Analog VLSI Implementation of Novel Hybrid Neural Network Multiplier Architecture

S. Venkatesh, P. Cyril Prasanna Raj

[1]M.Sc. (Engg.) Student, [2]Assistant Professor and Course Manager
VLSI System Design Centre
M.S.Ramaiah School of Advanced Studies, Bangalore 560 054

Abstract

Neural networks are suitable to resolve problems where conventional resolution methods fail. The multipliers form a basic and important block in realising a neural network and this is commonly known as "Synapse". Their roles are to multiply an input current with trained digital weights. Several research attempts to implement synapses. Some use numeric implementation whereas others use analogue circuit. Each of these implementations methods has its own advantages and drawbacks. Numeric multipliers are very suitable for applications that need high accuracy and precision results. Unfortunately it occupies a considerable silicon area. Analogue synapses are efficient-silicon area and are able to operate at high frequency. However synaptic weights are badly saved in their analogue form. To combine the advantages of these two implementation methods, a mixed implementation technique provides best performances. A mixed synapse named multiplier digital-to analogue converter (MDAC) is a multiplier bloc that multiplies an analogue reference by a binary coded synaptic weight .Current steering MDAC is the aim of this work for its capability to drive considerable charges at the output of synapses

The focus of the present thesis is on the Analog implementation of hybrid multiplier architecture where in a current is multiplied along with the digital weights. The Synapse or the multiplier includes 3 subcomponents: DAC, current steering circuits, and a current mirror circuit.The design of the Synapse is carried out in CADENCE VIRTUOSO and verified.
The circuit is tested for its performance with respect to accuracy and power. The functionality of the design is verified using AVAN WAVES. The layout is realised for the same along with the RC extracted view and also the GDSII for the proposed design is extracted. The INL and DNL of the MDAC is found to be 0.5LSB and 0.8LSB respectively, The area occupied is 176.4μm2.The MDAC is monotonic.

Keywords: Neural Networks, MDAC, Synapse, AVAN Waves

NOMENCLATURE

DNL	Differential Non linearity error
DRC	Design Rule Check
DAC	Digital to Analog Converter
INL	Integral Non linearity error
L	Length of the channel
ns	nano (10^{-9}) sec
NMOS	N-channel Mosfet
mm	millimetres
MHz	Mega Hertzs (Frequency)
mW	milli (10^{-3}) Watts
MDAC	Multiplying Digital to Analog Converter
PMOS	P-Channel Mosfet
TSMC	Taiwan Semiconductor Manufacturing Company
μA	microamperes
μs	micro (10^{-6}) sec
w	Width of the channel
W	Watts
VLSI	Very Large Scale Integration

1. INTRODUCTION

The artificial neural networks are inspired by the biological learning systems, which are built of a very complex network of web of neurons which are interconnected to each other [1]. Typically a human brain consists of 10^{11} neurons each with an average of $10^3 – 10^4$ connections. The immense computing speed of the of the brain is said to be the parallel and distributed computing performed by these neurons ,the transmission of these signals in biological neurons through synapse is a very complicated chemical process wherein specific transmitter substances are released from the sending side of the synapse[2].The effect is to rise or lower the electrical potential inside the body of the receiving cell. The neuron is said to fire once the threshold is said to be reached, this is said to be the characteristic of an artificial neuron model proposed by Mcculloh and Pitts. This neuron is widely used in artificial neural networks with the modifications required.

The important design challenge to be addressed in a neural network is the efficient multiplication circuits wherein the digital weights are multiplied with the input current, the digital weights are stored in a memory cell which are applied as one of the inputs is multiplied with the incoming current, this is normally known as "Synapse". Lots of multiplier architectures were used and the most commonly used architecture is the Gilbert cell multiplier.

2. WORKING OF ANALOG MULTIPLIER

Neural networks are defined as the nonlinear function of weighted sum of signals. The p-inputs of neuron, X0, X1… Xp-1 shown in Fig. 1, are multiplied by the p-synaptic weights, W0, W1…, Wp-1. The weighted sum is then forwarded to the neuron output via a nonlinear activation function S(.). Neuron output Y is then given by **(1)**.

$$Y = S\left(\sum_{i=0}^{p-1} X_i W_i\right)$$

(1)

From the previous equation, multiplication operation of XiWi and the addition ΣiXiWi are the two arithmetic operations are performed by the neuron. The implementation of the addition is easy if outputs of the synapse are currents.

