Implementations of OTAs for Modeling Neurons in the Artificial Neural Network

by

Tawfik Imtawbil

A thesis
presented to the University of Manitoba
in fulfillment of the
thesis requirement for the degree of
Master of Science (EE)

Electrical and Computer Engineering

Winnipeg, Canada 1990

ISBN 0-315-71944-3

IMPLEMENTATIONS OF OTAs FOR MODELING NEURONS IN THE ARTIFICIAL NEURAL NETWORK

BY

TAWFIK IMTAWBIL

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Abstract

This thesis studies the use of Operational Transconductance Amplifiers to model neurons in Artificial Neural Networks. The intrinsic properties of the OTAs is studied for VLSI implementations. Two types of OTAs are investigated for multiple input capabilities; the common and the cascaded. Two test circuits have been designed that contain 4 neurons modeled by the two types of OTAs and have a digital control circuit for updating and refreshing the weights.

ACKNOWLEDGMENT

The author wishes to express his sincere thanks to Professor H.K. Kim for his supervision and guidance throughout the thesis. Grateful acknowledgment is also due to the guys of the VLSI laboratory, who provided me with assistance with the design tools and software packages.

Contents

1 Introduction **1**

1.1 Neurobiological systems 3

1.2 Artificial neurons 5

1.3 OTA 6

1.4 Objective 7

2 Neurons Modeling With Typical OTAs **9**

2.1 The typical OTA operation 9

2.2 Output Stage of the OTA 11

2.3 Modeling synapses 14

2.4 Control of the synapse's weight 16

2.5 Multiple input structure 18

2.6 Evaluation of the common OTAs for modeling neurons 20

3 Neurons Modeling With Cascade Staged OTAs **22**

3.1 Modeling a synapse 22

3.1.1 Operation of the synapse 23

3.1.2 Input/Output impedance of the synapse 27
3.2 Current Mirrors 30
3.2.1 Simple current mirror 30
3.2.2 Wilson current mirror 32
3.2.3 Cascaded current mirror 32
3.3 Construction of neurons 38
3.3.1 Synaptic weight control 38
3.3.2 The output impedance 41
3.4 Evaluation of the cascaded stage OTA for modeling neurons . 42

4 Implementations of Neurons **44**
4.1 The Hopfield architecture 44
4.2 Implementations of neurons using cascade stages 46
4.2.1 The synapse model 46
4.2.2 The chip architecture 46
4.2.3 Simulations of the cross-bar arrangement 52
4.3 Implementations of neurons using common OTAs 53
4.3.1 The D/A converter used to control the weight 55
4.3.2 Simulations of the the cross-bar arrangement 58

5 Conclusions **59**
5.1 Future Work 62

A The Transistor Model used **66**

B The SPICE Parameters 68

C 8 Neuron Chip layout 71

List of Figures

1.1 Biological neurons. 3

1.2 Formal model of a biological neurons. 4

1.3 Artificial neurons. 5

1.4 Symbol for an OTA. 6

2.1 The simple OTA circuit. 10

2.2 Output current of the typical OTA as a function of differential input voltage. 12

2.3 Output current versus output voltage for a simple OTA. . . . 13

2.4 The wide range OTA. 14

2.5 The quality of the output stages of the simple OTA and wide range OTA. 15

2.6 Two ways of weight storage. 15

2.7 The multiple input OTA to model a neuron. 17

2.8 Variations of the I_o with input voltage. 18

2.9 Variations of g with V_b. 19

2.10 Simulations of one input. 19

2.11 Simulations of twenty inputs. 20

3.1 The synapse (cascade stage). 23

3.2 Variations of the output current with V_{in}. 25

3.3 Controlling the weight in the synapse. 26

3.4 (a) Small signal model, (b) The Capacitance involved in the cascade stage. 28

3.5 The output impedance of the cascade stage. 28

3.6 Synapses connected together to form multiple input 29

3.7 A simple current mirror. 30

3.8 Simulation results of the simple current mirror. 31

3.9 The Wilson current mirror. 33

3.10 Simulation results of the Wilson current mirror. 33

3.11 Cascaded current mirrors. 34

3.12 Simulation results of the cascaded current mirror ($\frac{W}{L} = 4$). . . 36

3.13 Simulation results of the cascaded current mirror ($\frac{W}{L} = 2$). . . 37

3.14 The connection of the OTAs 39

3.15 Variations of the output current with input voltage for different gain voltages. 40

3.16 Variations of transconductance with the gain voltage. 41

3.17 Comparisons between a number of synapses connected together. 42

3.18 The multiple input structure. 43

4.1 The cross-bar arrangement. 45
4.2 The Synapse model. 47
4.3 One neuron with 4 synapses. 48
4.4 The chip architecture. 49
4.5 Analog multiplexer. 49
4.6 The address decoder used to control the capacitor refreshing and updating. 50
4.7 The layout of the chip. 51
4.8 The control circuit. 54
4.9 The Synapse model. 54
4.10 The neuron and synaptic connections. 55
4.11 D/A 3 bit converter. 56
4.12 The layout of the chip. 57

A.1 The transistor small-signal model used. 67

C.1 8 Neuron chip. 72

Chapter 1

Introduction

In some special computational problems such as image processing, familiarity recognition, categorization and word identification, the biological systems out-perform an aggregation of super computers. Digital computer architecture is not amenable for analyzing fuzzy and ill-defined problems like face recognition.

Some differences [3] between computer systems and the neurobiological systems are observed below.

- The neurobiological systems are made of *wet* 3-dimensional cells , the others is made of inorganic flat 2-dimensional chips.
- The neurobiological systems are powered by biochemistry whereas the other are powered by rectifiers and transformers.
- The neurobiological systems have 100 millivolt level nerve impulses lasting nearly a millisecond while the others have 5 volt signal levels

switching at nanosecond intervals.

- The neurobiological systems communicates mostly in an analog manner in real time, while the other computes in a digital mode with hierarchical processing.

- The neurobiological systems are fault tolerant, so the loss of few cells will not affect the brain's performance. On the other hand the loss of 1 transistor might cause the disfunctionality of the whole conventional computer.

Those observations suggest the usefulness of analog VLSI (Very Large Scale Integrated circuits) technology in microelectronic simulation and emulation of neurobiological systems. An entirely new type of analog computer (using VLSI circuits) is sought to investigate neurobiological systems for further understanding. It will be capable of performing some tasks presently achieved only by the neurobiological systems. This type of analog processor is called an *Artificial Neural Network* (ANN). Analog VLSI networks are very attractive architectures to build ANNs for simulation of neurobiological systems consisting of a massive aggregates of regularly spaced neurons and synapses. It is estimated that the human brain consists of about 100 billion neurons and each neuron is typically connected to approximately 10,000 other neurons [19]. A neural system has two types of amplifiers, namely normal (excitatory) and inverted (inhibitory). It is also required that the ANNs be capable

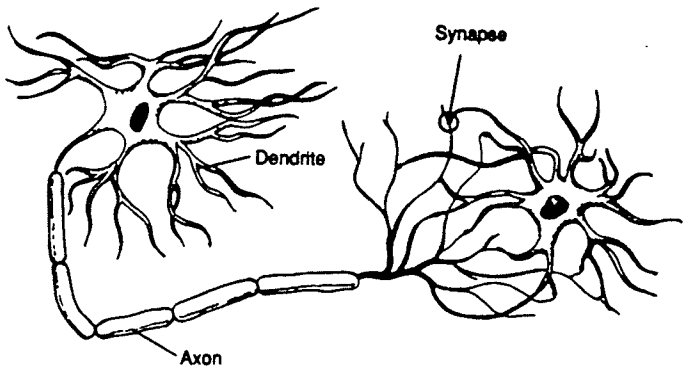

Figure 1.1: Biological neurons.

of implementing feedback. By using VLSI technology, a parallel architecture for real time computations can be accommodated. Processors and sensors can be integrated on the same chip. The biological systems have more layers available for wiring, as compared to 2-dimensional chips.The speed advantage of the chip may be used to offset the spatial problem by time-multiplexing the signals on a single wire.

1.1 Neurobiological systems

Neurobiological systems have inspired interest in ANNs. Fig. 1.1 shows two biological neurons in synaptic contact [2]. The system is formed of:

- The cell body, the large round central body of the neuron approximately 100 μm in diameter.

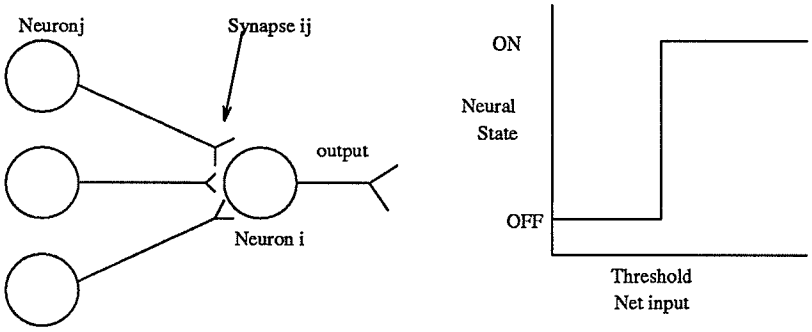

Figure 1.2: Formal model of a biological neurons.

- The axon, which is attached to the soma and electrically produces the pulses which are emitted by the neuron.
- Electrical signals travel through the axon to other neurons. These signals are transmitted to other neurons across a narrow gap between the cell membranes (Synapse).
- The synapse controls the conductance of the membrane.
- The dendrites are electrically passive and receive inputs from other neurons by the synaptic contacts.

Fig. 1.2 shows the formal model of a biological neuron. It sums inputs from other neurons and turns on (fires) by sending a series of voltage spikes down the axon when the net input is greater than some threshold.

The synapse conductance, which transmits these voltage spikes from the sending neuron to the receiving neuron [1], increases if the sending neuron

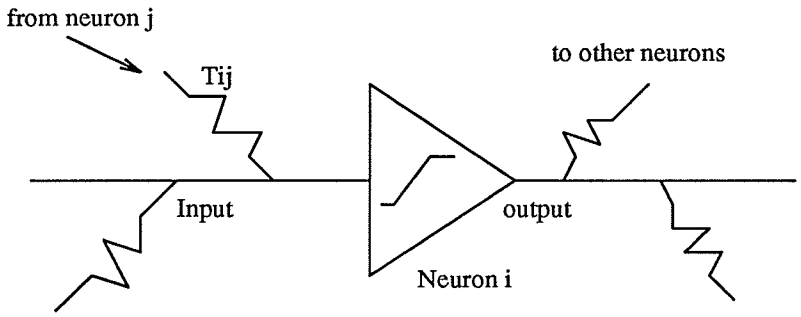

Figure 1.3: Artificial neurons.

repeatedly causes the receiving neuron to fire.

The human cerebral cortex is comprised of about 10^{11} neurons, with each one having roughly 1000 dendrites, that form some 10^{14} synapses, (the system operates at frequency $\approx 100Hz$), and functions at the rate of about 10,000,000 billion interconnections/sec. [2]. The human brain weighs approximately 1.5 Kgs, covers an area of $0.15m^2$ and 2 mm in thickness.

1.2 Artificial neurons

Artificial neurons are similar to the biological neurons, the simplest representation of artificial neuron is shown in Fig. 1.3. Neurons become processing elements, the axon and dendrites become wires, and synapses become variable resistors carrying weighted inputs that represent data from other neurons. An individual neuron does only very simple computations, but together they become very powerful parallel processing machines, Fig. 1.3 shows a processing node (amplifier) interconnected to other neurons by resistors. The output

of the node is a sigmoidal function and neuron (i) gets input from neuron (j) through a resistor with the conductance T_{ij} (known as weight)[8].

$$V_{outi} = f(\sum_{j=1}^{N} V_{outj} T_{ij}) \tag{1.1}$$

The advantage of analog networks is that a single resistor can perform a multiplication using ohm's law and currents summed according to Kirchhoff's law. Therefore, an analog circuit that computes sums of products can be built much more compactly than a digital circuit which requires multipliers and adders that take a large space of the silicon wafer.

1.3 OTA

The Operational Transconductance Amplifier (OTA) is a natural gain device of large gain bandwidth. Basically it is a voltage controlled current source and easily amenable to large scale integration. It has a transfer characteristic,

$$I_{out} = g(V_1 - V_2) \tag{1.2}$$

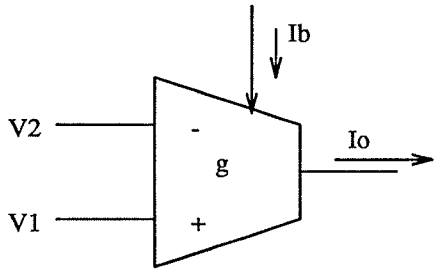

Figure 1.4: Symbol for an OTA.

where g is the transconductance controlled by bias current I_b. The common symbol used for OTA is shown in Fig. 1.4. The characteristics of an ideal OTA are similar to those of an op-amp except that the OTA has very high output impedance. Due to the fact that the output impedance is high, the output signal of the OTA is best described in terms of current that is proportional to the difference between the input voltages.

$$g = \frac{dI_{out}}{dV_{in}} \tag{1.3}$$

The unique features of the OTAs are the current output and the external control of the transconductance or gain. They are general purpose amplifiers suitable for a wide range of applications. The other important feature is that it is suitable for monolithic design. In spite of those attractive features, OTAs have not yet found wide spread applications. Due to their limited differential input linear range $\simeq$30mV, the output current saturates making the device nonlinear at small voltage values.

