

VLSI IMPLEMENTATION OF OLFACTORY  
CORTEX MODEL

By

SANJAY B. PATIL

Bachelor of Science

College of Engineering

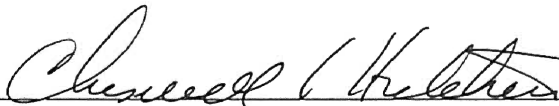
Poona, India

1987

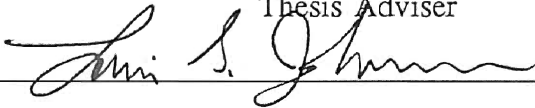
Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
MASTER OF SCIENCE  
May, 1993

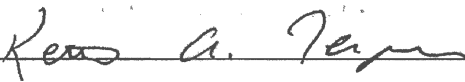
VLSI IMPLEMENTATION OF OLFACTORY  
CORTEX MODEL


Thesis Approved:

  
\_\_\_\_\_

Thesis Adviser

  
\_\_\_\_\_

  
\_\_\_\_\_

  
\_\_\_\_\_

Dean of the Graduate College

## PREFACE

This thesis attempts to implement the building blocks required for the realization of the biologically motivated olfactory neural model in silicon as the special purpose hardware. The olfactory model is originally developed by R. Granger, G. Lynch, and Ambros-Ingerson. CMOS analog integrated circuits were used for this purpose. All of the building blocks were fabricated using the MOSIS service and tested at our site. The results of this study can be used to realize a system level integration of the olfactory model.

I wish to express my gratitude to my major advisor, Dr. Chriswell Hutchens, for his guidance, inspiration, invaluable counsel, and financial support. I appreciate the endless time and effort he put in this work. I am also thankful to Dr. Louis Johnson, Dr. Teague, and Dr. Richard Cummins for serving on my committee.

I wish to thank the office of Naval Ocean System Center, San Diego, for the computing resources and financial support they provided for the project. I am also thankful to Dr. Ramesh Sharda for his guidance and support.

My special thanks goes to Dr. Patrick Shoemaker, for his invaluable assistance in this work. I would also like to acknowledge MOSIS fabrication services, for fabricating our circuits. I extend my thanks to my friends, David Born and Subbaraju Gadhiraaju for proof reading.

Finally, my deepest appreciation is extended to my parents, brother, and sister for

their love, support, moral encouragement, and understanding. This work is dedicated solely to my parents.

## TABLE OF CONTENTS

Chapter	Page
I. OLFACTION AND ELECTRONIC NEURAL NETWORKS . . . . .	1
Olfaction . . . . .	1
About Neural Networks . . . . .	3
"Abstract" Verses "Tightly Coupled" Neural Network Paradigm . .	6
Hardware Implementation of "Tightly Coupled" Computational Models: A Review . . . . .	8
Proposal for Hardware Implementation of GLA Olfactory Model . .	11
II. OLFACTORY MODEL AND ITS HARDWARE IMPLEMENTATION	14
The Bulbar-Cortical Model . . . . .	14
Olfactory Bulb . . . . .	15
Piriform Cortex . . . . .	18
Learning . . . . .	21
Multi-Sampling . . . . .	23
Hardware Implementation . . . . .	27
III. SYSTEM BUILDING BLOCKS . . . . .	34
Glomerulus Normalization . . . . .	34
AGC and Offset Combined Normalizing Function . . . . .	39
Transconductance Multiplier . . . . .	40
Simulations. . . . .	52
Testing. . . . .	54
Offset Circuit . . . . .	54
Simulations. . . . .	62
Testing. . . . .	64
Linear Limiter with AGC Normalization Function . . . . .	65
Square Law Bulb Normalization Function . . . . .	66
Approximate Sigmoidal Function . . . . .	67
Simulations. . . . .	71
Testing. . . . .	71
Mitral Patch . . . . .	75
Simulations . . . . .	78
Testing . . . . .	81

Chapter	Page
Bi-directional Voltage/Current Buffers . . . . .	81
Simulations . . . . .	88
Testing . . . . .	88
Weight Matrix . . . . .	93
Floating Gate Avalanche Injection MOS Memory . . . . .	99
Metal Nitrite Oxide Silicon Memory . . . . .	104
Dual Injector Floating Gate MOS Memory . . . . .	105
Floating Gate Analog Memory in Standard CMOS Process . . . . .	108
Memory Structure . . . . .	109
Field Enhanced Fowler-Nordheim Tunneling . . . . .	115
Programming . . . . .	116
Winner Take All . . . . .	124
Simulations . . . . .	127
Testing . . . . .	130
Tie Resolver . . . . .	133
Testing . . . . .	136
Dynamic Current Copier Integrator . . . . .	136
Background . . . . .	137
The Operating Principle of the Current Copier Integrator . . . . .	140
Circuit Design . . . . .	143
Upper Integration Limit . . . . .	143
Maximum Switching Frequency . . . . .	145
Minimum Switching Frequency . . . . .	147
Mechanisms of Errors . . . . .	149
Charge Injection . . . . .	149
Switch Feedthrough . . . . .	150
Cascode Configurations . . . . .	151
Simulations . . . . .	152
Testing . . . . .	152
Summary . . . . .	155
IV. CONCLUSIONS AND FUTURE PROSPECTS . . . . .	156
REFERENCES . . . . .	160

## LIST OF TABLES

Table	Page
I. Injector Structures . . . . .	112
II. Truth Table . . . . .	135

## LIST OF FIGURES

Figure	Page
1. Block Diagram of the Olfactory System . . . . .	16
2. Flowchart of the Multi-Sampling Process . . . . .	24
3. AGC and Offset Combined Linear Normalization Function . . . . .	36
4. Linear Limiter with AGC Normalization Function . . . . .	37
5. Square Law Normalization Function . . . . .	38
6. Transconductance Multiplier . . . . .	41
7. Demonstration of Class AB Principle with Two Transistors . . . . .	43
8. CMOS Equivalent of Single MOS Transistor . . . . .	45
9. Linear MOS Transconductor Principle . . . . .	47
10. Linear MOS Transconductor Using the MOS Equivalent Pair . . . . .	49
11. Double Pair Implementation of a Floating Voltage Source . . . . .	50
12. DC Transfer Characteristics of a Transconductance Multiplier . . . . .	53
13. Multiplier Output Voltage Obtained from Simulations . . . . .	55
14. Multiplier Output Voltage Obtained from Test Results . . . . .	56
15. Offset Summer Circuit . . . . .	57
16. Maximum Function Circuit . . . . .	60
17. Offset Summer DC Characteristics for Four Branches . . . . .	63
18. Approximate Sigmoidal Function . . . . .	68



Figure	Page
19. AC Response of Squashing Function . . . . .	72
20. DC Transfer Characteristics of Sigmoidal Function (Simulations) . . . . .	73
21. DC Transfer Characteristics of Sigmoidal Function (Test Results) . . . . .	74
22. Mitral Patch . . . . .	77
23. DC Response of Mitral Cells . . . . .	79
24. Transient Response of Mitral Cells . . . . .	80
25. Bi-directional Voltage/Current Buffers . . . . .	82
26. Block Diagram of the CC-II $\pm$ . . . . .	84
27. Bi-directional Voltage/Current Conveyors . . . . .	86
28. Transient Response of the Current Conveyor . . . . .	89
29. Current Conveyor AC Response . . . . .	90
30. DC Transfer Characteristics of the CC Obtained from Simulations . . . . .	91
31. DC Transfer Characteristics of CC Obtained from Test Results . . . . .	92
32. Weight Matrix . . . . .	95
33. Cross-section of the FAMOS Structure . . . . .	100
34. Cross-section of the DIFMOS Structure . . . . .	106
35. Electrical Equivalent Schematic of the Layout . . . . .	110
36. Test Setup for Testing Memory Cell . . . . .	117
37. Threshold Voltages of Un-programmed Devices . . . . .	119
38. Threshold Voltages of Programmed Devices . . . . .	120
39. Threshold Voltage Retention After 3 Hours . . . . .	122
40. Threshold Voltage Retention After 130 Hours . . . . .	123
41. Winner Take All Circuit . . . . .	125

Figure	Page
42. Demonstration of the Resolving WTA Inputs . . . . .	128
43. Ties in the WTA Outputs . . . . .	129
44. Effect of Mean on the Settling Time . . . . .	131
45. Effect of Difference Current on the Settling Time . . . . .	132
46. Tie Resolver . . . . .	134
47. Basic Current Copier . . . . .	139
48. Current Copier Integrator . . . . .	141
49. The P-Cell . . . . .	146
50. Transient Operation of the CCI . . . . .	153

## NOMENCLATURE

$(W/L)_x$	Width to length ratio of subscripted MOSFET $x$
$\alpha$	Coupling ratio
$\beta_x$	Transconductance parameter of subscripted MOSFET $x$
$\delta_w$	Synaptic increment of $w_{ijkl}$ per training episode
$\Theta_{ff}$	Piriform refractory frequency facilitation threshold
$\Theta_I$	Threshold to periglomerular to eliminate inhibition noise floor
$\Theta M_j$	Threshold of the $j$ th mitral cell
$\Theta_P$	Piriform cell threshold
$\lambda_x$	Channel length modulation of subscripted MOSFET $x$
$\phi_1$	System clock active in forward phase
$\phi_2$	System clock active in backward phase
$\phi_{11}$	Sub-phase of $\phi_1$ to initialize WTA
$\phi_{12}$	Sub-phase of $\phi_1$ to latch winning piriform cells into tie resolver
$\phi_{21}$	Sub-phase of $\phi_2$ to store feedback inhibition $\Gamma_i^*$
$\phi_{22}$	Sub-phase of $\phi_2$ to update feedback inhibition $\Gamma_i^*$
$A_{Vx}$	Voltage gain of subscripted operational amplifier $x$
$C_B$	Bootstrap capacitor
$C_{GB}$	Gate to body capacitance
$C_{GD}$	Gate to drain capacitance

$C_{GS}$	Gate to source capacitance
$C_{Gx}$	Gate capacitance of subscripted MOSFET
$C_{inj}$	Injector capacitor
COX'	Per unit oxide capacitance
$D_x$	Subscripted diode
$g$	Number of glomeruli in bulb patch
$G'_i$	Automatic Gain Controlled signal
$G'_{imax}$	Maximum value of element in $G'_i$
$G^*_i$	Glomerulus input (un-normalized)
$G^*_{imax}$	Maximum value of element in $G^*_i$
$g_{ds}$	Small signal drain transconductance
$G_i$	Normalized glomerulus output; mitral patch input
$gm_x$	Small signal channel transconductance of subscripted MOSFET
$g_s(x)$	Nonlinear mapping function that maps glomeruli activity
$h$	# of piriform cells per piriform patch
$I^*_i$	Aggregate un-thresholded inhibition to glomerulus $i$
$i$	Counting index for $g$
$I^*_{ij}$	Weighted inhibition on LOT line $ij$ in backward direction
$I_{Dx}$	Drain current of subscripted MOSFET $x$
$I_{FS}$	Full scale current
$I_i$	Integrated and thresholded inhibition signal into glomerulus $i$
$j$	Counting index for $m$
$k$	Counting index for $p$
$K_{1-4}$	Constants of non-linear function

$K_G$	Constant to set percentage activation of the glomerulus
$l$	Counting index for h
$m$	Number of mitral cells per glomerulus
$M_{ij}$	LOT from jth mitral cell of the ith glomerulus
$M_x$	Subscripted MOSFET
$O_i$	Olfactory sensor output; olfactory system input
$p$	Number of piriform patches in cortex patch
$P_{kl}^*$	Output of the weight matrix, cortex patch input to lth piriform cell of kth piriform patch
$P_{kl}$	Output of cortex patch from lth piriform cell in kth piriform patch
$PW_{kl}$	Winning cortex output; olfactory output
$V_{DD}$	Positive supply voltage
$V_{DS}$	Drain to source voltage of subscripted MOSFET x
$V_E$	Erasing voltage
$V_{GSx}$	Gate to source voltage of subscripted MOSFET x
$V_K$	Normalization scaling constant
$V_P$	Programming voltage
$V_{SS}$	Negative supply voltage
$V_{tun}$	Tunneling voltage
$V_{Tx}$	Threshold voltage of subscripted MOSFET x
$W$	Weight matrix
$w_{ijkl}$	Synaptic weight from LOT $M_{ij}$ to the piriform cell $P_{kl}$
$w_{max}$	Maximum value of the synaptic weight $w_{ijkl}$
$W^T$	Transpose of the weight matrix

CHAPTER I  
OLFACTION AND ELECTRONIC  
NEURAL NETWORKS

Olfaction

The current research surge in neural networks (NN) falls into essentially three broad categories. The first category is that of the mathematical description and analysis of the learning properties of neural networks, often working from biological and physiological exemplars [1,2]. The second, and perhaps the largest, research effort uses computer simulations to verify the validity of the neural network models in addition to demonstrating their applications [3,4]. Since the publication of John Hopfield's paper [5] on the prospect of compact and dense hardware implementation of neural networks in analog integrated circuit form, a third group of research topics, into which this thesis falls, has emerged. The researchers in this category attempt to implement neural networks in LSI/VLSI hardware [6,7,8,9,10,11].

The theory of biological neuron and the actual neural processing within the brain are complex and involved [12]. The physical and chemical processes in the nerve cells that are responsible for learning and memory are beginning to yield to experimental study by physiologists and anatomists. By and large, biological neural nets exhibit massive parallelism and parallel processing. The modulation of synaptic junctions has long been

regarded as the likely mechanism for learning and memory [13]. The long term potentiation (LTP) that is observed in the hippocampus, limbic system, and in some cortical structures of the brain, is believed to be similar to the mechanism used for learning [14]. The changes in the synaptic strength due to LTP are rather coarse compared to precise and graded weight changes that are offered by artificial neural networks. How a nervous system might respond to the computationally limited neural learning and neural processing that is used by artificial neural networks due to two dimensional connectivity [15], is a question. Extensive research is being carried out using computer simulations on such abstract neural network models to understand the effects of incorporated artificiality and also in an attempt to elucidate the organizational principle at the system level [1,2].

R. Granger, G. Lynch, and Ambros-Ingerson of U. C. Irvine have reported a potentially useful model, referred to as the GLA model henceforth, for the operation of the interacting neural networks of the olfactory bulb and piriform cortex that has been observed in rats [16,17,18]. Computer simulations of this model have demonstrated interesting computational properties, such as, (1) the ability to perform the hierarchical clustering of the input cues (odors) presented by the pattern of activity on the input lines from the olfactory receptors, (2) the extensibility to unsupervised learning, and 3) the ability to detect weaker stimuli when masked by a stronger one.

A central feature of this model is the periodic sampling of stimuli at the so-called theta rhythm, to which the network response is locked. Hierarchical clustering and unmasking operations proceed sequentially with successive sampling (sniffs) of the inputs. For example, if you are serious gardener, on the first sniff you might get a

response indicating the odor of flower, on the second a rose, and finally on the third, an "Oklahoma Orange Spirit."

The goal of this research work is to develop a simplified electronic realization of GLA olfactory model suitable for analog implementation in bulk CMOS circuitry. This realization will retain the essential clustering properties of the olfactory bulb (OB) and paleocortex. The dominance of the theta rhythm in GLA model suggests the suitability of the synchronous or clocked approach, but the actual computation between the two clock cycles is analog, asynchronous, and carried out in parallel. In the GLA model, categorization in the paleocortex is done through an iterative procedure of sniffs, usually less than 5, with each sniff leading to a specific clustered solution down in the hierarchy.

#### About Neural Networks

Before the problem is presented, some review of the basic concepts of neural networks may be useful. Even though, at the later stages of this chapter, we have tried to subtly distinguish the GLA olfactory model from the traditional "abstract" neural networks (since the olfactory model more closely mimics the nervous system), GLA model still uses many of the concepts exploited by abstract artificial neural networks. The interested reader is referred to the work of Patri K. Simpson [19] for history and more details of artificial neural networks. This section gives the reader a brief review of neural networks.

Artificial neural networks (ANN's) go by many different names, such as connectionist models, parallel distributed-processing models, or neuro-computers. The structure of artificial neural networks is based on the present understanding of the



biological nervous system. ANN's provide an alternative form of computation that attempts to mimic the neurophysiological functions. ANN's are composed of many nonlinear computational elements. These computational elements operate in parallel and arranged in patterns reminiscent of biological neural networks. Elements are connected via densely connected weights. Weights are typically adapted during use (learning) [3]. The information is held in these weights. The new information is captured by changing the strength of the connection. Contrary to Von Neumann's computer, which processes instructions sequentially, neural network models explore many hypothesis simultaneously using their massive parallel structures.

In its simplest form, a neuron sums weighted inputs and passes the result through a non-linearity. The neuron is characterized by an internal threshold or offset and by the type of non-linearity. The various types of non-linearities are hard and soft limiters, sigmoidal logistic non-linearities, and hyperbolic tangents [3]. The hyperbolic tangent is similar in shape to the logistic function. It is often used by biologists as a mathematical model of nerve-cell activation ( $OUT = \tanh(x)$ ). The most commonly used non-linearity is the sigmoidal logistic which is continuously differentiable.

Based on existing results, most neural networks adapt the connection weights over time to improve network performance. Adaption or learning is a major interest area of neural net research. An example of such adaption is speech recognition, where training data is limited. The new speakers, words, dialects, phrases, and contexts are continuously encountered. Traditional statistical techniques are not adaptive. They typically process all training data simultaneously before being used with new data. Neural net classifiers are non-parametric and they make weaker assumptions concerning the shapes of

underlying distributions than traditional statistical classifiers do. Such neural network adaptive systems are often described by energy functions and/or probability distributions.

The discussed neuron and the neural processing is a simplified version of biological neuron and neural processing. The biological neuron consists of a cell body called soma and an axon or nerve fiber that connects the cells to each other [20]. The junctions between neurons occur either on the cell body or on spin-like extensions of the cell body called dendrites. These junctions are referred to as synapses. Nerves and dendrites can be viewed as insulated conductors used for transmitting electrochemical signals to neurons. In the human nervous system, about  $10^{11}$  neurons participate in perhaps  $10^{12}$  interconnections over a transmission path that may range for a meter or more [20].

The neuron processing times are larger compared to today's advanced computer cycle times. The cycle time is the time taken to process a single piece of information from input to output. The cycle time of most advanced computers corresponding to one clock cycle for the CPU is on the order of 1 nanosecond. The average cycle time for a neuron in the brain is 2 milliseconds. The difference in speed is  $2 \times 10^6$ , yet due to brain's parallel nature, the brain is more time efficient than conventional computers.

Neural network models offer their greatest potential in areas such as speech processing, image recognition, and pattern classification. In such applications, many hypothesis are pursued in parallel, high fault tolerant computation rates are required, and the existing computer systems are far from equaling human performance. When compared to traditional computing methods, the benefits of neural networks extend beyond the high computation rates provided by massive parallelism. Degree of robustness or fault tolerance provided by neural networks is greater than fault tolerance provided by

Von-Neumann sequential computers. Because of the many processing elements, damage to a few neurons and synapses does not significantly impair overall performance. Like humans, trained neural networks recognize partial input information.

## "Abstract" Verses "Tightly Coupled"

### Neural Network Paradigms

This section focuses on distinguishing biologically coupled neural network paradigms (i.e. olfactory) from so-called traditional "abstract" neural networks. We believe that the subcategory of biologically mimicked or tightly coupled neural networks is necessary to highlight behavioral features and functions. Tightly coupled neural networks are significantly different in treatment when compared to many widely used abstract neural networks.

