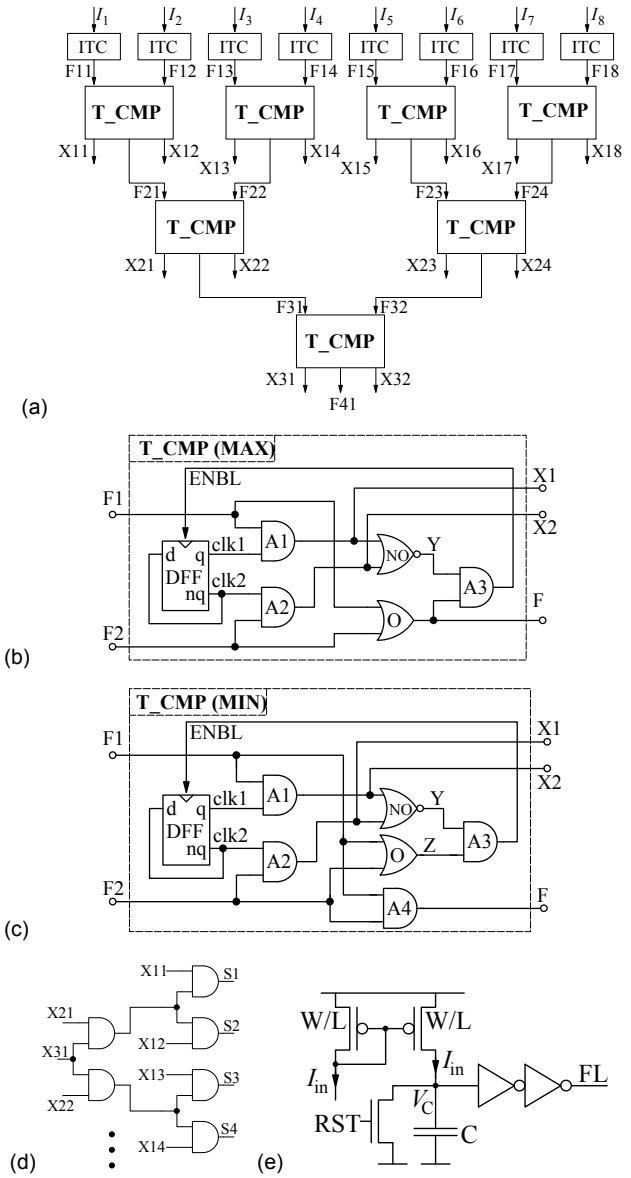

Fig. 2. The proposed WTA/LTA circuit: (a) the general structure, (b, c) the time domain comparator for the Max and the Min functions respectively, (d) the address detecting block (e) the current-to-time converter.

The second component in the proposed circuit is the time domain comparator T_CMP. Two versions of this block can be used, depending on type of the required function i.e. the Min or the Max. These circuits have been derived from the circuit proposed earlier in [14] for a different application. The main task of this circuit is to switch over the internal D-flip flop in such a way, to make the comparator pointing out (by use of the X signal) this input flag that 'came' as first or as second, depending on the configuration. The output flag becomes then the input signal to one of the competing units at the next layer in the tree. The flag signal in the Min and the Max circuit occurs at the output in different moments. In the Max circuit the output flag becomes the logical '1' just after at least one of the input flags becomes '1'. This allows to detect 'the fastest' input signal and then the fastest pairs

at particular layers. As fast as the flag at the output of the last layer with only one competing unit becomes '1', all ITC blocks are immediately reset and turned off to save energy.

In case of the Min mode the output flag becomes '1' only if both the input flags become '1'. In this case even if one of the flags is fast, a given pair loses the competition if the second flag is slower than flag signals in other pairs. The internal structure of both the Min and the Max circuits differ insignificantly i.e. one additional gate (A4) is required in the comparator in the Min circuit and the connection scheme of the X1 and X2 output signals is reversed in this case.

