3.3 Row Winner-Takes-All (ROW_WTA)

We name the analog block constituted of row amplifier (ROW_AMP) and Winner-Takes-All (WTA) cell as ‘ROW_WTA’. The schematic view of ROW_WTA can be seen in Figure 3.34. Transistors M1, M2, M5 and M6 form the row amplifier and transistors M3, M4 form the WTA cell.

Lazarro’s WTA network has been used as Winner-Takes-All network [31]. There are two reasons for this choice.

1. N being number of input signals of winner-takes-all network; network connections have order of N complexity.
2. Input signal is compatible with the current-based output signal of the block DCELL.

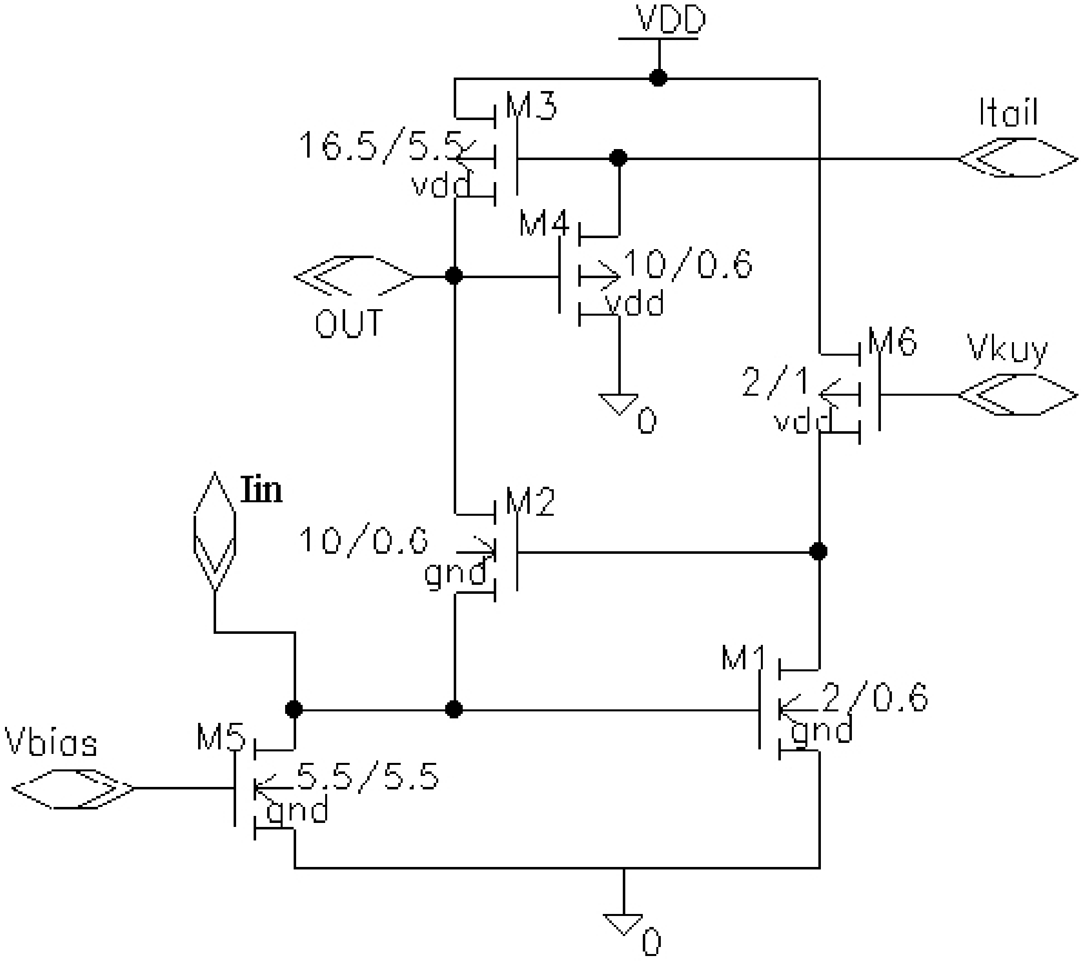

Figure 3.34 : The schematic view of ROW_WTA

WTA networks select the largest one among its input signals. Many different winner-takes-all networks have been presented in literature; each type of network has its own advantages and disadvantages. [2, 9]-[15-17].

WTA networks must satisfy two features:

1. Select one and only one winner for every moment t. (resolution)
2. Selected input must be the largest one among WTA's inputs for every moment t. (precision)

I will examine later to what extent ROW_WTA satisfies these criteria. Lazarro's WTA network can be seen in Figure 3.35. NMOS complementary realization is also possible but in this case, we have to mirror row amplifier output current. However, current mirrors cause precision loss because of the fabrication process noise and their finite output resistance.

In Figure 3.35, M1i and M2i constitute Lazarro's WTA cell number i. For N input signal, N WTA cell must be used. Node CSN is common for each cell, and it is connected to the source terminal of every M1i. The operation principle of the network can be summarized as follows:

