High-reliability programmable CMOS WTA/LTA circuit of $O(N)$ complexity using a single comparator

Y.-C. Hung and B.-D. Liu

Abstract: A high reliability complementary metal–oxide–semiconductor (CMOS) winner-takes-all/loser-takes-all circuit of $O(N)$ complexity with programmable capability is designed. Based on the proposed architecture, the precision of the circuit is independent of the number of inputs. This circuit is easily programmed for WTA or LTA function by an enable signal, without modifying the circuit structure or preprocessing the input variables. Since the circuit contains only simple logic gates and a single comparator, it is tolerant of VLSI process variations. The response time of the circuit increases linearly with the number of inputs. The input signal range of the circuit allows rail-to-rail (0–$V_{DD}$) operation. The supply voltage ranges from 2.7 V to 5 V. An experimental chip with six inputs was fabricated using 0.5-µm CMOS double-poly double-metal technology. The results show that a cell is either a winner or a loser if its input voltage is larger or smaller than the other cells by 10 mV.

1 Introduction

Winner-takes-all (WTA) and loser-takes-all (LTA) circuits are the major function blocks in pattern classification, optimisation problems, and self-organising neural networks [1, 2]. The function of a WTA (or LTA) circuit identifies the largest (or the smallest) input variable and inhibits the remaining ones. Many analogue circuits have been proposed, in which some circuits have current signals as input [3–8], whereas the others’ inputs are voltage signals [9–15].

In general, these traditional circuits can roughly be grouped as: (i) global-inhibition structure [2, 3], [9, 12–14], in which the connectivity increases linearly with the number of inputs; (ii) cell-based tree topology [5, 7]; (iii) excitatory/inhibitory connection [11]; (iv) serial cascade structure [15]. Figure 1 shows the conceptual diagrams of these topologies. In Fig. 1a, each cell receives the same global inhibition, and a common current $I_{com}$ or voltage $V_{com}$ is shared by all the cells. The cell represented in a square block is a nonlinear signal-processing element. In this architecture, the precision of the circuit is independent of the number of inputs. Since the operation of this circuit relies on the cells matching, a stable fabrication process is required for manufacturing a high-precision system. The complexity of the connectivity of the circuit is $O(N)$, where $N$ is the number of inputs. Figure 1b shows a cell-based tree topology, with $N–1$ cells arranged in a tree topology for $N$ inputs. Each cell receives two input variables to compare and outputs the larger (or smaller) of the two input signals. The backward digits in the bottom cell are then successively fed back to the first layer to identify the maximum (or minimum) input. The precision of the circuit is sensitive to cells matching. Figure 1c shows an excitatory/inhibitory connection with $O(N^2)$ complexity. Each cell receives the inhibited signals from other cells and an excitatory signal from itself. With this design, chip area increases with the square of the number of inputs. Based on ‘many comparators’ operation, Fig. 1d shows $N–1$ analogue comparison blocks and $N–1$ digital blocks cascaded in series. Within a comparison time $T_{comp}$, the largest magnitude input in each analogue block is sent to the next stage to compare with other inputs. The result of each comparison is then sent to the corresponding digital block, and a decision digit is fed back from right block to left block to identify the maximum input. As a result, the response time of the circuit is approximately $(N–1)·T_{comp}+T_{dig}$, where $T_{dig}$ is the total propagation time of the digital part. The precision of the architecture is dominated by the worst offset of the comparators. The major limitations of these conventional architectures are the fabrication process variations and the matching requirements of internal cells. The variations of a CMOS fabrication process include the transistor threshold voltage, actual device size, thickness of the gate oxide, and a variety of other factors. Analogue circuits are sensitive to these effects, especially for high-precision designs.

In this paper, a new high-reliability WTA/LTA architecture with wide input range and an adjustable supply voltage is proposed. The conceptual diagram of the circuit is shown in Fig. 2. In the proposed scheme, there are $N$ identical control cells and a single comparator for $N$ input variables. A comparator block multiplexes in time to achieve comparison of all inputs, and the comparator block itself is realised using an auto-zero technique. The WTA/LTA circuit is composed solely of basic elements; these include latches, inverters and logic gates. Since there are no critical elements in this design, performance of the circuit is stable with respect to variations in CMOS manufacturing.