The address block (ADDR) is used for an unambiguous indication of the Max or the Min input signal. In both cases the same circuit may be used as it uses the X signals.

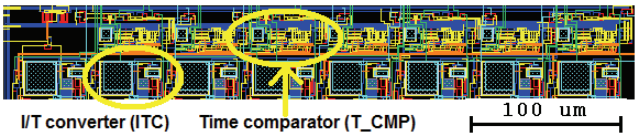

Fig. 3. Layout of the proposed Min / Max circuit in an example case with 8 inputs in TSMC CMOS 0.18 µm technology

Layout of an example Min circuit that is to be used in the next prototype of the Kohonen neural network is shown in Figure 3 for 8 independent analog inputs. The chip area occupied by this circuit equals 20.000 µm^2 i.e. 2500 µm^2 per a single input. Approximately 65% of this area is occupied by the analog ITC block.

Evaluation of the proposed circuit performance

To determine basic parameters of the proposed circuit let us consider equations (1) – (4). These equations may be derived on the basis of Figure 4 that illustrates charging the capacitor C for different input currents. The T_D time constant is the smallest delay between two input flags that allows the comparator to make a proper decision. This time is required to switch over the DFF element in the comparator. Precision of the Min / Max circuit depends on the total charging time. The longer is this time, the smaller differences between the input currents can be properly distinguished. On the other hand, we must remember that long charging time limits the maximum data rate, so there is a trade-off between these parameters. The charging time depends on the value of the input current but also on the value of the capacitor, C, and the supply voltage V_{DD}. All these parameters, as well as the assumed relative difference, n, between two input currents which are compared in a given comparator are combined in Equation (1) that allows calculating the maximum allowable value of the input currents for required circuit precision, which is the function of the n parameter. Equations (2) and (3) describe the overall minimum time of a single detection cycle for both the maximum and the minimum values of the input currents, respectively.

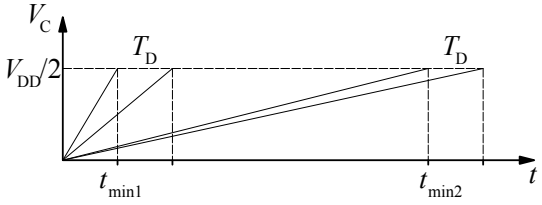

Fig. 4. The WTA / LTA circuit operation scheme.

(1) $$I_{max} = \frac{V_{DD}C}{2T_D}\frac{n}{(1-n)}, \quad n = \frac{I_1 - I_2}{I_1}$$

(2) $$T_{DEL\_max} = \frac{(1-n)T_D}{n} + \underbrace{\log_2 M \cdot T_{GATE}}_{T_{BT}} + T_{RST}$$

(3) $$T_{DEL\_min} = \frac{V_{DD}C}{2I_{min}} + T_{BT} + T_{RST}$$

(4) $$f_{max} = 1/T_{DEL\_min}$$

Since the T_{DEL_min} time can be considered to be the worst case, therefore this time should be used to calculate the maximum data rate, as it is described by Equation (4). The $T_{DEL\_max\ (min)}$ time of Equations (2) and (3) contains two components. One of them, T_{BT}, is a delay of the overall tree structure, while the second one, T_{RST}, is the time required to reset capacitors in the ICT blocks, which is performed after each detection cycle. The T_{BT} time component is linearly proportional to the number of the layers in the tree, which equals $\log_2 M$. The second component has been assumed to be equal to $10 \cdot T_{RC}$, where T_{RC} is a time constant of the discharge circuit in each of the ITC blocks ($T_{RC} = C \cdot R_{NMOS}$ → $T_{RC} \approx 50$ ps for CMOS 0.18µm technology).