Artificial neural networks may be classified according to learning algorithms, topologies, and node characteristics. Another factor that might be of paramount importance is the degree of biological plausibility of the network in question. Adaption or learning is the major process in neural networks and thus forms an important criteria for classification. Most neural networks adapt connection weights over time to improve performance. A number of widely used neural network paradigms feasible for parallel implementation are based on adaptations of conventional statistical and numerical techniques [1,21,22]. These neural paradigms are non-parametric. They make weaker assumptions concerning the shapes of underlying distributions than traditional statistical classifiers do. Such adaptive systems are described by their energy function and probability distribution. Examples of such networks are traditional, layered, and heavily

interconnected feed-forward architectures such as the multi-layer perceptron with back propagation learning [1], vector quantization [21], and probabilistic neural networks [22] i.e. Boltzmann machine. The reciprocally and symmetrically interconnected architectures described by Hopfield [5] and Boltzmann machine [23] are examples of physical systems. All of these networks [1,5,21,22,23] can be categorized as abstract neural networks. Abstract neural networks attempt to emulate the functionality of the brain and intelligence within it. These networks seem to be more heavily influenced by the underlying statistical distributions rather than being truly inspired by a straight forward one to one biological processing. The biological neural mechanisms at the neuron and synaptic level are considerably more involved and complex than those modeled by most widely used "abstract" neural networks. It is difficult to conclude which of the biological mechanisms to retain in the interest of computational efficiency. This is partially due to poor understanding of neural theories which are in part due to the extreme experimental difficulties encountered in biological network neuroscience. The answer depends upon whether one wishes to develop artificial computational models or to understand neurobiology. Our goal is certainly the former. However, we also believe that modeling a considerable part of the biological machinery is helpful in creating thinking machines. The theory that we wish to emphasize is that present artificial network models are too abstract to retain the computational efficiencies that are present in the biological world. Therefore in summary, we view the broad spectrum of neural networks models as spanning from "abstract" (perceptron) to the loosely coupled (Kohonen) to the more closely emulated (Grossberg) to the tightly coupled GLA olfactory model.

Thus, by way of contrast, the tightly coupled neural network models bear a the

straight forward structural relationship to a specific neural function within a nervous system. Tightly coupled neural networks are the subject of much current interest [24,25,26,37]. In this class, unfortunately, understanding of the collective function of neural networks in vertebrates is largely limited to sensory structures i.e. early processing. The sensory functions have been studied in the greatest depth and with most success. It appears that while most artificial neural networks are typically comprised of a densely connected layered network of simple neurons, tightly coupled networks employ sparsely connected networks of much more elaborate neurons in which substantial information processing occurs within a single neuron. The GLA olfactory paradigm is most certainly inspired by tightly coupled neural network philosophy. However, substantial biological complexity in these cases also is a result of constrained molecular properties (i.e. channel membrane transport). Therefore, ultimately a certain amount of abstraction (determined by application) must be justified in order to build silicon hardware models.

### Hardware Implementation of "Tightly Coupled"

#### Computational Models: A Review

Now that the particular subcategory (tightly coupled neural networks) under which the olfaction problem is to be studied is defined, the following text will review what type of the tightly coupled neural networks have been implemented in hardware, before the hardware implementation of our olfaction model is proposed. In spite of a substantially different (straight forward tightly coupled) computational approach from most abstract neural networks, the essential technologies and hardware techniques remain the same in both cases. The interested reader is referred to the extensive literature review done by

John Wagnon [27] for details on various abstract neural systems that have been implemented to date in the hardware.

The literature search connected with the hardware implementation of tightly coupled neural networks yielded only two papers detailing problems with the software simulations of olfaction [16,26]. To date, no parallel hardware implementation of olfaction has been reported. J. Bailey and D. Hammerstrom [28] have proposed the serial implementation of the GLA olfactory model. However, some researchers, most notably Carver Mead, have attempted to build silicon models of a biologically plausible early processing structures for sensory inputs [29,30,31].

Usually, the required real time auditory signal processing burden is too high to be handled by artificial speech recognition systems due to computational limitations. Computationally efficient special purpose hardware in analog integrated VLSI circuitry can be used to handle the large signal processing burden, thus forming an efficient solution to the problem of computational limitation. Carver Mead and co-workers have reported a working analog VLSI chip that implements a stereausis model of biological early auditory processing in the brain [32]. The chip essentially is an artificial cochlea that analyzes a sound wave and detects a fundamental note missing from the harmony. The binaural information exploited by the stereausis algorithm improves speech intelligibility in noisy environments compared to the monaural audio signal processing exhibited by most artificial speech recognition systems due to their computational limitations. The chip is based on the stereausis model of biological auditory processing that encodes bi-neural cross-correlation and spectral auto-correlation information by deriving a two dimensional representation of binaural sound waves from two sound inputs

(ears). Their algorithm has also demonstrated the ability to naturally segment monaural signals into distinct spectral regions. Although, to some extent the responses are found sensitive to the noise in the data, output patterns have demonstrated feature extraction capability for speech signals, specifically the spectral information of various sound waves and their location. The chip is comprised of 10,000 transistors using two micron analog CMOS technology and fabricated from MOSIS.

An interesting real time hardware implementation of vertebrate retinal inhibitory behavior has recently been presented by Mead and Mahowald [33]. The processing relies on the lateral inhibition to adapt the system to a wide ranges of viewing conditions, and to produce an output that is independent of the absolute illumination level. Such processing is a direct result of initial *inhibitory* analog stage in retinal processing. The secondary effect of the lateral inhibition mechanism is enhancement of spatial edges in the image. Their silicon model implements the first stage of retinal processing on a single chip where the logarithm of the incident light is computed by a photoreceptor. The output of a photoreceptor is further spatially smoothed by a resistive network (grid). The amplitude difference between photoreceptor output and its smoothed counterpart is amplified to form a second order spatial filter. They have performed the experiments on a 48x48 array of silicon pixels on one quarter of a square centimeter chip in CMOS technology. Compared to the entire biological visual system, even though the system is realized at very low level, it creates the true biological representation upon which higher level processing stages can be built. The mathematical analysis of the network is presented by J. G. Taylor [34] which allows the extension of the results to a general class of resistive grids and inhibitory feedbacks.

Along similar lines, a neural network approach to the color consistency problem has been reported [35]. Color consistency is the ability to judge the reflectance of an object under different illumination conditions, since illumination elucidates the same object under different lighting conditions. The system is based on the Land's retinex theory. Land's retinex theory is inspired by mammalian neurobiology and human psychophysics [35]. This algorithm models our ability to see colors intensities roughly constant as light varies. Their computer simulations have confirmed validity of their implementation of the Land's model. They have implemented Land's algorithm in subthreshold analog CMOS VLSI using a two micron process from MOSIS. The chip is comprised of about 60,000 transistors and is reported to be operative at video rates.

H. C. Card and W. R. Moore report that the learning and memory behavior at neuron and synaptic levels can be best understood in simple invertebrate animals such as worms and insects [36]. To demonstrate neuron and synaptic level memory behavior, they chose the well-studied marine specimen, mollusc *Aplysia*. These small animals exhibit the same learning patterns as vertebrates while keeping neural processing relatively simple. They have also proposed analog CMOS circuitry that explores both; the associative and non-associative learning mechanisms of habituation and sensitization in *Aplysia*.

### Proposal for Hardware Implementation of GLA Olfactory Model

We propose the simplified hardware implementation of GLA olfactory model in a two micron, p-well, double poly, double metal bulk CMOS process from MOSIS. The implementation will retain the essential clustering properties of the GLA olfactory model.



The GLA model inherently possess many favorable features to aid simple hardware implementation. These features are: (1) mixed mode processing instead of a pure analog, (2) current and voltage mode processing, (3) rhythmic clocking for synchronization, 4) discrete, course, and unidirectional learning leading to simplified learning algorithm, and 5) single quadrant multiplication instead of a four quadrant multiplication to obtain scaling closer to that exhibited by a synapse.

The electronic implementation of the GLA olfactory model involves the integration of various mathematical functions in silicon as integrated sub-components. The system level GLA olfactory architecture can best be realized by developing such mathematical functions separately in the form of building blocks to allow for simplified testing, and then incorporating all such blocks onto a single substrate. This thesis specifically addresses the design, simulation, layout, fabrication, and testing of these basic building blocks. The system level realization is beyond the scope of this thesis.

As opposed to traditional voltage mode analog signal processing, in which inherently current signals are transferred to the voltage domain before any analog signal processing takes place, the current mode analog signal processing approach is taken here. The use of current rather than voltage as an active parameter can result in higher gain, accuracy, and wider bandwidth due to the reduced voltage excursion at dynamic nodes [38].

Simulations are performed using SPICE on a personal computer. Layouts are accomplished by using MAGIC on Sun work-stations. All circuits are fabricated using fabrication service from MOSIS. Finally, the testing is performed.

The following, II chapter, will describe in detail our interpretation of GLA olfactory model and proposed electronic implementation. Chapter III will focus on the design,

simulation, and testing of all building blocks. Chapter IV will offer conclusions based on the results and suggestions for the future work connected with investigation of the this proposed hardware implementation of olfactory model.

CHAPTER II  
OLFACTORY MODEL AND ITS HARDWARE  
IMPLEMENTATION

The Bulbar-Cortical Model

Modeling of olfaction is a difficult task since olfaction theories are still in the developmental stages. On the one hand, a computer simulation of a too detailed anatomical olfactory model may result in large volumes of hard to analyze data, while on the other hand, too much abstraction and simplification of the anatomical olfactory model may altogether lose its relevance to biology with a potential loss of computational power for the anatomical model. Thus, the efforts towards the development of the moderately abstracted olfactory model is necessary. Such a model helps to understand the model as well as preserves the essential features of the model. A moderately abstracted mid level [17] GLA olfactory model has been proposed by Granger, Lynch, and Ambros-Ingerson. The interested reader is referred to the work of Granger et al. for details [16,17,18]. In this section, we will focus on our interpretation of the essential features of the GLA model, leading to our simplified olfactory architecture suitable for a proposed hardware implementation. Throughout the course of the discussion, we will justify various assumptions and simplifications which are essential to keeping the implementation simple yet practical. These assumptions have resulted in a slightly modified architecture.

Our architecture of the bulbar-cortical (BC) is shown in the Figure 1. The model basically consists of the olfactory bulb (OB) and the piriform cortex (PC). The olfactory nomenclature is given in the preliminaries.

### Olfactory Bulb

The olfactory bulb receives the input via the olfactory nerve (ON). Olfactory nerves originate from the olfactory receptor sheet and project onto the periglomerular in a topographic fashion. The receptor cells, which are most responsive to the particular chemical stimuli, project their axons to a delimited area of the olfactory bulb referred to as the glomerulus. The receptor cells fire with higher frequency for higher concentrations of odorant. The concentration of the odorant is modeled by the magnitude of a real positive number. This number reflects aggregate firing frequency and represents ON input to the corresponding glomerulus.

The olfactory bulb is organized into a number of glomeruli  $g$ . Each glomeruli consists of  $m$  mitral/tufted cells. Each glomerulus receives excitatory input from an ON collectively forming system input vector  $O_i$ . It also receives inhibitory input vector  $I_i$  from the PC feedback through weights which are set during the developmental period to be discussed later. The excitatory inputs are summed with the inhibitory feedback signal forming the net un-normalized input activity  $G_i^*$  to the glomerulus. This un-normalized glomerulus activity is given by:

$$G_i^* = \max(O_i - I_i, 0) \quad (1)$$

The resulting net inputs are then subjected to non-linear, scaled, and global

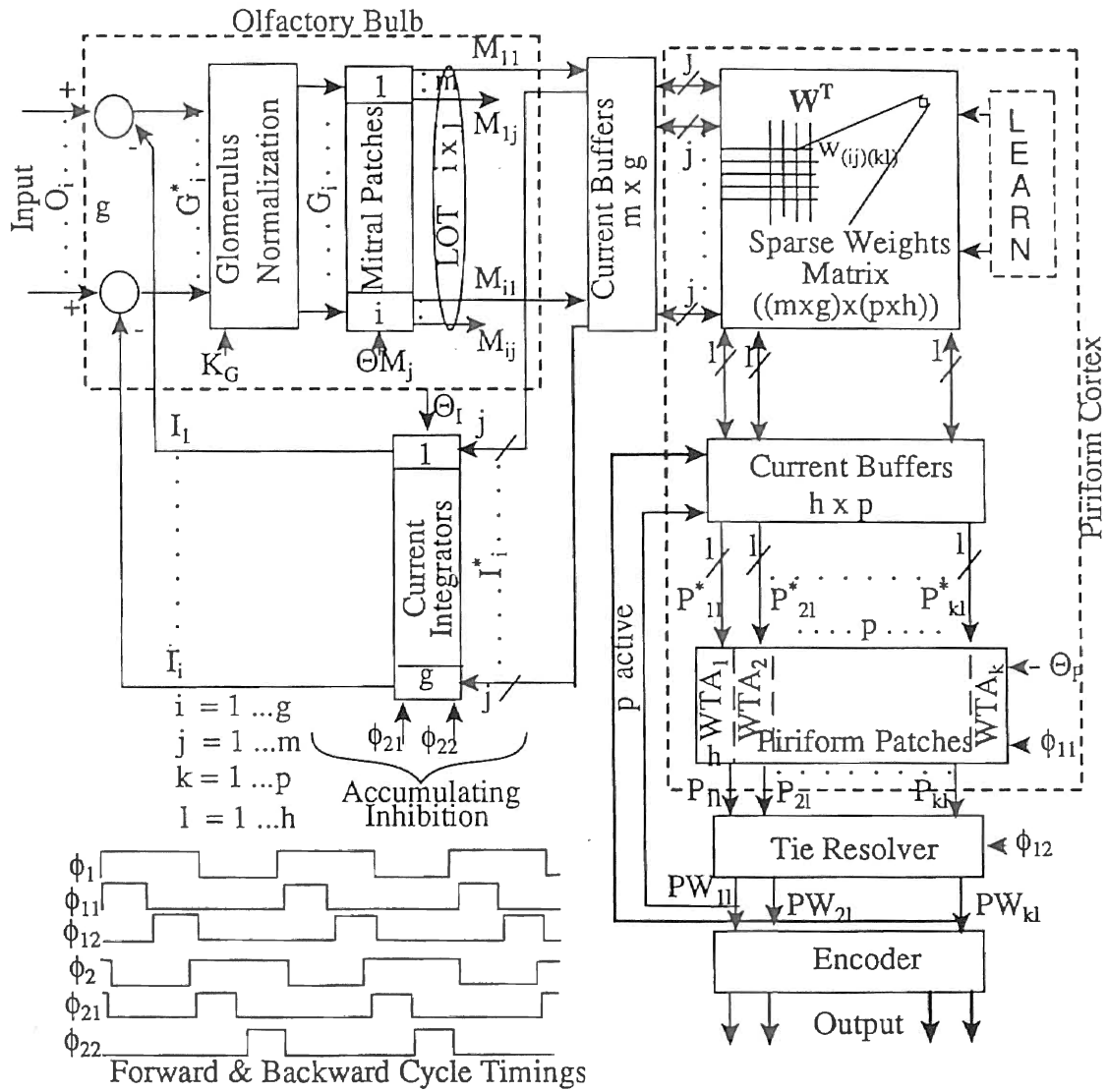


Figure 1. Block Diagram of the Olfactory System

normalization mediated by the interaction between the excitatory and inhibitory cells of OB. This serves to normalize the output of the bulb by keeping the total number of mitral cells that are activated constant across the stimuli for different intensities and compositions. In this normalization scheme, the sum of the normalized glomeruli activity is maintained at a constant level. The normalization is obtained in such a way that the sum of the non-linearly mapped and scaled normalized activity remains nearly constant. Mathematically, the normalized glomerulus activity is given by:

$$G_i = g_s(V_K G^*_i) \quad (2)$$

In the above equation, the scaling constant,  $V_K$ , is the smallest positive value that satisfies

$$\sum_{i=1}^g g_s(V_K G^*_i) = K_G \quad (3)$$

where  $K_G$  is the glomerulus activity constant and  $g_s(\cdot)$  is a non-linear mapping function that maps glomeruli activity into number of the activated mitral cells. Mathematically,

$$g_s(x) = K_1 \left[ K_2 K_3 x - \frac{(K_3 x)^2}{2} \right] (1 + K_3 K_4 x) \quad \text{for } K_2 \geq K_3 x$$

$$= \frac{K_1 K_2^2}{2} (1 + K_3 K_4 x) \quad \text{for } K_2 \leq K_3 x \quad (4)$$

where  $K_1$ ,  $K_2$ ,  $K_3$ , and  $K_4$  are circuit constants. The variable,  $x$ , is the input to non-linear mapping function.

The intensity of the normalized glomerulus activity  $G_i$  is linearly reflected in the number of mitral cells within the particular glomeruli that it activates. The mitral cells have increasing thresholds,  $\Theta M_j < \Theta M_{j+1}$  ( $0 \leq j \leq m$ ), where  $\Theta M_j$  is the activation threshold of the  $j$ th mitral cell in a glomerulus. Thus an increasing amount of normalized

glomerulus activation results in a greater number of mitral cells being fired. The mitral cells are modeled as two state devices (active or inactive) or McCullough-Pitts neurons which are either high (logic 1) or low (logic 0) with glomerulus activity either above or below its threshold respectively. Mathematically,

$$M_j = \begin{cases} 1 & \text{if } G_j \geq \theta M_j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Thus, the overall processing within the olfactory bulb in the absence of inhibitory feedback is as follows. The nonlinear normalization and the constraint on the total normalized glomeruli activity results in the accentuation of insignificant components in the odorant while attenuation of stronger components, which is intuitively pleasing. The normalized glomerulus activity is spatially thermometer-coded via the mitral patch where the increasing amount of normalized glomerulus activity is reflected in greater number of mitral cells being triggered. The constant level of glomerulus output activity serves to keep the total number of mitral cells that are activated reasonably constant across stimuli with different concentrations and compositions. This means that even though the same odorant at different concentrations results in different activations of the same glomeruli, the normalization can result in an identical thermometer code. This makes GLA model insensitive to odor concentration, i.e.,  $O_i$  amplitudes.

### Piriform Cortex

The main features of the piriform cortex are the sparse, and forward projections of the mitral cells onto piriform cells [17] via the lateral olfactory tracts (LOT), and backward inhibition feedback to the OB.

The outputs of the mitral cells in the OB,  $M_{ij}$ , are projected on to the piriform cells in the piriform cortex via the LOT lines, forming a connection matrix between the OB and the PC in layer Ia of the PC. The excitory synapses  $W_{(ij)(kl)}$  in the piriform cortex are sparse, meaning synapses are made at random with a sparseness on the order of 10%. In our model, we assume a uniform distribution of synapses. However, in the GLA model, the sparsity decreases (tapering) as one travels from the rostral to the caudal region of the piriform cortex [17], i.e., from closer to OB to further away. Further, our model does not consider synapses that are present from PC to PC (thickening synapses) in layer Ib of PC, which are present in the GLA model. This assumption was necessary to simplify the winner take all (WTA) structure leading to a saving in the silicon area.

The excitory piriform cells  $P_{kl}$  are arranged into  $p$  disjoint piriform patches with  $h$  piriform cells per patch. The indice  $k$  indicates the patch while indice  $l$  indicates the cell number within the piriform patch. The total input activation to the piriform cell is

$$P^*_{kl} = \sum_{i=1}^g \sum_{j=1}^m M_{ij} W_{(ij)(kl)} \quad (6)$$

At each operating cycle, due to the strong local inhibition, the piriform patches exhibit a winner take all competition within a patch, which results in only the strongest or few near strongly activated piriform cells to fire, while the rest of the piriform cells remain quiescent. The winner take all competition is exhibited due to the presence of inhibitory interneuron within the layer II (the stellate cells). Stellate cells are activated by the most strongly activated piriform cell, producing strong local inhibition to all other piriform cells except the strongly activated piriform cell within the patch. Thus, the strongly activated piriform cell tries to become more activated while the activation of the



other piriform cells is suppressed. The piriform patch compartments makes this event local and thus competitive by stronger local inhibition.

The winning piriform cell is declared activated only if the corresponding input activation to the piriform cell is equal to or greater than a fixed piriform cell threshold  $\Theta_p$ . The output of the piriform cell is given by:

$$PW_{kl} = \begin{cases} 1 & \text{if } P^*_{kl} \geq \Theta_p, \text{ and } P^*_{kl} \geq P^*_{kj} \text{ for all } 1 \leq j \leq h \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The GLA model states that, in addition to a fixed threshold, piriform cells also have frequency facilitation (ff) and refractory states. In the ff state, the  $ff_{kl}$  is increased or decreased by one every time a cell activates or remains quiescent. The refractory state starts when  $ff_{kl}$  exceeds the threshold  $\Theta_{fr}$ . The refractory state of previously active piriform cell ( $\Theta_{fr}=1$ ) assures distinct piriform bulbar output code in each minor cycle. The non-refractory restriction to winning ensures that the piriform cells which won in the previous cycle will never win in the present cycle. However, our model does not implement ff and refractory states.