1.4 Objective

The objective of this research is to reevaluate and analyze OTAs for use in modeling the multiple input neuron in artificial neural networks. The intrinsic properties of the OTA will be closely investigated for possible dense arrays to generate a large number of subcircuits in a cross-bar arrangement to be useful for VLSI implementations. Two types of OTAs that are under

investigation: namely the typical differential OTAs and the cascaded OTAs.

A new way of modeling neurons using the typical differential OTA with variable weight and small synaptic sizes that are suitable for VLSI implementations were suggested and investigated. A chip that contained 4 fully connected neurons and a weight control circuit was designed. The CAD tools used for the VLSI design was Electric 4.05B and simulations were carried out using SPICE3B and HSPICE.

Multiple input neurons using cascaded stages OTAs was reevaluated, analyzed and tested. Synaptic sizes that are suitable for VLSI implementations were also suggested. Another chip that contained 4 fully connected neurons and weight control circuit was fabricated by Canadian Microelectronics Corporation's fabrication service. The CAD tools and simulators were the same as above.

The possibility of implementing a number of controlled inputs using both types of OTAs was developed and tested within the context of OTAs.

Chapter 2

Neurons Modeling With Typical OTAs

OTAs are very useful building blocks for analog networks since amplifiers may be designed in CMOS (Complementary Metal Oxide Semiconductors) technology. This makes OTAs very suitable for VLSI implementations.

2.1 The typical OTA operation

The simple differential amplifier (or OTA) in CMOS is very compatible with analog VLSI technology. The main objective of the OTA is to *amplify the difference between the input signals.* The simple OTA circuit consists of 5 CMOS transistors as shown in Fig. 2.1.

T_1 and T_2 form the differential pair, and are designed to have the same threshold and size. T_3 and T_4 are of the same size and form the pMOS current mirror. They are used to generate an output current that represents the difference between the input voltages. T_5 is the current sink of the differential

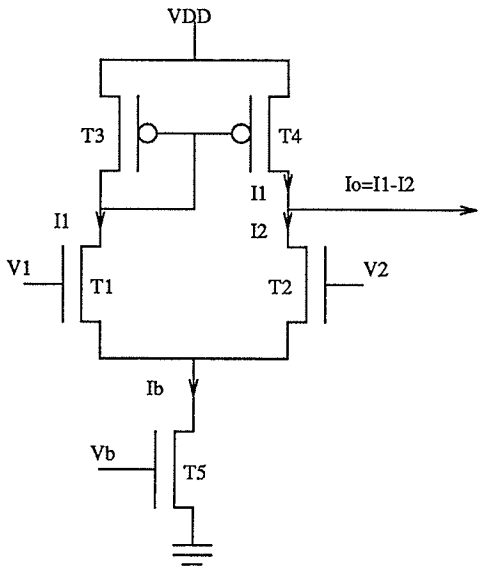

Figure 2.1: The simple OTA circuit.

pair and it controls the gain (weight) of the OTA.

The output current is derived as follows [13]:

$$I_b = I_1 + I_2 \tag{2.1}$$

$$V_{dif} = V_1 - V_2 = \sqrt{\frac{I_1}{\beta}} - \sqrt{\frac{I_2}{\beta}} \tag{2.2}$$

Where $\beta = \frac{\mu C_{ox} W}{2L}$

μ = mobility, C_{ox} = Oxide thickness, W = transistor width, L = transistor length,

Substituting (2.1) into (2.2) yields

$$I_1 = \frac{I_b}{2} + \frac{I_b}{2}\sqrt{\frac{2\beta V_{dif}^2}{I_b} - \frac{\beta^2 V_{dif}^4}{I_b^2}} \tag{2.3}$$

$$I_2 = \frac{I_b}{2} - \frac{I_b}{2}\sqrt{\frac{2\beta V_{dif}^2}{I_b} - \frac{\beta^2 V_{dif}^4}{I_b^2}} \tag{2.4}$$

$$I_o = I_1 - I_2 \tag{2.5}$$

$$I_o = I_b\sqrt{\frac{2\beta V_{dif}^2}{I_b} - \frac{\beta^2 V_{dif}^4}{I_b^2}} \tag{2.6}$$

Simulations of this circuit using Spice parameters [20] are shown in Fig. 2.2, where the output current saturates when there is enough voltage difference between the two inputs. The gain (transconductance) is the slope of the output current.

The transistor sizes are as below:

Transistor	$W(\mu m)$	$L(\mu m)$
T1	9	7
T2	9	7
T3	7	7
T4	7	7
T5	7	19

2.2 Output Stage of the OTA

The simple OTA has a limited output voltage range in which it operates correctly [3]. There is an upper limit near the supply voltage VDD where the output current decreases rapidly and a lower limit near V_b where the current increases rapidly, as shown in Fig. 2.3.

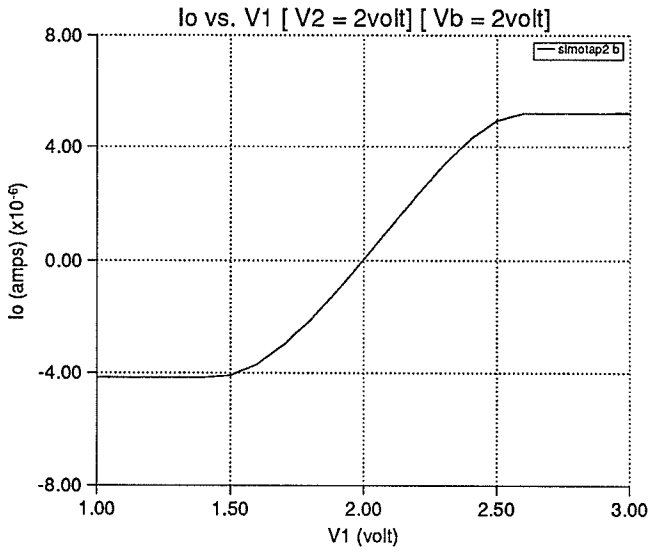

Figure 2.2: Output current of the typical OTA as a function of differential input voltage.

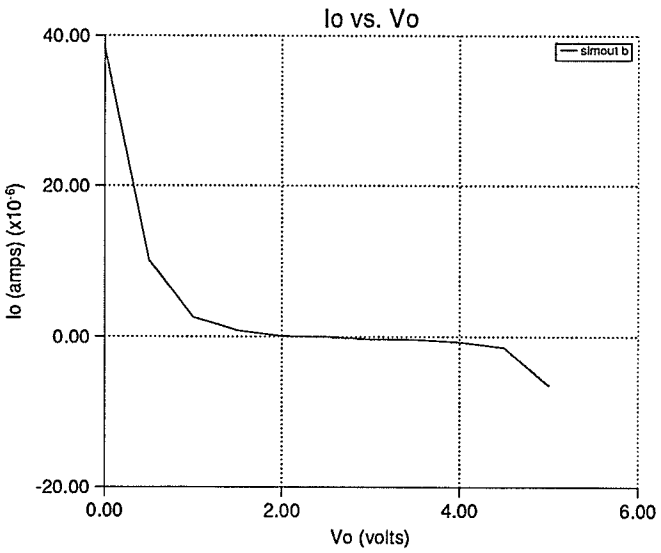

Figure 2.3: Output current versus output voltage for a simple OTA.

To improve the output stage, the wide range OTA made of 9 CMOS transistors as shown in Fig. 2.4 is used. Fig. 2.5 compares the quality of the output stage of the simple OTA and the wide range OTA. One can observe that the output current of the wide range OTA does not vary a great deal with the output voltage. The input and output voltages can vary from $0 \rightarrow VDD$. The output transistors of the wide range OTA yield the gain higher than those of the simple OTA.

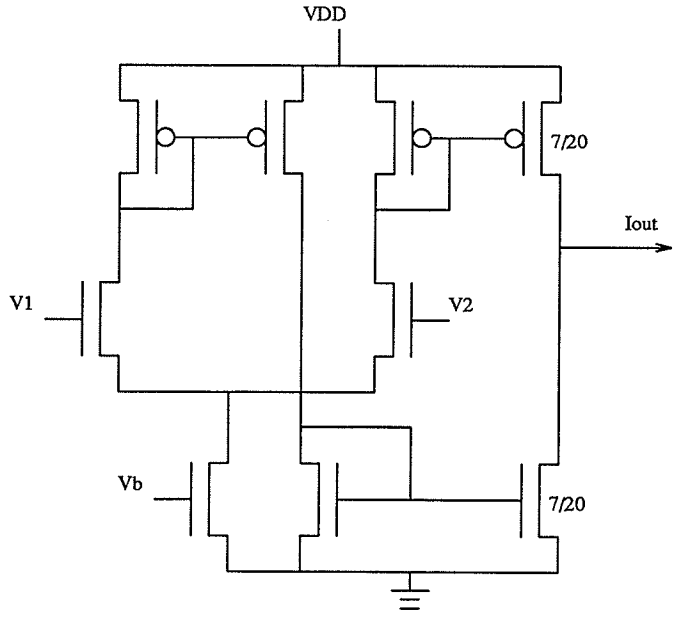

Figure 2.4: The wide range OTA.

2.3 Modeling synapses

The synapse can be modeled by a 3 transistor multiplier. The differential pair is used to form the inputs while the weight is controlled by the current sink transistor.

One method is to use V_1 and V_2 as inhibitory and excitatory inputs, where one of these inputs is grounded or used as a reference. Therefore each input would be inhibitory or excitatory only. The weight is controlled by V_b and the charge is stored on a capacitor as shown in Fig. 2.6(a).

Another method is to use V_b as the input while storing weights on the two capacitors [8]. The input then is multiplied by the voltage difference on the

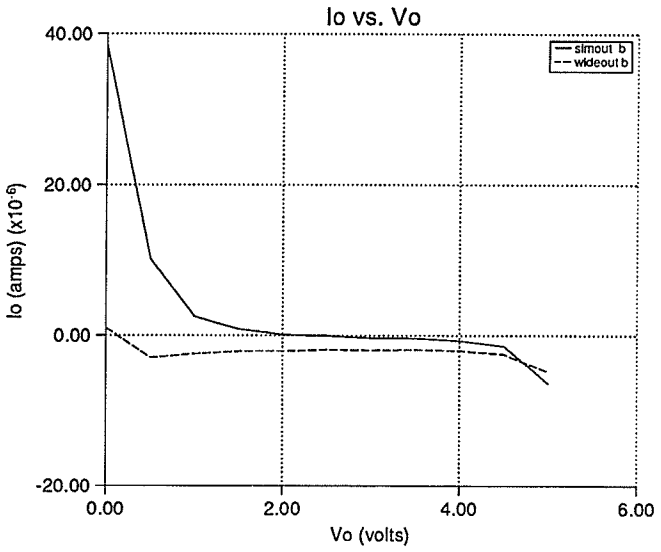

Figure 2.5: The quality of the output stages of the simple OTA and wide range OTA.

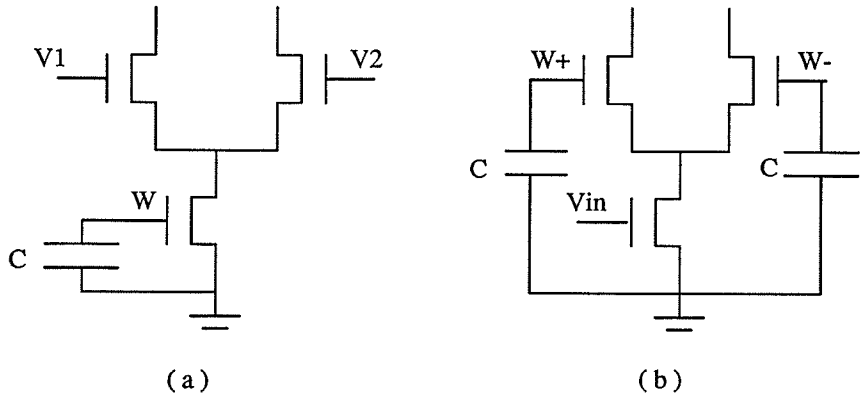

Figure 2.6: Two ways of weight storage.

two capacitors as shown in Fig. 2.6(b). This method requires 2 capacitors. Several of these synapses can be connected together to form a multiple input neuron as shown in Fig. 2.7. The pMOS and nMOS current mirrors perform the addition and subtraction of the excitatory and inhibitory inputs, while the gain (or weight) of each input (or Synapse) is individually and continuously controlled by V_b.

$$I_o = \sum_{i=1}^{n} I_{b_i} \sqrt{\frac{2\beta V_{dif_i}^2}{I_{b_i}} - \frac{\beta^2 V_{dif_i}^4}{I_{b_i}^2}} \tag{2.7}$$

The area of the pMOS and nMOS mirrors is $84 \times 101 \mu m$ and each synapse takes an area of $57 \times 75 \mu m$. The synapse density is $234 synapses/mm^2$. These small sizes suggest that implementation is very feasible for VLSI technology.