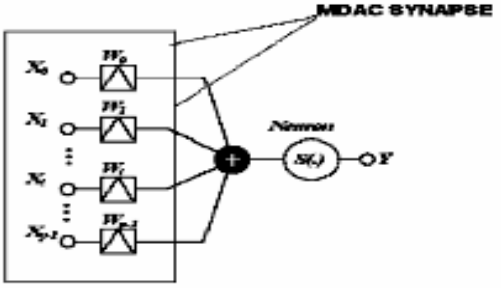

Fig. 1 Neuron Cell [1]

It is performed when synapse outputs are connected together according to Kirchhoff's Current law (KCL). The difficulty lies in the implementation of the multiplication operation. The MDAC synapse designed performs the multiplication of the analog input with the digital weights represented as W1,W1 to Wp-1 , The analog input is in the form of the current and is multiplied accordingly with the value of the Digital weights represented by D3 to D0 and the current as Iin, the current Iin is given through the current mirror circuit. A cascoded current mirror circuit for the higher output impedance, so that it can drive large circuits for the same current values, the weighted current steering approach along with the current mirror circuit acts as the input for the DAC[3][4][5]. And also other constant multiplication factor can also be added with this approach.

3. SYSTEM MODEL

In this section a MDAC synapse is designed and demonstrated for different values of input current and the digital weights. The hybrid multiplier is designed and checked for its performance and is carried out from schematic till the GDSII for the new multiplier architecture is also retrieved. The power and the area for the MDAC architecture is reported along with its accuracy. The proposed MDAC architecture has a NMOS transistors connected in the form of R – βR, wherein a control is given to the circuit to either pass the outputs directly or through the current mirror circuit. The proposed MDAC has a better accuracy.
It has two main blocks:

a) Weighted Current Steering Circuit
b) MDAC Architecture

3.1 Weighted Current Steering Circuit

The current mirror used here is the folded cascode current mirror. The main purpose of this current mirror here is to replicate the reference current irrespective of the load[6]. The designed weighted current steering circuit produces current of 32uA for an input reference current of 16 uA, and this is achieved by doubling the widths of the successive transistors. This is based on the current equation of the MOS transistors in saturation region which states current is directly proportional to the width. This current is given as the input for the MDAC. The weighted current steering circuit is shown in Fig.2.

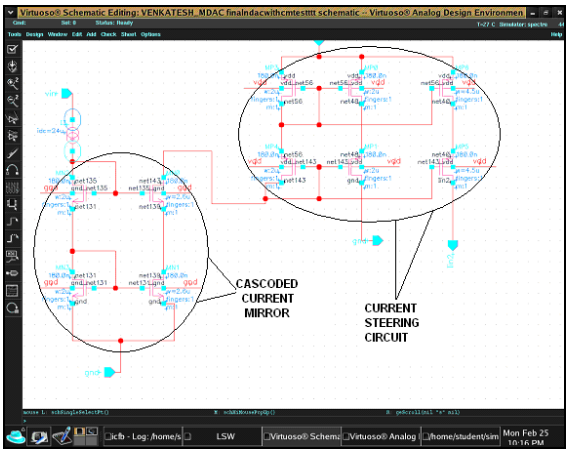

Fig. 2 Weighted Current Steering Circuit Schematic from Virtuoso

3.2 MDAC Architecture

This DAC basically functions as a Digital to Analog converter with an additional feature of the memory attached to it. The MDAC functions with respect to the equation (2).

$$IOUT = Din * In \quad (2)$$

Where Din is the digital weights inputs

$$Din = (1)^{Ssign} \sum_{I=0}^{I=3} Si/2.3^{i} \quad (3)$$

The proposed MDAC works on the principle of R-βR ladder network shown in Fig.3 is modified from the R-2R ladder network such that a more accurate current divider circuit can be obtained under the following conditions.

$$\beta > 2(1+\varepsilon)/(1-\varepsilon)^2 \quad (4)$$

where ε denotes the error on the resistance resulted from the effect of the device mismatch[7] , for a TSMC 0.18µm CMOS technology , error range is taken to be
$5\% \leq \varepsilon \leq 5\%$.
Substituting ε = 5% into eqn 4, we get β = 2.3