2 Operating principle and circuit design

2.1 Operating principle

A symbol $\text{COMP}(V_{inp}, V_{out}) \ (1 \leq j, k \leq N)$ is defined such that the $i$th comparator cell receives two input variables ($V_{inp}$

---

© IEE, 2004

IEE Proceedings online no. 20040654
doi:10.1049/ip-cds:20040654

Paper first received 7th November 2003 and in revised form 14th April 2004. Originally published online: 29th November 2004

The authors are with the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan 70101, Republic of China
and $V_{\text{sink}}$ to compare in magnitude at time $t$, and the output $Z_t$ of the cell is the larger variable or a binary value. For $\text{COMP}_A(V_{\text{inj}}, V_{\text{sink}})$ operation, $Z_t$ is defined as

$$Z_t = \begin{cases} 1 & \text{or } V_{\text{inj}} > V_{\text{sink}}, \\ 0 & \text{or } V_{\text{sink}}, \text{ otherwise} \end{cases}$$

Therefore, the tree topology of Fig. 1b, WTA mode, is represented as

$$t_1 : \text{COMP}_{11}^1(V_{\text{in}1}, V_{\text{in}12}), \text{COMP}_{12}^2(V_{\text{in}1}, V_{\text{in}4}), \text{COMP}_{12}^{N/2}(V_{\text{in}(N-1)}, V_{\text{in}N})$$

$$t_2 : \text{COMP}_{12}^{N/2}(Z_{t1}^1, Z_{t1}^2), \text{COMP}_{12}^{N/2+1}(Z_{t1}^2, Z_{t1}^3), \ldots$$

$$t_{(\log_2 N)1} : \text{COMP}_{12}^{(N-1)}(Z_{t1}^{N-3}, Z_{t1}^{N-2})$$

After time $O(\log_2 N)$, the maximum input variable is obtained. $N-1$ identical comparators are necessary for this operation. Figure 2 shows the proposed topology and uses a single comparator to accomplish the whole operation. The operating procedures are described as follows:

$$t_1 : \text{COMP}_{11}^1(V_{\text{in}1}, V_{\text{in}2})$$

$$t_2 : \text{COMP}_{12}^2(Z_{t1}^1, V_{\text{in}3})$$

$$\vdots$$

$$t_{(N-1)} : \text{COMP}_{11}^N(Z_{t1}^{N-1}, V_{\text{in}N})$$

Since the required minimum time in each comparison step is different, each time interval must be designed as the worst value among all comparisons. The strategy adopted to find the maximum/minimum among a set of variables is that two variables are first compared; then the result of this comparison is compared with the next input variable using the same comparator. The procedure continues until comparisons of all input variables are completed. The single comparator operates in sequential time steps to compare different input levels. The comparison mode of the architecture is thus changed to a discrete time operation. Conceptually, circuit operation is similar to serial comparison. Unlike the traditional architectures that require $N-1$ comparators, this architecture requires only a single comparator. Using the algorithm described earlier, the LTA function is easily obtained by reversing the output state $Z_t$ in the same architecture.

### 2.2 Comparator

A comparator design from [16] is used herein; the schematic diagram is shown in Fig. 3. Transistors $M_{\text{sw1}}, M_{\text{sw2}}, M_{\text{sw3}}$ are used as switches. The circuit operates in two phases, the auto-zero phase and the comparison phase. Initially, in the auto-zero phase, clock $V_{\text{az}}$ becomes high to sample $V_{\text{in}1}$. Meanwhile, switch $M_{\text{sw3}}$ is closed; therefore the first inverter is biased at $V_b$. At this time, the amount of electric charge $Q_{\text{b1}}$ at node B is expressed as

$$Q_{\text{b1}} = (V_b - V_{\text{in}1}) \cdot C_s + V_b \cdot C_p + V_b \cdot C_m \quad (1)$$

where $C_s$, $C_p$, and $C_m$ are the sampling capacitance,
parasitic capacitance of the bottom plate of \( C_s \), and input capacitance of inv-1, respectively.

In the comparison phase, assuming the voltage at node B at this time is \( V_x \), \( V_x \) is the meta-stable voltage for the comparator. The amount of electric charge \( Q_{B2} \) at node B is expressed as

\[
Q_{B2} = (V_x - V_{a2}) \cdot C_s + V_x \cdot C_p + V_x \cdot C_{in}
\]  

(2)

Based on charge conservation, \( Q_{B1} \) in (1) and \( Q_{B2} \) in (2) must be equal, so that

\[
V_x = V_x + (V_{a2} - V_{in}) \cdot \frac{C_s}{C_s + C_p + C_{in}}
\]

where \( \Delta V_x \) represents the level difference of the two inputs, and \( z = \frac{C_s}{C_s + C_p + C_{in}} \), which represents a degrading factor. The function of the N-latch is to sample the voltage at node D as \( latch_{clk} \) sets high, and to hold the comparison result as \( latch_{clk} \) sets low. The max/min selector signal modifies the polarity of the compared result; therefore, without the need for structural modification, this circuit possesses a win/lose configurable capability. The dashed block in Fig. 3 shows the comparison block, which is used for all comparison procedures.