Finally, the output pattern formed by the winning piriform cells is regarded as the spatially encoded output of the bulbar-cortical system. Intuitively, it is clear that these winning piriform cells happen to have a relatively large number of their synapses from the active mitral cells.

The glomeruli in the OB are enervated by the inhibitory feedback generated by the winning piriform cells of the PC. This inhibition is weighted through synapses. Synapses are modulated over different input cues during the developmental period according to a correctional or Hebb rule. Feedback inhibits those glomeruli which are most responsible

for firing the corresponding winning piriform cells, thus attempting to deactivate those winning piriform cells which had generated the output of PC competitions in the previous cycle. The weighted inhibition on LOT line  $ij$  in the backward direction is given by:

$$I^*_{ij} = \sum_{k=1}^p \sum_{l=1}^h PW_{kl} W_{(kl)(ij)} \quad (8)$$

Inhibition on consecutive  $m$  LOT lines in the backward direction (from where the respective forward LOT lines were originated) is summed by grouping them together forming aggregate un-thresholded inhibition  $I^*_i$  to the glomeruli. Inhibition is given by:

$$I^*_i = \sum_{j=1}^m I^*_{ij} \quad \text{for } i=1, \dots, g \quad (9)$$

The feedback inhibitory signal into glomerulus is obtained by thresholding on  $\Theta_I$  as

$$I_i = \sum_y I^*_i \quad \text{where,} \quad (10)$$

$$I^*_i = \begin{cases} I^*_i & \text{if } I^*_i \geq \Theta_I \\ =0 & \text{otherwise} \end{cases}$$

where,  $y$  is the indice for each minor clustering cycle.

### Learning

Only one type of learning mechanism has been modeled. After long term potentiation (LTP), the active synapses  $W_{(ij)(kl)}$  project from active mitral cells in OB onto the winning piriform cells in PC. The weight matrix  $W$  consists of such a sparsely placed synapses. The learning involved in these synapses is referred to as adult plasticity. The weights of these synapses are non-decremental, incremented in discrete steps  $\delta w$  ( $\sim 10\%$  of their maximum weight), and saturated beyond maximum value  $w_{\max}$  ( $\sim$  two to three

times their naive weights). Mathematically, learning can be described as

$$W_{(y)(kl)} = \begin{cases} \min[(W_{(y)(kl)} + \delta_w w_{\max})] & \text{if } W_{(y)(kl)} \neq 0, \text{ and } M_{ij} > 0, \text{ and } P_{kl} > 0 \\ W_{(y)(kl)} & \text{otherwise} \end{cases} \quad (11)$$

From the above equations it is clear that the synaptic alterations take place only in physically existing synapses and only if pre and post synaptic sites are active.

In our model, since we do not implement synapses that are present from PC to PC (thickening) of layer Ib of GLA model, we do not implement the learning associated with these synapses. The learning involved in these synapses is similar to adult plasticity described above [13].

The anatomical model calls for a distinct forward path from OB to PC and a feedback path from PC to OB. Forward excitatory synapses (adult) are trained according to the rules given by equation 11 and backward inhibitory synapses are trained by a correlative Hebb rule [] during the developmental phase prior to its use for actual hierarchical clustering. To facilitate area efficient electronic implementation, at this stage we propose common adult and developmental plasticities. Common adult and developmental plasticity allows use of a single time multiplexed weight matrix W in feed forward and backward cycles.

In the feedback path, the active synapses projected from winning piriform cells in the PC onto the glomerulus in the OB are strengthened over different input cues during the simulated developmental period. For each input sample on the olfactory nerve (ON), feedback synapses projected on the glomeruli and co-activated by both, the ON input and the piriform feedback are strengthened while the remaining synapses are unchanged. However, since in any particular feedback path, feedback correlations arise as a direct

consequence of the given connectivity and strength of the forward excitory synapses in the corresponding column of the weight matrix  $W$ , the same effect can be obtained by using the transpose  $W^T$  of weight matrix  $W$  to compute bulbar inhibition. Architecturally, this implies that a single weight matrix with time multiplexing can be used to compute the weighted excitory bulbar input to the PC in the forward phase, followed by weighted inhibitory feedback from winning piriform cells to the OB in the backward phase resulting in improved area efficiency. Equation 8 gives the glomerulus inhibition using  $W^T$ .

### Multi-Sampling

The computational properties of the coordinated operation of the entire bulbar-cortical structure can best be described by a so-called multi-sampling process. The flowchart of multi-sampling process is shown in Figure 2. It is observed that activity in the various brain regions of small mammals is synchronized to their sniffing rate at the so-called theta rhythm (4-5 Hz,  $\sim$  200 ms) [17]. The GLA model states that the role of the theta rhythm is for synchronization. This eliminates the potential for oscillations due to feedback. Such a synchronization permits the entire OB to operate in rhythmic synchronization with the brain, where upon reaching the thresholds, the mitral/tufted cells fire in synchrony at the theta clock. The input to the piriform cortex arises due to the synchronous bursting of the mitral cells, yielding to the cyclic activity of the reciprocal process of feed-forward excitation of the PC by the OB followed by feedback inhibition of the OB by the PC at the theta rhythm.

As the animal sniffs a single odor, the following sequence of events takes place in

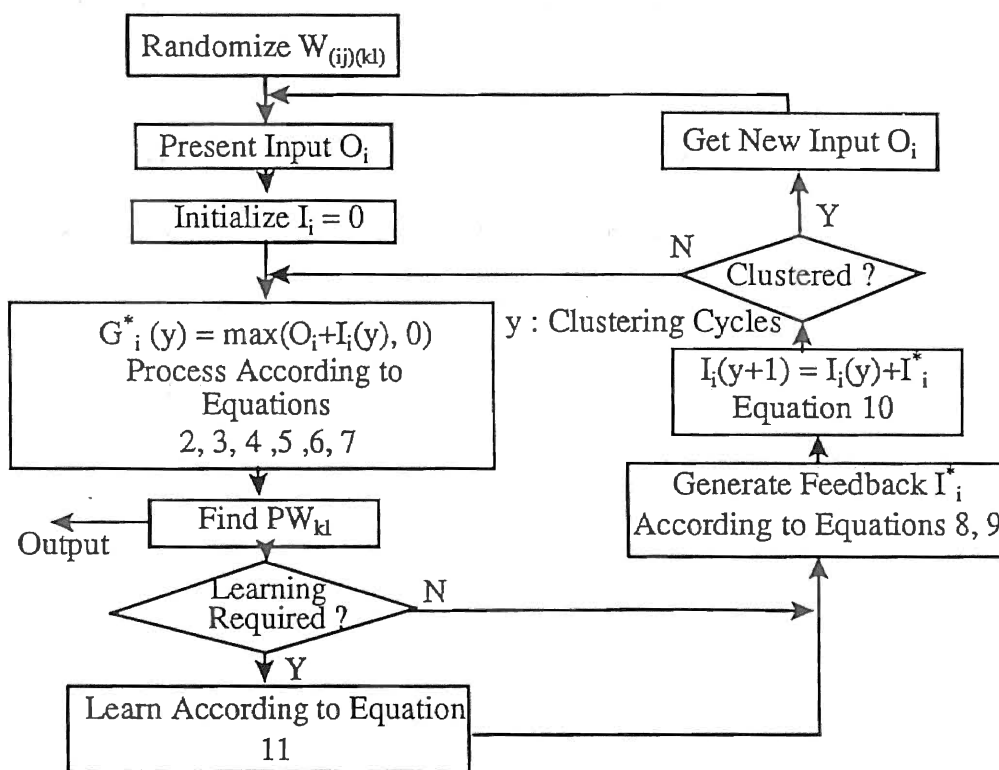


Figure 2. Flowchart of the Multisampling Process

the naive network. After the first sniff ( cycle 1), depending on the input composition, the OB output triggers the most active piriform cell in each patch of the PC based on the discussed operating rules and random connectivity. The winning piriform cells in the PC produce a feedback signal to the OB. Once the feedback signal from the PC crosses the feedback threshold  $\Theta_p$ , the glomerulus with the most significant input components are strongly inhibited for remaining cycles via the phenomena called "long lasting inhibition" that is observed in the OB. In the subsequent sniffs (cycles 2,3,..y), the normalized activity of the uninhibited glomeruli increases (according to the normalization property which attempts to keep total glomerulus activity at a constant level) in order to compensate for the inhibition of the strongest components in the previous cycle. This allows weaker components in the input vector to be expressed. As a result, the thermometer code or the spatial pattern of the mitral cells activity differs significantly from the spatial patterns in the previous cycles. Mitral cells associated with the glomeruli that are now inhibited do not fire while a larger number of mitral cells fire from glomerulus whose normalized activity has been increased. A different patterns of activation from the bulb at each step assures a distinct bulbar-cortical output codes.

The process of obtaining distinct cortical responses by successively inhibiting components of the original stimuli is referred to as multi-sampling. This multi-sampling process is repeated until the bulb is sufficiently inhibited to be largely quiescent, meaning every component in the input stimuli, no matter how weak it is, is given a chance to be expressed in the hierarchical clustering process.

The naive network can be trained on a training set containing noiseless versions of selected dissimilar odorant (vectors) according to the learning rules discussed above. The

flowchart of the learning process is shown in Figure 2. The effect of learning in the network, is to cluster essentially random PC responses into nearly equivalent estimates of the input vectors. These vectors are sufficiently close in space to the ones used in the training set. Thus, learning develops the ability in the network to cluster the sufficiently close input vectors. The similar vectors in the training set form one category while dissimilar vectors form distinct categories giving rise to a new class every time an input vector is found dissimilar to all other vectors. Any input vector similar to some vectors in the training set is accommodated in that category whereas a novel vector dissimilar to all vectors in training set gives rise to a new category.

A performance comparison study [17] of untrained and trained networks for dissimilar noisy odors concludes that the trained network enhances the overall overlap of patterns obtained for noisy instances of the same odorant. Also, it reduces the overlap of all pairs of patterns obtained for noisy instances of a different odorant. This indicates an inclination towards accommodating all noisy instances of the original odor under the same category. It also indicates that after training, the set of cortical responses are largely distinct.

After the first cycle, the overlap between the sequence of the cortical responses in the subsequent cycles becomes progressively lower for different cues, increasingly distinguishing a given input cue and thus producing a unique encoding for an individual odorant. During the first cycle, the network responses are nearly identical to the input cues which are sufficiently close in space, grouping them together in a sub-cluster. At the same time, it maintains extremely low overlap between two sub-clusters such that during the second cycle, responses are nearly identical for the members of the sub-clusters

while different responses for the vectors that were not the members of the sub-clusters. The responses in the third cycle are nearly unique producing unique encoding for individuals. Thus, during the multi-sampling process, a hierarchical clustering takes place where initial output codes indicate broad class or cluster membership, and subsequent codes indicate sub-clusters within clusters, and finally individuals within those sub-clusters. Cluster and sub-cluster breadth in the input vector space are influenced by the weight increment size, the ratio of saturated to naive weight values, and dimension of the input vectors in the training set.

### Hardware Implementation

This section focusses on our hardware implementation of GLA olfactory model. Our simplified olfactory architecture is suitable for the hybrid implementation in the MOSIS two micron, p-well, double poly, double metal bulk CMOS process. The GLA model inherently possess many favorable features to aid such simple hardware implementation.

Out of the numerous possible architectures, one potentially feasible olfactory architecture is shown in Figure 1. The hierarchical clustering at the theta rhythm in the GLA model necessitates the synchronous or clocked approach, rather than truly analog continuous parallel processing.

The input cues, analog *current* input vectors  $O_i$ , are assumed to be generated by suitable sensory structure (receptor in anatomical model) which are sampled periodically at an artificial theta rhythm. For each cycle in theta rhythm, there are two major non-overlapping phases: activation of the OB and feed-forward excitation of the PC indicated by forward phase  $\phi_1$ , followed by feedback inhibition of the OB by the PC indicated by



$\phi_2$ . Each phase,  $\phi_1$  and  $\phi_2$ , is further subdivided into two non-overlapping sub-phases,  $\phi_{11}$ ,  $\phi_{12}$  and  $\phi_{21}$ ,  $\phi_{22}$ , respectively. The timing diagram of the olfactory system is shown in Figure 1. Prior to using the network for hierarchical clustering, the network is trained over a set of the input cues by updating the forward (excitatory) nonvolatile weights in parallel according to the adult plasticity rule discussed in learning section. Even though system controls are derived from the clocks, the actual computation between two clocks is truly analog, concurrent, and carried out in parallel. Clocks are merely for multi-sampling, and synchronization purposes.

The following sections discuss the overall operation of our architecture, the different building blocks used by the architecture, and the top level architectural issues together with their relevance to the anatomical model. The essential blocks and their functions in the proposed architecture are: (1) The glomeruli normalizer within the OB to normalize the glomerulus activity at a constant level. (2) The mitral patch within glomeruli to generate the LOT lines or to thermometer encode the net normalized input. (3) The sparse weight matrix to scale and sparsely expand the LOT line activity onto the PC via the modifiable synapses. (4) The WTA piriform patches within the PC to exhibit the winner take all competition. (5) The tie resolver to digitally resolve the potential ties which occur among winning piriform cells within a piriform patch, and (6) the current copier integrator (CCI) to provide the thresholded, collateral, and cumulative feedback to the OB.

The analog input *current* vectors  $O_i$  ( $1 \leq i \leq g$ ) generated by the receptors are sampled periodically at an artificial theta clock. In the OB, the net input  $G^*_i$  to the glomerulus is formed by summing the real positive input vector  $O_i$  point by point with the negative

inhibitory feedback *current* vector  $I_i$  (equation 1).

The  $G_i^*$  is then subjected to the global nonlinear normalization by a glomeruli normalizer. Several alternatives for normalization have been developed. Essentially all normalization schemes are implemented with a closed feedback loop circuit similar to that used in automatic gain control (AGC).  $K_G$  is the constant to which the sum of the normalized activity (equation 3) is maintained (20%).

Each normalized glomerulus signal  $G_i$  is thermometer coded by the  $m$  mitral cells per mitral patch. Mitral cells have increasing equidistance thresholds, i.e.,  $\Theta M_j < \Theta M_{(j+1)}$  ( $0 \leq j \leq m$ ), where  $\Theta M_j$  is the activation threshold of  $j$ th mitral cell. Mitral thresholds are generated globally by a capacitive ladder. The mitral cells are modeled as two state devices (active or inactive) or McCullough-pitts neurons by the two stage comparators which are either high (logic 1) or low (logic 0) with glomerulus activity  $G_i$  either above or below threshold  $\Theta M_j$ , respectively. Electronically, this is equivalent to front end of a flash A/D convertor.

The binary *voltage* output of the mitral cells  $M_{ij}$  in the OB is spatially projected onto the  $h \times p$  piriform cells in the piriform cortex via  $m \times g$  LOT lines, forming the synapses between the OB and the PC. The synaptic weights  $W_{(ij)(kl)}$  are realized by a floating gate, non-volatile, analog programmable memory. The memory is used in conjunction with a MOS transistor operating either in the triode or saturation region. The conductance of a MOS transistor is modulated by the charge on the floating gate. The weights are non-decremental, incremented in discrete steps ( $\sim 10\%$  of their maximum weight), and saturated beyond the maximum value of  $w_{\max}$  ( $\sim$  two to three times their naive weights). The excitory synapses  $W_{(ij)(kl)}$  are sparse rather than topographic, that is, they are randomly

distributed within the PC with a sparseness on the order of 10%. The sparse weight matrix  $W_{(m \times g)(h \times p)}$  consists of sparsely placed synapses. Synapses are randomly arranged in the 4x5 sub-matrices. Restricting the PC random interconnections to a small local area is biologically unsupported. However, the choice of a 4x5 sub-matrix area was selected for fabrication convenience without any biological formulation. Each sub-matrix receives 4 consecutive LOT lines (rows) and five consecutive piriform lines (columns) resulting in the 20 cross junctions. The 10 percent sparse random connectivity within the sub-matrix is achieved by establishing two randomly chosen connections and placing a weighing transistor at these cross junction. Within the sub-matrix, any LOT line may be interconnected with any piriform input line, with the exception that double interconnections between a pair of lines is excluded. During layout, the location of the metal contact to the weighing transistor will be derived by executing a macro that generates a randomized connection between LOT and piriform line. Local grouping of interconnects minimizes interconnection and routing area, resulting in 10 to 20 percent area saving [15].

Simplification of the weight matrix (specifically the local interconnect) architecture results in the loss of certain statistical independence of the connectivity exhibited in the anatomical model. The architecture also results in the uniform distribution of weights as opposed to the increasingly tapered distribution from caudal to rostral region in the anatomical model. Further, due to the restrictions imposed on the connectivity of the sub-matrix, there exist a zero probability for forming some particular pattern of connectivity within a sub-matrix, where as in the absence such restrictions, corresponding probabilities would have some finite values. The architecture would seem to be less prone to these

effects in networks with a sufficiently wide input vector, since according to the central limit theorem, with increasing LOT lines the constrained distribution in the sub-matrix tend to be very similar to the unconstrained interconnection patterns of the anatomical model.

As discussed earlier, we use a common adult and developmental plasticities which allows the use of the single weight matrix  $W$  in the forward and the backward cycles. This requires that weight matrix  $W$  be must time multiplexed to compute the weighted excitatory bulbar output *currents* into the PC in the forward phase, and the weighted inhibitory feedback *currents* from winning piriform cells into the OB in the backward phase. The use of a common weight matrix results in a significant area saving since the weight matrix dominates the total silicon area as the weight matrix area grows in a square while the input/output dimensions grow linearly.

Current conveyor (CC) based bi-directional voltage/current buffers (BiVI) permit bi-directional use of  $W$ . They provide the dual functions of voltage drivers and current sources/sinks to isolate the  $W$  matrix in the forward and backward mode. During the feed forward cycles, the BiVI buffers on the mitral side and the piriform side act as the voltage controlled voltage sources and the current controlled current sources respectively. Their roles are reversed in backward cycle. The detailed, time-multiplexed BiVI buffer operation is described in chapter III.

The currents produced by the inner-products between the LOT activity and the sparse weights, are summed on the columns of  $W$  according to Kirchoff's current law. The weight matrix columns are organized into  $p$  patches with  $h$  neighboring columns/patch. The resulting inner-product analog currents  $P_{ki}^*$  are amplified/scaled by the BiVI buffers

and fed into the PC. In the PC, the excitory piriform cells  $P_{kl}$  are arranged into  $p$  disjoint winner-take-all piriform patches with  $h$  piriform cells per patch. The indice  $k$  and  $l$  indicate the piriform patch and the cell number within the piriform patch, respectively. Each column feeds only one piriform cell. During the sub phase  $\phi_{11}$ , the piriform patches exhibit a winner take all competition within a patch which results in only the piriform cell associated with highest input current to become logic high while the rest of the cells remain at a logic low. The winning piriform cell is declared activated only if the corresponding input current to the piriform cell is equal or greater than the piriform cell threshold  $\Theta_p$ .

The output  $P_{kl}$  of WTA ideally should have only  $p$  winners. But due to the finite resolution of the WTA circuit, it is not possible to avoid ties among the highest and the few near highest input currents. A tie resolver circuit has been added to do post WTA processing during phase  $\phi_{12}$  thereby digitally resolving the ties. The vectors  $P_{kl}$ ,  $PW_{kl}$  are unresolved input, and resolved output respectively. During the multi-sampling process, resolved WTA outputs produce a distinct output code. This output code is used for clustering as well as forming the basis for feedback inhibition.