In the transistor level R-βR ladder network each resistor is implemented by an NMOS transistor biased in triode region [3]. An NMOS transistor operated at triode region can be characterized by

$$Id = \mu nCoxW/L[(Vgs-Vtn)Vds-Vds^2/2 \quad (5)$$

Where Id is the drain current, μn is the mobility of electrons W the channel width, L the channel Length. Vgs is gate to source voltage Vtn is threshold voltage and Vds is drain to source voltage.
The channel resistance R of an NMOS transistor is derived as

$$Rn = (\partial Id/\partial Vds)^{-1} = 1/(\mu nCox(W/L)(Vgs-Vtn) \quad \alpha \quad (L/W) \qquad (6)$$

Apparently, the channel resistance is inversely proportional to the aspect ratios of transistors. Hence the R-βR ladder network can be easily implemented by specifying the aspect ratios of the transistors satisfying the relationship given by [7]

$$(W/L)R = 2.3(W/L)\beta R \qquad (7)$$

Where (W/L)R and (W/L)βR represent the aspect ratio of the NMOS transistors implementing the resistance R and βR respectively.

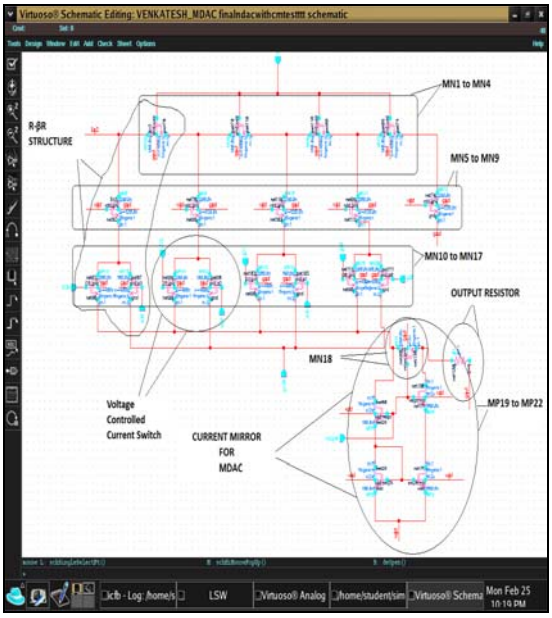

Fig. 3 MDAC Architecture Schematic from Virtuoso

3.3 Voltage controlled current Switch

A switch can be realized in several ways, an NMOS or a PMOS transistor is inherently a pass transistor that can be realized as a switch, NMOS transistor connected with a PMOS transistor in parallel to be a transmission gate which can also be realized as a switch as shown in Fig.4. However there are some shortcomings in the pass transistor switch. Those shortcomings are:

a) The inability of the PMOS to transfer good logic zero and NMOS to transfer good logic ones cause failures.

b) During the application of the control signal to its gate terminal switched between logic zero and ones ,the charge injection offset will arise which makes big impact on the current mode circuits.

In MDAC circuit, the voltage controlled current switch designed in differential topology as shown in Fig.4.

A NMOS or PMOS is a pass transistor, realizes a switch. An NMOS connected with PMOS transistor in parallel to form a transmission gate which also realizes as a switch. The ON and OFF state for a transmission gate is between drain terminal and source terminal controlled by complementary signals applied to the PMOS transistor. The transmission gate or the pass transistor does not conduct current when it acts as closed switch. It will not completely block the flow of current when it play as open switch due to the existence of leakage current.

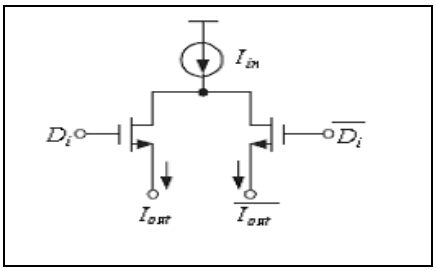

Fig.4 Voltage Controlled Current Switch [6]

3.4 Operation of Voltage Controlled Current Switch

Voltage controlled current switch operates in the manner of If D='1' then, Iout = Iin, and Ioutb = 0 Else Iout = 0 and Ioutb = 1

That is the current is switched completely from one pass transistor to another according to the voltage level of the Weight signal "D".

The current mirror circuit with an input sign acts as a memory wherein if the voltage level is 1.8 or high the values pass to the output directly, if the current mirror is activated the current is stored here. When the input of the current mirror is reversed then the output current from the current mirror circuit, this also acts for negative weights wherein the current is made to flow into the PMOS current mirror circuit. The length and the widths of the transistors are shown in Table.1.