2.4 Control of the synapse's weight

Fig. 2.8 shows the variations of I_o with respect to V_1 for four different values of V_b. At smaller values of V_b the output current saturates at smaller differences of input voltages. Fig. 2.9 shows calculated variations of g with V_b. They have an almost linear relationship. g varies from $2 \rightarrow 14 \mu A/volt$ as V_b varies from $1 \rightarrow 2.5V$. If power consumption is a problem it could be operated in the subthreshold mode of the gain transistor, which means that the gain transistor has to be little wider to realize a nanoamp current. The only

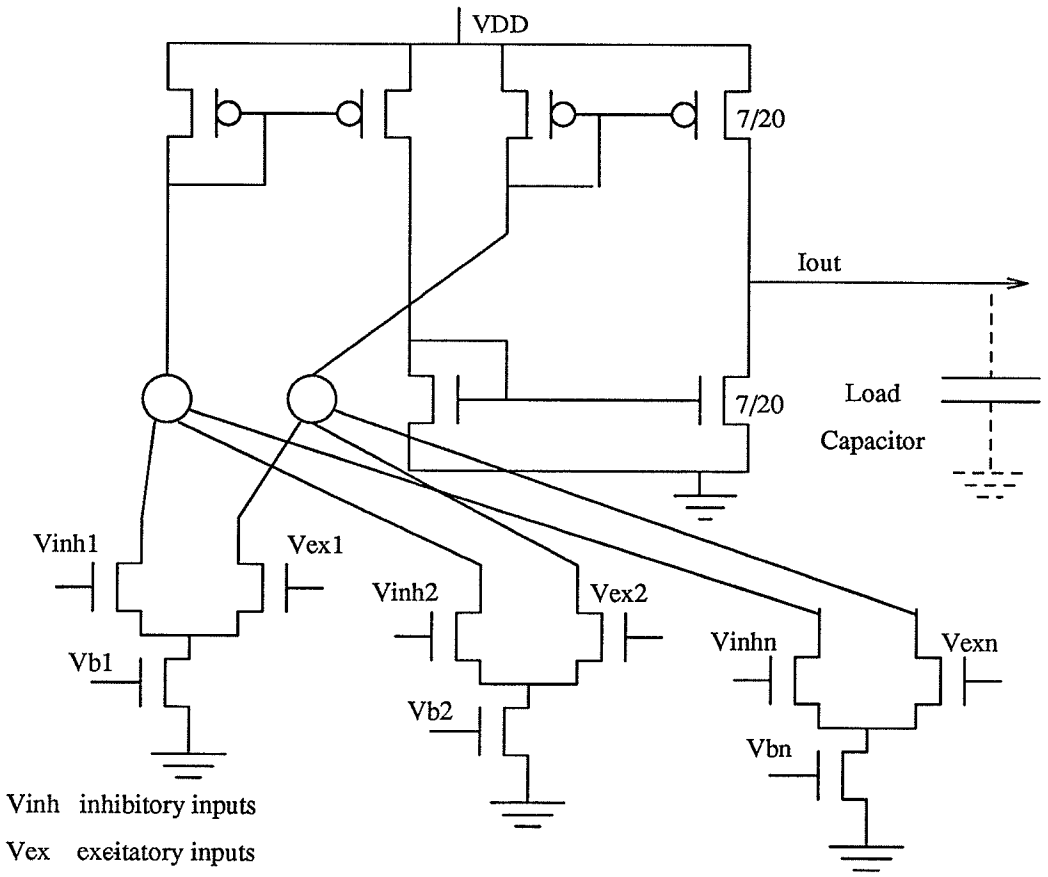

Figure 2.7: The multiple input OTA to model a neuron.

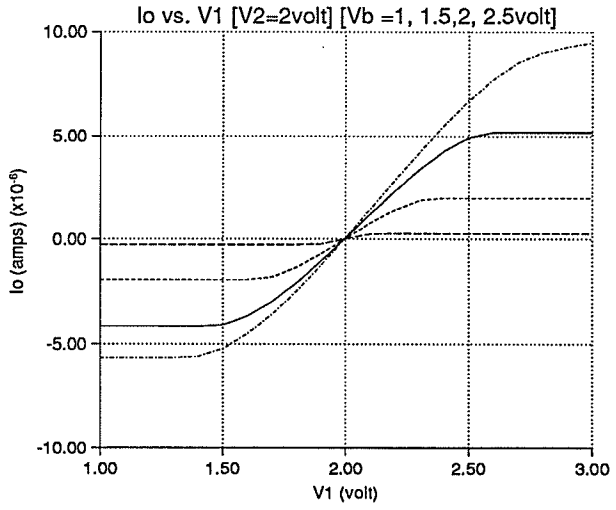

Figure 2.8: Variations of the I_o with input voltage.

problem is that I_d increases exponentially with V_b in this mode. In some applications this might be required [3].

2.5 Multiple input structure

To test the capability of multiple inputs, SPICE simulations of one input and of twenty inputs were performed. Fig. 2.10 shows results of one input when V_b= 1V, and Fig. 2.11 shows the simulation of twenty inputs having the same gain, to show that they indeed do multiply, add and subtract. The multiple input structure suggested is shown in Fig. 2.7.

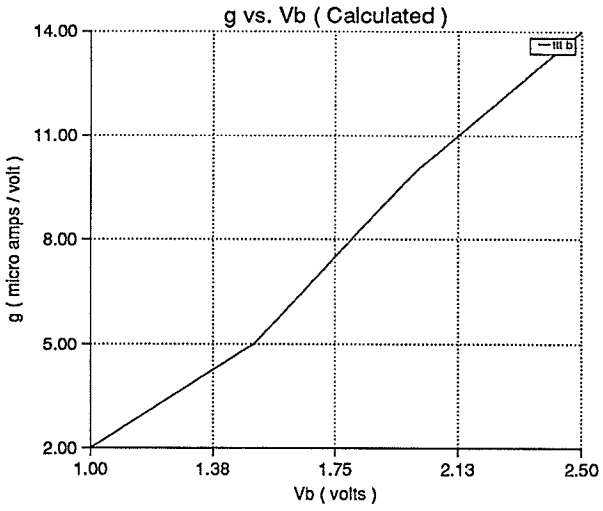

Figure 2.9: Variations of g with V_b.

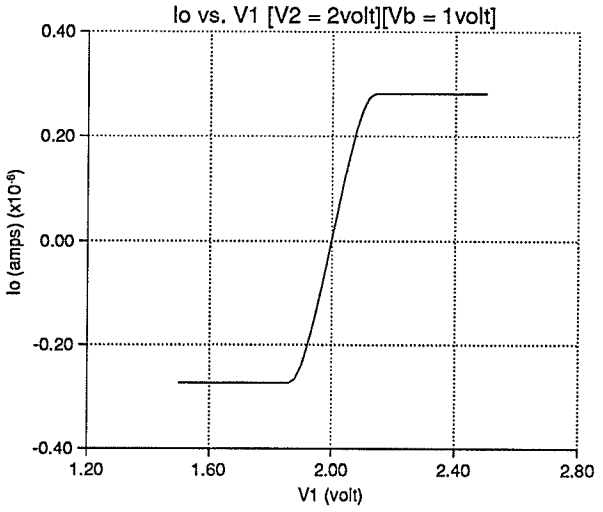

Figure 2.10: Simulations of one input.

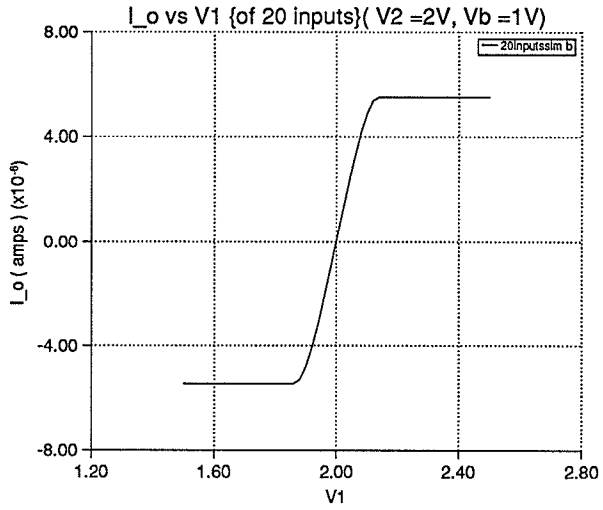

Figure 2.11: Simulations of twenty inputs.

2.6 Evaluation of the common OTAs for modeling neurons

If the synapses are to be operated in the subthreshold of the gain transistor then there is no limit to the number of the synapses that could be connected to the current mirror.

Simulations showed that when the common OTA is used above threshold, and has more than 38 synapses there were large differences between the positive input and negative input limiting currents (i.e., when $V_b \geq 1.7$ volts). The negative inputs limiting current exceeded the positive inputs limiting current by more than 40%. Typically the difference between the positive and negative limiting currents is 20% [3]. To solve this problem a dendrite

of one current mirror and 38 synapses could be used to collect the weighted inputs which are then are easily summed by the neuron. If this OTA is to be used above threshold (i.e., if $V_T = 0.7$ volt) then it is recommended that $0.7 \leq V_b \leq 1.5$ volts.

Chapter 3

Neurons Modeling With Cascade Staged OTAs

The cascaded stage OTA is based on the cascade stage to control the gain and input voltage, while a current mirror is used to produce a single output that is a function of the weighted sum of inputs. These inputs have a wide linear range and this type of OTA has the capability to accept a large number of inputs.

3.1 Modeling a synapse

Fig. 3.1 shows the synapse which represents one weighted input to the neuron. It sinks a current I_o which is a linear function of V_{in} and has a weight (transconductance) controlled separately by the bias voltage V_b (i.e., the gain of the block represented by the transconductance g).

T_1 operates as VCCS *(Voltage controlled current source)* of *common source configuration* and it regulates the output current. T_2 *(common gate tran-*

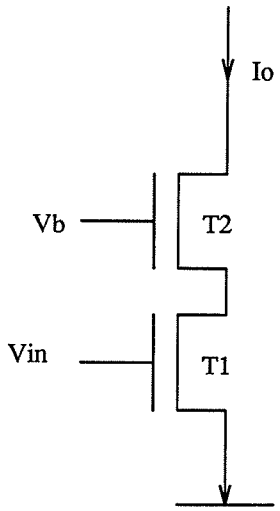

Figure 3.1: The synapse (cascade stage).

sistor) increases the output resistance and decreases the input capacitance [10].

3.1.1 Operation of the synapse

T_1 is operating in the active region. Assume that the Threshold voltages are equal $V_T = V_{T1} = V_{T2}$, and V_{bs} effects are ignored [7], then

$$V_{gs1} - V_T \geq V_{ds1}$$

$$I_1 = \mu C_{ox} \frac{W_1}{L_1} [(V_{gs1} - V_T) - \frac{V_{ds1}}{2}] V_{ds1}$$

$$\beta = \mu C_{ox} \frac{W_1}{L_1}$$

Where μ = mobility, C_{ox} = oxide thickness, W = transistor width, L = transistor length.

When $\frac{W_2}{L_2} \gg \frac{W_1}{L_1}$ and T_2 is biased in the saturation region

$$V_{gs2} - V_T \leq V_{ds2}$$

$$V_{ds1} \simeq V_b - V_T \tag{3.1}$$

Since V_b is a constant voltage,

$$I_1 = \beta(V_b - V_T)[V_{in} - \frac{V_b - V_T}{2}] \tag{3.2}$$

$$I_1 = g(V_{in} - \frac{V_b - V_T}{2})$$

Therefore, the transconductance(gain) of the cascade stage is determined by the transconductance of T_2,

$$g = \beta(V_b - V_T) \tag{3.3}$$

where $\beta = \mu C_{ox} \frac{W_1}{L_1}$. The relationship between V_{in} and I_o is linear, but when $V_{gs1} \leq V_b$, the response becomes nonlinear and I_o approaches 0 as shown in Fig. 3.2. Thus the input must be $\geq V_b$. Fig.3.3 shows the relationships between I_o and V_{in} for different values of V_b. When V_b goes from 2 to 3V, the transconductance changes from 1.3 to 3 $\mu A/V$. g is linear function of V_b. The linearity is dependent on the ratio of $(\frac{W_2}{L_2}/\frac{W_1}{L_1})$, where large ratios give a more linear response [7]. In Neural Networks linearity is not a critical factor, therefore the ratios will be smaller to be useful for VLSI implementations. (All simulations in this work have used SPICE parameters listed in appendix A.)

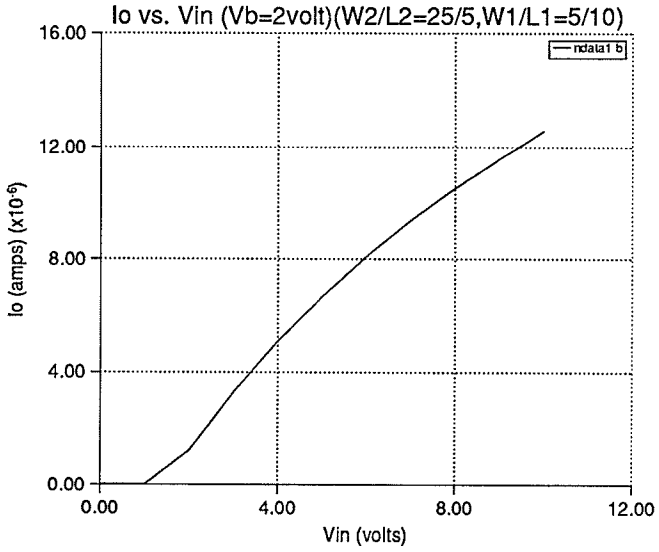

Figure 3.2: Variations of the output current with V_{in}.

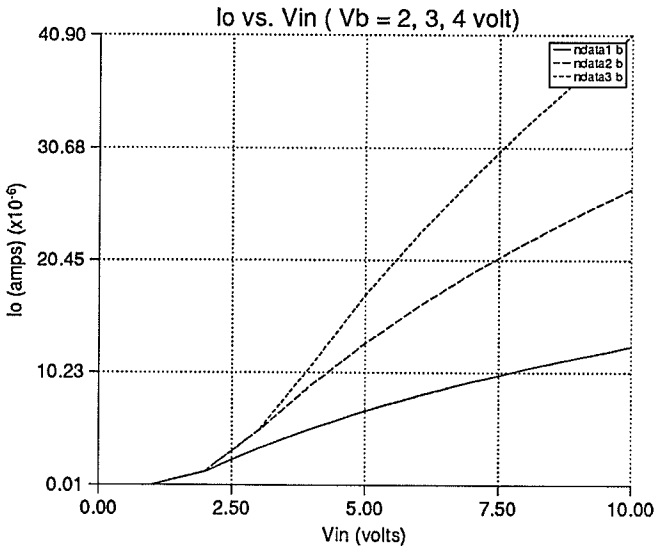

Figure 3.3: Controlling the weight in the synapse.