To implement feedback inhibition to the OB by the PC during the backward phase  $\phi_2$ , binary outputs of the resolved winning piriform cell  $PW_{kl}$  are latched and reciprocally applied via the caudal BiVI buffers to the multiplexed  $W^T$  matrix. This generates the inhibitory currents on the respective LOT lines configured for sinking the currents. The resulting inhibitory currents are amplified/scaled by the rostral BiVI buffers. By Kirchoff's current law, inhibition on consecutive  $m$  caudal BiVI or backward LOT lines is summed by switching them together. Thus, forming an aggregate un-thresholded

inhibition  $I_i^*$  (equation 9) to the glomeruli, from which the respective forward LOT lines are originated. The multiplexed operation of the weight matrix together with BiVI buffers is discussed in chapter III.

The CCI provides the function of accumulative collateral feedback inhibition from the active piriform patches. If the corresponding LOT line was active in the feed-forward phase,  $I_i^*$  is sampled and stored in each feedback cycle by the CCI circuit. The CCI runs in two phase  $\phi_{21}$  and  $\phi_{22}$ , for storing and for updating  $I_i^*$  respectively. During the multi-sampling process, of the feedback phase, inhibition  $I_i$  is applied to the glomerulus at the end of each minor cycle. Inhibition persists to be used during the next cycle in the forward phase. During the successive cycles, all of the inhibition currents that are generated in the backward phase are sampled and added to previously stored inhibition. In this way, as the multi-sampling proceeds, cumulative inhibition up to present minor cycle is applied to the glomerulus to inhibit the stronger input patterns. Thus, making the remaining (weaker) patterns more significant and allowing them to take an active part in the overall clustering process.  $\Theta_i$ , the inhibitory threshold imposed on  $I_i^*$  (equation 10) is necessary to eliminate the effects of floor noise on inhibition.

## CHAPTER III

### SYSTEM BUILDING BLOCKS

This chapter focuses on the design, simulation, and testing of the basic building blocks. Simulations were performed using SPICE. Some auxiliary circuits such as current sources/sinks, digital control pulses, and triggering circuits had to be bread-boarded before actual blocks could be tested for their functionality. Testing is mainly performed to check DC response. Future test of transient response for speed verification will require that the testing circuitry be fabricated on-chip along the functional block being tested. This avoids the external capacitance contribution to the block due to the set up itself. Fabrication of such on-chip testing circuitry is beyond the scope of this thesis. However, we have reported transient responses as measured by an oscilloscope at the pad pins. This obviously adds a parasitic setup capacitance to the circuit nodes. Therefore, extrapolation of transient results is required to estimate the internal bandwidth. Once again, considering the vast and more important topics ahead, we leave this topic for the future. The system level integration of the olfaction system is also beyond the scope of this study.

#### Glomerulus Normalization

In real world artificial intelligence problems, such as pattern recognition, natural language processing, and olfactory clustering, signals in the input vector on the multiple

channels convey useful information both, in the position and amplitude ratios of the vector elements. The signal processing burden in such cases is high. The absolute level of these signals may have minimal bearing on the final outcome of the classification or clustering of related observations. In such applications, simple analog circuits can be used to perform the task of signal normalization. That is, to generate an output array in which each element is proportional to the corresponding element in the input array when normalized by a suitably-derived metric of the overall magnitude of the input, generally, the largest element or sum of all the element in the output array.

Several signal normalization techniques based on the translinear principle have been reported and realized in monolithic form [39]. These circuits exhibit undesirable pattern sensitivity because they don't scale with respect to a suitably-derived metric of the overall magnitude of the element in input array, i.e., scaling becomes a function of the number of input elements. Also being bipolar, they can not be used in the bulk CMOS process. The concept behind normalization circuits is the prospect of performing massively parallel and truly concurrent signal processing.

In this section, various schemes to achieve olfactory bulb normalization are presented. The normalization scheme shown in Figure 3 consists of two feedback loops across all of the bulb inputs. An AGC function controls gain and an offset function ensures that the activity of normalized outputs is large in amplitude to strongly activate the mitral patch. The other normalizing schemes illustrated in Figure 4 and Figure 5 do not incorporate an offset function but use a linear and nonlinear (square law) functions respectively, to process the bulb inputs in the feedback loop. The following sub-sections elaborate each normalizing scheme in detail.



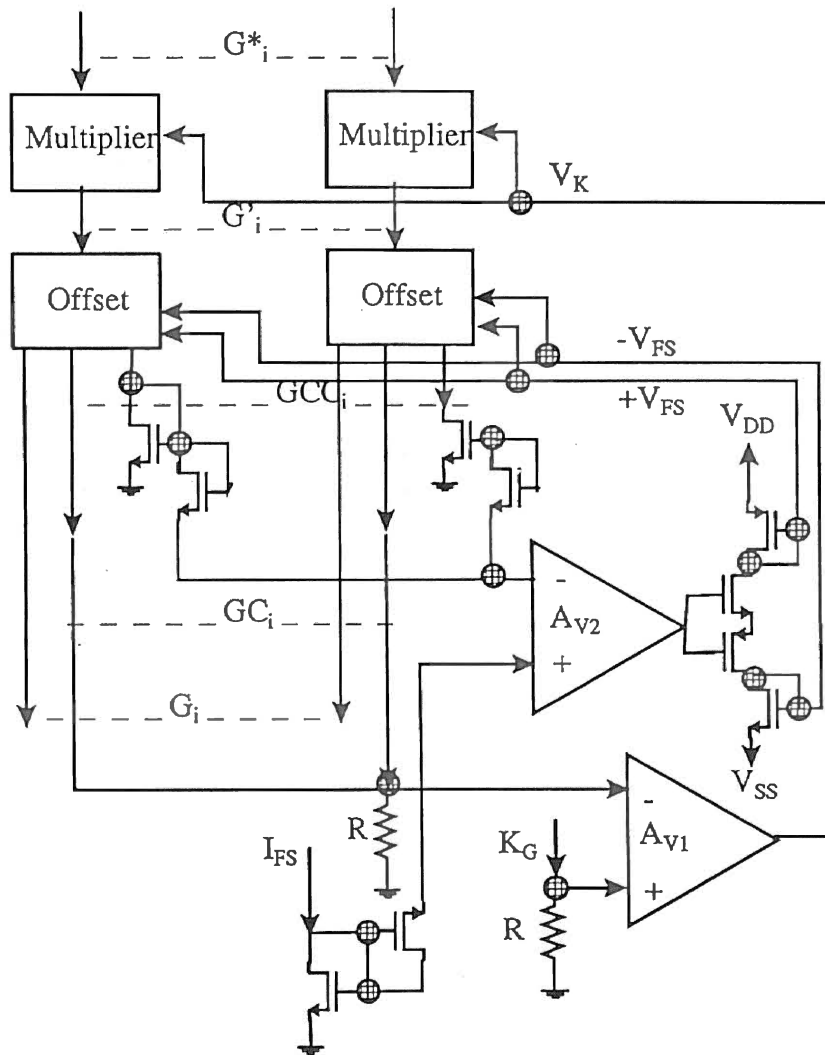


Figure 3. AGC and Offset Combined Linear Normalization Function

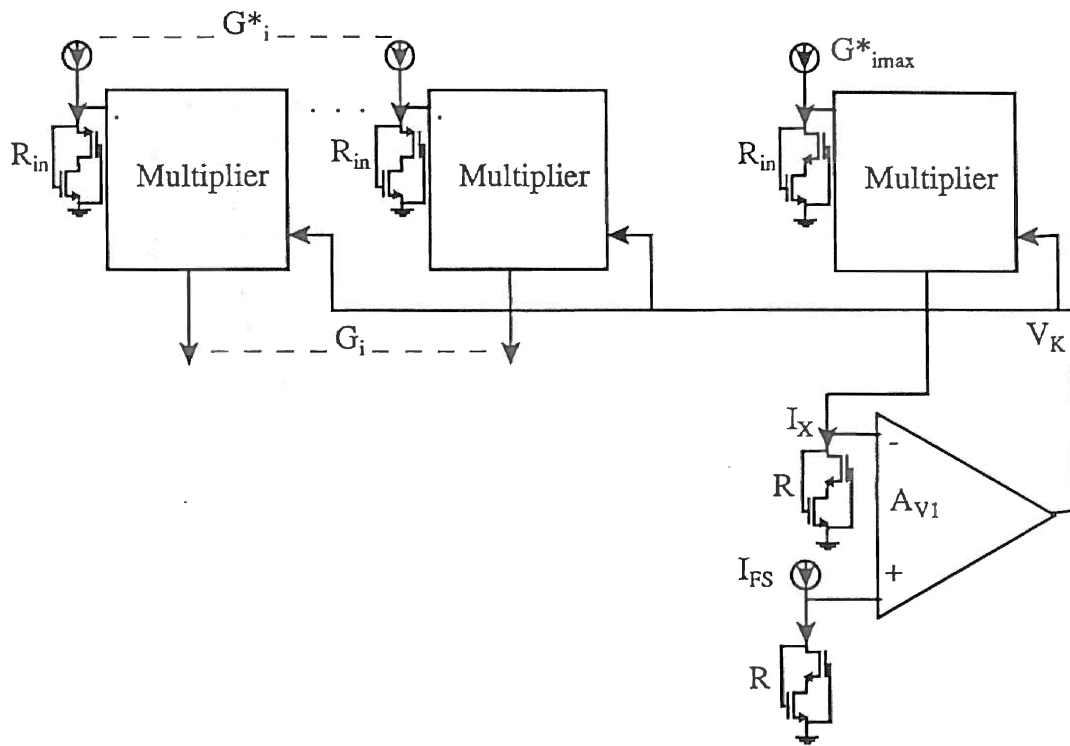


Figure 4. Linear Limiter with AGC Normalization Function

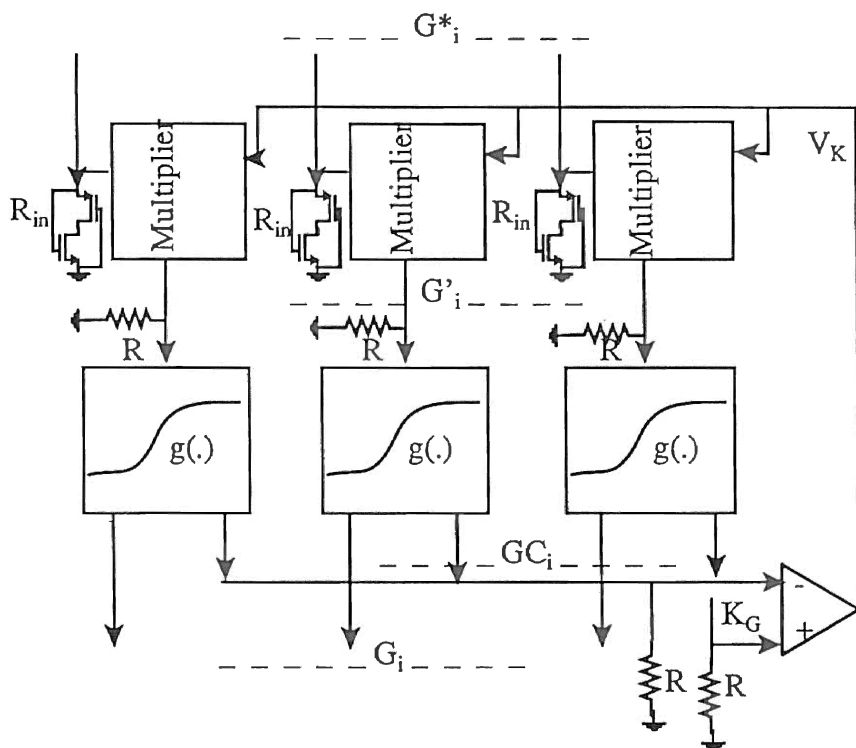


Figure 5. Square Law Normalization Function

### AGC and Offset Combined Normalizing Function

The block diagram of the AGC and offset combined linear normalization function is shown in the Figure 3. The normalization block consists of two feedback loops across the inhibited bulb input  $G_i^*$ . The basic building blocks involved are the multiplier, offset summer, and operational amplifier. The off-chip operational amplifiers will be used to simplify conceptual block testing.  $G'_i$  is the signal obtained after AGC while  $G_i$  is the normalized offset output signal. The multiplier in the closed loop ensures a constant level,  $K_G$  percent, of mitral activity. The offset summer in the closed loop detects the maximum value of element  $G'_{imax}$  in  $G'_i$  and adds the difference between the  $G'_{imax}$  and the set point  $I_{FS}$ , to all the elements in  $G'_i$ . This difference between the set point  $I_{FS}$  and  $G'_{imax}$  is referred to as an offset. The offset activity can best be illustrated mathematically,

$$Offset = I_{FS} - G'_{imax} \quad (12)$$

and output is given as

$$G_i = G'_i \pm |Offset| \quad \begin{array}{l} + \text{ if } G'_{imax} < I_{FS} \\ - \text{ if } G'_{imax} > I_{FS} \end{array} \quad (13)$$

The multiplier activity is given by

$$G'_i = gm (G_i^* \times V_K) \quad (14)$$

where  $gm$  is the linear transconductance of the multiplier and scaler  $V_K$  is the smallest value that satisfies

$$\sum_{i=1}^g G_i = K_G \quad (15)$$

where  $K_G$  is glomerulus activity constant. AGC closed loop activity sets  $V_K$  at

$$V_K = \left( K_G - \left( \sum_{i=1}^g G_i \right) \right) R A_{VI} \quad (16)$$

The multiplication in equation 14, and offsetting of the automatic gain controlled vector in equation 12 and 13, is accomplished by the multiplier and offset summer circuits of Figure 6 and 15, respectively. The following sections discuss each functional building block in detail.

### Transconductance Multiplier

The schematic diagram of the cross-coupled double quad CMOS transconductance multiplier used in the normalization circuit is shown in Figure 6. The circuit is composed of the linear CMOS transconductor and its biasing circuitry. The transconductance is a crucial component of the design since it may limit the multiplier linearity, frequency response, and noise performance. High linearity for large input signals, low noise, no dominant internal poles, large transconductance, and low quiescent power dissipation are the desired properties of any transconductance circuit. Several techniques for improving the linearity of the MOS transconductance elements have been proposed [40]. Most of the differential transconductance schemes can be broadly classified into four categories: adaptive biasing, class A-B, source degeneration, and current differencing. Some combine two or more of these techniques to achieve linearization. Detailed information can be found elsewhere [40].

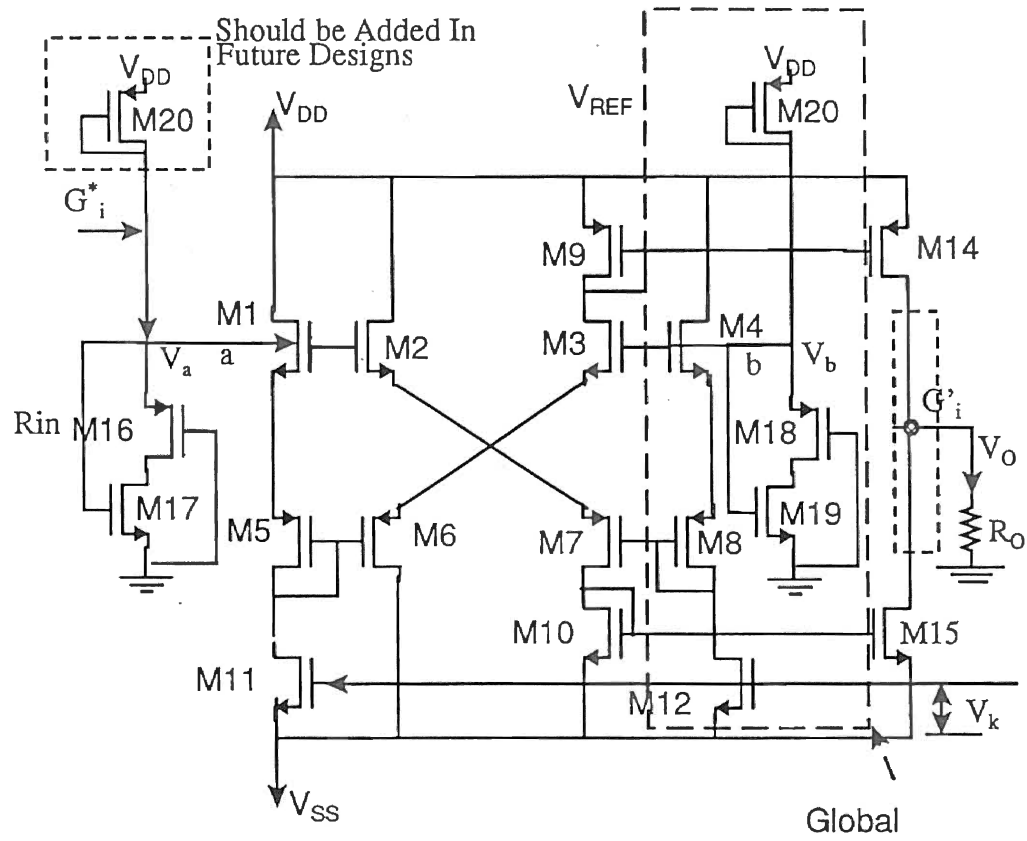


Figure 6. Transconductance Multiplier

The proposed circuit comes under the class A-B transconductors. Transconductors in which the maximum output is greater than the quiescent bias current ( $\eta > 100\%$ ), generally operate in the class A-B mode. Class A-B transconductors typically exploit the square law characteristics of an MOS transistor in the saturation region to achieve linearization [41]. In this section, we will discuss the fundamental principle and design of the transconductance multiplier circuit.

The fundamental principle of the class A-B transconductors can be understood by examining the two transistor configuration shown in Figure 7. Assuming both transistors are perfectly matched and operating in the saturation region, the differential output current is given by:

$$I_{diff} = I_{D1} - I_{D2} = \frac{\beta}{2} (V_{GS1} - V_{GS2})(V_{GS1} + V_{GS2} - 2V_T) \quad (17)$$

Above equation states that a linear transconductance can be achieved by ensuring that the sum of the gate-to-source voltage is constant. With the sum constant, if  $V_{id}$  is equal to  $V_{GS1} - V_{GS2}$  then equation 17 reduces to:

$$I_{diff} = \beta (V_{CM} - V_T) V_{id} \quad (18)$$

where  $V_{CM} = (V_{GS1} + V_{GS2})/2$  is the common mode input level. The transconductance  $gm = \beta(V_{CM} - V_T)$  is linear and may be varied electronically by adjusting the common mode input level. From the above analysis, the fundamental principle of class A-B operation can be defined as: "Under the conditions of a constant sum of gate to source voltages, two matched MOS transistors operating in the saturation region display a linear relationship between the difference of the gate-to-source voltages and the difference of the drain

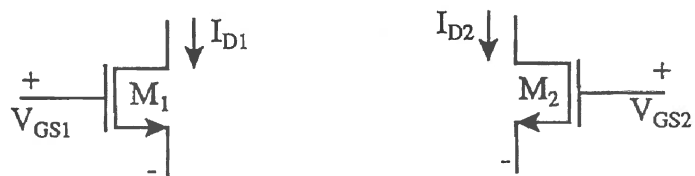


Figure 7. Demonstration of Class AB Principle .  
With Two Transistors



currents" [40]. This applies in the operating range:

$$|V_{id}| \leq 2(V_{CM} - V_T) = \sqrt{\frac{4I_{DC}}{\beta}} \quad (19)$$

In this region current may vary as

$$\begin{aligned} -2I_{DC} &\leq I_{diff} \\ I_{DC} &= \beta(V_{CM} - V_T)^2 \end{aligned} \quad (20)$$

where  $I_{DC}$  is the total dc bias current.