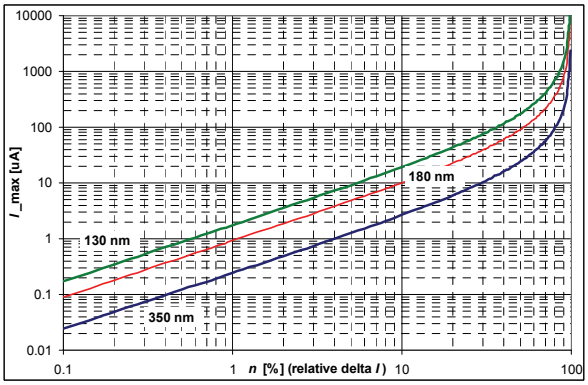

Fig. 5. The maximum value of the input signal for a given precision. The precision means here the smallest difference between any two input signals that can be properly distinguished by the circuit. This is the comparative study for different technologies.

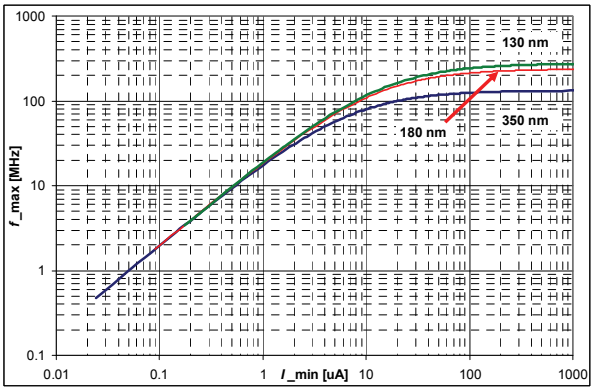

Fig. 6. Maximum data rate as a function of the minimum value of the input current I_{min} (comparative study for different technologies).

To better illustrate the performance and the parameters of the proposed circuit, two diagrams are presented in Figures 5 and 6. The results shown in these diagrams have been obtained for the capacitors C of 100 fF.

Figure 5 illustrates the maximum allowable input current for a given assumed relative difference n for the CMOS 0.35 µm, 0.18 µm and 0.13 µm processes. The difference between particular cases results from different values of the

T_D time constant, which depends on the minimal value of the transistor channel length in a given technology. The example values of this time constant have been determined to be equal to 2.04 ns, 0.57 ns and 0.3 ns in the 0.35μm, 0.18μm and 0.13μm processes, respectively. These values are for the supply voltage of 1.5 V for 0.35 μm process and 1 V in case of the other two. For other values of the supply voltage the T_D time constant will have different values.

The minimum value of the input current, for a given value of C, determines the maximum data rate of the overall circuit, as shown in Figure 6. To explain this, let us assume a required circuit precision of 99% that means that currents that differ by 1% (n=0.01) should be properly distinguished. In case of the 0.18 μm technology the maximum current is equal to about 1 μA in this case (Figure 5). The minimum current, I_{min}, has been selected to be 5 time smaller i.e. 200 nA. As a result, the maximum data rate is equal to 3.5 MS/s (for one layer). In the 0.35 μm technology, for the same precision, the maximum current is equal to 250 nA, while the I_{min} current, which also has been assumed to be 5 times smaller equals 50 nA. Two negative effects are visible in this case. Lower value of the I_{min} current reduces the maximum data rate to only 1 MS/s (for one layer). The other problem is that the operating point of transistors is in this case moved to the week inversion region that reduces the circuit precision. The first effect can be slightly reduced by increasing the value of the I_{min} current, although this reduces the dynamic range of the signal, since I_{max} can not be increased for a given value of C. On the other hand, increasing the value of C allows for increasing the I_{max} current, but this increases the energy consumption and the chip area. This shows that there exists a trade-off between described parameters. The most efficient way to improve the circuit performance is by decreasing the value of the T_D parameter, which is possible by either using the newer technology or by increasing the value of the supply voltage V_{DD} in comparators. In the second case this enlarges the energy consumption. The diagram below illustrates different trade-offs that exist in the proposed circuit. The symbol (↑) means increasing, (↓) decreasing, while (c) keeping const the value of the given parameter.