### 2.3 Architecture

The architecture of the N-input circuit is shown in Fig. 4, in which the \( N \) cells control_cells (1 \( \leq N \leq N \)) are identical. These cells are arranged in serial form, and \( N \) cells are required for \( N \) input variables. Each cell contains a status block, control_switch block, and latch blocks. The status block is designed to indicate whether the corresponding input variable is the winner or the loser. A following-inhibit signal ("fol_inhibit") from the following cell is used to reset the status block, and an output inhibit signal ("inhibit") is propagated to the preceding cell. The control_switch block receives an input variable and feeds this variable to the comparison block at the proper moment. Two input signals from following cell (the following-inhibit-switch signal

---

Fig. 3  Comparison block and control signals

Fig. 4  Architecture of the whole circuit
(fol_inhibit_sw) and the following-register signal (fol_reg)) are used to generate a sampling period. Two output inhibiting signals (inhibit_sw and own_reg) are used to inhibit or retain the winner's status in the preceding cell. Figure 5 shows clocks for the whole circuit.

2.4 Whole operation
The operation of the entire circuit is described using the circuit architecture in Fig. 4 and the clock waveform in Fig. 5. First, at t1, the reset signal is used to initiate the status blocks, control_switch blocks, and latch blocks. The N-latch in the status block and R01, R02, ..., R0N are reset to zero by the reset signal. Based on the max/min selector signal, MOS transistors Ms1, Ms2, Ms3, and Ms4 preset the initial sampling voltage (0 V or VDD) at node cap_comn. Regardless of the magnitude of the input-1 variable, it must be a winner during the initial interval for a serial comparison. The initial sampling voltage at node cap_comn is thus set as 0 V when the max/min selector signal is set to logic 1 for WTA operation, and vice versa.

Then, at t2, Vsel clock becomes high (auto-zero phase) to sample the initial voltage (0 V or VDD) at node cap_comn. Next, at t3, R01 becomes high to sample voltage Vin1. At this time, clock Vsel becomes low (comparison phase) to compare Vin1 with the initial sampling voltage, and the result of the comparison is stored in the N-latch of the first status block. The state of the N-latch is logic 1 if the variable is the winner. At t4, the present winner Vin1 is sampled again. At t5, a new comparison between previous winner Vin1 and Vin2 is made. At t6, the winner (the result of the Vin1 and Vin2 comparison) is sampled again. Then, a new comparison between the present winner and Vin3 is performed. The procedure continues until comparison of all the input voltages is completed. Ultimately, only one state Vosn (n = 1, ..., N) in these cells is at logic 1 for WTA/LTA indication; all others are at logic 0. Therefore, a WTA or LTA operation has been accomplished.

2.5 Block circuit description
Status block: Figure 6 shows the status block. The N-latch is used to store the result of the comparison (binary voltage Vos). Vos logic high means that the corresponding input variable is currently either a winner or a loser for WTA/LTA operation. The shaded part of the circuit is designed for a two-layer architecture, and will be discussed in Section 4.

Control_switch block: Figure 7 shows the control_switch block. It receives an input variable and controls the transmission gate to sample input variable. The shaded part in the figure is used in a two-layer architecture. Latch Block: Figure 8 shows the latch block [17]. The latch circuit is composed of an N-latch and a P-latch. By using a single-phase clock reg_clk, there is no clock skew in the circuit.
2.6 Design consideration

The accuracy of the comparator cell dominates circuit precision. The accuracy of the comparator is dependent on two factors: one is the clock feedthrough error and charge-injection error in transistor $M_{on}$, shown in Fig. 3 and the other is the degrading factor in (3). The charge-injection error is a complicated function of substrate doping concentration, load capacitor, input voltage, clock voltage, clock falling rate, MOS channel dimension, and threshold voltage. Therefore, this error is difficult to completely eliminate. Charge $Q_{overlap}$ induced by the lateral diffusion of channel length, and charge $Q_{inversion}$ induced by the channel inversion layer, are the primary error sources. First-order analysis of the error voltage gives

$$E_{err} \propto \frac{Q_{overlap} + Q_{inversion}}{C_s + C_{in} + C_p} = \frac{V_{DD} \cdot C_{in} \cdot W \cdot L \cdot \left( V_{DD} - V_{th} \right) \cdot C_{ik} \cdot W \cdot L}{C_s + C_{in} + C_p}$$