The second fundamental principle involved in the design of cross-coupled double quad transconductance is replacement of the single transistor  $M_1$  or  $M_2$  with the CMOS double pair as shown in Figure 8. This overcomes the matching problems associated with n-channel class A-B operation. Using the saturation region equation of the MOS device results into,

$$V_{GS} = V_T + \sqrt{\frac{2I_D}{\beta}} \quad (21)$$

The equivalent gate-to-source voltage,  $V_{GSeq} = V_{GSN} + V_{GSP}$  of the double-pair is then given by:

$$\begin{aligned} V_{GSeq} &= \left( \frac{1}{\sqrt{\beta_N}} + \frac{1}{\sqrt{\beta_P}} \right) \sqrt{2I_D} + V_{TN} + V_{TP} \\ &= \sqrt{\frac{2I_D}{\beta_{eq}}} + V_{Teq} \end{aligned} \quad (22)$$

where,

$$V_{Teq} = V_{TN} + V_{TP} \quad (23)$$

and

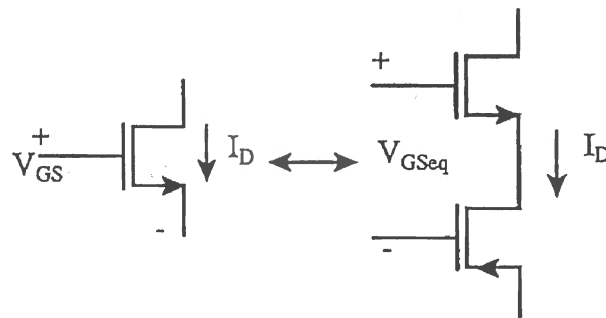


Figure 8. CMOS Equivalent of Single MOS Transistor

$$\beta_{eq} = \frac{\beta_N \beta_P}{(\sqrt{\beta_N} + \sqrt{\beta_P})^2} \quad (24)$$

This concludes that a pair of opposite polarity MOS transistors acts as a single transistor with an equivalent threshold voltage and transconductance given by equations 23 and 24, respectively.

In equation 18, the transconductance is perfectly linear. Although it has excellent linearity and efficiency, some of the class A-B implementations suffer from limitations such as the requirement of fully balanced signals for non-linearity cancellation and poor common mode rejection. The class-AB double quad circuit, which overcomes most of these problems, is shown in Figure 9. From the fundamental principle of class AB transconductance and equation 18, the sum of the gate-to-source voltage of  $M_1$  and  $M_2$  must be constant for non-linearity cancellation to occur. Applying KVL around the loop

$$V_{GS1} + V_{GS2} = 2(V_B + V_T) \quad (25)$$

and

$$V_{GS1} - V_{GS2} = 2 V_{id} \quad (26)$$

In the above equation,  $V_{id}$  need not be a balanced input. Current source biasing can be ratio. As a result, the common mode input level no longer affects the transconductance or the linear range. Substituting equations 25 and 26 into equation 17 results in

$$I_{diff} = 2\beta V_B V_{id} \quad (27)$$

The transconductance  $gm = 2\beta V_B$  is perfectly linear and can be tuned by changing  $V_B$ . It should be noted that the differential current flows through the floating voltage sources.

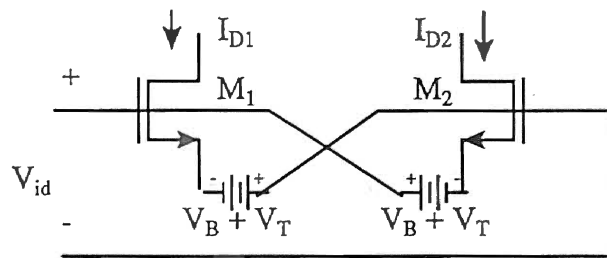


Figure 9. Linear MOS Transconductance Principle

The magnitude of these voltage sources should remain constant regardless of the current flowing through them. A better solution can be realized by replacing a single transistor ( $M_1$  &  $M_2$ ) with its CMOS equivalent double pair as described previously. This does not change the circuit behavior since the CMOS double pair acts like a single transistor, except  $\beta$  in equation 27 is replaced by  $\beta_{eq}$  and  $V_T$  by  $V_{Teq}$  as given by equations 23 and 24 (see Figure 10). In this configuration, drain current no longer flows through the floating voltage sources. Thus the required floating voltage sources can now be achieved by the diode connected CMOS pairs biased with the current sink as shown in Figure 11. From equation 22, the bias voltage  $V_B$  is given by:

$$V_B = \sqrt{\frac{2I_B}{\beta_{eq}}} \quad (28)$$

Combining this biasing network with Figure 10 results in the final transconductor as shown in Figure 6. The differential output current is obtained by incorporating differential mirroring pairs consisting of  $M_9, M_{14}$  and  $M_{10}, M_{15}$ . Finally, from equation 27 and 28

$$I_{diff} = \sqrt{8\beta_{eq}I_B} V_{id} \quad (29)$$

for the differential range

$$V_{id} \leq \sqrt{\frac{2I_B}{\beta_{eq}}} \quad (30)$$

The transconductance of equation 29 is perfectly linear and can be tuned with the bias current  $I_B$ . Since both, the quiescent current and the maximum linear output currents are  $4I_B$ , the maximum efficiency is 100%. The efficiency can be increased above 100% by

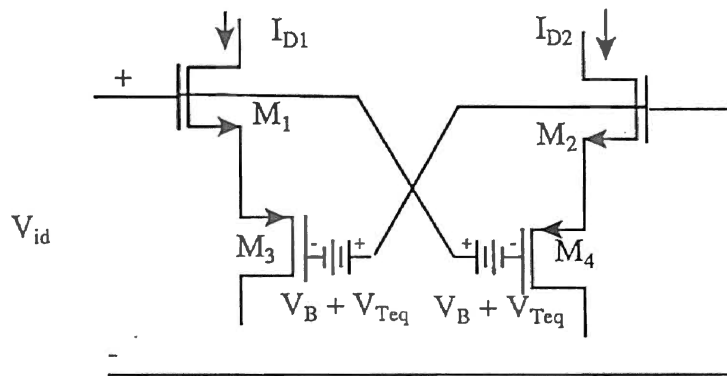


Figure 10. Linear MOS Transconductance Using the MOS Equivalent pair

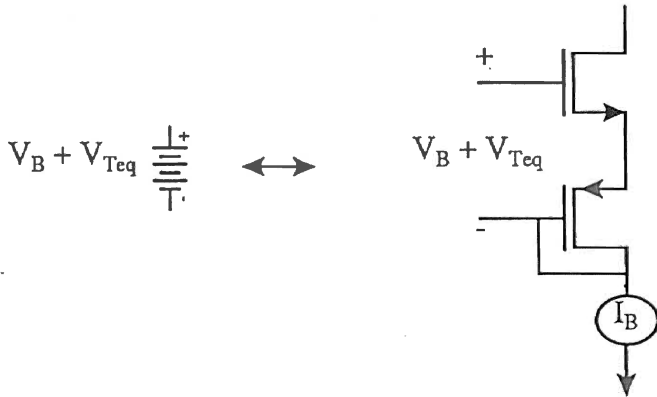


Figure 11. Double Pair Implementation of a Floating Voltage Source

decreasing the W/L ratio of the inner quad transistors with respect to the outer quads. The differential output current is obtained by incorporating the differential mirroring stage in series with the inner quads. From Figure 6 and equation 29

$$I_{diff} = I_1 - I_2 = 2 \frac{\sqrt{(\beta_2 \beta_7 \beta_{11})}}{(\sqrt{\beta_2} + \sqrt{\beta_7})} (V_K - V_{T11}) V_{id} \quad (31)$$

It should be clear from the equations 29 and 31 that by making the bias current  $I_B$  a function of one variable,  $V_K$ , the transconductance circuit can be used as a linear multiplier. However, it is essential that both bias current sources remain in saturation over the entire operating range to be a linear function of the gate voltage  $V_K$ . In order to maintain symmetry of operation in both the quads, as well as to achieve good efficiency, the following geometrical relations can be determined by inspection. For the same bias voltage  $V_B$ :  $(W/L)_1 = (W/L)_4$ ,  $(W/L)_5 = (W/L)_8$ , and  $(W/L)_{11} = (W/L)_{12}$ . To have symmetrical operation of the inner quads:  $(W/L)_2 = (W/L)_3$ , and  $(W/L)_6 = (W/L)_7$ . To have identical biasing resistance for both quads:  $(W/L)_{16} = (W/L)_{18}$ ,  $(W/L)_{17} = (W/L)_{19}$ ; and for proper differential current mirroring:  $(W/L)_9 / (W/L)_{14} = (W/L)_{10} / (W/L)_{15}$ . Moreover, to simplify the circuit design, let  $\beta_2 = \beta_7 = \beta_3 = \beta_6 = \beta$  for the inner quad, and  $\beta_1 = \beta_5 = \beta_4 = \beta_8$  for the outer quad. The inner quad is made geometrically half of the outer quad to achieve greater than 100% efficiency, i.e.,  $\beta_2 = \beta_1 / 2$ . With these simplifications, equation 31 becomes

$$I_{diff} = \sqrt{(\beta \beta_{11})} (V_K - V_{T11}) V_{id} = gm (V_K - V_{T11}) V_{id} \quad (32)$$

where  $gm = (\beta \beta_{11})^{1/2}$  is the linear transconductance. As pointed out earlier, the input does



not have to be fully balanced. Keeping one end (point b) of the differential input  $V_{id}$  at ground potential, the differential signal can be made single ended. But, then  $V_{DS}$  at 5 V is not sufficient to keep  $M_8$  and  $M_{12}$  in saturation for higher values of  $V_K$ . Thus  $M_{12}$  falls into triode region and the multiplier from suffers linearity degradation. Therefore, the common mode range of the differential signal should be increased by  $V_T$  or  $2 V_T$ , allowing increased  $V_{DS}$  headroom to keep  $M_8$  and  $M_{12}$  in saturation over the entire operating range of  $V_K$ . This is achieved by two identical complementary linear resistors  $R_{in}$ , each made of transistors  $M_{16}$ ,  $M_{17}$  and  $M_{18}$ ,  $M_{19}$  connected in a back to back fashion as shown in Figure 6. Finally,

$$\begin{aligned} G'_i &= I_{diff} \\ &= gm (G^* R_{in}) \Delta V_K \end{aligned} \quad (33)$$

where,  $\Delta V_K = V_K - V_{T11}$ . The resistance  $R_{in}$  is given by:

$$R_{in} = \frac{1}{\beta_{17}(V_a - V_T)} \quad \text{for } V_a \leq 2V_T \quad (34)$$

Clearly, the output current is the function of input current  $G^*_i$  and scaler  $V_K$ .

Simulations. The SPICE simulations of the DC transfer characteristics of the multiplier are shown in the Figure 12. A family of curves is obtained by ramping inhibited receptor currents  $G^*_i$  from 0 to 250  $\mu A$  for the different closed loop voltage,  $V_K$ , varied over the 1 V to 2.4 V range in steps of 0.2 V. The output currents are sampled via  $R_O$ .

The transconductance obtained from the DC transfer characteristics is approximately linear and satisfies equation 32. The non-linearity at lower values of inputs is due to the

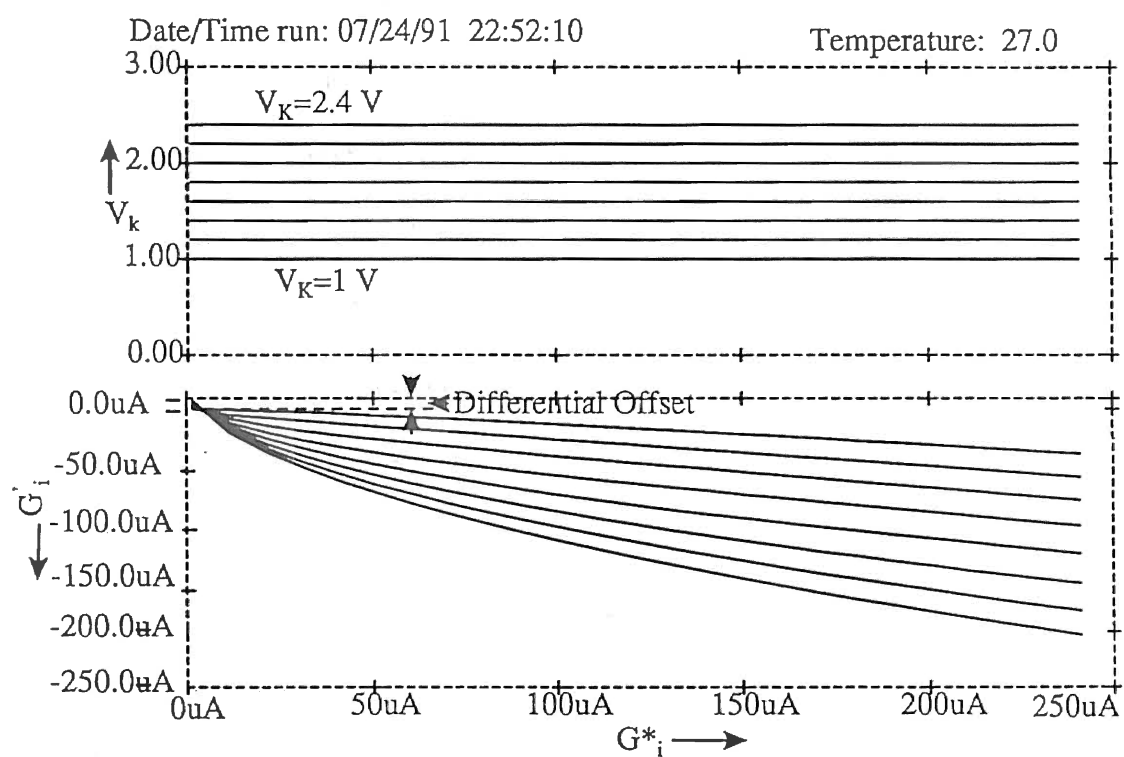


Figure 12. DC Transfer Characteristics of a Transconductance Multiplier

non-linear  $R_{in}$  (equation 34). The non-zero output current at  $G_i^*=0$  is a result of the differential offset that is present due to the finite amount of current that is required to flow to provide  $V_b$ .

Testing. Due to the difficulty in obtaining the ramped DC sinking current  $G_i^*$ , saw tooth voltage  $V_a$  is used instead of  $G_i^*$ . This requires that the bias voltage  $V_b$  (see Figure 6) must be known to confirm the zero crossing of the output currents. The SPICE simulations performed on the extracted file are shown in Figure 13.  $V_b$  was found to be 2.25 V. Keeping this in mind, a family of curves is obtained by ramping  $V_a$  from 1 V to 3.5 V for different values of  $V_k$  varied over a 1 V to 2.5 V range. Test results are shown in the Figure 14. During the testing, two quadrant operation of the multiplier is exhibited due to the fact that  $V_b$  is biased to a positive voltage (2.25 V) instead to a zero. Thus, the sign of the differential input voltage ( $V_{id} = V_a - V_b$ ) and output current changes (equation 32) when  $V_a$  is varied from below  $V_b$  to above  $V_b$ . The output current is sampled in terms of the voltage drop  $V_o$  across the 10 Kohms precision resistor  $R_o$ .

The test results are compared with the results that are obtained from the SPICE simulations performed under the identical conditions. The multiplier behaves linearly within the operating range. The percentage difference between the output currents obtained from simulation and testing is found to be below 25 percent. This linearities is present due to the nonlinear resistance given by equation 34.

### Offset Circuit

The offset circuit for a one bit wide input vector is shown in Figure 15. The circuit

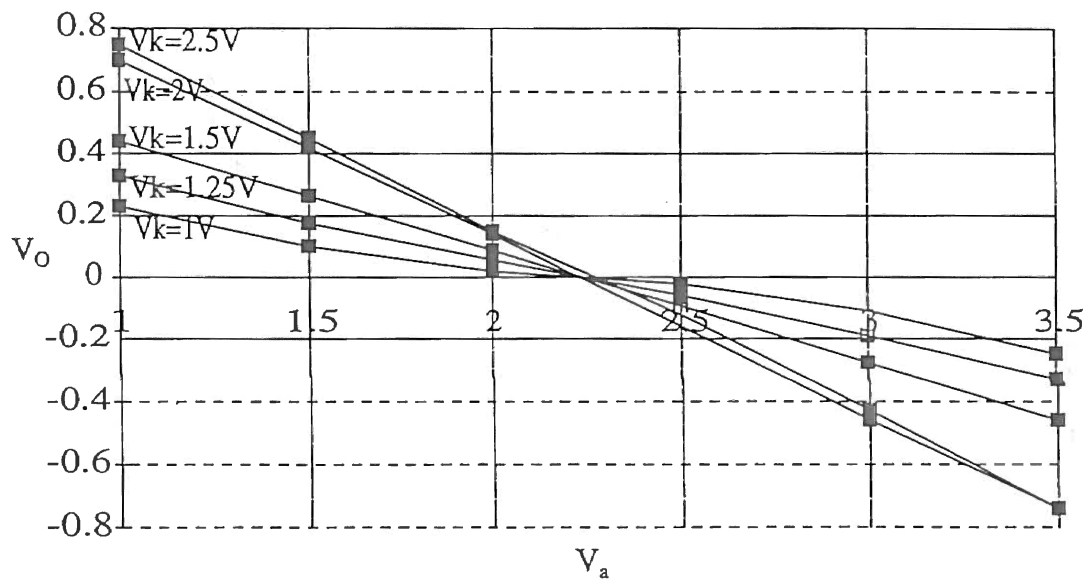


Figure 13. Multiplier Output Voltage Obtained from Simulations

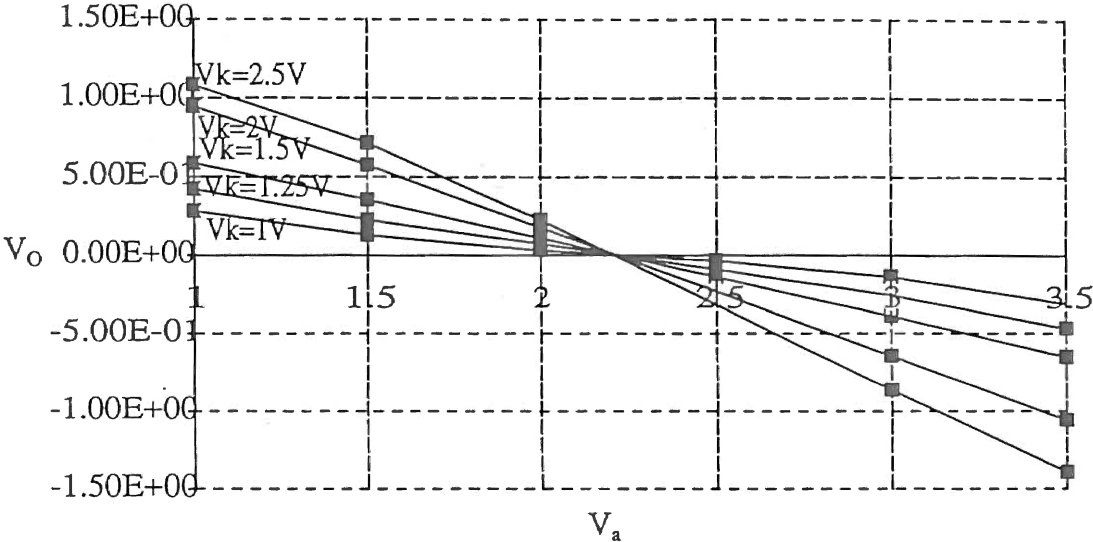


Figure 14. Multiplier Output Voltage Obtained from Test Results

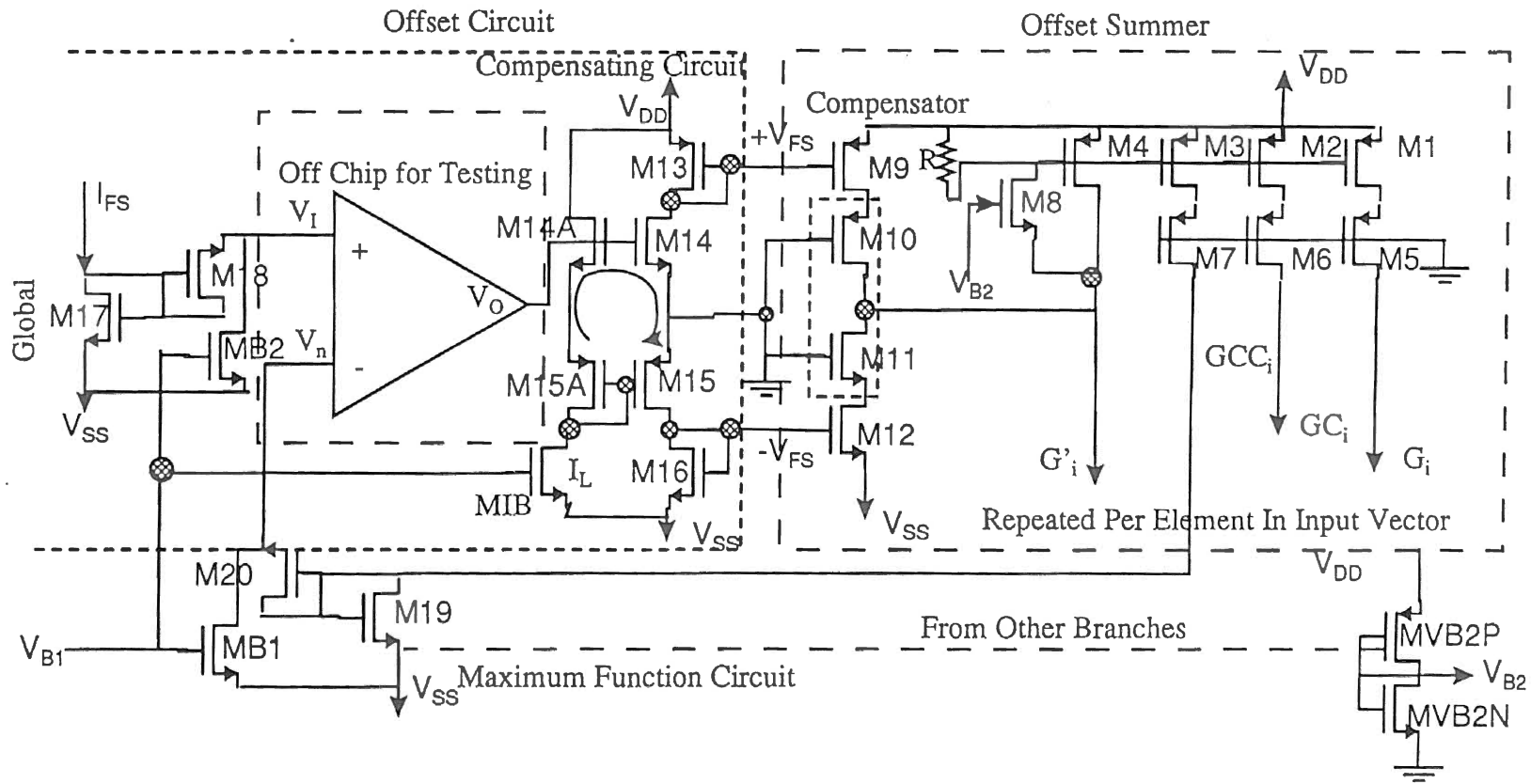


Figure 15. Offset Summer Circuit

is divided into two parts: the global off-setter and the offset summer that is repeated per element of the input vector. The simulations consider four bit wide input and output analog current vectors,  $G'_i$  and  $G_i$ , respectively.