For a given constant value of C

to (↓) the assumed relative difference n → (↓) I_{max} → (↓) I_{min} (to keep dynamic range const) → (↓) data rate & (↑) gain error of CM in ITC (due to larger influence of V_{TH} mismatch)

to (↓) gain error of CM & (c) relative difference n → (↑) V_{DD} & (↑) I_{max} & (↑) I_{min} → (↑) power dissipation

By increasing the value of C →

(↑) I_{max} & (c) relative difference n → (↓) gain error of CM

(↑) I_{max} → (↑) power dissipation

(↑) circuit precision → (↑) power dissipation

Postlayout simulation results

Selected simulation results are presented in Figures 7 and 8 for the following parameters: C = 100 fF, V_{DD} = 1 V and M = 8. The assumed smallest relative difference n equals to 0.01 (1%). For such parameters both the T_{BT} and T_{RST} time constants can be neglected, since their values are below 1% of delay of a single ITC block. The I_{max} current has been selected to be equal to 1 μA, while the I_{min} as 200 nA.

The results presented in Figure 7 are for the Max circuit. Figure 7 (top) illustrates the case, in which all input signals are close to the lower range and differ by 1%. Data rate can properly be calculated only in the worst case scenario, in which all comparators are forced to switch over during a single detection cycle. To ensure such conditions, the input currents between two adjacent cycles change the values from I_i = {200, 201, 202, 203, 204, 205, 206, 207} to I_i = {207, 206, 205, 204, 203, 202, 201, 200} [nA]. Figure 7 (bottom) illustrates a typical situation, in which all the input signals are spread over the entire input data range. This allows for better illustration of the performance of particular components. The top three panels illustrate flag signals at the outputs of particular T_CMP blocks at particular layers in the tree. The 4th panel illustrates operation of the ADDR block, while the 5th panel illustrates the operation of the ITC blocks. The supply current I_{DD} is shown in the 6th panel. Figure 8 illustrates the similar results for the Min circuit.

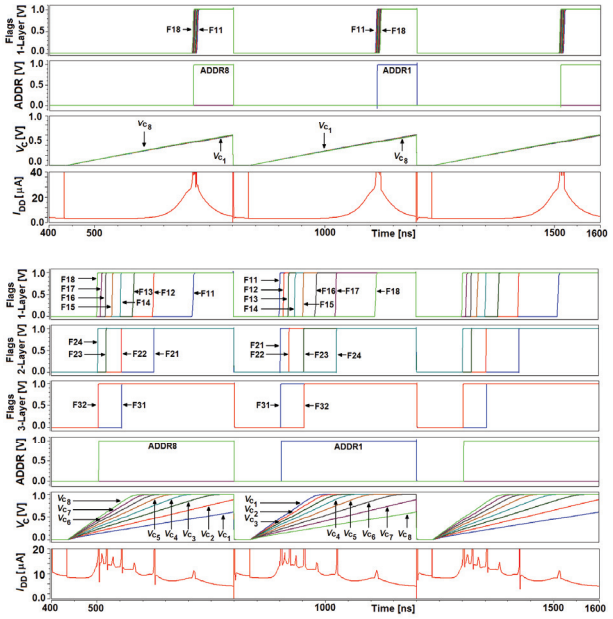

Fig. 7. Simulation results for the Max circuit: (top) the worst case i.e. for all input signals being close to the lower range, (bottom) a typical situation for all input signals being spread over an entire input range.

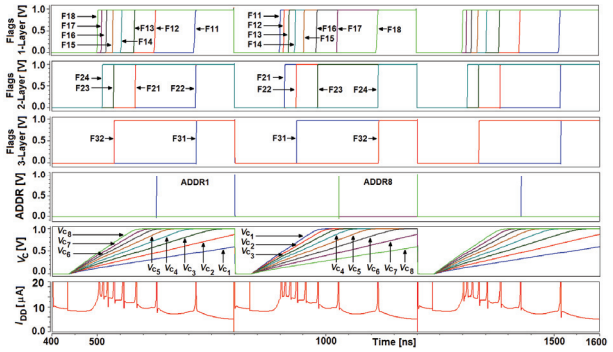

Fig. 8. Simulation results for the Min circuit configuration for a typical situation of all input signals being spread over an entire input range.