(4)

where $C_{in}$, $W$, $L_D$, and $\gamma$ are the gate oxide capacitance per unit area, channel width, lateral diffusion length, and fractional part of the charge leaving the MOS transistor $M_{on}$, respectively [18, 19]. In general, a complementary clock, transmission gates, and dummy transistors are adopted for switch realisation to reduce the error. Based on (3), the parasitic and input capacitances degrade the effective magnitude of the differential voltage ($V_{in2} - V_{in1}$). Therefore, circuit resolution decreases as parasitic capacitance $C_p$ and input capacitance $C_{in}$ increase. Accordingly, precision can be improved by increasing the sampling capacitance $C_s$ and reducing the dimensions of the first inverter (Fig. 3). In contrast with accuracy considerations, the settling time of the input signal during the sampling phase can be reduced by using a transmission gate ($M_{on}$ in Fig. 3) having smaller on-resistance or smaller sampling capacitance ($C_s$ in Fig. 3) to improve the speed. The trade-off is between circuit accuracy and circuit operating speed.

3 Simulation and measurement results

WTA/LTA functions, supply-voltage range, and Monte Carlo analysis of transistor variation were tested by simulation to verify the performance of the circuit. An experimental chip with six inputs was fabricated using 0.5 μm CMOS double-poly double-metal technology.

3.1 WTA/LTA functions

To test the functioning of the circuit, each example takes ten input voltages for WTA/LTA operation. For supply voltage $V_{DD} = 3.3 \text{V}$, the input variables ($V_{in1}$, $V_{in2}$, ..., $V_{in10}$) are (0.003, 0.006, 1.000, 0.997, 2.000, 2.003, 2.000, 3.297, 3.300, 3.297 V) to test WTA function, and (3.294, 2.000, 1.997, 2.000, 1.000, 0.997, 0.006, 0.009, 0.003 V) to test LTA function. During WTA operation, the logic $V_{out}$ of each cell at each time slice becomes:

$$V_{out} = \begin{cases} \frac{1}{10}, & V_{in}=0.003, 0.006, 1.000, 0.997, 2.000, 2.003, 2.000, 3.297, 3.300, 3.297 \text{V} \\ 0, & \text{otherwise} \end{cases}$$

When all comparisons are completed, logic ($V_{out1}$, $V_{out2}$, $V_{out3}$, ..., $V_{out10}$) = (0, 0, 0, 0, 0, 0, 0, 1, 0). Therefore, among these ten inputs, input voltage $V_{in10}$ is the maximum. Figure 9 shows the results of HSPICE simulation for WTA operation. The period of the latch clock (top trace) is 100 ns. For LTA operation, the final logic ($V_{out1}$, $V_{out2}$, $V_{out3}$, ..., $V_{out10}$) is (0, 0, 0, 0, 0, 0, 0, 0, 1, 0), and the input voltage $V_{in10}$ is the minimum variable. Choice of the above test voltages was based on the following: (i) input voltages of neighbour cells should be as close as possible to test discrimination capabilities; (ii) input voltages are distributed from 0 V to 3.3 V to test over a wide input range.

3.2 Supply voltage range

All circuit parameters, such as transistor dimensions, clock periods and sampling capacitance $C_s$ are held constant. Supply voltage $V_{DD}$ varies from 2 V to 5 V. Simulation results show that the circuit operates successfully with 3-mV discrimination over a supply voltage range from 2.7 V to

![Fig. 9 Simulation results of the circuit for WTA operation](image-url)
5 V. The unit operates under various commonly used supply voltages without any rescaling of device size.

3.3 Monte Carlo analyses of transistors – dimension and threshold-voltage variations
A statistical distribution of manufacturing parameters always occurs during CMOS fabrication. Threshold voltage, channel widths, and channel lengths of all MOS transistors were set to nominal values with ±5 % variation at the 3σ level, and each transistor was given an independent random Gaussian distribution. After thirty Monte Carlo iterations, HSPICE results indicate that circuit precision and speed are not degraded over this range. In addition, to verify the circuit with multi-technology support capability, circuit performance was simulated for different CMOS fabrication parameters. Results show that even for different fabrication processes the circuit functions correctly without tuning device dimension.