Table 1 Length and Widths of Transistors for MDAC

Transistors	Widths for 180 nm technology with length equals 180nm
MN1 to MN4	400nm
MN5 to MN9	520nm
M10 to M18	400nm
MP19 to MP20	2μm
MP21 to MP22	2.8μm

The transistors which form the R- βR structure have widths of 400nm, 520nm. The voltage controlled current switch has width of 400n each. The current mirror used has two top PMOS transistors of widths 2μm each and bottom 2 transistors have widths of 2.8 μm each. The length used is 180nm.

From the size of the widths the total comes to 920nm for the the R-βR realization of the circuit The transistors MP19 to MP22 acts as a current mirror circuit which has an input of VSIGN which is used as the select input for the output to pass through the circuit directly to the output or the output can be delayed for a while by passing through the current mirror, negative weights can also be realized using this input by changing the input of the VSIGN input [8].

3.5 Characteristics for MDAC Architecture

a) Resolution

The resolution is defined as the number of unique analog levels corresponding to different digital words. Thus an N-Bit resolution implies that converter can resolve 2^n distinct analog levels, resolution doesn't require to be an indication of the accuracy of the converter, It refers to the number of inputs.

b) Offset and gain error

The offset error, Eoff which is defined as the output that occurs for the input code of zero. Mathematically, Eoff = output/Vlsb|0000

The gain error is defined to be the difference at the full scale value between the ideal and actual curves when the offset error is reduced to zero, for D/A converter the gain error is defined as the difference at the full scale value between the ideal and actual curves. The error gain is given by [9][10]

$E(gain) = \{[Vout/Vlsb]|1...1-[Vout/Vlsb]|0\ldots0\}-(2^n-1)$ (6)

c) Accuracy

The absolute accuracy of a converter is defined to be the difference between the expected and actual transfer responses. The absolute accuracy includes the offset, gain, and linearity errors.

d) Monotonicity

The monotonicity of a converter is one in which the output increases as the input increases, if the maximum DNL error is less than 1LSB,however many circuits have maximum INL more than 1 LSB then the circuit is said to be monotonic, similarly a circuit is said to be monotonic if the INL is less than 0.5LSB.

e) Integral Non linearity error [11]

This is defined as the maximum deviation from the expected ideal response. The INL for the proposed architecture 0.5 LSB

f) Differential Non linearity error

This is defined as the maximum deviation from 1 LSB. The DNL for the proposed architecture is 0.8LSB [12].

4. ANALYSIS OF RESULTS

MDAC proposed architecture results are shown below. For different values of input weights and the reference current given as 16µA.

The transient response on the MDAC is performed wherein the weighted input is given as 1000 and the expected output is calculated to be 16µA. This is clearly represented in the Fig.6, wherein the x axis represents the time and the y axis represents the current in terms of micro amperes.

Table.2 gives the table of the different values of the output currents for the different digital inputs which are known as the weights, the digital weights is changed from 0000 to 1111 for this 4 bit architecture, the D0 is the MSB and the D3 is the LSB. The differential error is also calculated and tabulated in the Table.2.