The off-setter is comprised of an operational amplifier (op-amp) connected in the negative feedback loop, and the offsetting circuit ( $M_{13-16}$ ,  $M_{14A}$ ,  $M_{15A}$ , and  $M_{1B}$ ). The op-amp may locally be integrated on-chip or it may be connected off-chip at the expense of the bandwidth. For testing, op-amp is connected off-chip. Simulations are performed by considering an ideal op-amp, while for testing purposes it is replaced by a discrete off-chip amplifier. The voltage drop across  $M_{17-18}$  and  $M_{B2}$  corresponding to the full scale input current  $I_{FS}$  forms the inverting input to the op-amp.

The offset summer is comprised of the compensator ( $M_{9-12}$ ), a current mirror ( $M_{1-8}$ ) and the maximum function circuit ( $M_{19-20}$ ,  $M_{B1}$ ). The bias voltage  $V_{B2}$  required by the current mirror is generated on-chip by the voltage reference circuit ( $M_{VB2P}$  and  $M_{VB2N}$ ).

$G'_i$ , where  $i=1,2, \dots, g$ , forms the input vector to the offset summer. Let  $G'_{imax}$  be the maximum value of the input current among the elements of the  $G'_i$ . This circuit achieves the offsetting of the input vector  $G'_i$  by adding the difference between the full scale current and the maximum input current ( $I_{FS}-G'_{imax}$ ) to all the elements of the input vector including  $G'_{imax}$ . All of the elements in the output vector  $G_i$  are offset by an equal amount. This is due to fact that the compensating circuit, along with the op-amp generates in parallel the same global feedback  $+V_{FS}$  and  $-V_{FS}$  to all the individual elements. The appeal of this circuit lies primarily in the prospect of performing massive parallel signal normalization. Note that the bandwidth is limited by slew rate of the operational amplifier and the  $C_{GS}$  load of  $M_{10-11}$ .

Two copies,  $GC_i$  and  $GCC_i$ , of the output vector  $G_i$  are generated by the current mirror. Cascodes  $M_{5,7}$  minimize the copying error that is present due to channel length modulation.  $G_i$  forms the input to the mitral patch for further processing of the signal.  $GC_i$  is used in a closed loop to maintain the input activity to  $K_G$  percent while  $GCC_i$  feeds the maximum function circuit that is used to detect the maximum value of the element  $G'_{imax}$  in the input vector  $G'_i$ . The multiple input-single output maximum function circuit is comprised of several single input-single output sub circuits connected in parallel, and repeated for each element of  $G'_i$ . Each sub-circuit ( $M_{19-20}$ ) is comprised of two MOS devices connected in diode fashion. The circuit diagram ( $g=4$ ) of the maximum function circuit is shown in the Figure 16.  $M_{B1}$  is a long channel transistor necessary to provide the leakage current to bias  $M_{20A}$ ,  $M_{20B}$ ,  $M_{20C}$ , and  $M_{20D}$ . Nodes A, B, C, and D possess different potentials depending on the corresponding mirrored input currents that are flowing through  $M_{19A}$ ,  $M_{19B}$ ,  $M_{19C}$ , and  $M_{19D}$  respectively. The node "-" acquires the  $\max\{V_A, V_B, V_C, V_D\}$  corresponding to the maximum current. This voltage reverse biases all other diodes except the diode in the branch with maximum current, thereby detecting the maximum potential corresponding to the  $G'_{imax}$ . The value of the  $G'_{imax}$  can be estimated by comparing  $\max\{V_A, V_B, V_C, V_D\}$  with the drop across an identical structure ( $M_{17-18}$ ) due to known full scale current. The voltage drop across the identical sub-circuit  $M_{17-18}$  forms the inverting input of the op-amp.  $I_{FS}$  being a single element,  $M_{18}$  and  $M_{B2}$  are unnecessary. They maintain symmetry for minimizing the input offset.

The output of the op-amp  $V_o$  together with the compensating circuit, provides global feedback to a compensator via two buses:  $+V_{FS}$  and  $-V_{FS}$ . Applying a KVL around the loop shown in Figure 15, we get:



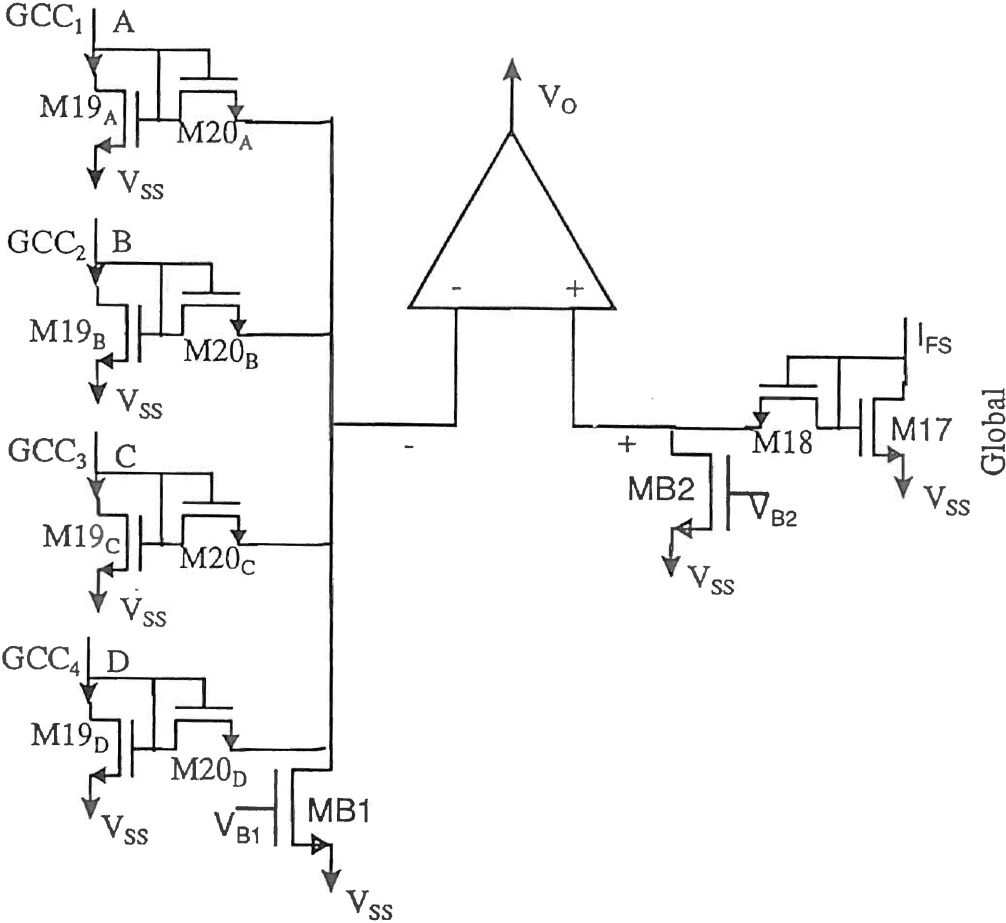


Figure 16. Maximum Function Circuit

$$V_{GS14} - V_{GS15} + V_{GS15A} - V_{GS14A} = 0 \quad (35)$$

Assuming  $M_{14A}$  and  $M_{15A}$  are operating in saturation,

$$V_{GS14} - V_{GS15} + \sqrt{\frac{2I_L}{\beta_{15A}} + V_{T15A}} - \sqrt{\frac{2I_L}{\beta_{14A}} - V_{T14A}} = 0 \quad (36)$$

If  $\beta_{15A} = \beta_{14A}$ , then  $V_{GS14} = V_{GS15} = V_O$ . The long channel device  $M_{IB}$  is necessary to provide a leakage current  $I_L$  to keep  $M_{14A}$  and  $M_{15A}$  conducting in the subthreshold region.

The sign and value of the differential input determines sign and value of  $V_O$ . As  $V_O$  increases in the positive direction,  $V_{GS14}$  and  $V_{GS15}$  increase.  $M_{14}$  pulls down bus  $+V_{FS}$  while PMOS  $M_{15}$  loses pull-up action therefore bus  $-V_{FS}$  also decreases. The converse is true if  $V_O$  decreases in negative direction. At any point in time, the difference between  $+V_{FS}$  and  $-V_{FS}$  remains constant, but the mean changes. In other words,  $+V_{FS}$  and  $-V_{FS}$  vary in the same direction by an equal amount.

The operating principle of the offset summer circuit can best be illustrated by considering the circuit operating in a closed loop configuration. Assuming that  $G'_{imax} > I_{FS}$ , the following sequence of operations takes place. The current mirror generates two copies,  $GC_i$  and  $GCC_i$ , of the output vector  $G_i$ . Each one is used for a specific purpose as described earlier. Since  $G'_{imax} > I_{FS}$ ,  $V_n$  becomes greater than  $V_T$ , thus  $V_O$  increases in the positive direction resulting in a decrease in both  $+V_{FS}$  and  $-V_{FS}$ .  $M_9$  starts conducting while  $M_{12}$  shuts off. If  $G'_{imax}$  has to be normalized to  $I_{FS}$  then the extra current,  $G'_{imax} - I_{FS}$ , must come from  $M_9$ . Thus, only  $I_{FS}$  flows through  $M_4$  which after mirroring is available as an offset version of the input  $G'_i$ . Since  $+V_{FS}$  and  $-V_{FS}$  are common to all the elements in the input vector, the same offset,  $G'_{imax} - I_{FS}$  is added to every element in the input

vector. Since  $G'_i$  is a constant, only  $G'_i - (G'_{imax} - I_{FS})$  flows through the corresponding  $M_4$  which when mirrored is available as  $G_i$ . The identical but reverse action takes place if  $G'_{imax} < I_{FS}$ .  $M_9$  shuts off and the shortfall  $(I_{FS} - G'_{imax})$  flows through  $M_{12}$ , thus the element corresponding to  $G'_{imax}$  gets offset to become  $G'_{imax} + I_{FS} - G'_{imax} = I_{FS}$  and all other elements are offset to  $G'_i + I_{FS} - G'_{imax}$ .

Simulations. The SPICE simulations of the DC transfer characteristics of the offset summer circuit are shown in the Figure 17. For simplicity, a four bit wide input vector  $G'_{1-4}$  is considered. The DC transfer curves are obtained by holding  $G'_{2-4}$  at the constant levels, i.e.,  $44 \mu A$ ,  $34 \mu A$ , and  $24 \mu A$  respectively, whereas  $G'_1$  is ramped from  $100 \mu A$  to 0. The full scale current,  $ID(M_{FS})$ , is held at  $64 \mu A$ . The offset output current is sampled via  $M_4$  in each sub-circuit which, when mirrored, is available as normalized output vector  $G_{1-4}$ . Normalization is analyzed at the four discrete points A, B, C, and D. These points are shown on the plot.

At point A,  $G'_1 = G'_{imax}$ . According to previous discussion, when  $G'_{imax} > I_{FS}$ , offset is  $I_{FS} - G'_{imax}$ , i.e.,  $-36 \mu A$  in this case is added to all the elements in the input vector. Mathematically, the normalized currents  $G_{1-4}$  become  $64 \mu A$ ,  $8 \mu A$ ,  $-2 \mu A$ , and  $-12 \mu A$  respectively. The normalized currents obtained at point A in the plot are off by  $4 \mu A$  since the maximum element gets normalized to  $68 \mu A$  rather than to  $64 \mu A$ . This error is shown on the plot. The error is attributed primarily to the copying fidelity of the current mirrors.

At point B,  $G'_1 = 64 \mu A$ . Thus the offset reduces to zero. Mathematically, the normalized currents  $G_{1-4}$  should possess their original values of  $64 \mu A$ ,  $44 \mu A$ ,  $34 \mu A$ ,

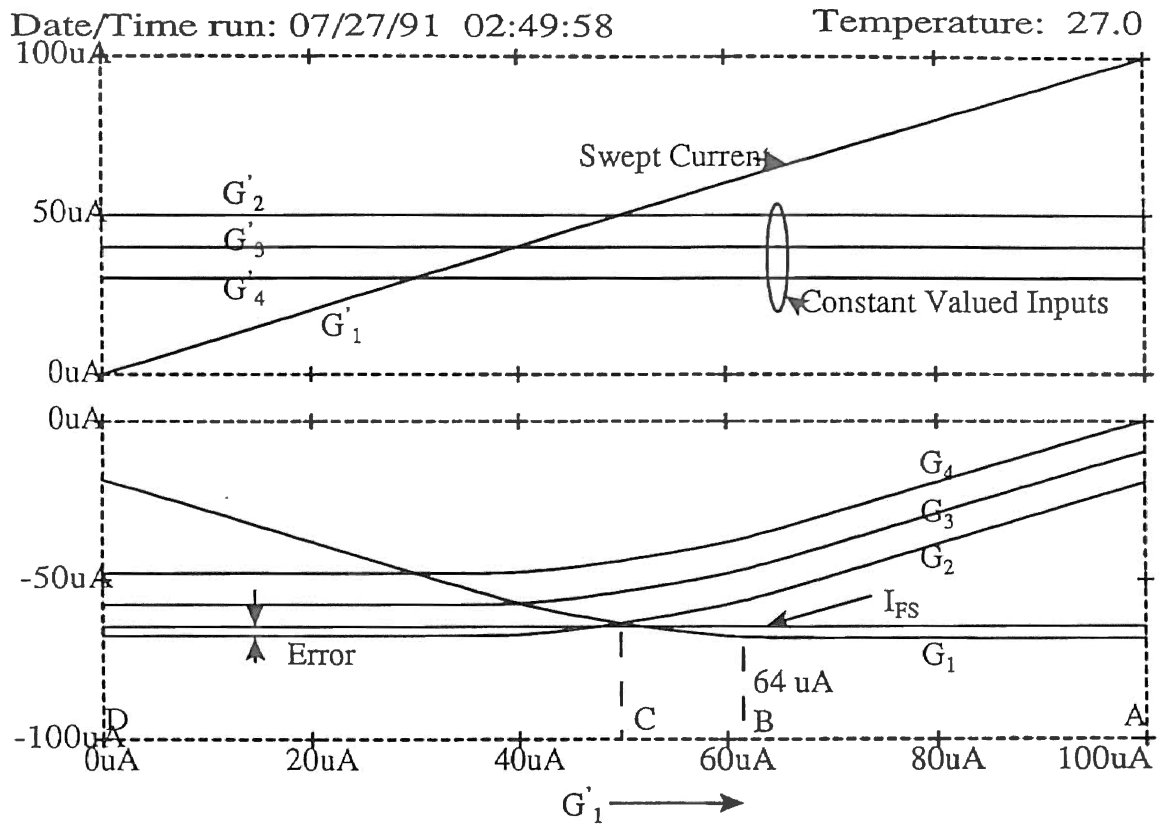


Figure 17. Offset Summer DC Characteristics for Four Branches

24  $\mu\text{A}$  respectively. Currents obtained are once again off by 4  $\mu\text{A}$  because of the previously stated reasons.

At point C,  $G'_1=G'_2$ . Any further reduction in  $G'_1$  makes  $G'_2=G'_{\text{imax}}$ . Thus, from this point forward, a constant offset ( $I_{\text{FS}}-G'_2$ ), 20  $\mu\text{A}$  in this case, is added to  $G'_{1,4}$ . Mathematically, at point D, the normalized currents  $G_{1,4}$  should possess 20  $\mu\text{A}$ , 64  $\mu\text{A}$ , 54  $\mu\text{A}$ , and 44  $\mu\text{A}$  respectively. The currents obtained from the plot at point D verify these values.

Testing. The testing of the entire offset circuit connected in a closed feedback loop didn't lead to conclusive results. To locate the fault, each sub-circuit was tested separately.

The common node formed by gates of  $M_{1,4}$  is a high impedance node. The leakage resistance  $R$  in Figure 15 is essential to provide the bias current to  $M_g$ . Without  $R$ , even though correctly biased by  $V_{B2}$ ,  $M_g$  fails to configure  $M_{1,4}$  in the current mirror mode, thus no feedback in the closed loop is made available to the maximum function circuit. The value of  $R$  is of the order of one meg ohm.

CMOS technology inherently does not offer area efficient way to realize linear high on-chip resistances. Realizing  $R$  internally by using a poly resistor is not an area efficient solution. Alternatively, a common gate can be made available externally so that the external resistance can be used. The later needs  $i$  pins for  $i$  bit wide input vector. Our failure to implement  $R$  by either means restricted us from testing the entire circuit in a closed loop. However, the effect of  $V_O$  on the  $+V_{\text{FS}}$  is observed. The  $V_O$  is ramped linearly from 0 to 3 V. From Figure 15,

$$+V_{FS} = V_{DD} - \sqrt{\frac{\beta_{14}}{\beta_{13}}} (V_O - V_{T14}) + V_{T13} \quad (37)$$

The testing results closely follow the above equation.

### Linear Limiter with AGC Normalization Function

The block diagram of the linear limiter with AGC normalizing function is shown in Figure 4. It is essentially identical to the AGC and offset combined normalization function, except that it uses different normalization parameters and does not have an offset function. It consists of an AGC feedback loop across all the inhibited bulb inputs to perform the task of signal normalization, that is, to generate an output array in which each element is proportional to the corresponding element in the input array divided by the largest element of the input array.