3.4 Measurement result
Figure 10 shows a micrograph of the chip. Input variables \((V_{i1}, V_{i2}, ..., V_{in})\) were set at \((1.00, 1.01, 2.00, 2.01, 3.29, 3.30)\) V to test WTA function. For simplicity, Fig. 11 shows the clock \(\text{reg:\_clk}\) and output responses \(V_{o4} - V_{o6}\). Obviously, \(V_{o6}\) is the winner among the input voltages. Input variables \((V_{i1}, V_{i2}, ..., V_{in})\) of \((3.00, 2.99, 1.00, 0.99, 0.01, 0.00)\) V were used to test LTA function. The measurement shows that \(V_{o6}\) is the loser. Table 1 summarises the characteristics of the circuit. Table 2 lists the results of comparisons with various circuits in the open literature.

4 Two-layer architecture and applications

4.1 Two-layer architecture
To further enhance performance, multiple WTA/LTA circuits can be expanded to a two-layer architecture. \(N\) inputs are divided into \(K\) groups to reduce the comparison time, each group having \(M\) input elements. Figure 12 shows a two-layer architecture consisting of a cascade of \(K\) WTA/LTA circuits at layer 1 and a single WTA/LTA circuit at layer 2. The \(K\) WTA/LTA circuits at layer 1 locate the local maximum or minimum variable among \(M\) inputs, and the result of layer 1 is sent to layer 2 to locate the global maximum or minimum. After the comparison result of layer 2 is fed back to layer 1, the maximum or minimum of voltage \(V_{o_{glb}}\) \((n = 1 \text{ to } N)\) is found. Since these \(K\) circuits operate in parallel, the response time of the circuits at layer 1 is \(M \cdot T_{\text{unit}}\) (excluding initiation time), and time \(K \cdot T_{\text{unit}}\) is required for layer 2 operation where \(T_{\text{unit}}\) is a unit of time for each variable comparison. Therefore, time \(T\) for the whole circuit to accomplish WTA/LTA operation is expressed as

\[
T = M \cdot T_{\text{unit}} + K \cdot T_{\text{unit}}
\]

(5)

Furthermore, if a total number \(N\) of input variables is required, the number of circuits at layer 1 must satisfy the

Table 1: Chip characteristics

<table>
<thead>
<tr>
<th>Process technology</th>
<th>0.5 (\mu)m CMOS double-poly double-metal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>WTA/LTA configurable</td>
</tr>
<tr>
<td>Supply voltage</td>
<td>2.7 V–5 V</td>
</tr>
<tr>
<td>Sampling capacitance</td>
<td>3 pF</td>
</tr>
<tr>
<td>Resolution</td>
<td>5 mV–10 mV</td>
</tr>
<tr>
<td>Latch clock</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Response time</td>
<td>linearly increasing with number of inputs</td>
</tr>
<tr>
<td>Dynamic range</td>
<td>rail-to-rail</td>
</tr>
<tr>
<td>Power dissipation</td>
<td>850 (\mu) at 3.3 V, 10 MHz (including I/O PAD)</td>
</tr>
</tbody>
</table>

Fig. 10 Micrograph of the chip

Fig. 11 Measured results of WTA operation
In order to reduce time $T$, the optimum number of inputs $M$ in each WTA/LTA obtained by differentiating (5) with respect to $M$ and using (6) to obtain

$$M = \sqrt{N}$$

The total response time $T$ of the two-layer architecture is

$$T = 2 \cdot \sqrt{N} \cdot T_{\text{unit}}$$

Therefore the time complexity of a two-layer WTA/LTA operation is improved from $O(N)$ to $O(\sqrt{N})$.

### 4.2 Two-layer applications

Based on the max/min-selector signal setting, different two-layer configurations are set for various requirements.

(i) WTA–WTA: All the max/min-selector signals at layer 1 and layer 2 are set as logic 1. This indicates that the function of each layer is set as WTA, and the result is a global WTA function.

(ii) LTA–LTA: All max/min-selector signals at two layers are set at logic 0 to attain a global LTA function.

(iii) WTA–LTA (LTA–WTA): The max/min-selector signals of layer 1 are set at logic 1, and the layer 2 signal is set at logic 0. The output of the whole system identifies the minimum among the local maxima. An LTA–WTA...
that 5 mV was obtainable at the cost of reduced operating speed.

6 Acknowledgments

This work is in part supported by the Chip Implementation Center, and in part by the MOE Program for Promoting Academic Excellence of Universities under Grant EX-92-E-FA09-5-4.

7 References