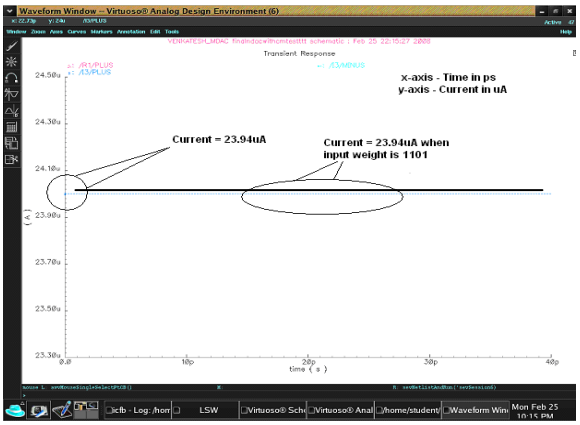

Fig. 5 Transient Response for D3, D2, D1, D0=1, 1, 0, 1

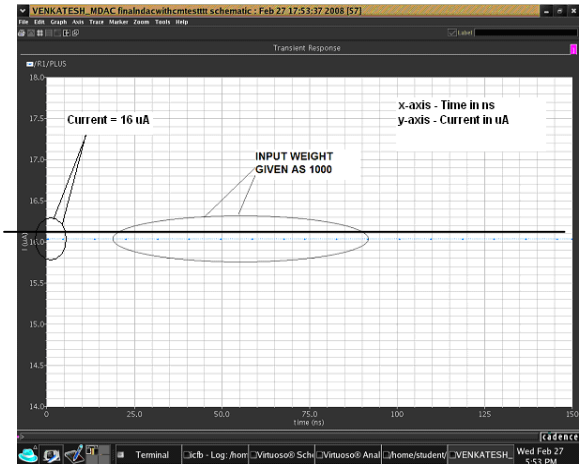

Fig. 6 Transient Response for D3, D2, D1, D0 =1, 0, 0, 0

5. LAYOUTS FOR MDAC

MDAC layout is designed using Virtuoso tool. The design is developed on 0.18µ technology. The MDAC layout is shown as in Fig.7. Layout is designed using Virtuoso. After designing layout first step is design rule check. When there are no DRC errors the design is checked for Layout versus Schematic check i.e. the layout is verified w.r.t the schematic whether the connections made in the layout are same and correct compared to schematic. Here the fingering concept is used.
The size of the transistor increase as the height of the transistor also increases. To fit the transistors in the defined area, fingering concept is used. After LVS compare warnings and extract warnings are removed and the RC extraction is carried out.

Table 2 Comparison of Theoretical, Practical Outputs with the Error for Proposed Architecture

Digital Inputs D0,D1,D2,D3	Theoretical Outputs in Amps	Practical Outputs in Amps	Error
0000	0	2.3e-12	2.3e-12
0001	1.3µ	1.7 µ	0.4
0010	3.02µ	3.6 µ	0.5
0011	4.3µ	5.2 µ	0.9
0100	6.9µ	7.5 µ	0.6
0101	8.25µ	8.7 µ	0.4
0110	9.96µ	10.5 µ	0.5
0111	11.3µ	12.1 µ	0.8
1000	16µ	16 µ	0
1001	17.3µ	16.7 µ	0.6
1010	19.02µ	18.8µ	0.1
1011	20.3µ	20µ	0.2
1100	22.9µ	23µ	0.1
1101	23.8µ	23.9µ	0.1
1110	25.9 µ	26.8µ	0.9
1111	28.2µ	28.6µ	0.4

The GDSII extracted view of the proposed MDAC is as shown in the Fig.8. This is performed using the CADENCE VIRTUOSO tool, The GDS II file extracted represents or shows that the proposed MDAC architecture is physically realizable in the form of a chip, Typically extraction from the layout is carried after the post layout simulations wherein the circuit is checked for its functionality after the interconnect resistance and capacitances are added.

6. CONCLUSION

A MDAC is realized in the 0.18µ technology with the PMOS transistors as the main component in the form of R-2R network, whereas the same is proposed using NMOS transistors with R-βR network wherein the errors recorded were low compared to the R-2R network architecture. The comparisons are tabulated and the layout and RC extractions for the R-βR MDAC structure is drawn and the GDSII also being extracted.

a) The multiplier proposed has the gain error of 0.5LSB offset error of 2.2 fA the DNL and INL are 0.8LSB and 0.5LSBs respectively.

b) The MDAC designed is checked for the functionality with the existing architecture for the functionality.

c) The proposed architecture follows the R-βR for digital to analog conversion.

d) The input for the DAC is given by the weighted current steering circuit for better accuracy, the comparative output data is tabulated and the error is calculated.

e) The proposed architecture has less error which is noted.

f) The layout for the MDAC, The RC extraction and the GDSII for the MDAC structure is realized.

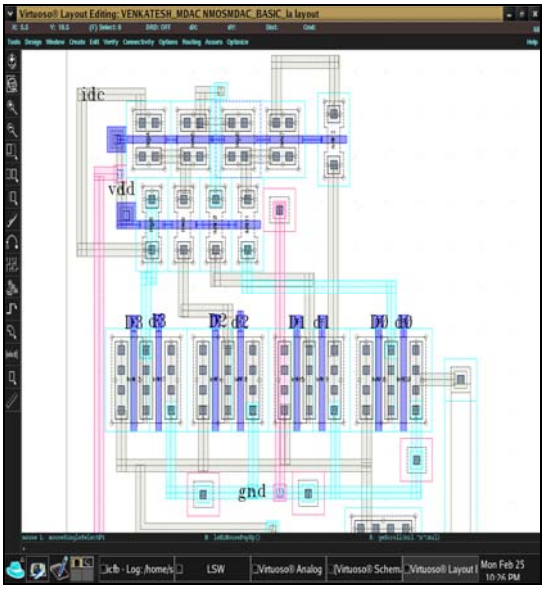

Fig. 7 Layout for MDAC Architecture

REFERENCES

[1] Koosh, V.F.and Goodman, R., "*VLSI neural network with digital weights and analog multipliers*", IEEE International Symposium on Circuits and Systems, vol.3, no.1, pp. (233-236), 6th to 9th May, 2001.