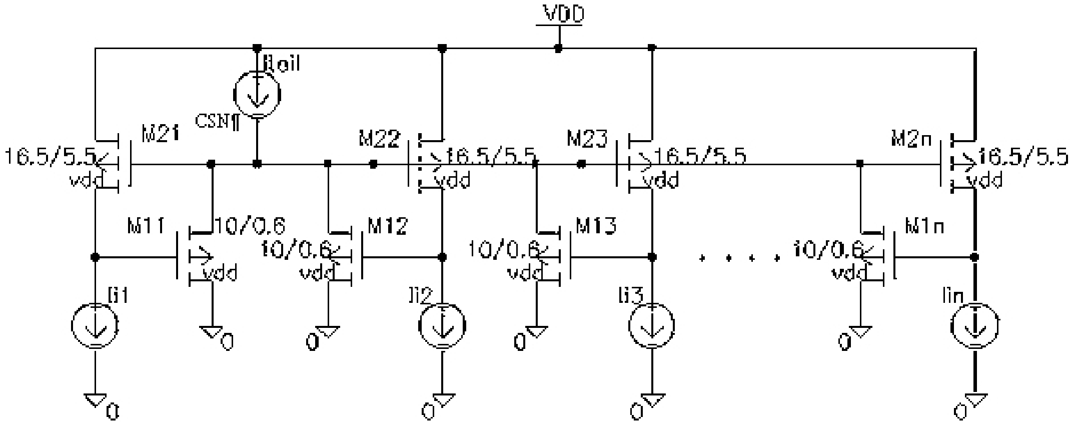

Figure 3.35 : Lazarro's WTA network

Voltage CSN is determined by transistor M2 of WTA cell that has the largest input current. If the drain current of a WTA cell's M2 transistor is less than its input current at a moment, voltage CSN must decrease for equaling drain current of cell's transistor M2 to cell's input current. Thus firstly, voltage of cell's output node start to decrease, cell's transistor M1 sinks more current and voltage CSN start to decrease until M2’s drain current is equal to cell’s input current (new equilibrium). If the cell that we examine is one of the looser-cells, cell’s M2 transistor tries to source more current (equal to the input current of the winner-cell) than cell’s input current. Thus, voltage of cell’s output node starts to increase and cell’s M1 transistor enters in the cut-off region. Finally, to equalize its drain current and input current, cell’s transistor M2 enters in the linear operation region. Only the winner-cell’s M1 transistor is conducting. Thus, tail current flows through the winner-cell’s M1 transistor. Tail current, input current and aspect ratios of M1 and M2 transistors determine winner-cell’s output voltage level.

Let me now examine stability conditions of the block ROW_WTA. As we can see clearly from Figure 3.36. , ROW_WTA contains two independent feedback loops. Each of them has two poles. Thus, we can easily state that ROW_WTA is unconditionally stable. To prevent ringing, non-dominant pole must be far enough from dominant pole in each feedback loop. Transistors M1 and M2 constitute first feedback loop. The dominant pole is on node `K` and non-dominant is on node `IN`. Dominant and non-dominant poles frequencies can be expressed as follows:

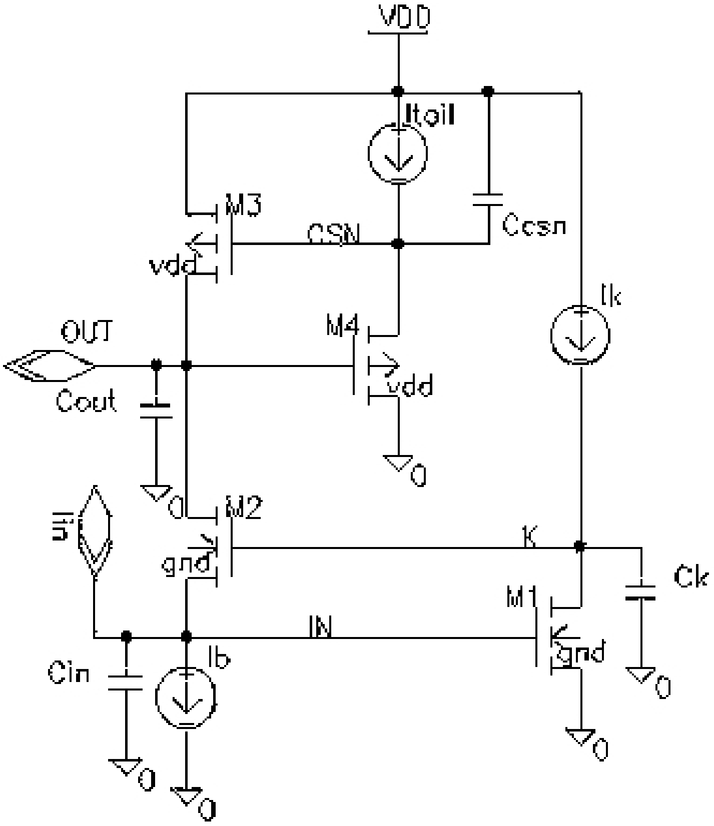

Figure 3.36 : Schematic view of ROW_WTA block with load capacitors on each node.

$$f_{P(K)} = \frac{1}{2\pi C_K R_K}$$

$$f_{NP(IN)} = \frac{1}{2\pi C_{IN} R_{IN}} \qquad 3.64$$

Gain bandwidth product is the product of feedback loop gain and dominant pole frequency. Thus,

$$GBW = K_L f_{PK} \qquad 3.65$$

K_L is the feedback loop gain.

$$K_L \cong gm_1 R_K$$

$$GBW = gm_1 R_K \frac{l}{2\pi C_K R_K} = \frac{gm_1}{2\pi C_K} \qquad 3.66$$

$$= \frac{\sqrt{2\beta_1 I_K}}{2\pi C_K}$$

$$C_K \cong \frac{2}{3} W_{M2} L_{M2} Cox$$

To prevent ringing, following condition must be satisfied

$$f_{NP(IN)} > 3GBW$$

$$\frac{1}{2\pi C_{IN} R_{IN}} > 3\frac{\sqrt{2\beta_1 I_K}}{2\pi C_K} \qquad 3.67$$

$$R_{IN} \cong \frac{1}{gm_2} \quad \text{body effect is negligible.}$$

$$gm_2 = \sqrt{2\beta_2 I_{D(M2)}} \qquad 3.68$$

$$\frac{1}{3}\frac{\sqrt{2\beta_2 I_{D(M2)}}}{\sqrt{2\beta_1 I_K}} > \frac{C_{IN}}{C_K}$$