3.1.2 Input/Output impedance of the synapse

Fig. 3.4 shows the capacitances involved. They present no problem (no load) when operated at low frequencies [9]. When T_1 is biased in the active region the total capacitance is

$$C_{tot} = C_{gd} + C_{gs} \simeq C_{ox}WL \tag{3.4}$$

The input capacitance can be made small by keeping T_1 small.

The output impedance of the OTA should be very high (ideally should be ∞). Fig. 3.5 shows the simulations of I_o and V_{out}, where $V_b = 3V$, $V_{in} = 8V$, $W_1/L_1 = 5/10$, $W_2/L_2 = 25/5$, where W and L in μm.
$Z_o \simeq$33 MΩ for $4V \leq V_{out} \leq 8V$. Fig. 3.5 also shows that for the output impedance to be high the output voltage should be:

$$V_o \geq V_b - V_T \tag{3.5}$$

From Fig. 3.5 I_o is relatively independent of V_o, therefore output nodes of several cascade stages could be connected together to provide multiple input stages as shown in Fig. 3.6. The output current is then the algebraic sum of the individual currents.

$$I_o = I_{o1} + I_{o2} + I_{o3} + I_{o4} + \ldots \tag{3.6}$$

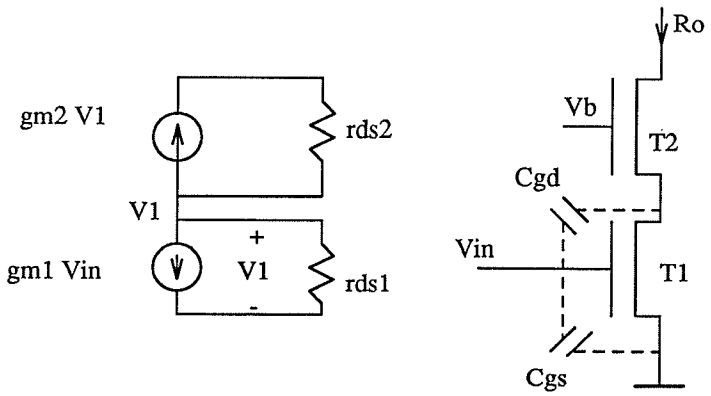

Figure 3.4: (a) Small signal model, (b) The Capacitance involved in the cascade stage.

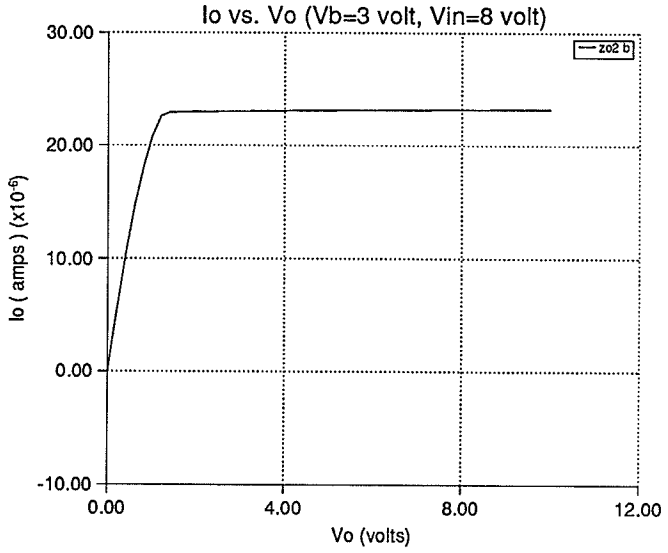

Figure 3.5: The output impedance of the cascade stage.

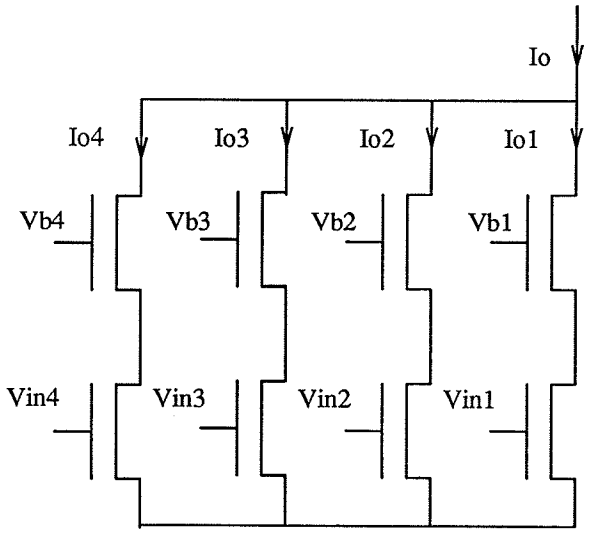

Figure 3.6: Synapses connected together to form multiple input

The weight of each synapse is separately and individually controlled by V_{bi}.

$$I_o = \sum_{i=1}^{N} g_i(V_{in_i} - \frac{V_{b_i} - V_T}{2}) \tag{3.7}$$

where, as before, g_i is

$$g_i = \beta(V_{b_i} - V_T)$$

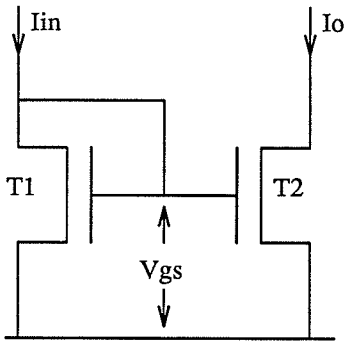

Figure 3.7: A simple current mirror.

3.2 Current Mirrors

Current mirrors are very important building blocks in MOS analog circuit design. They determine the offset voltages, thus if matching is not correct there will be an error in the output current. Their design is very important since low output resistance will be influenced by the changes in the output voltage. This in turn change the current ratio $\frac{I_o}{I_i}$. The nMOS current mirrors will be discussed for sake of simplicity. The circuit performance and small-signal output resistance are similar to that of the pMOS [13].

3.2.1 Simple current mirror

Fig. 3.7 shows the simple current mirror where both transistors have the same gate source voltage [11].

$$\frac{I_o}{I_i} = \frac{L_1}{W_1} \times \frac{W_2}{L_2} \tag{3.8}$$

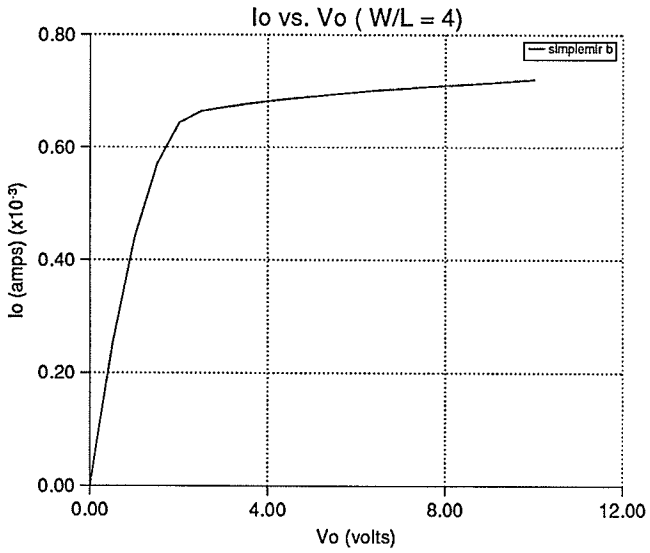

Figure 3.8: Simulation results of the simple current mirror.

Thus the current ratio $\frac{I_o}{I_i}$ is determined by the aspects ratio of the transistors, for identical transistors the ratio is unity that is $I_i = I_o$.

For the simple two transistor current mirrors it is quite impossible in the real world for the two drain-source voltages to be the same. Therefore, the current ratio has an error due to the channel shortening effects when drain-source voltages are not equal [11].

Fig. 3.8 shows the simulation results of the simple current mirror when both transistors have identical sizes $W = 20\mu m$ and $L = 5\mu m$. It is obvious that the value of I_o is dependent on V_{out}. The small-signal output resistance of

the simple current mirror is proportional to the slope of the curve shown in Fig. 3.8. $Z_O = 105K\Omega$ for $3V \leq V_{out} \leq 5V$. The current variation was $50\mu A$ for $3V \leq V_{out} \leq 10V$. When the ratio of $\frac{W}{L}$ was reduced to 1, the current variations decreased to $12\mu A$ which is still too large for this application. The output resistance of the simple current mirror is low and should be increased, because a low output resistance means that the output current will be affected by the changes in the output voltage.

The problem can be solved by using the Wilson current mirror as shown in Fig. 3.9, or by using the cascaded current mirror of four transistors as shown in Fig. 3.11.

3.2.2 Wilson current mirror

The Wilson current mirror circuit is shown in Fig. 3.9. The output resistance is increased with the use of negative feedback. Simulation results are shown in Fig. 3.10, where $W_3 = 15\mu m$, $L_3 = 5\mu m$ and $W_1 = W_2 = 10\mu m$, $L_1 = L_2 = 5\mu m$. One can observe that the output resistance has increased when compared to the simple current mirror output resistance. $Z_O = 10M\Omega$ for $3V \leq V_{out} \leq 5V$. The current variation was $78nA$ for $3V \leq V_{out} \leq 10V$, which is low compared to the simple current mirror.

3.2.3 Cascaded current mirror

The cascaded current mirror is shown in Fig. 3.11. In the cascaded current

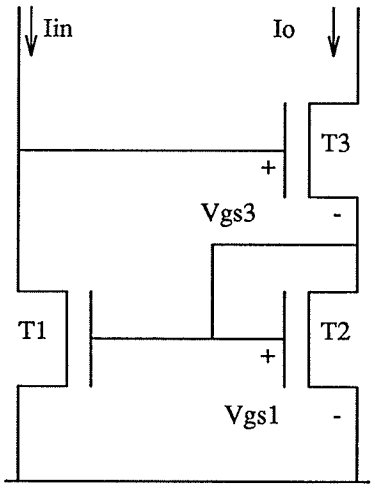

Figure 3.9: The Wilson current mirror.

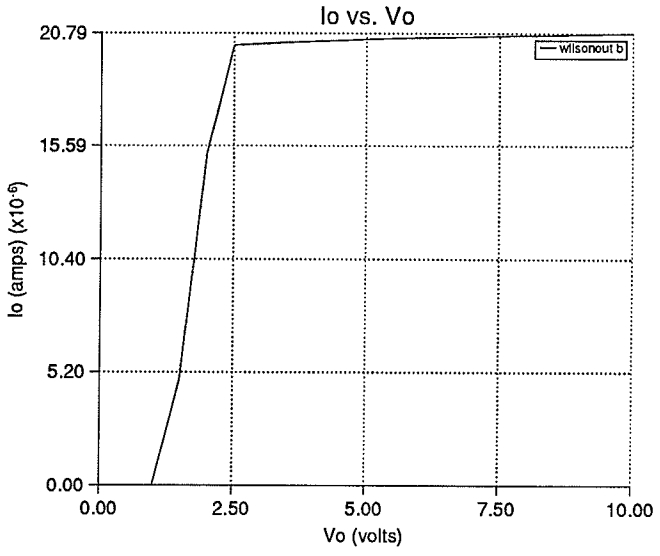

Figure 3.10: Simulation results of the Wilson current mirror.

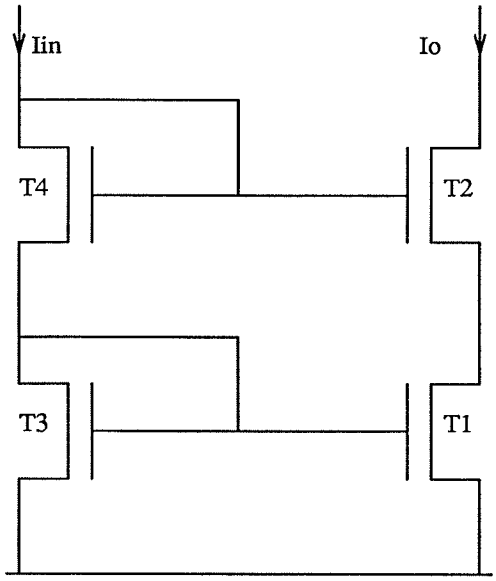

Figure 3.11: Cascaded current mirrors.

mirror T_2 shields T_1 from any variations in voltage that might occur at the output node. This mirror causes I_o to be less dependent on V_o.

The small signal output resistance is [9]:

$$Z_o = \frac{V_o}{I_o} = r_{ds1} + r_{ds2} + g_2 r_{ds1} r_{ds2} \tag{3.9}$$

$$Z_o \approx r_{ds1}(g_2 r_{ds2})$$

where r_{ds} = The small signal output resistance and g = the transconductance.

Fig. 3.12 shows the simulation results of the cascaded current mirror, where $W = 20\mu m$ and $L = 5\mu m$. Clearly this shows a much better performance and higher output impedance than the previous current mirrors.

$Z_o = 500M\Omega$ for $3V \leq V_{out} \leq 5V$. The current variation was 10 nA for $3V \leq V_{out} \leq 10V$.