First, using the previously described maximum function circuit, the maximum element in the inhibited input vector  $G^*$ , is detected. From Figure 4 and equation 33

$$I_X = gm(G^*_{imax} R_{in}) \Delta V_K \quad (38)$$

where  $gm$  and  $R_{in}$  have their usual meanings. Also from Figure 4

$$\Delta V_K = R(I_{FS} - I_X) A_{V1} \quad (39)$$

where  $\Delta V_K = V_K - V_{T11}$ . Substituting equation 38 into equation 39 results in

$$\Delta V_K (1 + G^*_{imax} A_{V1} R_{in} R gm) = A_{V1} R I_{FS} \quad (40)$$

If  $A_{V1}$  is sufficiently large such that  $1 + G^*_{imax} A_{V1} R_{in} R gm \approx G^*_{imax} A_{V1} R_{in} R gm$  then,

$$\Delta V_K \approx \frac{I_{FS}}{G^*_{imax} R_{in} gm} \quad (41)$$

Note that,

$$G_i = gm (G^*_i R_{in}) \Delta V_K \quad (42)$$

Finally, combining equation 41 and equation 42 results in

$$G_i = G^*_i \frac{I_{FS}}{G^*_{imax}} \quad (43)$$

The maximum element in the input vector is always normalized to some predetermined full scale value, while all other elements are ratioed corresponding to their absolute values with respect to a maximum value. Note that this scheme represents the normalized vector in terms of ratios of relative value of the individual elements with respect to maximum element in the input vector.

#### Square Law Bulb Normalization Function

The block diagram of the square law bulb normalization function is shown in Figure 5. Conceptually, this scheme is similar to schemes described previously except for some important features. GLA model [17] calls for the scaling of un-normalized glomerulus activity  $G^*_i$  by a suitable scaler  $V_K$ , such that the sum of the *non-linearly* processed and scaled un-normalized glomerulus activity is constant (equation 2 and equation 3). The effect of such a nonlinear normalization on an overall clustering process is discussed in multi-sampling section. The nonlinear sigmoid-like transfer function is mathematically characterized by equation 4.

The normalization scheme is comprised of: the multiplier, the approximate sigmoid function  $g_s(\cdot)$ , and on or off chip operational amplifier. From Figure 5 and equation 33, the AGC scaled vector  $G'_i$  is

$$G'_i = gm (G^* R_{in}) \Delta V_K \quad (44)$$

and the normalized activity is

$$G_i = g_s (R G'_i) \quad (45)$$

where  $gm$  and  $R_{in}$  carry their usual meaning,  $g_s(\cdot)$  is the approximate sigmoid transfer function given by equation 4, and scaler  $V_K$  is the smallest value that satisfies

$$\sum_{i=1}^g g_s(R G'_i) = K_G \quad (46)$$

In closed loop,  $V_K$  settles at

$$V_K = \left( K_G - \sum_{i=1}^g g_s(R G'_i) \right) R A_{VT} \quad (47)$$

The following sections discuss the electronic realization of the approximate sigmoidal transfer function.

### Approximate Sigmoidal Function

The squashing cell is shown in Figure 18 [42]. It takes advantage of the inherent nonlinear drain to source I-V characteristics of a MOS device to generate the continuously differentiable and gain programmable transfer function. The cell is versatile, is extremely simple to design, and provides independent voltage or current programmable control of the gain. From an analog electronic system perspective, the sigmoidal non-linearity can



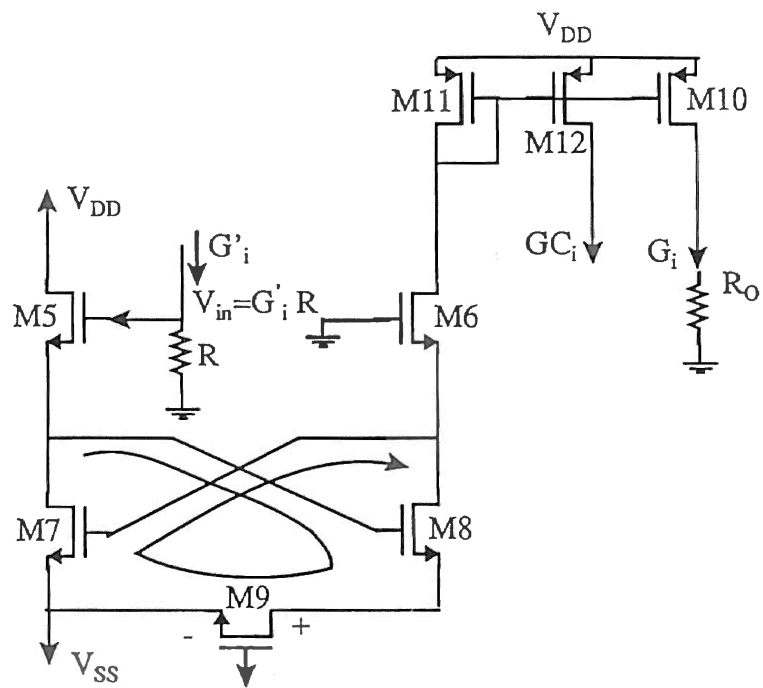


Figure 18. Approximate Sigmoidal Function

be thought of as an amplifier with a nonlinear transconductance. This results in nonlinear DC transfer characteristics. The gain is the slope of the output-input curve at a specific input excitation level. It varies from a low value at large positive or negative excitations (flat portions of the curve in Figure 20), to a maximum value at zero excitation. This non-linear transconductance normalizes the input activity. In this novel cell, the inherent nonlinear drain to source I-V characteristics of the MOS device are utilized to generate high gain near zero crossover using the triode region and low gain using the saturation region at high excitations.

In Figure 18,  $G_i R$ ,  $\Theta_G$ ,  $+V_C$ , and  $G_i$  are the input voltage, the threshold controlled voltage, the sigmoidal gain control voltage, and the normalized output currents respectively.  $G_i$  forms the input to the mitral patch used for further processing of the signal, whereas  $GC_i$  in a closed loop is used to set the output activity to  $K_G$  percent. The geometries of  $M_{5,6,7,8}$  are designed such that all of the MOS devices operate in the saturation region. Applying KVL around the loop shown, results in the following voltage loop equation:

$$\begin{aligned} G_i R &= V_{GS5} + V_{GS8} + V_{DS9} - V_{GS7} - V_{GS6} \\ &= \sqrt{\frac{2I_{D5}}{\beta_5}} + V_{T5} + \sqrt{\frac{2I_{D8}}{\beta_8}} + V_{T8} + V_{DS9} - \sqrt{\frac{2I_{D7}}{\beta_7}} - V_{T7} - \sqrt{\frac{2I_{D6}}{\beta_6}} - V_{T6} \pm \theta_F \end{aligned} \quad (48)$$

From Figure 18,  $I_{D5}=I_{D7}$ , and  $I_{D6}=I_{D8}$ . Selecting  $(W/L)_5=(W/L)_7$  and  $(W/L)_6=(W/L)_8$  results in  $\beta_5=\beta_7$  and  $\beta_6=\beta_8$ . Since they are all n type devices, it is assumed that the threshold voltages of all of the devices are matched. However, because of different body potentials there will be slight mismatch in the threshold voltages. Assuming matched  $V_T$ 's, equation 48 simplifies to

$$G'_i R = V_{DS9} \pm \theta_F \quad (49)$$

Noting that  $G'_i R$  is impressed across  $M_9$ , and using an accurate strong inversion model [43] of an n-channel MOS transistor operating in the triode and saturation regions, the drain or output current of the NMOS device operating in the triode and saturation regions is modeled as

$$\begin{aligned} I_{D9} &= K_n \left( \frac{W}{L} \right)_9 \left[ (V_C - V_{T9}) G'_i R - \frac{(G'_i R)^2}{2} \right] (1 + \lambda G'_i R) \quad ; \quad V_C - V_{T9} \geq G'_i R \pm \theta_F \\ &= \frac{K_n}{2} \left( \frac{W}{L} \right)_9 (V_C - V_{T9})^2 (1 + \lambda G'_i R) \quad ; \quad V_C - V_{T9} \leq G'_i R \pm \theta_F \end{aligned} \quad (50)$$

where  $\lambda$  is the channel length modulation parameter. In general,

$$\begin{aligned} G_i &= I_{D9} \\ &= g_s(G'_i) \end{aligned} \quad (51)$$

It is important to note that in the transition between the triode and saturation regions, commonly referred to as the moderate inversion, the MOS model neither fits into the triode nor the saturation model. In many treatments, no moderate-inversion is defined. Sometimes this region is considered as the lower part of strong inversion. Such models can lead to large errors.

Note that the described squashing function operates in a single (1st) quadrant. The symmetrical two quadrant (1st and 3rd) operation can be achieved by incorporating the complementary equivalent into the circuit in Figure 18. At any instance, only one quadrant is operative depending on the polarity of the input voltage  $V_{in}$ . Symmetry in two quadrants is maintained by the proper selection of device geometries in their respective parts. The simulations and fabrication are based on the two quadrant squashing function.

Note that in Figure 18, the R can be replaced by a MOS transistor operating in linear region.

In summary, the squashing circuit transfers the input voltage across the drain to the source of the transistor  $M_9$ . Over the supply voltage range, this MOSFET has a continuously differentiable I-V characteristics. In the triode ( $|V_C - V_{T9}| > G_i' R$ ) and saturation ( $|V_C - V_{T9}| < G_i' R$ ) regions, the transconductance gain,  $g_{ds}$ , is  $\beta(V_{GS} - V_T - V_{DS})$  and  $\lambda I_D$ , respectively. The resulting nonlinear drain current is mirrored by  $M_{11-12}$ . The current is linear at small values of input voltage and saturates, as the input voltage increases and the transistor  $M_9$  enters into the saturation region.

The input voltage  $G_i' R$  is "squashed" into a nonlinear current that is made available as an output after current mirroring. The important features are independently programmable control of the sigmoidal gain and offset  $\Theta_G$ . The saturation knee point can be placed anywhere simply by proper combination of gate voltages and geometry of  $M_9$ . Finally, it is important to point out that this circuit achieves voltage to current conversion (transconductance).

Simulations. Figure 19 shows the AC response of the cell. The cell achieves a bandwidth on the order of 10 Mhz into a one Megaohm load. The SPICE simulations of the DC transfer characteristics of the cell are shown in Figure 20. The family of curves is obtained by ramping the input voltage  $V_{in}$  from -3 V to 3 V for different sigmoidal gain voltages,  $V_C$ . The output current  $G_i$  is sampled via  $R_O$ .

Testing. The DC transfer characteristics obtained from the experimental data are shown in the Figure 21. When compared with the simulation results for the same  $V_{in}$  and

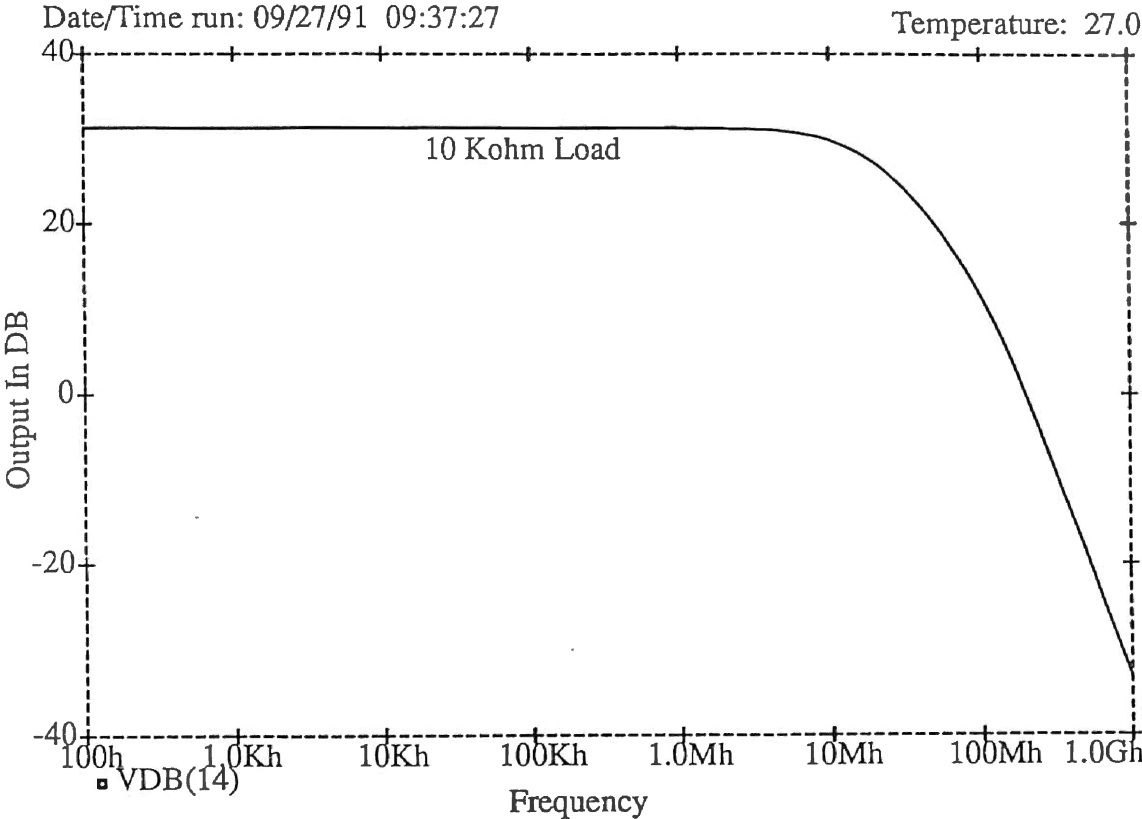


Figure 19. AC Response of Squashing Function

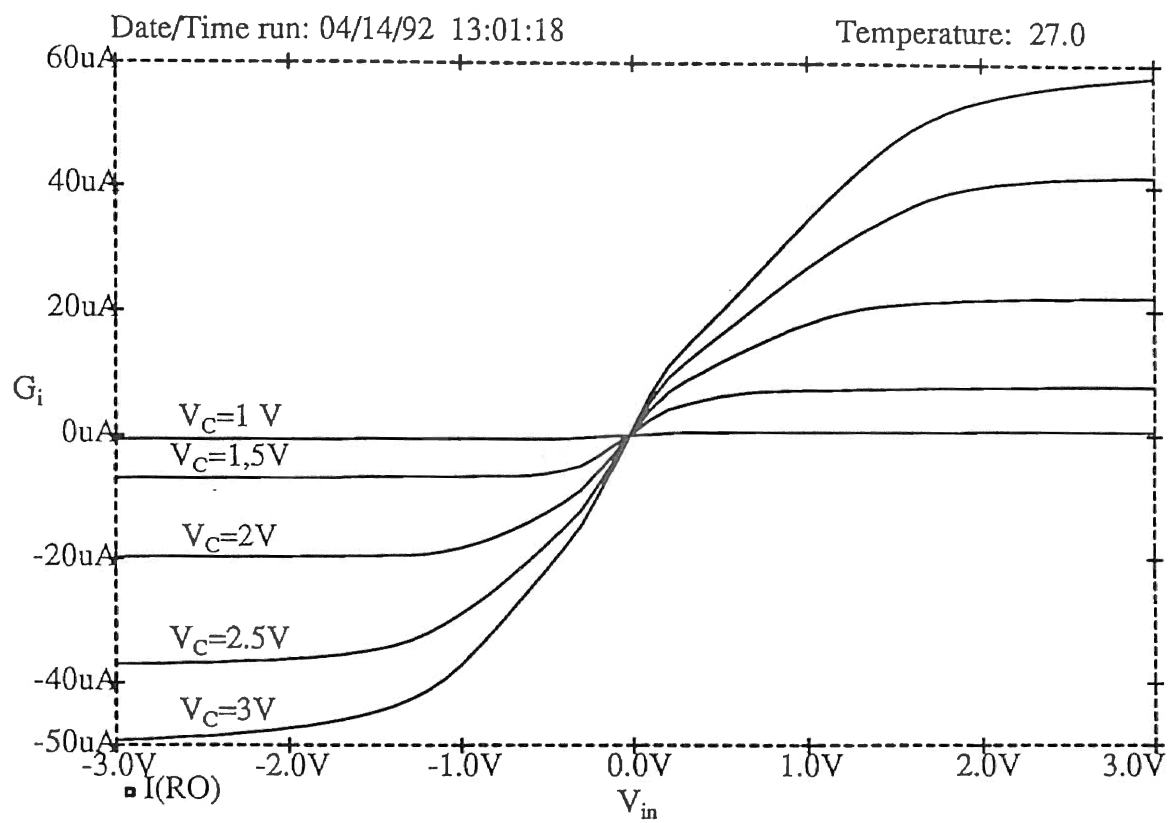


Figure 20. DC Transfer Characteristics of Sigmoidal Function Obtained From Simulations

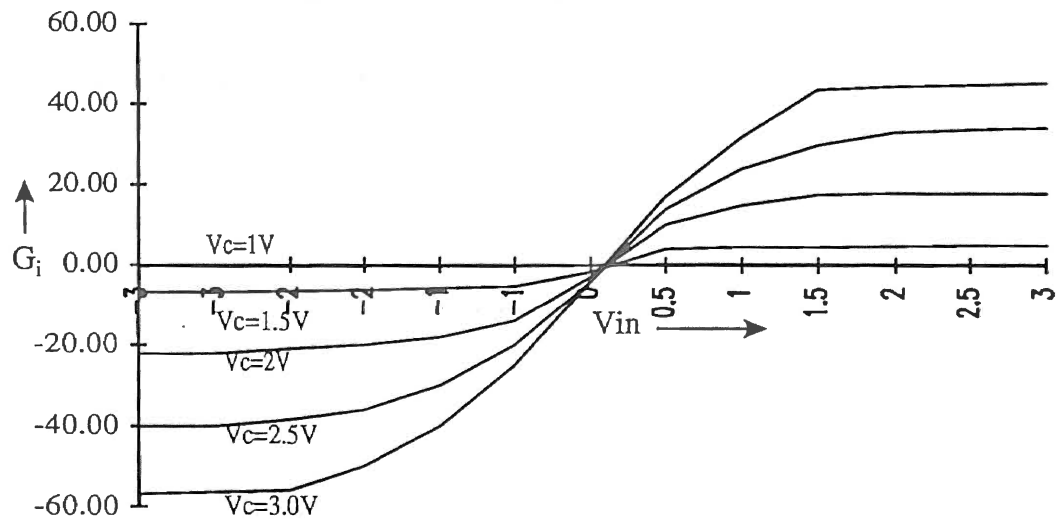


Figure 21. DC Transfer Characteristics of sigmoidal Function (Test Results)

$V_C$ , current in the 1st quadrant due to n the devices is lower than in the 3rd quadrant due to p devices. This is due to the threshold and beta mismatches between the n and p devices. With the  $V_C$ 's set at low values, the threshold mismatch was found to be 0.266 V.  $+V_C$  and  $-V_C$  were adjusted for threshold mismatch before recording the test data. The current mismatch at higher values of  $V_C$ 's is mainly due to the beta mismatch and the channel length modulations for large voltages of n and p devices. The testing data closely follows equation 50.

The small signal transient step response rise and fall times with a 10 k $\Omega$  output resistor and 20 pF oscilloscope probe capacitance plus test fixture capacitance are found to be 2.5  $\mu$ s and 2.25  $\mu$ s respectively.

### Mitral Patch

The bulb simulations consider  $m \times g$  projections (mitral cells). Projections are divided into  $g$  separate groups (mitral patches). Each mitral patch is excited by the normalized input from one group of peripheral receptors. The normalized output ensures input activity of mitral patch significantly large in amplitude to strongly activate the mitral patch and to keep the total number of mitral cells that are activated reasonably constant across the input vectors with different intensities and compositions. Within the mitral patch, the intensity of a normalized input is thermometer coded by the number of active cells. In other words, a thermometer code is an output representation, in which input activity is linearly coded by the increased number of units being triggered for the increased input activity. Thus, depending on the input activity, each mitral patch is spatially expanded from 1 to  $j$  thermometer coded LOT lines which project onto the



pyriform neurons. The mitral patch implements A/D conversion with a logical thermometer code.

The circuit diagram of the mitral patch is shown in Figure 22. It is comprised of a global capacitor reference ladder that sets the full scale current into  $m$  equidistance global thresholds,  $\Theta M_j$ . These thresholds are then compared to the input by  $m$  comparators in each mitral patch. The output currents,  $G_i$ , of the normalizing glomerulus function are used as an input to mitral patch. The scheme is equivalent to front end of the  $m$  level flash A/D converter, generating the thermometer coded digital LOT lines.