In worst case

$$I_{D(M2\min)} = I_B - I_{IN\max}$$

$$\frac{C_{IN}}{C_K} < \frac{1}{3}\sqrt{\frac{(w/l)_2}{(w/l)_1}}\sqrt{\frac{I_B - I_{IN\max}}{I_K}} \qquad 3.69$$

$$C_{IN} = \frac{2}{3}W_{M1}\,L_{M1}\,Cox + 2M\,C_{DB(CELL)}$$

M is the number of DCELL on the DCELL row. $C_{DB(DCELL)}$ is the drain-bulk diffusion capacitance of the transistors which flow output current in DCELL blocks. Transistors M3 and M4 constitute the second feedback loop. During the stability analysis of the second feedback loop, we must consider that only the transistor M4 of winner-cell is conducting after it is chosen. Dominant pole is on node `OUT' and non-dominant is on node `CSN`. Dominant and non-dominant poles frequencies can be expressed as follows:

$$f_{P(OUT)} = \frac{1}{2\pi C_L\,R_{OUT}}$$
$$f_{NP(CSN)} = \frac{1}{2\pi C_{CSN}\,R_{CSN}} \qquad 3.70$$

Gain bandwidth product is

$$GBW = K_L\,f_{P(OUT)}$$
$$K_L \cong gm_3\,R_{OUT} \qquad 3.71$$

K_L is the second feedback loop gain. Thus,

$$GBW = gm_3\,R_{OUT}\frac{1}{2\pi C_L\,R_{OUT}} = \frac{gm_3}{2\pi C_L} = \frac{\sqrt{2\beta_3 I_{D(M3)}}}{2\pi C_L}$$
$$C_L = \frac{2}{3}W_{M4}\,C_{M4}\,Cox + C_O \qquad 3.72$$

Where C_O is the output load capacitance. To prevent ringing, following condition must be satisfied:

$$f_{NP(CSN)} \qquad > 3GBW$$

$$\frac{1}{2\pi C_{CSN} R_{CSN}} > 3\frac{\sqrt{2\beta_3 I_{D(M3)}}}{2\pi C_L}$$

$$R_{CSN} \cong \frac{1}{gm_{M4}} \qquad \text{body effect is negligible.}$$

$$gm_{M1} = \sqrt{2\beta_4 I_{TAIL}} \qquad 3.73$$

$$\frac{1}{3}\frac{\sqrt{2\beta_4 I_{TAIL}}}{\sqrt{2\beta_3 I_{D(M3)}}} > \frac{C_{CSN}}{C_L}$$

Worst case,

$$I_{D(M3\max)} = I_B - I_{IN\min}$$

$$\frac{C_{CSN}}{C_L} < \frac{1}{3}\sqrt{\frac{(w/l)_4}{(w/l)_3}}\sqrt{\frac{I_{TAIL}}{I_B - I_{IN\min}}} \qquad 3.74$$

$$C_{CSN} = N.\frac{2}{3}W_{M3}.L_{M3}.Cox$$

N is the number of input signals of WTA network. With respect to conditions (3.69) and (3.74), we determine biasing conditions and aspect ratios of ROW_WTA block.

In the fabrication step, mismatch is inevitable between WTA cells because of the process noise. The cell that has smaller input current may be assigned as winner-cell instead of the cell that has the largest input current due to the mismatch. I will model the effect of the process noise and try to find out the critical design parameters that must be taken into account to minimize it.

When we analyze closely the WTA network, we can easily remark that the node CSN is the common for each cell. That means after the voltage of the node CSN is determined, network is in steady state. Thus, the function determining the winner is the transfer function from the input current to the voltage of the node CSN. Once we find this transfer function, we can easily find out the sensitivity of it to process noise, with a straightforward calculation.

To find out the transfer function, let us consider a simple WTA cell in Figure 3.37(a). We start our analysis by opening the feedback loop as in Figure 3.37(b). We will calculate the open loop transfer function and then we will close the feedback loop and find out closed loop transfer function.

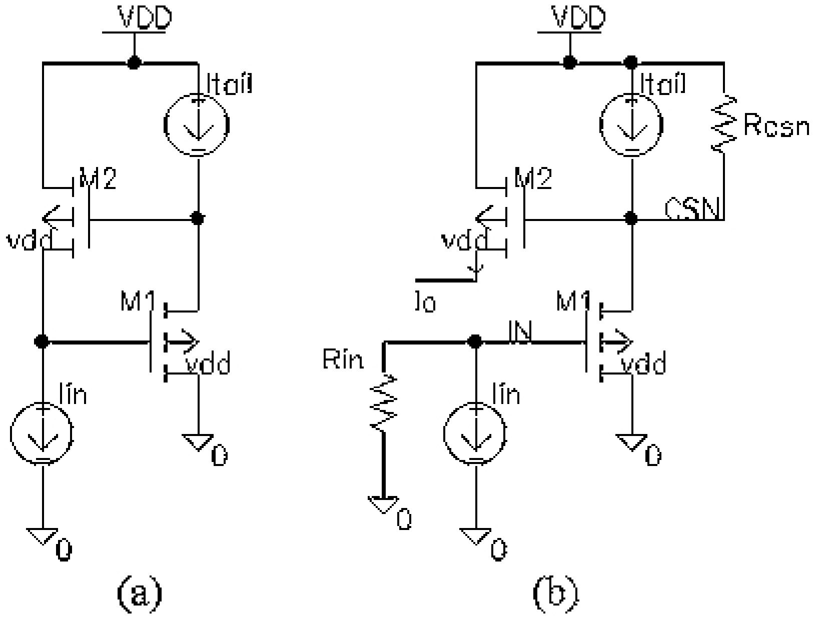

Figure 3.37 : (a) Simple WTA cell (b) The WTA cell that the feedback loop is opened.

The resistors R_{IN} and R_{CSN} in Figure 3.37.b are the total resistance on node IN and CSN respectively. The small signal equivalent circuit of the cell in Figure 3.37.b can be seen in Figure 3.38.