When the transistor ratio $(\frac{W}{L})$ was reduced to 2, that is $W = 10\mu m$ and $L = 5\mu m$, the Z_o increased to $1000M\Omega$. The current variation was $5nA$ for the same range as shown in Fig. 3.13.

The cascaded current mirror is clearly the best choice for this application because of:

- lower offset voltage
- high output resistance

which are required for the OTA, since the output resistance is the parallel combinations of Z_o of the current mirror and Z_o of the cascade stage.

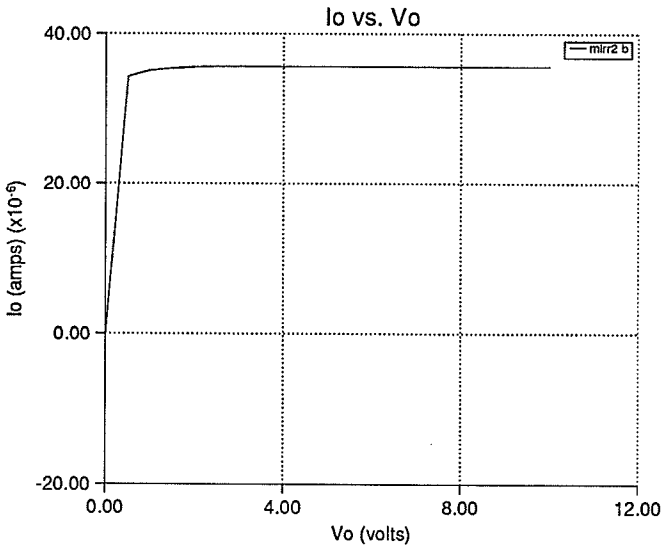

Figure 3.12: Simulation results of the cascaded current mirror ($\frac{W}{L} = 4$).

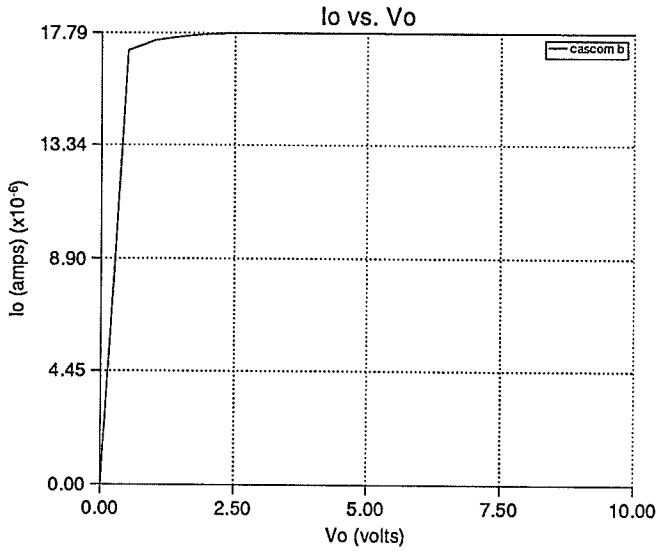

Figure 3.13: Simulation results of the cascaded current mirror ($\frac{W}{L} = 2$).

3.3 Construction of neurons

At was discussed previously, each cascade stage (synapse) has a separate bias voltage V_b and if $V_{in} \leq V_b$ the input transistor is not in the active region. In order to cancel the offset voltage, the cascade stage must be connected in pairs as shown in Fig. 3.14 and the transistors must be of identical size. The output current will be the difference between the two currents that have the same offset. (If T_1 and $T_1^{'}$ have the same bias, size and threshold their offsets will be equal and will cancel each other.) If T_1 is used as an input then $T_1^{'}$ will be used as reference and adjusted to account for the offset, and vice versa. Another alternative is to model each synapse by one cascade stage connected as excitatory or inhibitory input to the neuron and an extra cascade stage to be used to cancel the offset of all the inputs to that neuron.

Each synapse will therefore have 2 nMOS transistors as was shown in Fig. 3.1, where $W_1 = 5\mu m$, $L_1 = 10\mu m$ and $W_2 = 25\mu m$, $L_2 = 5\mu m$. The actual synaptic density will be 378 synapse/mm^2. The transfer characteristic of this OTA is shown in Fig. 3.15.

3.3.1 Synaptic weight control

The gain is controlled by V_b as shown in Fig. 3.15. Fig. 3.16 shows the calculated transconductance plotted against V_b. One can see that g varies quite linearly with V_b, increasing from $0.66 \rightarrow 12\mu A/V$ as V_b varies between $-3.5V \rightarrow -0.5V$.

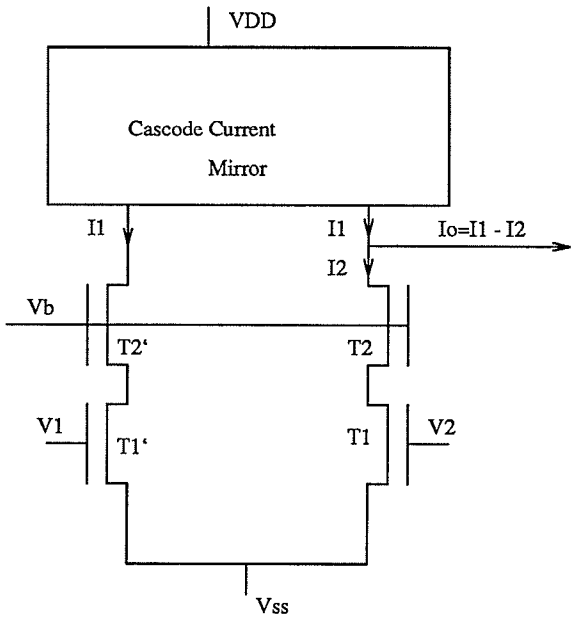

Figure 3.14: The connection of the OTAs

To test the multiple input capability a test on two inputs and a ten input OTA were performed with $V_b = -1.5V$ as shown in Fig. 3.17.

The multiple input neuron structure is shown in Fig. 3.18, the inhibitory inputs are connected on one side and the excitatory inputs are connected on the other side of the cascaded current mirror. The output current is the weighted difference between the inhibitory and excitatory inputs [7]. Since the devices will have the same size, the expression for I_o will be

$$I_o = \sum_{i=1}^{N} \pm g_i (V_{in} - \frac{V_{b_i} - V_T}{2}) \tag{3.10}$$

Where the $\pm$ sign of the above equation depends on whether the input is excitatory or inhibitory. If an op-amp with a feedback resistor is used to

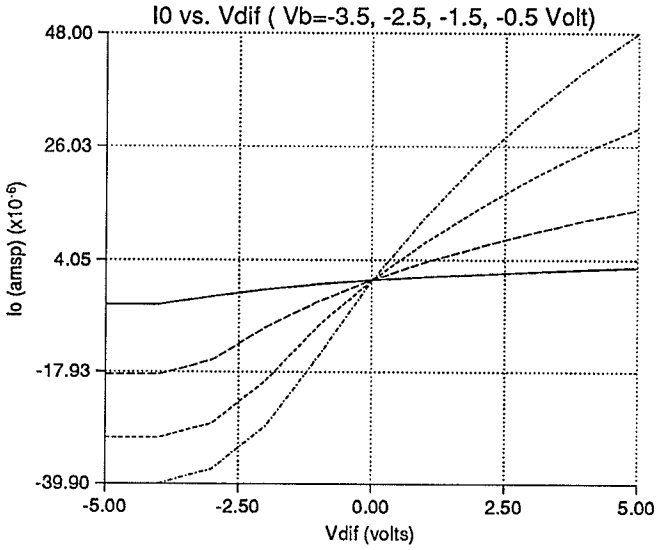

Figure 3.15: Variations of the output current with input voltage for different gain voltages.

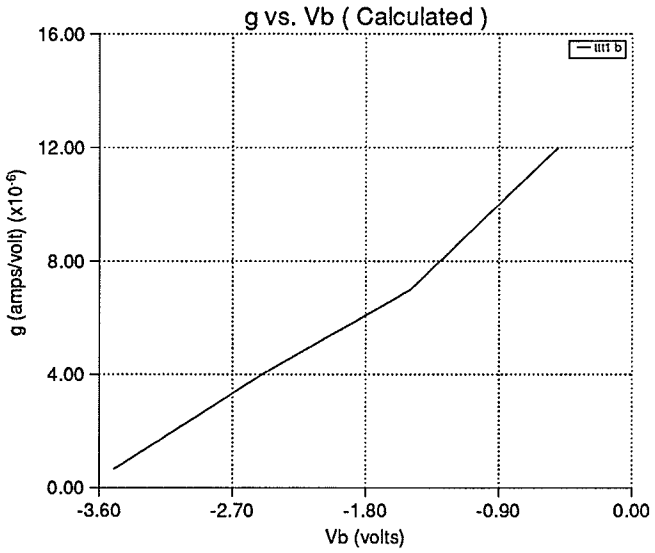

Figure 3.16: Variations of transconductance with the gain voltage.

convert the output current to output voltage (transresistance amplifier) then

$$V_o = RF(I_o) \tag{3.11}$$

When I_o is greater than some value V_o will saturate at the supply voltage which will give it sigmoidal shape.

3.3.2 The output impedance

The output impedance of this OTA is the parallel combination of the output impedance of the inverting cascaded stage and the cascaded current mirror. Therefore, for a high output impedance the cascaded current mirror and T_2 of the cascade stage must be in saturation, when $V_o \geq V_{DD} - 2V_T$ the cascaded

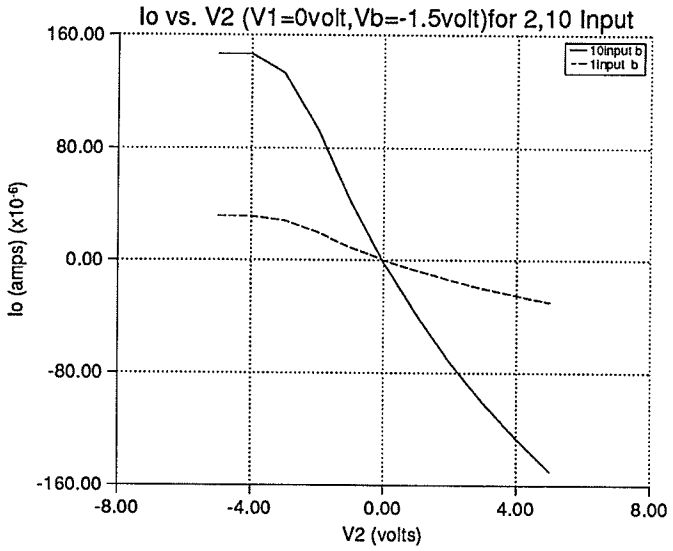

Figure 3.17: Comparisons between a number of synapses connected together.

current mirror will not be in saturation and when $V_o \leq V_b$ the cascade stage will not be in saturation. The output impedance of both is going to be maximum whenever $V_b \leq V_o \leq V_{DD} - 2V_T$.

3.4 Evaluation of the cascaded stage OTA for modeling neurons

This type of OTA should be used at low gain to minimize the power consumption and to increase the wide linear range of the inputs. Each synapse could be modeled by two transistors one for the input and one for the gain. Synapses are connected directly as an inhibitory or an excitatory input. The

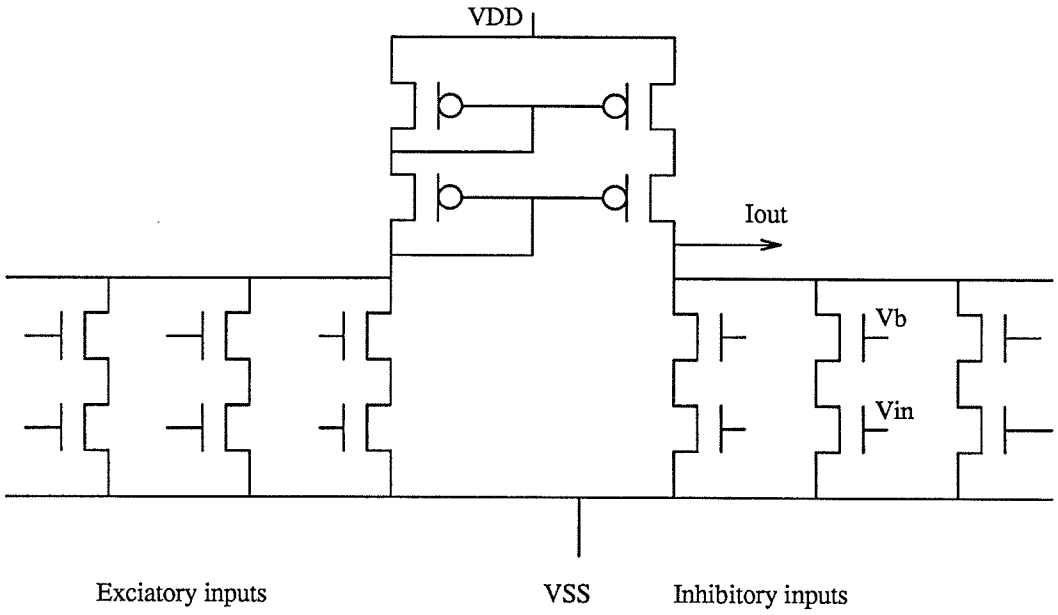

Figure 3.18: The multiple input structure.

input has to exceed the gain voltage to have a linear variation between I_o and V_{in}. In theory this cascaded OTA can take a large number of inputs, but simulations have shown that if more than 20 inputs are connected to the same neuron the negative limiting current seems to exceed the positive limiting current. Therefore, it should be used as dendrite if a great number of inputs are required. Then each dendrite will have a cascaded current mirror which takes an area of $65 \times 57\ \mu m$ to sum the weighted inputs. The output of the dendrites is the current which is then summed by the neuron and converted to voltage.