A very interesting feature of the proposed circuit in comparison to other BT circuits is that the relative difference n between the input currents is almost independent on the values of the compared currents. This means that currents, for example, of 1 μA and 0.99 μA, as well as 200 nA and 198nA (n=0.01 in both cases) will be properly distinguished. This looks like the offset is scaling down together with the signals, which means that the contrast remains const. This allows for realization of very sensitive neural networks.

The proposed Min / Max circuit has been compared with several binary tree solutions reported in the literature. This comparison is presented in Table 1 as well as in Figure 9. The proposed circuit is suitable for moderate data rates, but offers the smallest energy consumption per one input signal per detection cycle, which is one of the main advantages. The achieved precision is comparable to other solutions.

Table 1. Performance comparison between several Min / Max binary tree circuits reported in the literature.

Ref	Proc.	V_{DD}	No. inputs	P_{1_in}	f_S	Input range	Precision (1-*n*)	E_{1_in}
		[V]		[μW]	[MHz]	[μA]	[%]	[pJ]
[4]	2.4μm	5	8	200	13.8	100	99.00	14.5
[6]	0.5μm	3.3	8	106	5	3.3	99.80	21.2
[7]	0.35μm	3.3	8	70	1	10	99.00	70
[8]	0.8μm	6	8	120	2.8	50	N/D	43
[9]	N/D	3	8	284	20	70	98.57	14.3
This work	0.18μm	1	8	1.25	2.5 3 layers	0.2 - 1	99.5 [a] 97.9 [b]	0.5

a. For the strong inversion region of operation

b. For the weak inversion region of operation

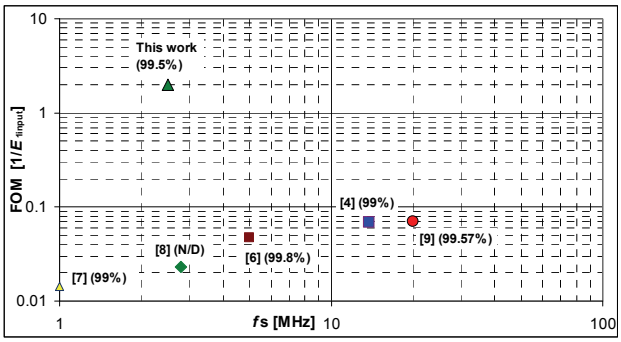

Fig. 9. Figure-of-Merit, defined as a reverse of energy consumption of a single input, as a function of data rate for reported binary tree solutions.

The proposed circuit is very robust against the process, supply voltage and temperature (PVT parameters) variation. The careful corner analysis, which typically is performed in commercial applications has been performed for different transistor models (TT, SS, FF), for temperature ranging between -40 and 90 ºC and the supply voltage that is in the range between 0.8 and 1.8V (in CMOS 0.18 μm process). The PVT parameters have the influence on the T_D, T_{RST} and T_{GATE} time constants but this does not considerably change the maximum data rate, since for small input currents of about 1 μA the main delay is introduced by the ICT block.

Conclusions

A novel simple binary-tree, current-mode Min / Max circuit has been proposed, implemented in the CMOS 0.18μm process and additionally verified in the CMOS 0.13μm and 0.35μm technologies for the comparison.

The circuit features a very simple structure, with very small chip area equal to 2500 μm^2 per a single input.

The only analog component (ITC block) in each channel consists of a single current mirror and one capacitor which due to connection in a very simple configuration, makes the proposed circuit robust against process, supply voltage and temperature (PVT) variation. The only influence of these parameters is on the detection time, which is insignificant, if some time reserve is assumed.

The circuit consumes a very low energy, which is the lowest among the reported binary tree solutions.

This circuit is addressed to ultra-low power neural networks, for application in portable medical equipment e.g. for online analysis of ECG signals for diagnostic purposes.