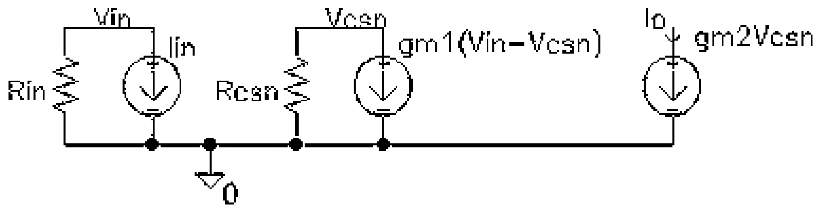

Figure 3.38 : Small signal equivalent circuit of the cell in Figure 3.37.b

Node equations are

$$V_{IN} = R_{IN} I_{IN} \qquad 3.75$$

$$V_{CSN} = gm_1 R_{CSN}\left(V_{IN} - V_{CSN}\right) \tag{3.76}$$

We substitute (3.76) in (3.75). Thus, open loop transfer function can be expressed as follows:

$$\frac{V_{CSN}}{I_{IN}} = G = \frac{gm_1 R_{IN} R_{CSN}}{1 + gm_1 R_{CSN}} \tag{3.77}$$

In order to find the close loop transfer function, we now close the feedback loop. Once the feedback is closed, we can model the circuit as in Figure 3.39.

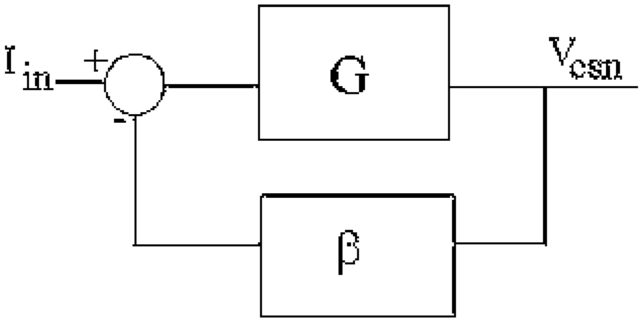

Figure 3.39 : Model of the feedback loop

In the Figure 3.39, the feedback circuit is

$$\beta = gm_2 \tag{3.78}$$

Thus, the close loop transfer function can be expressed as follows [36]:

$$Tf = \frac{V_{CSN}}{I_{IN}} = \frac{G}{1+\beta G} = \frac{\frac{1}{gm_2}}{1 + \frac{1}{gm_2 R_{IN}}\left(1 + \frac{1}{gm_1 R_{CSN}}\right)} \tag{3.79}$$

Obviously, under the assumption that the output resistance of the input current source is infinite, the resistance R_{IN} is equal to the output resistance of the transistor M2. Let us now calculate the relative sensitivity of the transfer function to the transconductances gm_1 and gm_2 that are varying with the process noise. The relative sensitivity of the transfer function to the transconductance gm_1 can be expressed as follows:

$$\int_{gm_1}^{Tf} = \frac{\delta Tf}{\delta gm_1}\frac{gm_1}{Tf} = \frac{1}{1 + gm_1 R_{CSN} + gm_1 gm_2 R_{CSN} R_{IN}} \qquad 3.80$$

The relative sensitivity of the transfer function to the transconductance gm_2 can be expressed as follows:

$$\int_{gm_2}^{Tf} = \frac{\delta Tf}{\delta gm_2}\frac{gm_2}{Tf} = \frac{1}{1 + \frac{1}{gm_2 R_{IN}}\left(1 + \frac{1}{gm_1 R_{CSN}}\right)} \qquad 3.81$$

Sensitivity analysis shows us that the transfer function is more sensitive to the parameter gm_2 rather than parameter gm_1. With respect to our previous analysis, we can draw the following conclusions to achieve better precision:

1. The parameter gm_2 must be designed insensitive to the process noise as far as possible. Thus, the channel area of the transistor M2 in Figure 3.37 must be as big as possible. In order to increase the denominator of the fraction in (3.81), the value of the parameter must be chosen as small as possible.
2. The parameter gm_1 must be chosen as big as possible.

Another important parameter for WTA networks is the resolution. To formulate this parameter, let me consider the WTA network constituted by two Lazarro`s WTA cells, in Figure 3.40.

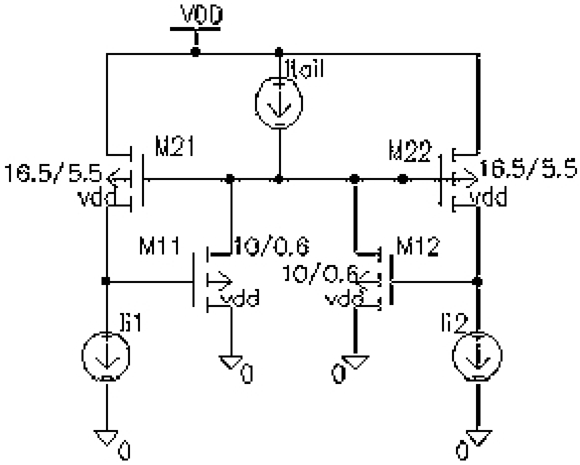

Figure 3.40 : Schematic of WTA network constituted by two Lazarro`s WTA cells

Gain of the feedback loop constituted by M1i and M2i is finite. To select correctly the winner-cell, that means all tail current flows through winner-cell’s M1 transistor, difference between winner-cell’s input current and others input currents must be greater than a minimum value determined by the finite feedback loop gain. To formulate this effect we must take into account output resistance of transistor M2i. Drain current of a PMOS transistor in saturation region can be expressed as follows:

$$I_D = \frac{\beta}{2}(V_T - V_{GS})^2 (1+\lambda|V_{DS}|) \qquad 3.82$$

The drain current of M21 is

$$I_{D(M21)} = I_{i1} = \frac{\beta_{21}}{2}(V_{TP(M21)} + V_{DD} - V_{CSN})^2 (1+\lambda_{21}(V_{DD} - V_{OUT1})) \qquad 3.83$$