Chapter 4

Implementations of Neurons

The architecture is based on the cross-bar arrangement of synapses (or Hopfield). The synapses are connected directly as inhibitory or excitatory inputs.

4.1 The Hopfield architecture

The Hopfild network consists of an array of fully connected synapses. All neurons are connected to each other through programmable connections either inhibitory or excitatory. The output of each neuron is fedback into the network. Each synapse will compute a new output value according to the output of the control neuron and the charge stored on it. Each neuron sums the output of all the synapses, then this sum determines the neuron's activity and fixes the neuron's output through its nonlinear transfer function, which is sigmoidal in shape. This type of network is used in pattern recognition and associative memories. This model suits the OTA networks discussed in the previous chapter.

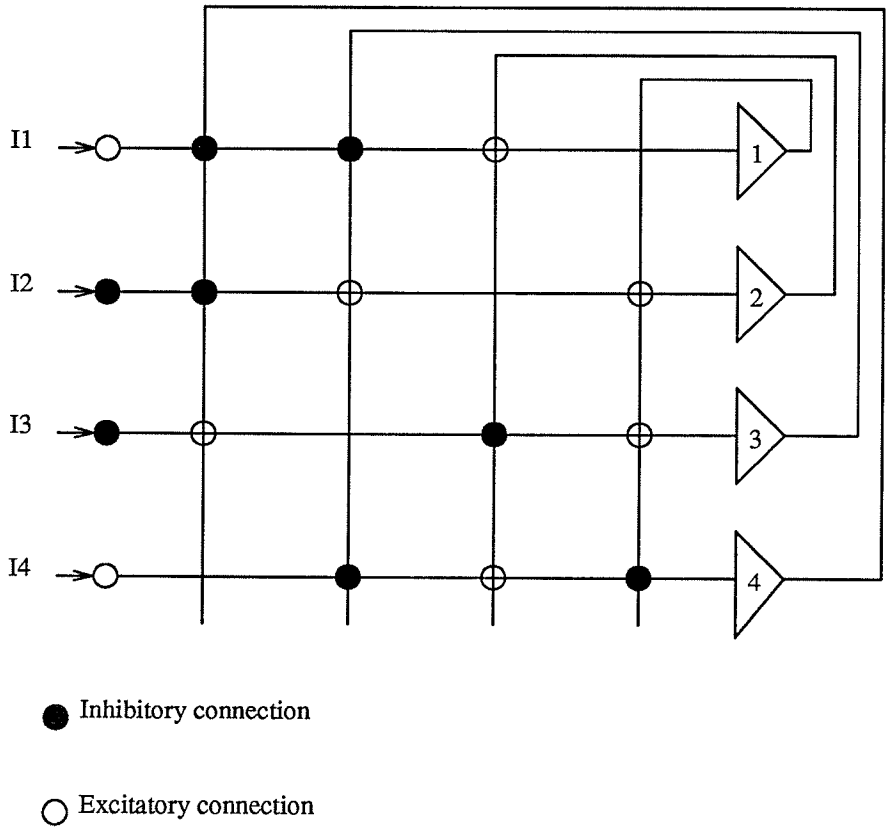

Figure 4.1: The cross-bar arrangement.

4.2 Implementations of neurons using cascade stages

The cascade stage is used to model a synapse and the cascaded current mirror is used to sum and subtract currents of the synapses. Since this OTA is going to be operated between -5 and +5 Volts, a range of 10V the minimum device size should be at least 5 or 6 μm [14].

4.2.1 The synapse model

Each synapse is modeled as shown in Fig. 4.2, the strength of the synapse is controlled by the amount of charge stored on the capacitor, G is a long transistor used to control the weight decay if needed and discharge the capacitors. The capacitor, the cascade stage and the long transistor take up an area of 0.0052 mm^2.

4.2.2 The chip architecture

The neurons are modeled after the cross-bar arrangement shown in Fig. 4.1 The neuron layout connection is shown in Fig. 4.3.

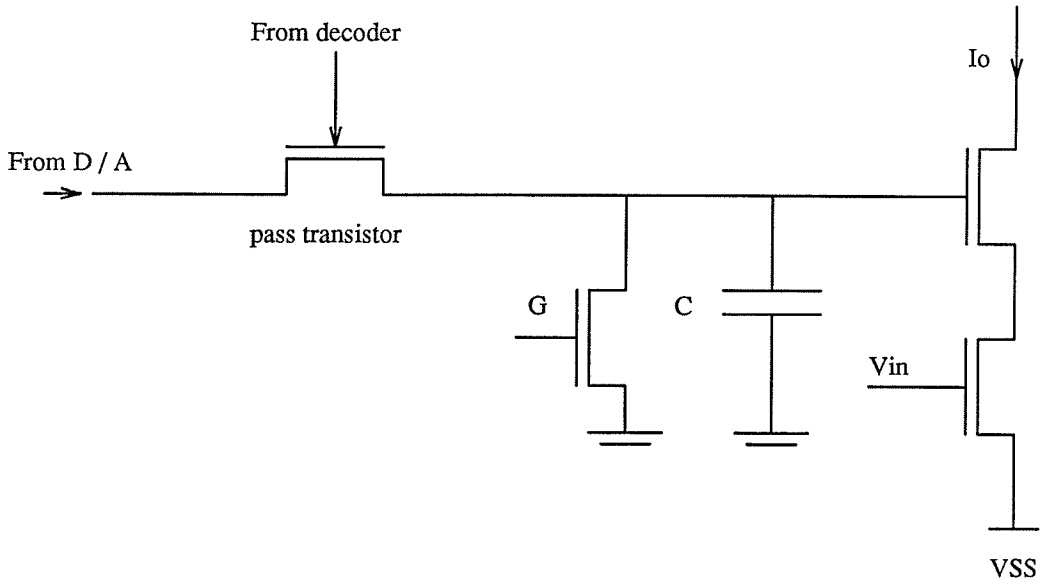

Figure 4.2: The Synapse model.

The transistor sizes of Fig. 4.3.

Transistor	$W\mu m$	$L\mu m$
T1	25	5
T2	5	10
T3	20	5
T4	20	5
T5	20	5
T6	20	5

The weights form a matrix W.

$$\begin{bmatrix} W_{11} & W_{12} & W_{13} & W_{14} \\ W_{21} & W_{22} & W_{23} & W_{24} \\ W_{31} & W_{32} & W_{43} & W_{34} \\ W_{41} & W_{42} & W_{43} & W_{44} \end{bmatrix}$$

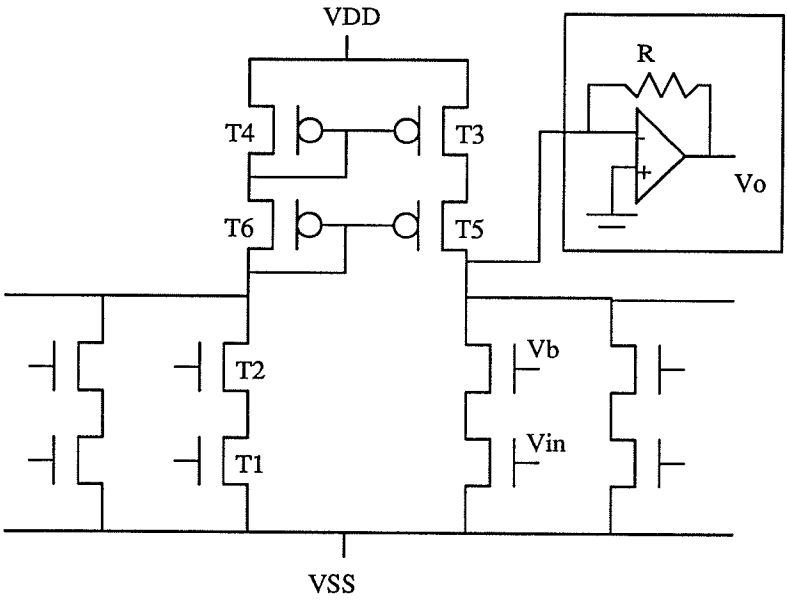

Figure 4.3: One neuron with 4 synapses.

This type of connection has been shown to be stable if the matrix is symmetrical with zeros on its diagonal axis (no connections), that is $w_{ij} = w_{ji}$ and $w_{ii} = 0$ [17].

The control circuit is the combination of the address decoder and the analog multiplexer. The control circuit will handle the refreshing and updating of the weights on the capacitors as shown in Fig. 4.4. An external D/A converter could be used to set the amount of the weight. The chip's layout is shown in Fig. 4.7.

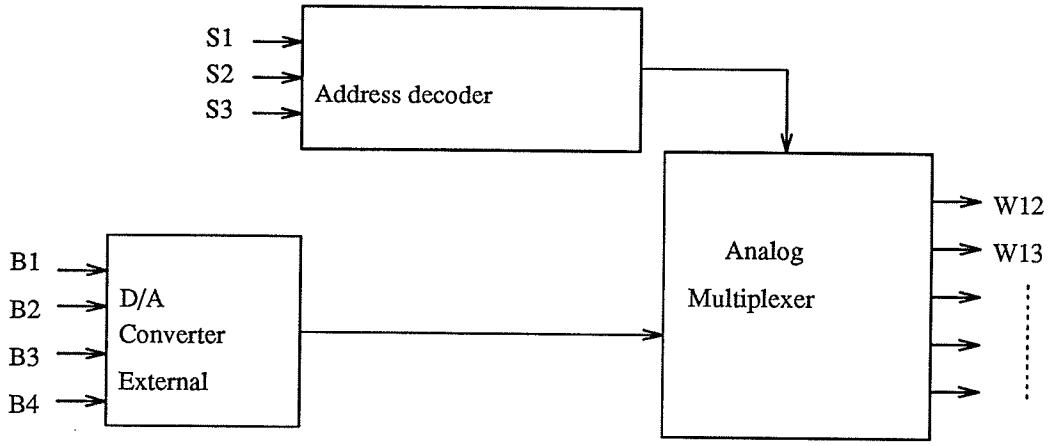

Figure 4.4: The chip architecture.

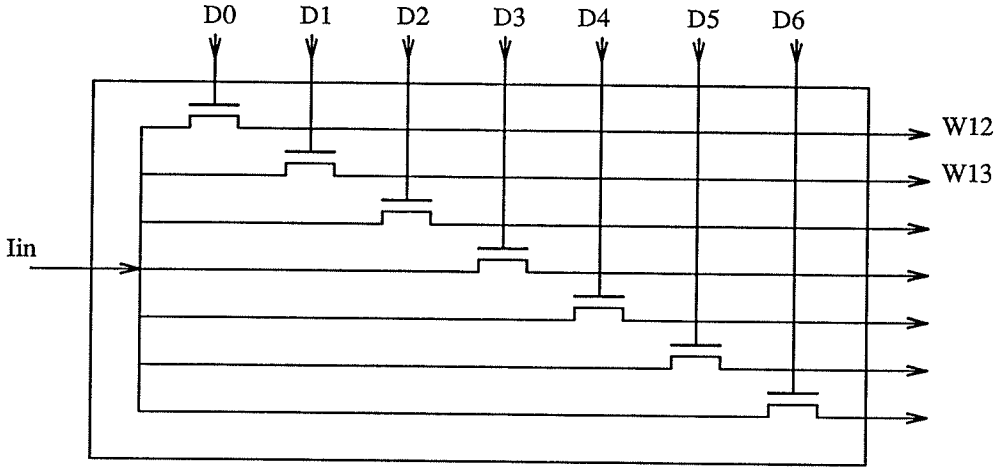

Figure 4.5: Analog multiplexer.

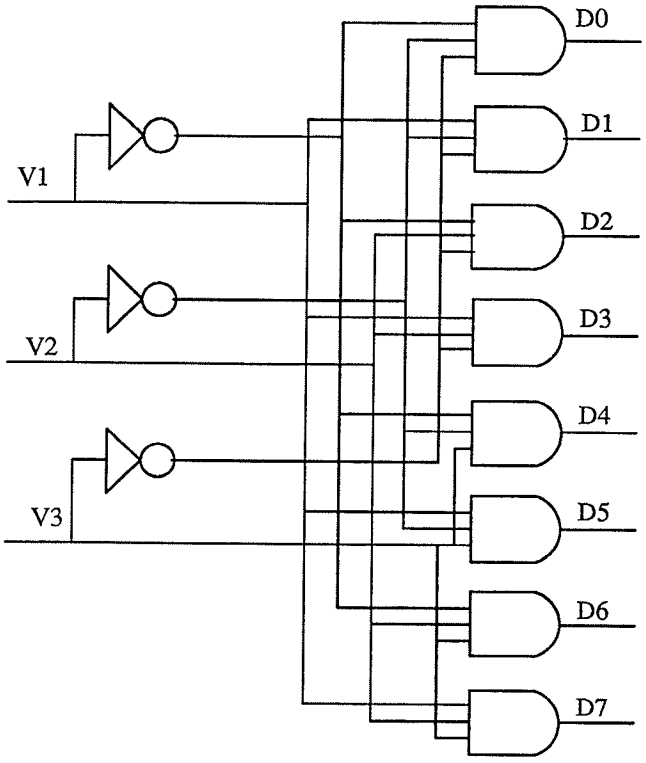

Figure 4.6: The address decoder used to control the capacitor refreshing and updating.