[2] Ryan J. Kier, Reid R. Harrison, and Randall D. Beer, "*An MDAC Synapse for, Analog Neural Networks*", IEEE proc.Circuits and Systems, vol.5, no.1, pp. (752-755), 23rd-26 May, 2004.

[3] David A. John and Ken Martin, "*Analog Integrated Circuit Design*" ISBN 9814-12-647-0, John Wiley & Sons.

[4] Vittoz, E.A.and Arreguit,X., "*Linear networks based on transistors*", IEEE proc. in Transistors Circuits, vol.29, no.3, pp. (297-299), 4th Febrauary,1993.

[5] Chunhong Chen and Zheng Li, "*A low-Power CMOS Analog Multiplier*", IEEE Transactions on Circuits and Systems –II, vol.53, no.2, February,2006.

[6] Chuen-Yau Chen, Chi-Jung Cheng and Chien-Cheng Yu, "*Design of Current-Mode Digital-to-Analog Converter in Hybrid Architecture*", IEEE-NEWCAS Conference, Vol.1, no.1, pp.(231-234), 19th-22nd June, 2005.

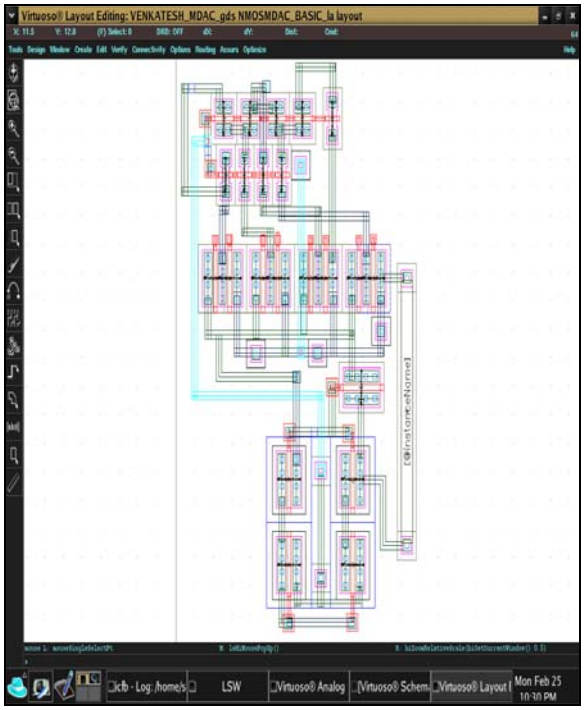

Fig. 8 GDSII Extracted Layout for MDAC Architecture

[7] Al-Nsour,M. and Abdel-Aty-Zhody,H., "*ANN digitally programmable analog synapse*", IEEE Circuits and Systems, vol.1, no.1, pp.(489-492).

[8] Shuo-Yuan Hsiao and Chung-Yu Wu, "*A 1.2 V CMOS Four-Quadrant Analog Multiplier*", IEEE Interational Symposium on Circuits and Systems, vol.1, no.1, pp.(241-244), 9-12th, June, 1997.

[9] George A. Hadgis and P.R.Mukund, "*A Novel CMOS Monolythic Analog Multiplier with Wide Input Dynamic Range*", 8th International Conference on VLSI design, pp. (310-314), January, 1995.

[10] J.Ramirez Angulo, R.G.Carvajal and J. Martinez Heredia, "*1.4V Supply, Wide Swing, High Frequency CMOS Analogue Multiplier with High Current Efficiency*", IEEE International Symposium on Circuits and Systems, pp. (533-536), 28-31st, May, 2000, Geneva, Switzerland.

[11] Gianluca Colli and Montecchi,F., "*Low Voltage Low Power CMOS Four-Quadrant Analog Multiplier for Neural Network Applications*", vol.1, no.1, pp. (496-499), 1996.

[12] Gowda, S.M., Sheu, B.J., Choi, J., Hwang, C.-G. and Cable, J.S., "*Design and characterization of analog VLSI neural network modules*", IEEE journal for Solid-State Circuits, vol28, no.3, pp. (301-313), March, 1993.