We consider first cell is the winner-cell. Thus, transistor M12 is in cut-off region.

$$I_{i1} > I_{i2} \qquad 3.84$$

Gate-source voltage of transistor M12 is

$$V_{GS(M12)} = V_{TP(M12)} = V_{OUT2} - V_{CSN} \qquad 3.85$$

Thus, output voltage of transistor M22 is

$$V_{OUT2} = V_{CSN} - V_{TP(M12)} \qquad 3.86$$

Drain current of transistor M22 is

$$I_{D(M22)} = I_{i2} = \frac{\beta_{22}}{2}(V_{TP(M22)} + V_{DD} - V_{CSN})^2 (1+\lambda_{22}(V_{DD} - V_{OUT2})) \qquad 3.87$$

Since cell1 is the winner cell, all of the current I_{TAIL} flows trough M11. Thus drain current of M11 is:

$$I_{D(M11)} = I_{TAIL} = \frac{\beta_{11}}{2}(V_{TP(M11)} + V_{CSN} - V_{OUT1})^2 \qquad 3.88$$

Winner cell output voltage level can be expressed as

$$V_{OUT1} = V_{CSN} + V_{TP(M11)} - \sqrt{\frac{2\,I_{TAIL}}{\beta_{11}}} \qquad 3.89$$

Let us assume followings

$$\begin{matrix} V_{TPM21} = V_{TPM22} = V_{TP2} & \lambda_{11} = \lambda_{22} = \lambda & \beta_{11} = \beta_{12} = \beta_1 \\ V_{TPM11} = V_{TPM12} = V_{TP1} & & \beta_{21} = \beta_{22} = \beta_2 \end{matrix} \qquad 3.90$$

With respect to previous equations and assumptions, cells' input currents can be written as follows:

Input current of winner cell is

$$I_{i1} = \frac{\beta_2}{2}\left(V_{TP2} + V_{DD} - V_{CSN}\right)^2 \left(1 + \lambda\left(V_{DD} + \sqrt{\frac{2\,I_{TAIL}}{\beta_1}} - V_{CSN} - V_{TP1}\right)\right) \qquad 3.91$$

Input current of loser cell is

$$I_{i2} = \frac{\beta_2}{2}\left(V_{TP2} + V_{DD} - V_{CSN}\right)^2 \left(1 + \lambda\left(V_{DD} - V_{CSN} - V_{TP1}\right)\right) \qquad 3.92$$

The ratio of input currents is

$$\frac{I_{i1}}{I_{i2}} = \frac{1 + \lambda\left(V_{DD} + \sqrt{\frac{2\,I_{TAIL}}{\beta_1}} - V_{CSN} - V_{TP1}\right)}{1 + \lambda\left(V_{DD} - V_{CSN} - V_{TP1}\right)} \qquad 3.93$$

We can simply express (3.84) as follows

$$\frac{I_{i1}}{I_{i2}} = 1 + \frac{\lambda\sqrt{\frac{2\,I_{TAIL}}{\beta_1}}}{1 + \lambda\left(V_{DD} - V_{CSN} - V_{TP1}\right)} \qquad 3.94$$

Assuming that parameter λ is too small with respect to 1, (3.94) becomes:

$$\lambda << 1$$

$$\frac{I_{I1}}{I_{I2}} \approx 1 + \lambda \sqrt{\frac{2I_{TAIL}}{\beta_1}} = 1 + \Delta I \qquad 3.95$$

ΔI represents the extra input current ratio needed for correctly choosing winner-cell. With respect to (3.94), if M2i transistors` output resistance goes to infinity, ΔI goes to zero. Also output resistance decreases with the increase of frequency. Thus, ΔI increases with frequency. DC ΔI of the implemented ROW_WTA is approximately equal to 1.6m. To achieve to higher output resistance, we can use cascode, regulated cascode or self-cascode configurations[35]. Regulated cascode configuration of ROW_WTA can be seen in Figure 3.41. Cascode configuration of ROW_WTA can be seen in Figure 3.42. Output resistance of regulated cascode configuration of ROW_WTA can be seen in Figure 3.43. Output resistance of cascode configuration of ROW_WTA can be seen in Figure 3.44. Output resistance of ROW_WTA can be seen in Figure 3.45.