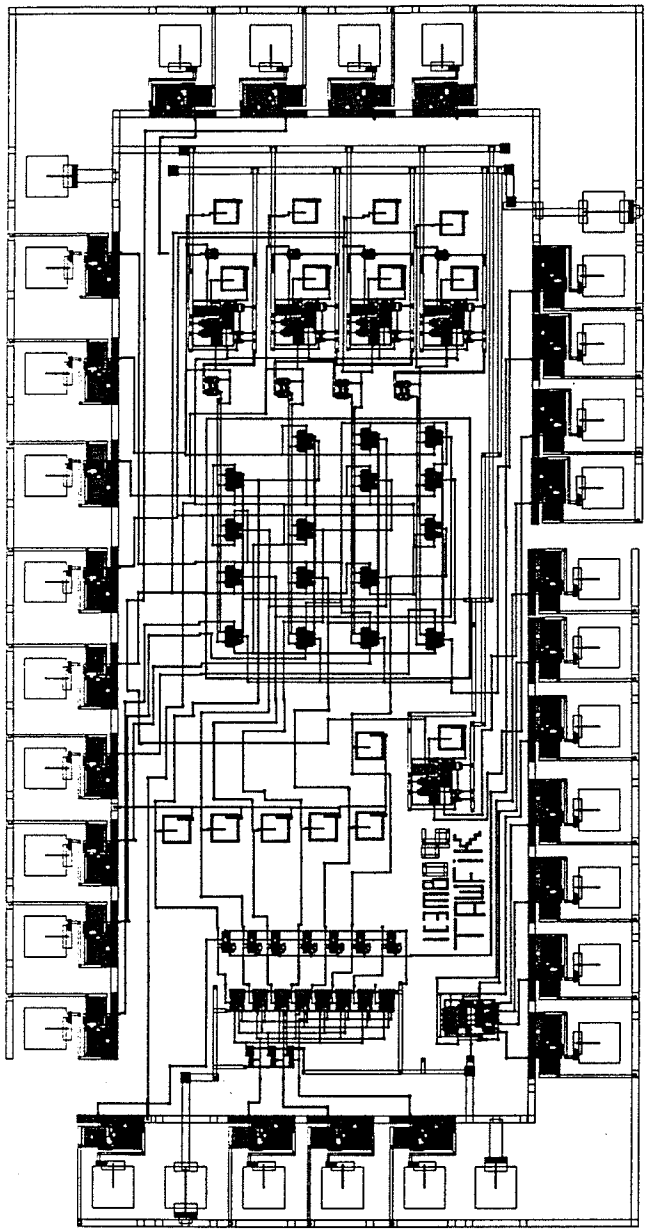

Figure 4.7: The layout of the chip.

4.2.3 Simulations of the cross-bar arrangement

Simulations were performed on the cross-bar arrangement of 4 neurons that was shown in Fig. 4.1. Each synapse is modeled by a cascade stage. Test was performed within the context of OTAs for random values of inputs. An inverter was used to convert the output current to voltage. The output voltage varies between 5 and -5 volts.

The weight matrix is: $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

where 1 represents -2 volt

The input and output voltage of the test circuit:

Vin(1)	Vin(2)	Vin(3)	Vin(4)	Vo(1)	Vo(2)	Vo(3)	Vo(4)
2	2	2	2	-5	-5	2	1.94
2	0	0	2	-5	-5	1.98	-5
0	0	2	2	1.5	-3	-3.9	-1.36
0	2	2	2	5	-5	-5	5

4.3 Implementations of neurons using common OTAs

This design uses the control circuit configuration that is shown in Fig. 4.8. Synapses are modeled using a 3 transistor multipliers as shown in Fig. 4.9. The capacitor, the 3 transistor multiplier and the long transistor take up an area of 0.0068 mm^2. The Weight is stored on the capacitor, which is refreshed and updated by an address decoder and an analog multiplexer. The neuron consists of an inverter to form the gain stage followed by the output buffer as shown in Fig. 4.10. On chip 3-bit D/A converter is used to set the weight value.

I_o is proportional to V_b, of the weight transistor, and [$V_{in} - V_{ref}$] which is applied to the differential pair. V_{in} is applied as inhibitory or excitatory. The synaptic and neuron circuit is shown in Fig. 4.10. Using V_H and V_L as the maximum and minimum output voltage swings, adjusted as required, and R_f is the feedback resistor which determines the closed loop gain (it could be implemented by transmission gate) [18]. The chip's layout of 4 neurons is shown in Fig. 4.12.

If power dissipation is a problem then the operation in the the subthreshold mode is required. (In dense array circuits usually power dissipation becomes a problem.) This circuit could be modified to suit the operation in subthreshold mode. The synaptic and neuron configuration are as shown in Fig. 4.10. Since the output of the neuron is fed to another synaptic connection the output

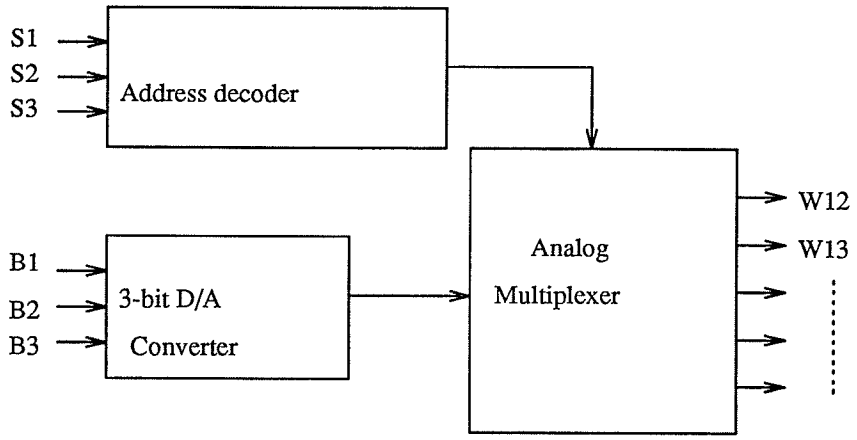

Figure 4.8: The control circuit.

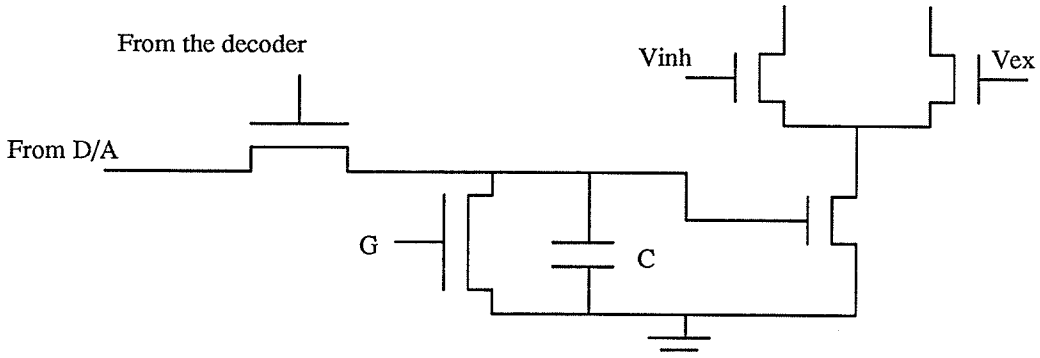

Figure 4.9: The Synapse model.

voltage has to be modified to fit the operating range of the subthreshold mode of the synapse. Therefore the output stage must be at a low power supply and $[V_H - V_L]$ is the maximum voltage swing of the outputs.

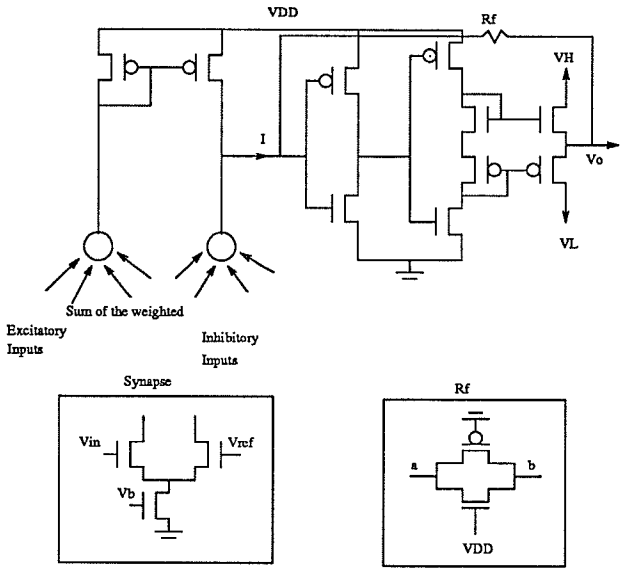

Figure 4.10: The neuron and synaptic connections.

4.3.1 The D/A converter used to control the weight

The circuit of the D/A is shown in Fig. 4.11 it takes an area of 0.1883 mm^2 [11]. The data for the 3-bit resistive string digital to analog converter, where $V_{ref} = 2V$.

A B C	V_o volt
0 0 0	0.0
0 0 1	0.250
0 1 0	0.499
0 1 1	0.749
1 0 0	0.999
1 0 1	1.250
1 1 0	1.500
1 1 1	1.750

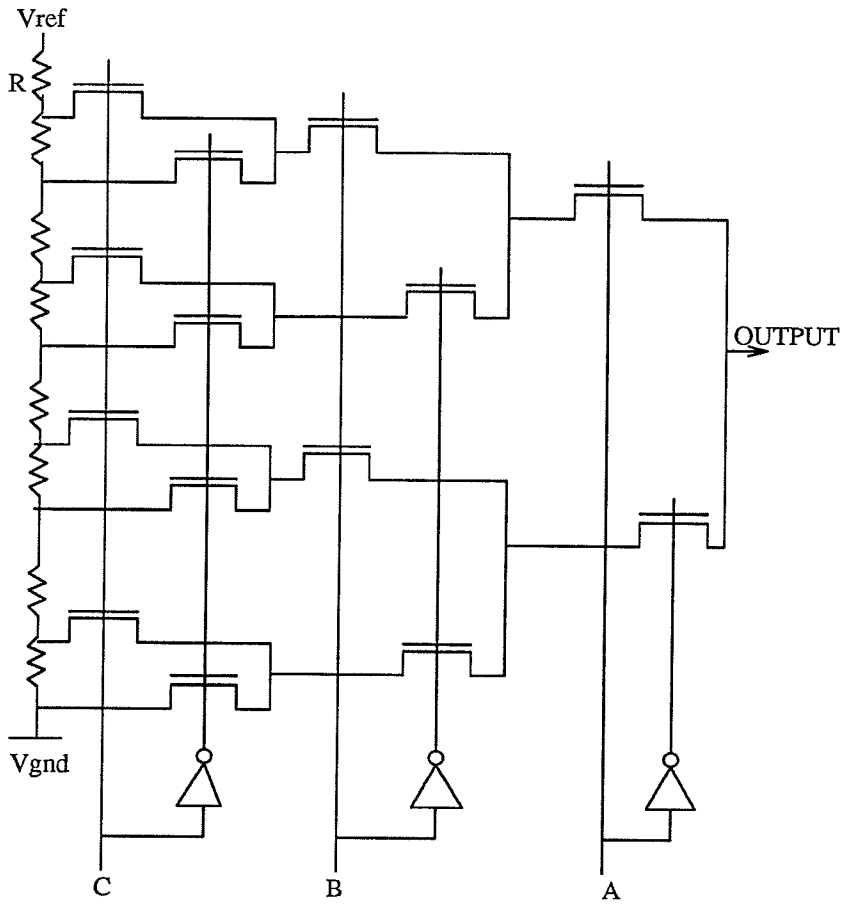

Figure 4.11: D/A 3 bit converter.

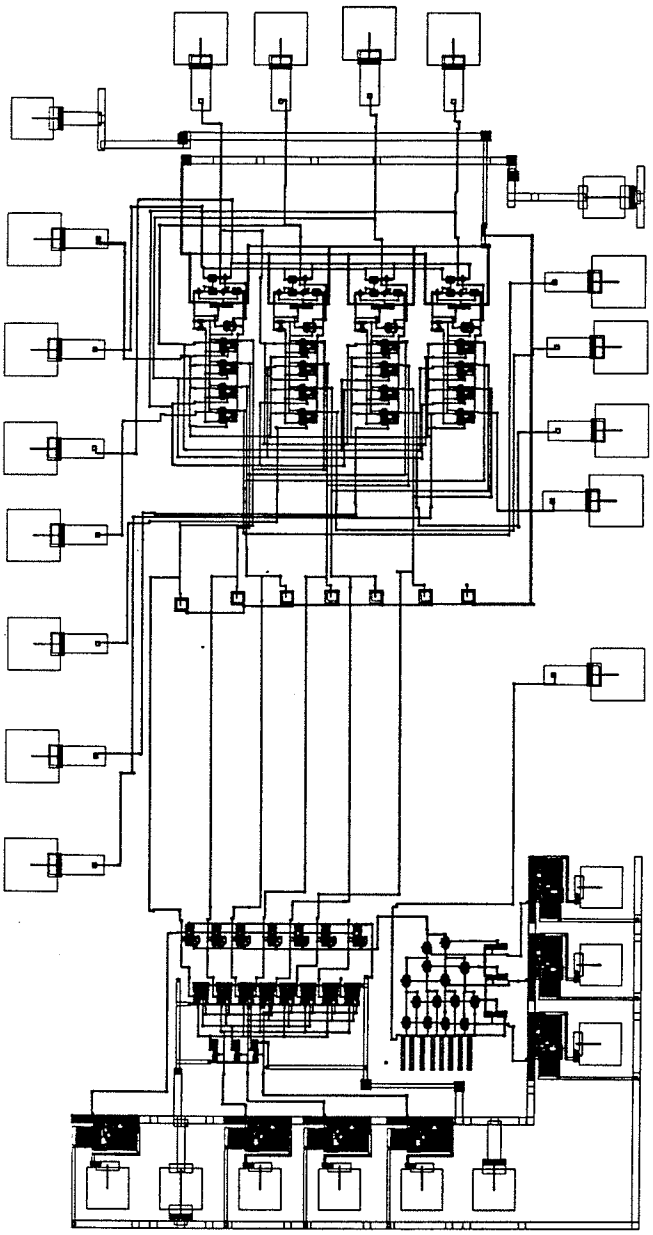

Figure 4.12: The layout of the chip.