REFERENCES

[1] KohonenT., Self-Organizing Maps, *Springer Verlag*, Berlin, 2001

[2] Vemis M., Economou G., Fotopoulos S., Khodyrev A., The Use of Boolean Functions and Logical Operations for Edge Detection in Images, *Signal Processing*, 1995, Vol. 45, 161-172

[3] Jackway P. T., Deriche M., Scale-Space Properties of the Multiscale Morphological Dilation-Erosion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, Vol. 18, No. 1, 38-51

[4] Demosthenous A., Smedley S., Taylor J., A CMOS analog Winner-Takes-All network for large-scale applications, *IEEE Transactions on circuits and systems-I: Fundamental theory and Applications*, 1998, Vol. 45, No. 3, pp.300-304

[5] Ramirez-Angulo J., Molinar-Solis J. E., Gupta S., Carvajal R., Lopez-Martin A., A High-Swing, High-Speed CMOS WTA Using Differential Flipped Voltage Followers, *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2007, Vol. 54, No. 8, 668-672

[6] Hung Y. C., Liu Bin-Da, High-reliability programmable WTA/LTA circuit of O(N) complexity using a single comparator, *IEE Proceedings-Circuits Devices Systems*, 2004, Vol. 151, No. 6

[7] Chien-Cheng Yu, Yun-Ching T., Liu Bin-Da, Design of high performance CMOS current-mode winner-take-all circuit, *5th International Conference on ASIC*, 2003, Vol. 1, 568-572

[8] Wawryn K., Strzeszewski B., Current mode AB class WTA circuit, *The 8th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2001, Vol. 1, 293–296

[9] Tomatsopoulos B., Demosthenous A., Low power, low complexity CMOS multiple-input replicating current comparators and WTA/LTA circuits, *European Conference on Circuit Theory and Design (ECCTD)*, Ireland, 2005, Vol. 3, 241-244

[10] Długosz R., Talaśka T., A Low Power Current-Mode Binary-Tree WTA / LTA Circuit for Kohonen Neural Networks, *16th International Conference Mixed Design of Integrated Circuits and Systems (MIXDES)*, Poland, 2009, 201-204

[11] Długosz R., Talaśka T., Pedrycz W., Wojtyna R., Realization of a Conscience Mechanism in CMOS Implementation of Winner Takes All Neural Networks, accepted for publication in *IEEE Transactions on Neural Networks*, 2009

[12] Croon J. A., Rosmeulen M., Decoutere S., Sansen W., Maes H. E., An Easy-to-Use Mismatch Model for the MOS Transistor, *IEEE Journal of Solid-State Circuits*, 2002, Vol. 37, No. 8, 1056-1064

[13] Conti M., Betta G. D., Orcioni S., Soncini G., Turchetti C., Zorzi N., Test structure for mismatch characterization of MOS transistors in subthreshold regime, *IEEE International Conference on Microelectronic Test Structures*, 1997, Vol. 10, 173-178

[14] Długosz R., Asynchronous Front-End Asic For X-Ray Medical Imaging Applications Implemented In CMOS 0.18μm Technology, *15th International Conference Mixed Design of Integrated Circuits and Systems (MIXDES)*, Poland, 2008, 627-632

Authors: *dr inż. Rafał Długosz, dr inż. Tomasz Talaśka, Faculty of Telecommunication and Electrical Engineering, University of Technology and Life Sciences, ul. Kaliskiego 7, 85-796 Bydgoszcz, E-mail: rafal.dlugosz@epfl.ch, talaska@utp.edu.pl*

dr inż. Rafał Długosz, Institute of Microtechnology, Swiss Federal Institute of Technology in Lausanne (EPFL), A.L.Breguet 2, CH-2000, Neuchatel, Switzerland and Poznań University of Technology, Department of Computer Engineering, ul. Piotrowo 3A, 60-965, Poznań Poland

The correspondence address is: e-mail: rafal.dlugosz@epfl.ch