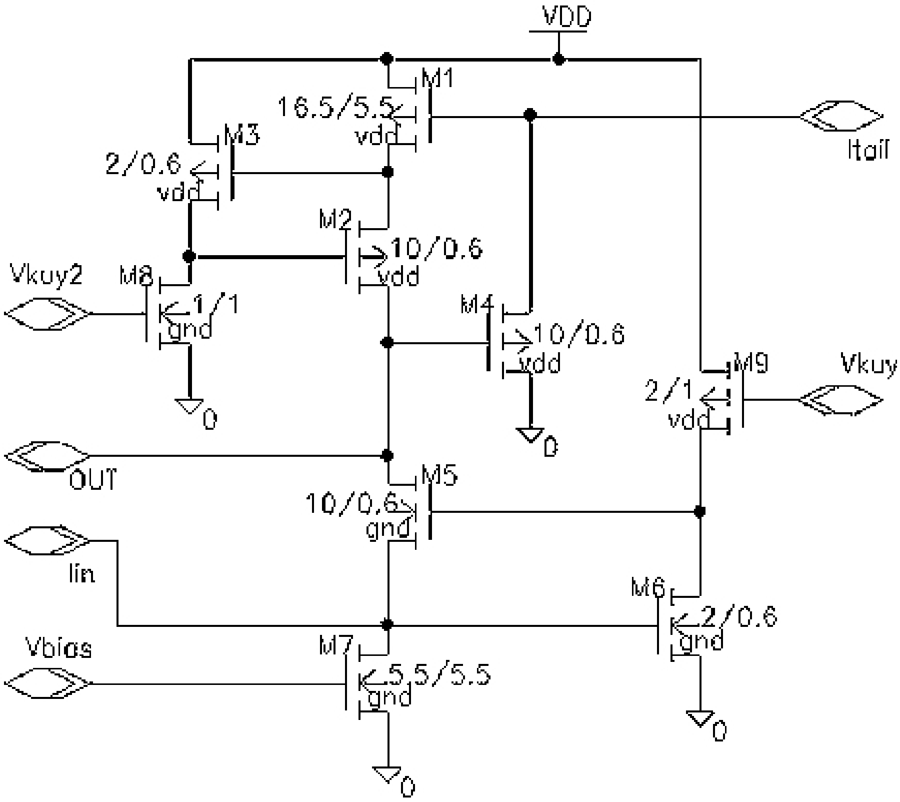

Figure 3.41 : Schematic view of regulated Cascode Configuration of ROW_WTA

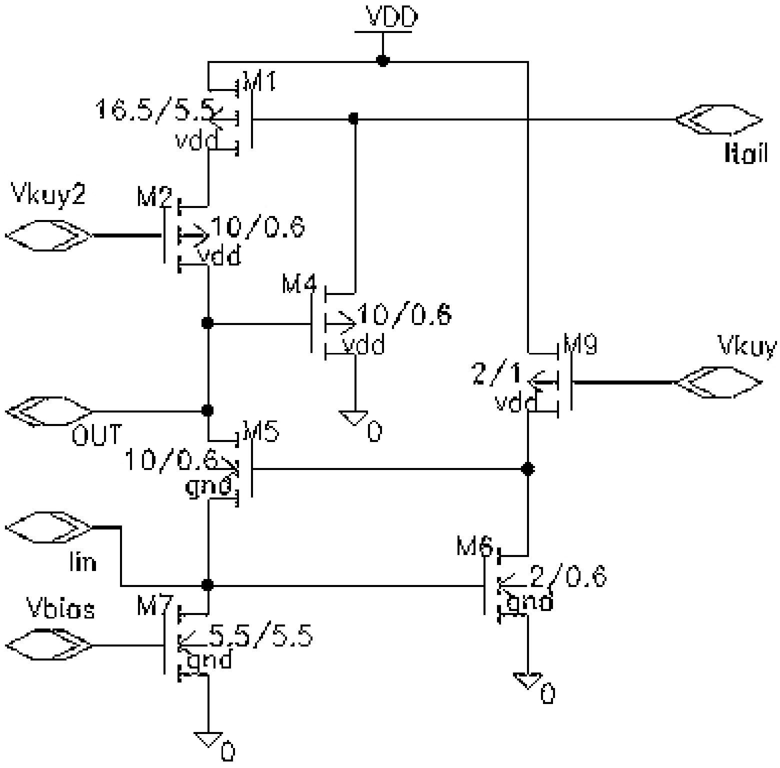

Figure 3.42 : Schematic view of Cascode Configuration of ROW_WTA.

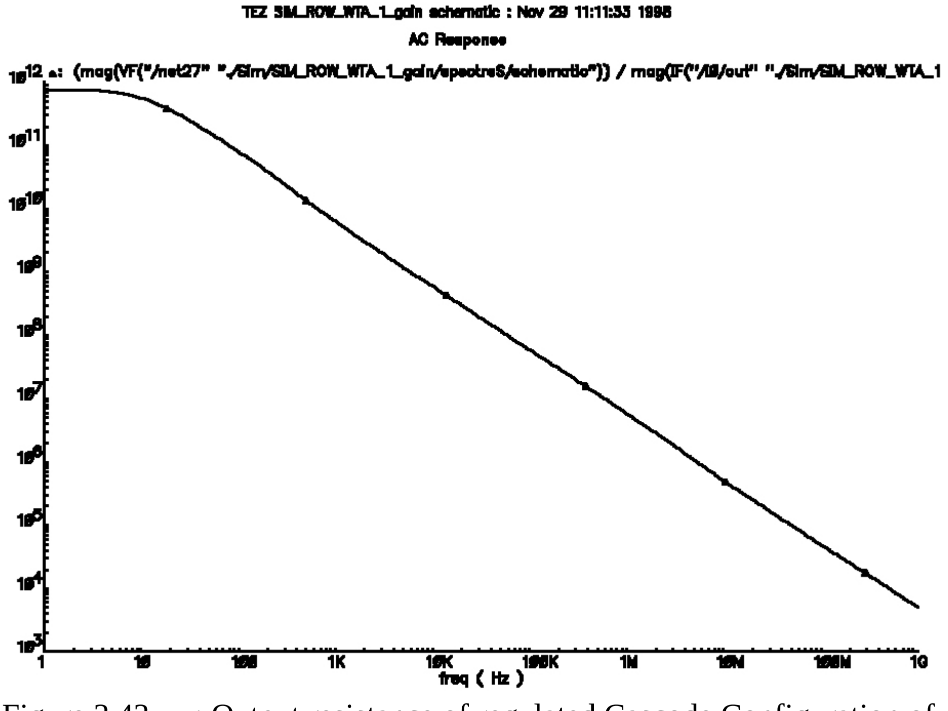

Figure 3.43 : Output resistance of regulated Cascode Configuration of ROW_WTA.