4.3.2 Simulations of the the cross-bar arrangement

Simulations were performed on the test circuit of 4 neurons. $VDD = 5V$, $V_H = 2.6$ and $V_L = 2.4$. The maximum voltage swing is $V_H - V_L = 0.2$ volts. The results of the simulations are as follows:

The weight matrix is: $\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$

where 1 represents 1 volt.

The input and output voltage of the test circuit:

Vin(1)	Vin(2)	Vin(3)	Vin(4)	Vo(1)	Vo(2)	Vo(3)	Vo(4)
2.6	2.6	2.4	2.6	2.4	2.6	2.59	2.4
2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.4
2.4	2.4	2.4	2.6	2.6	2.4	2.4	2.59
2.4	2.4	2.6	2.6	2.6	2.4	2.6	2.6

The weight matrix is: $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

The input and output voltage of the test circuit:

Vin(1)	Vin(2)	Vin(3)	Vin(4)	Vo(1)	Vo(2)	Vo(3)	Vo(4)
2.6	2.6	2.4	2.6	2.6	2.59	2.4	2.4
2.6	2.6	2.6	2.6	2.6	2.4	2.6	2.4
2.4	2.4	2.5	2.6	2.6	2.49	2.4	2.4

Chapter 5

Conclusions

The intrinsic properties of OTAs have been examined for possible VLSI simulations and emulations of neurons. The OTA is a natural gain device that could be very useful for modeling neurons. The current output and the voltage tunability of the transconductance provide a good method of controlling the gain.

Two types of OTAs have been investigated for multiple input capabilities: the typical differential OTA and the Cascaded OTA. Those OTAs were implemented in cross-bar arrangement that contained 4 neurons. Simulations were performed on both circuits within the context of OTAs. The sizes of the OTAs were compatible for analog VLSI implementations. The common OTA current saturates at small values of differential voltages, therefore it could be used as a thresholding neuron that gives an output of 0 or 1. The cascaded OTA has a wider linear range of input voltage.

The advantage of using these OTAs is that the weight multiplication of the

inputs and summation of the weighted inputs are done by a small number of transistors. In the common OTA multiplication of the input with the gain is done by 3 nMOS transistors and summation of the weighted inputs is done by a pMOS current mirror. In the Cascaded OTA the multiplication is done by 2 nMOS transistors and summation of the weighted inputs is done by a cascaded current mirror.

The weights are of analog values. The OTAs discussed are suited for cross-bar arrangement of synapses. Inputs are connected directly as inhibitory or excitatory and a control circuit would handle the weight refreshing and updating.

If the common OTA is used in the subthreshold mode there is no limit to the number of synapses that could be connected to the current mirror. In subthreshold mode the synaptic current will be low which means that power consumption will be low. Simulations have shown that when the common OTA is used above threshold, and has more than 38 synapses the limiting currents of the positive inputs tends to exceed the limiting currents of the negative inputs. To over come this problem one could use one current mirror and 38 synapses as dendrite. The output of the dendrites is current which is easily summed by the neuron. If the common OTA is used above threshold ($V_T = 0.7$ volt), it was recommended that $0.7 \leq V_b \leq 1.5$ volts.

The problem with using the cascade stage OTA is that the input has to

exceed the gain voltage in order to have linear variation between the input voltage and output current. Also it must cancel the offset of the inputs. One may cancel this offset by either dedicating a synapse or by connecting the synapses in pairs.

5.1 Future Work

For future work the following might be done:

- Synapses were designed using $3\mu m$ CMOS technology. If smaller technology is used (i.e., $1.2\mu m$ technology) more synapses per area may be achieved.

- This technique of modeling neurons could be used as the hardware basis to model neurons in a feedforward network that employs some algorithm for changing the weight. Also it should include some system that changes the sign of the synaptic connection either inhibitory or excitatory as required.

- Some modification could be made so that one can have many chips that have the same circuit design and connect them together to have large number of neurons.

References

[1] D.O.Hebb, *The Organization of Behavior.* New York:Wily, 1949.

[2] Dr.J.C.Lupo, *"Defense Applications of Neural Networks,"* IEEE Comm. Magazine, pp.82-88 , Nov. 1989.

[3] Carver Mead, *Analog VLSI and Neural Systems.* Reading, MA, Addison Wesley, 1989.

[4] Jack I.Raffel, *Electronic Implementations of Neuromorphic Systems.* Proc. IEEE 1988 Custom Integrated Circuits Conference. pp.10.1.1-10.1.7 .

[5] M.C.Maker, S.P.Deweerth, M.A.Mahowald and C.A.Mead, *"Implementing neural architectures using analog VLSI circuits,"* IEEE Trans. Circuits and Syst., pp.643-652, May 1986.

[6] D.W.Tank and J.J.Hopfield. *"Simple 'neural' optimization networks,"* IEEE Trans. Circuits and Syst., pp.533-541, May 1986.

[7] R.D.Reed and R.L.Geigier, *"A multiple input OTA circuit for neural networks,"* IEEE Trans. Circuits and Syst., pp. 767-770, May 1989.

[8] Hans P.Graf and L.D.Jackel. *"Analog electronic neural network circuits,"* IEEE Circuits and Devices Magazine, pp.44-49, july 1989.

[9] Paul R.Gray and Robert G.Meyer. *Analysis and Design of Analog Integrated circuits.* 2nd edition,John Wily & Sons,New York,1984.

[10] Y.Tsividis and P.Antognetti. *Design of MOS VLSI circuits for telecommunications.* Prentice-Hall,Inc., 1985.

[11] M.R. Haskard and I.C. May, *Analog VLSI Design nMOS and CMOS.* Prentice-Hall of Australia, 1988.

[12] Paul R. Gray, David A. Hodges and Robert W. Brodersen. *Analog MOS Integrated circuits.* IEEE Press, 1980.

[13] P.E Allen and E. *Sánchez – Sinencio, Swiched Capacitor Circuits.* New York, NY, Van Nostrand Reinhold Company Inc., 1984.

[14] Carruthers Hall, Queen's University, *"Guide to the integrated circuit implementation services of the Canadian Microelectronic Corporation".*

[15] T.X Brown, *" Neural Networks for Switching,"* IEEE Comm. Magazine, pp.72-81, Nov. 1989.

[16] R.P Lippmann. *"An Introduction to Computing with Neural Nets,"* IEEE ASSP Magazine, pp.4-22, April 1987.

[17] P.D Wasserman. *Neural Computing theory and practice.* New York, NY, Van Nostrand Reinhold, 1989.

[18] D.D. Caviglia, M. Valle and G.M. Bisio, *"Effects of Weight discretization on the BP learning method: Algrithm design and hardware realization,"* Proceedings of IJCNN international joint conference on neural networks, San Diego, California, pp.II-631-637, 1990.

[19] S.Y. Foo, L.R. Anderson and Y. Takefuji, *"Analog components for the VLSI of neural networks,"* IEEE Circuits and Devices, Vol.6, no.4, pp. 18-26, July 1990.

[20] Meta-Software, *HSPICE H8801 Users' Manual,* January 1988, Campbell, CA.

[21] N.Weste and K.Eshraghian. *Principles of CMOS VLSI Design a system perspective.* Reading, MA, Addison Wesley, 1985.

Appendix A

The Transistor Model used

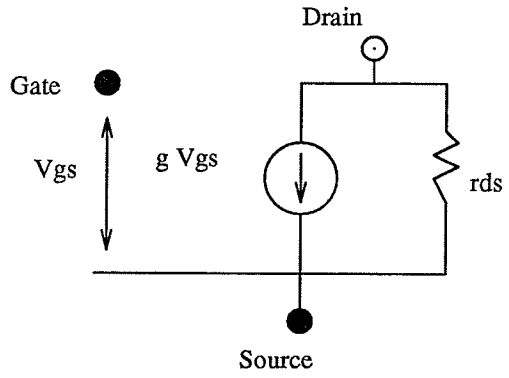

Figure A.1: The transistor small-signal model used.

The small-signal model of the MOS transistor

The ideal equations that describe the behavior of the nMOS transistor device are as follows [21]:

In the cut-off region $V_{gs} - V_t \leq 0$

$$I_{ds} = 0 \tag{A.1}$$

In the linear or active region $0 \leq V_{ds} \leq V_{gs} - V_t$

$$I_{ds} = \beta[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}] \tag{A.2}$$

In the saturation region $0 \leq V_{gs} - V_t \leq V_{ds}$

$$I_{ds} = \frac{\beta}{2}(V_{gs} - V_t)^2 \tag{A.3}$$

Appendix B

The SPICE Parameters

The SPICE parameters used:

Parameter	nMOS Model	pMOS Model
LEVEL	3	3
UO	726.1	246.8
VTO	0.702 V	-0.769 V
NFS	1.541E+11	4.121E+11
TPG	1.0	1.0
TOX	5.048E-08	5.048E-08
NSUB	2.022E+16	3.843E+15
VMAX	2.306E+05	4.667E+07
XJ	1.132E-07	3.091E-07
LD	2.693E-07	1.686E-07
DELTA	0.235	0.463
THETA	0.110	0.189
ETA	0.616	2.701
KAPPA	1.048	2.000E-17
PB	0.800	0.800
IS	1.000E-16	1.000E-16
JS	1.000E-04	1.000E-04
CJ	4.090E-04	1.440E-04
MJ	0.498	0.621
CJSW	4.780E-10	3.360E-10
MJSW	0.363	0.434
CGSO	2.910E-10	2.370E-10
CGDO	2.910E-10	2.370E-10
FC	0.500	0.500

The parameters of the of HSPICE

Parameter	nMOS Model	pMOS Model
LEVEL	5	5
UB	700	245
VT	0.7 V	-0.7 V
FRC	0.05	0.25
DNB	1.6E16	1.3E15
XJ	1.2	1.2
LATD	0.7	0.9
CJ	0.13	0.09
PHI	1.2	0.5
TCV	0.003	0.002
TOX	800	800

Appendix C

8 Neuron Chip layout

This is a chip that contains 8 neurons as shown in Fig. C.1. Neurons modeled using common OTAs. Two address decoders are used to control the synaptic weight and refreshing the weights that charged on capacitors. It also contain a D/A converter. This chip contains 8 neurons and 64 synapses.

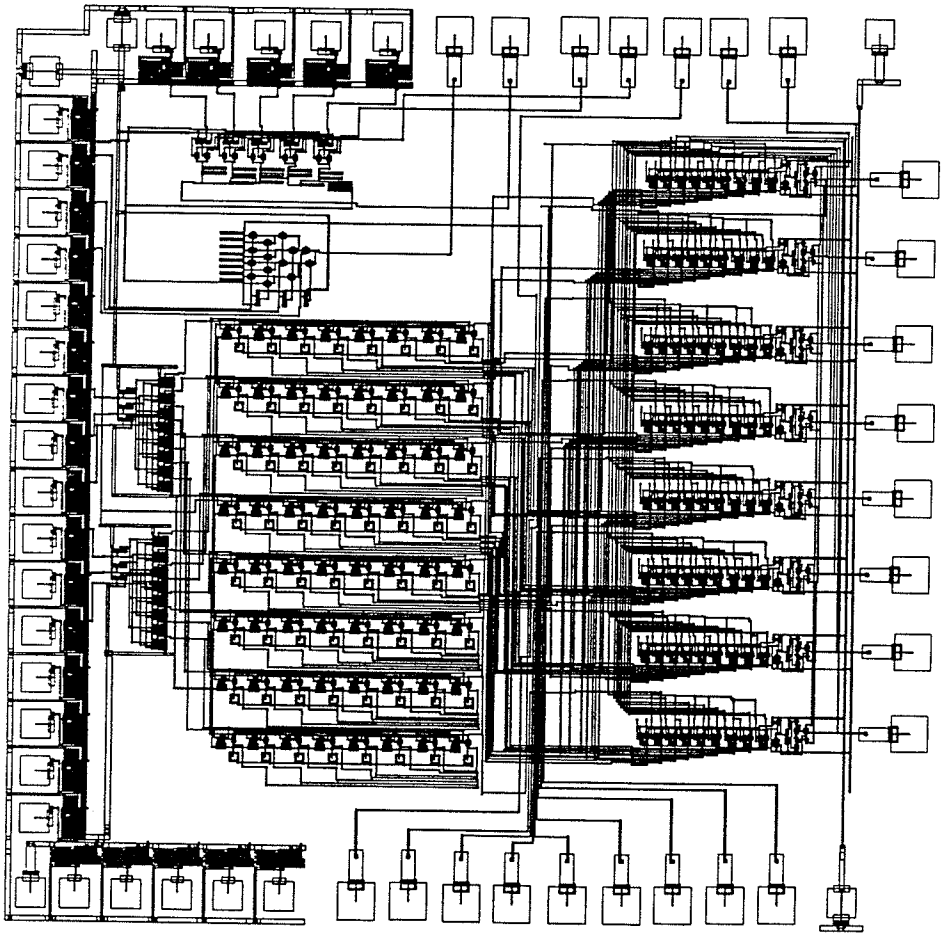

Figure C.1: 8 Neuron chip.