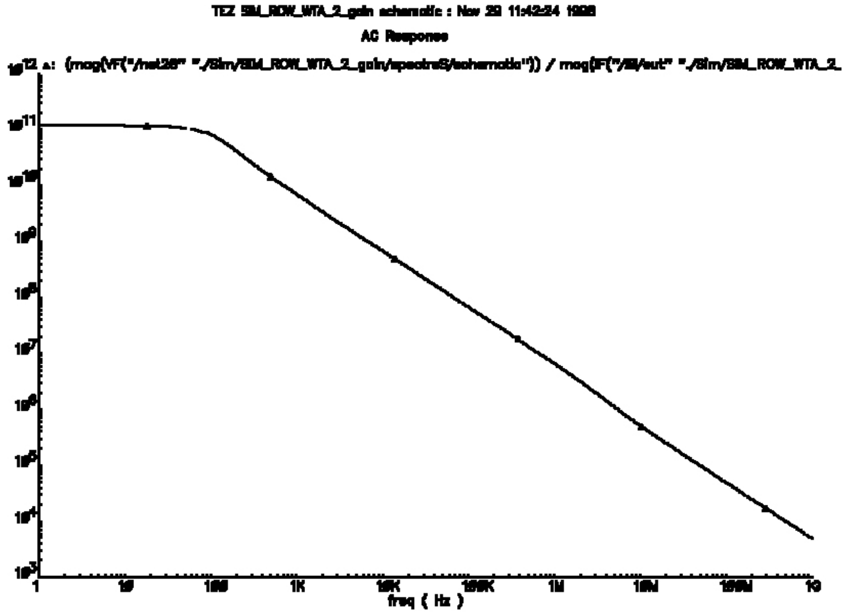

Figure 3.44 : Output resistance of Cascode Configuration of ROW_WTA.

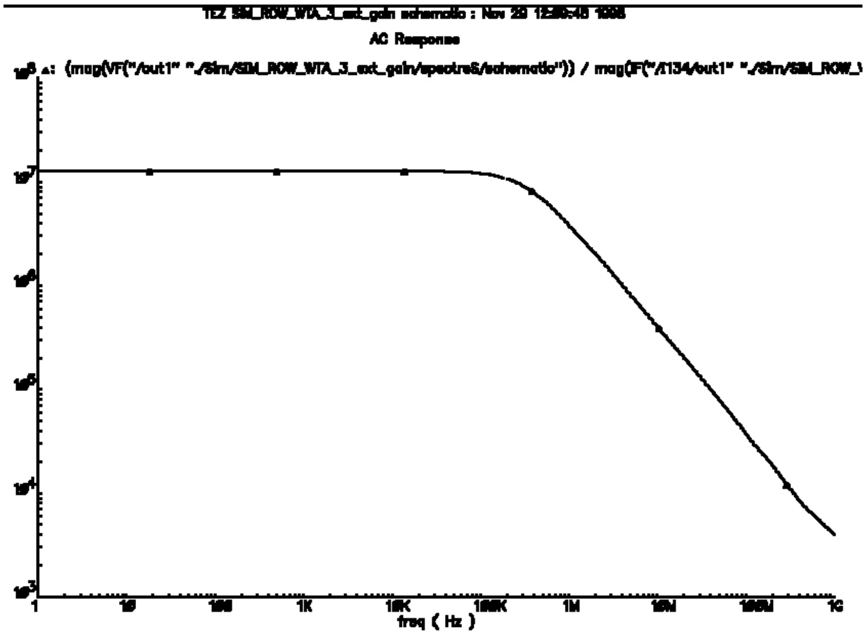

Figure 3.45 : Output resistance of ROW_WTA.

For regulated cascode configuration, we must reexamine the stability conditions because of the third pole introduced in the feedback loop. Both regulated cascode or cascode configurations require an extra biasing circuit.

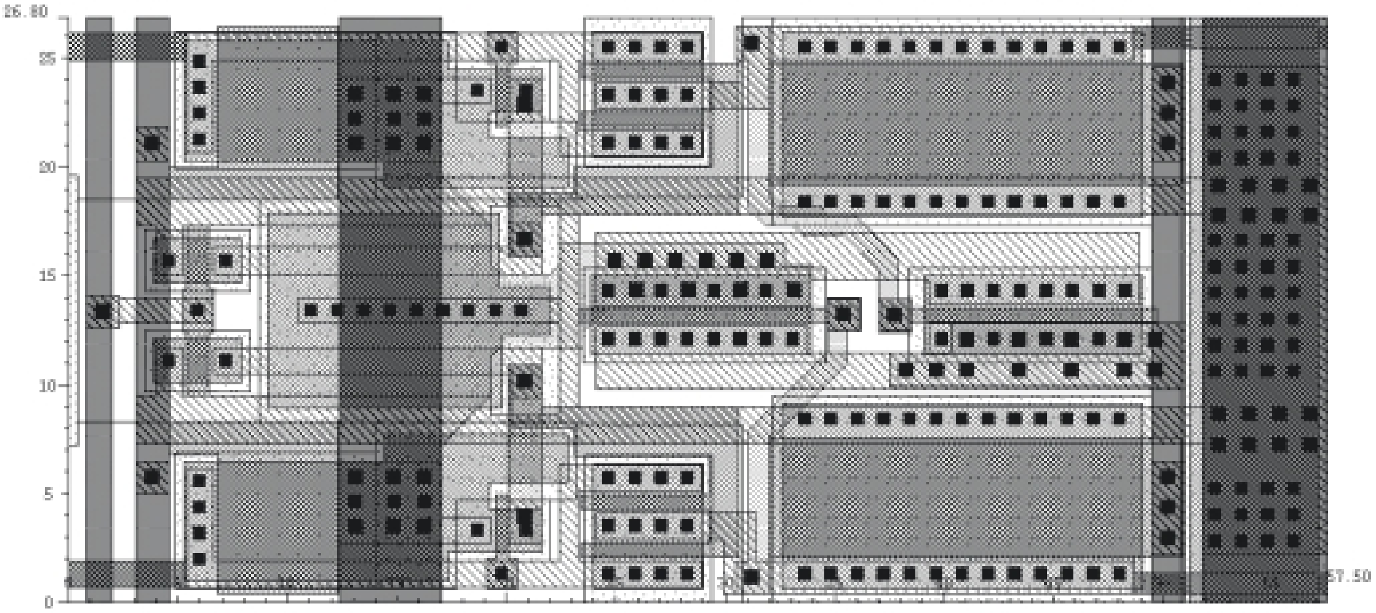

Figure 3.46 : Layout view of ROW_WTA.

Layout view of ROW_WTA can be seen in Figure 3.46. Alike the layout of the block DCELL, this layout contains two ROW_WTA blocks. The height of ROW_WTA is 26.8μm and the width of the block is 57.5μm. Thus, it consumes a silicon area of 7.705e-4 mm^2.

The block is optimized to response within the 100ns for an input current difference of 1μA. The list of transistors sizes can be found in Table 3.3.

Table 3.3 : Transistors sizes of the block ROW_WTA

Trans. Name	W (μm)	L (μm)
M1	2	0.6
M2	10	0.6
M3	16.5	5.5
M4	10	0.6
M5	5.5	5.5
M6	2	1

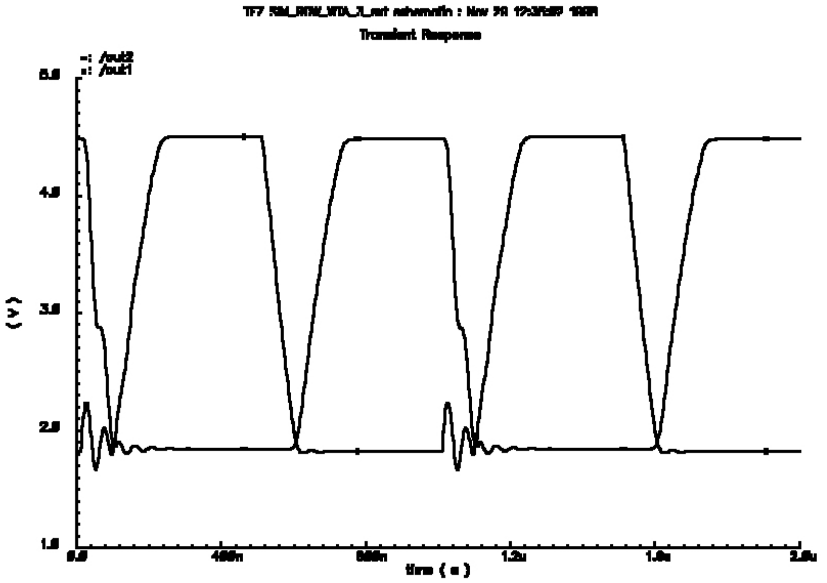

Figure 3.47 : Transient simulation result of WTA network which contain 64 ROW_WTA

Transient simulation result of WTA network which contain 64 ROW_WTA cells for 1μA input signal difference between winner-cell and others can be seen in Figure 3.47.

The power consumption of a matrix which contain NxM DCELL and M ROW_WTA can be calculated as follows:

$$P = (M(I_B + I_K) + I_{TAIL})V_{DD} \qquad 3.96$$

In (3.96), V_{DD} represents power supply voltage. I_{Tail} represents the current used for biasing node CSN in WTA network. M represents the total row count. I_B represents the constant current in ROW_AMP. I_K represents the constant current for biasing regulated cascode transistors in ROW_AMP. Although the parameter N does not appear in (3.96), it is used while determining the bias current I_B.

A behavioral AHDL code of ROW_WTA is developed for decreasing system level simulation time. Behavior code models a current differentiation function, output voltage levels and a simple maximum current selector circuit. The behavioral AHDL code can be found in Appendix G. AHDL code simulation result can be seen in Figure 3.48. The input stimulus is the same of the simulation in Figure 3.47.

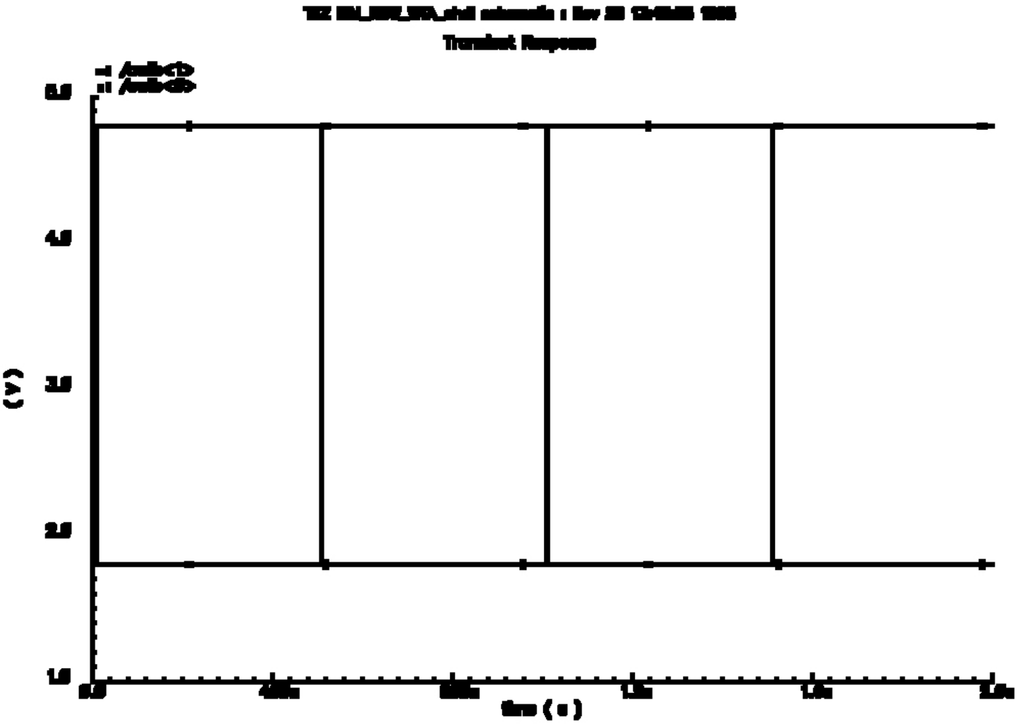

Figure 3.48 : AHDL code simulation result.