After mirroring through  $n$  mirror stage  $M_{R3-4}$  and  $p$  mirror stage  $M_{R1-2}$ , the full scale current  $I_{FS}$  is dropped across the active load ( $M_{R5-6}$ ) creating a full scale voltage reference. This voltage reference is impressed across the ladder of  $m$  identical MOS capacitors producing  $m$  voltage levels. The poly-1 to poly-2 unit capacitance has a tolerance of  $\pm 6$  Ff/ $\mu\text{m}^2$  with a typical value of 50 Ff/ $\mu\text{m}^2$ . Looking into the comparator, if  $C_{GS}$  is the gate to source capacitor of  $M_2$  then, the  $j$ th mitral cell threshold voltage is given by:

$$V_j = \left[ \frac{V_{j+1} + V_{j-1}}{2 + \frac{C_{GS}}{2C}} + V_{in} \frac{C_{GS}}{2C} \right] \quad (52)$$

$$\approx \frac{V_{j+1} + V_{j-1}}{2} \quad \text{for } C_{GS} \ll 2C$$

Nonidentical step capacitances result in non-equidistance threshold levels, if  $C \gg C_{GS}$ , then

$$C_1 = C \quad \& \quad (53)$$

$$C_j = \frac{C}{j}$$

Thus step threshold voltages can be approximated as,

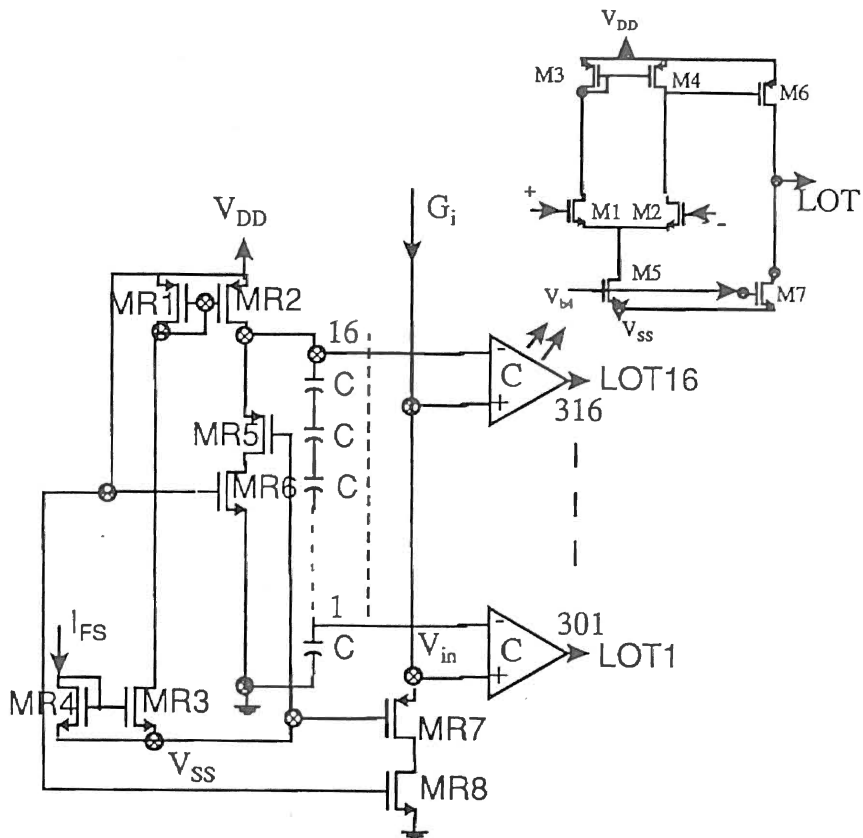


Figure 22. Mitral Patch

$$\Theta M_j = V_{FS} \frac{j}{m} \quad (54)$$

where  $j$  is the mitral cell index and  $m$  is the total number of mitral cells per mitral patch.

The output current of each glomerulus is equilibrated across an identical active load  $M_{R7-8}$ . The resulting voltage drop is compared with threshold voltages using a series of comparators. The two stage comparator is shown in the Figure 22 [44]. The low gain of the differential stage is augmented by the gain of the current sink inverting stage. The problem associated with such a comparator is a poorly predicted trip-point voltage.

### Simulations

The SPICE simulations of the DC and transient characteristics for the mitral patch circuit are shown in the Figures 23 and 24, respectively. With the full scale current set at  $64 \mu\text{A}$ , the DC characteristics are obtained by ramping the normalized bulb input currents  $G_i$ , from 0 to  $64 \mu\text{A}$ .  $V_{1-16}$  are the threshold voltages applied to the inverting input of the comparator,  $V_{301-316}$  are the digital output voltages of comparators and  $V_{in}$  is the non-inverting input of the comparator. As  $V_{in}$  crosses threshold voltage, the output of the corresponding comparator is driven high. In this manner, the intensity of a normalized input is thermometer coded by the number of cells that normalized input activates.

The transient step response reveals that the LOT lines that are thresholded near the full scale current are slower than those that are thresholded near ground. This is due to the changing input differential voltages as a function of ladder position. The differential voltage is a maximum at lowest threshold  $\Theta M_1$ , and minimum at highest threshold  $\Theta M_m$ .

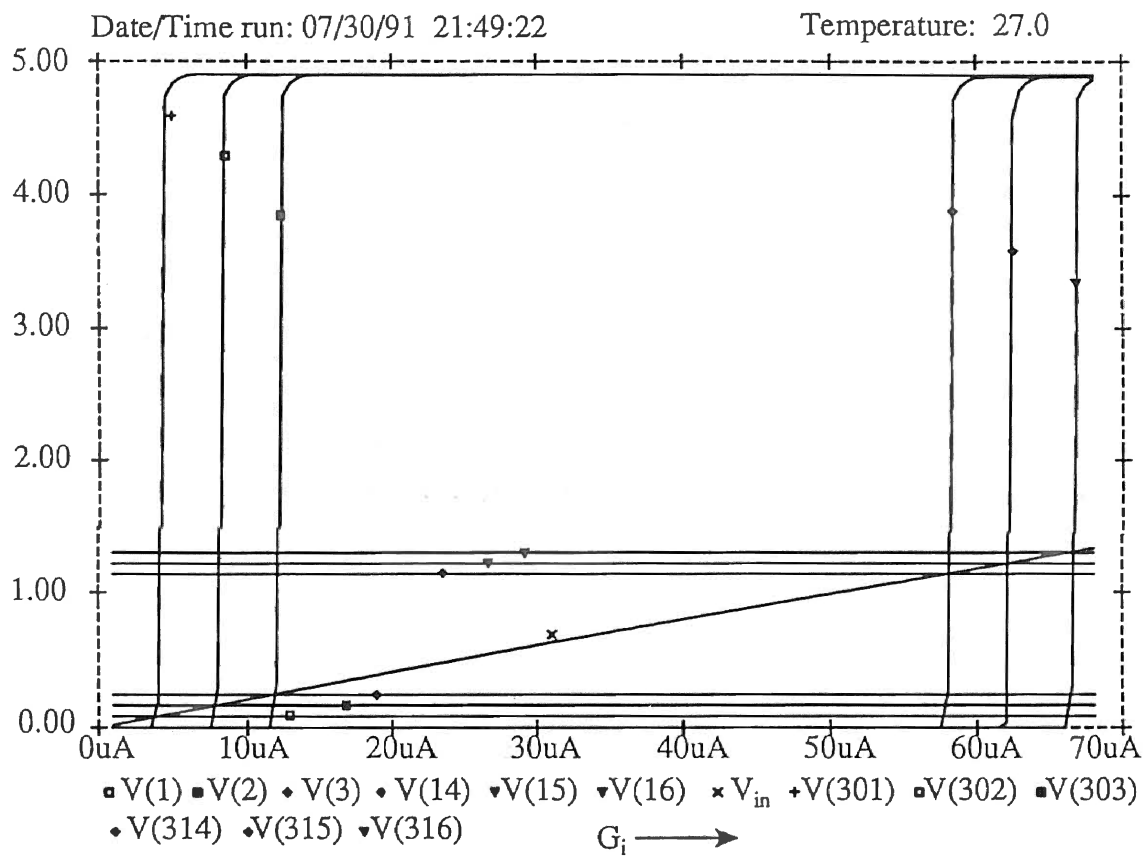


Figure 23. DC Response of Mitral Cells

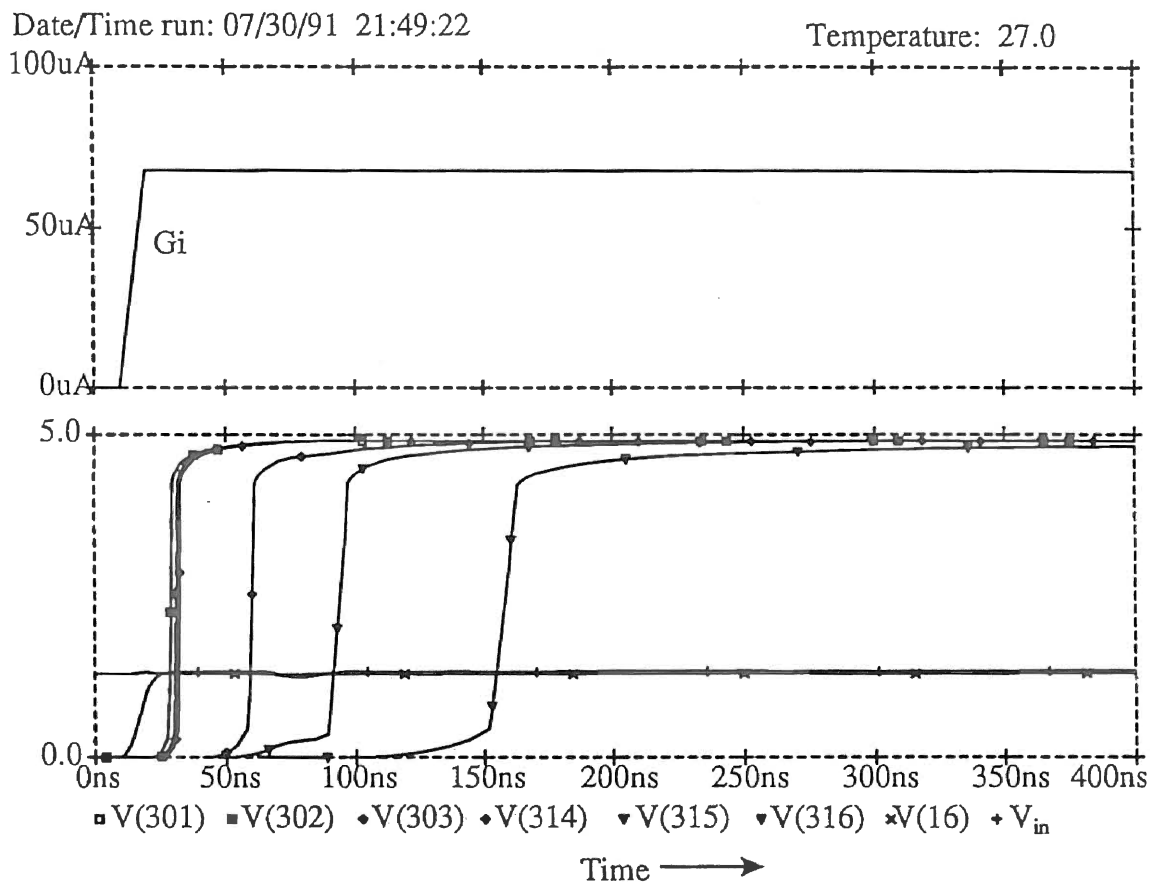


Figure 24. Transient Response of Mitral Cells

## Testing

The limitation on the package pins restricted external access to only a few LOT lines. To confirm the proper functionality of the capacitive ladder, LOT<sub>1</sub>, LOT<sub>15</sub>, and LOT<sub>16</sub> are connected to the pad-frame. With  $I_{FS}$  set to a known positive value,  $G_i$  is varied from zero to  $I_{FS}$   $\mu$ A and the state of the LOT lines is observed. The global capacitive ladder is suppose to set the full scale current into  $m$  (16 in this case) equidistance global thresholds. With  $I_{FS}$  equal to 64  $\mu$ A, theoretical toggling levels for LOT<sub>1</sub>, LOT<sub>15</sub>, and LOT<sub>16</sub> are 4  $\mu$ A, 60  $\mu$ A, and 64  $\mu$ A, respectively. The corresponding toggling levels recorded from testing data are 4.74  $\mu$ A, 40  $\mu$ A and 43  $\mu$ A.

The large signal step transient response agrees with the theoretical conclusion, i.e., LOT lines that are thresholded near  $I_{FS}$  are slower compared to those that are thresholded near ground. The rise time of LOT<sub>1</sub> is 4  $\mu$ s, LOT<sub>15</sub> is 10  $\mu$ s, and LOT<sub>16</sub> is 12  $\mu$ s.

### Bi-directional Voltage/Current Buffers

The bi-directional voltage/current (BiVI) buffers that are based on the current conveyor concept are shown in Figure 25. These buffers provide the dual functions of voltage drivers and current sources/sinks to isolate the  $W$  matrix in the forward and backward mode. During the feed forward cycle, the BiVI buffers on the mitral side are configured as voltage controlled voltage sources, and the buffers on the pyriform side are configured as current controlled current sources. During the inhibition in the backward cycle, their roles are reversed. Bi-directional operation is achieved by switching  $S_1$  and  $S_2$ , at the  $Y$  inputs of the current conveyors, to either a reference voltage  $V_{ref}$  or to the

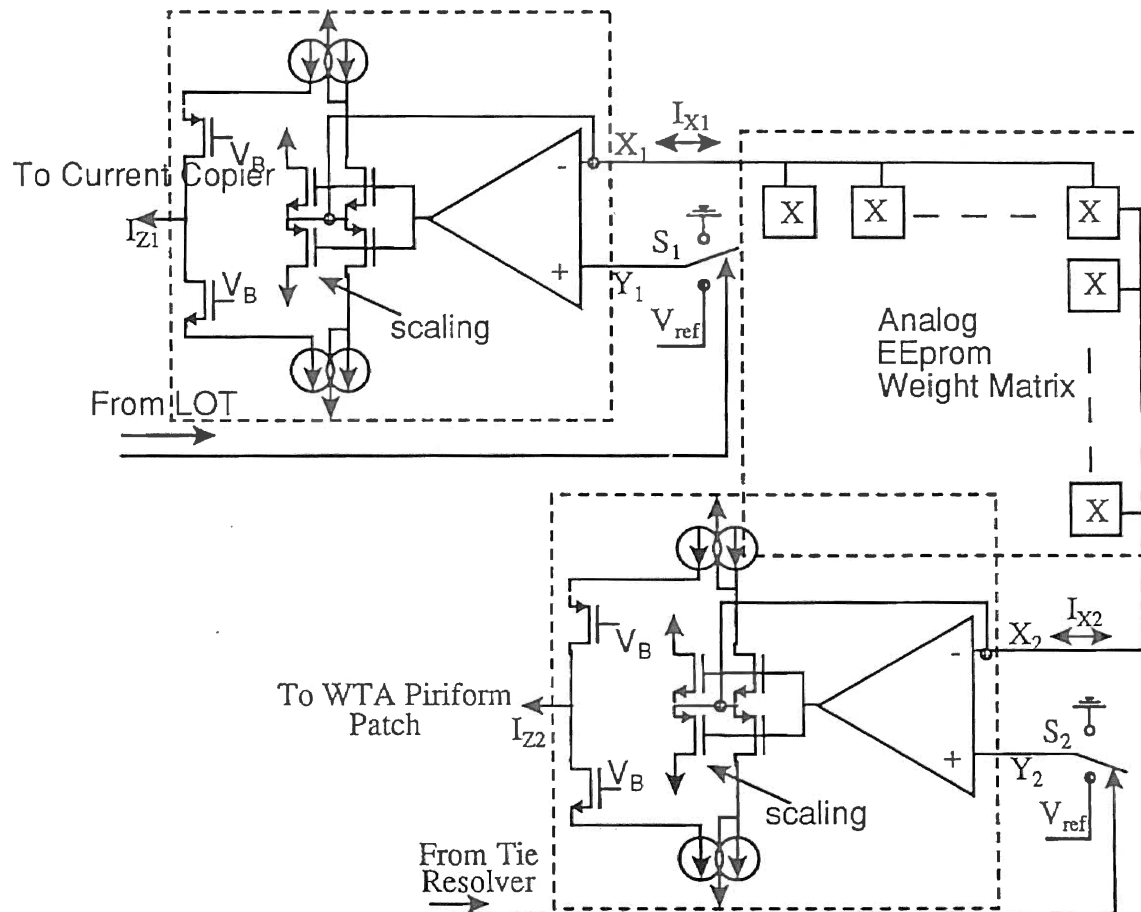


Figure 25. Bi-directional Voltage/Current Buffers

ground potential. When the Y input of one buffer is at ground, the other may be either at ground, or at  $V_{ref}$  causing a current flow proportional to the charge on the floating gate and  $V_{ref}$  to flow.

Current conveyor circuits began to emerge as an important class of circuits during the early 70's. They have proven to be functionally flexible and versatile, gaining acceptance as both a theoretical and a practical building block that offers an alternative way of abstracting complex functions. Current conveyors offer several advantages over conventional operational amplifiers. They provide higher gain over a greater signal bandwidth [46].

The block diagram of a CC is shown in Figure 26. Class-I (CCI $\pm$ ) and class-II (CCII $\pm$ ) conveyors have defined properties [45]. A CCII $\pm$  can be expressed in the following hybrid equations:

$$\begin{bmatrix} I_Y \\ V_X \\ I_Z \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \pm 1 & 0 \end{bmatrix} \begin{bmatrix} V_Y \\ I_X \\ V_Z \end{bmatrix} \quad (55)$$

The above equation states that no current flows into terminal Y, thus terminal Y exhibits an infinite input impedance. If the voltage is applied to input terminal Y, an equal voltage appears on the input terminal X, thus X exhibits a zero input impedance. Finally, an input current  $I_X$  on terminal X is conveyed to high impedance output terminal Z. The positive sign denotes that at any instant both,  $I_X$  and  $I_Z$  flow into or away from the conveyor signifying CCII+ while the minus sign denotes the opposite directions of the currents signifying CCII-.

The CCII may be viewed as an ideal MOS transistor [45]. The ideal behavior of the



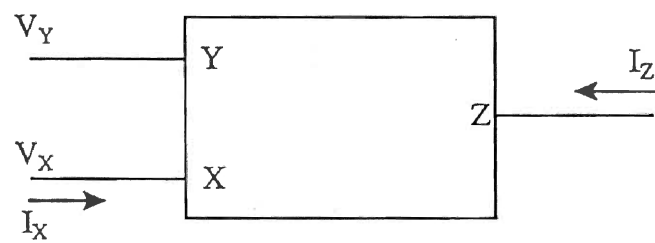


Figure 26. Block Diagram of the CC-II+

NMOS ( $M_{FN}$  in Figure 27) transistor can be achieved by incorporating transistor in the negative feedback loop of the operational amplifier. In which case, the current is restricted to flowing away from the X terminal. Similarly, with the PMOS ( $M_{FP}$ ) transistor incorporated in the feedback loop, current is restricted to flowing into the X terminal. Bi-directional current flow can be achieved by using a complementary pair of MOS transistors ( $M_{FN}$  and  $M_{FP}$ ) in the op-amp feedback loop. When mirrored by complementary mirrors, this current can be made available on the output node Z. Thus the input current  $I_x$  is conveyed to output current  $I_z$  (assuming  $ID(M_{7,8}) = 0$ ). The scaling of the input current can be obtained by designing proper mirror ratio or by providing an alternate parallel path for the current via branch  $M_{7,8}$ . Thus, allowing only a portion of input current to flow through the mirrors. This is a CCII+ realization since both,  $I_x$  and  $I_z$  simultaneously flow into or away from the conveyor.

The CMOS folded cascode op amp shown in the Figure 27 has been integrated on-chip to be used as the CCII+ op-amp. In the design of the op-amp, the locations of dominant poles are decided by high impedance nodes that are responsible for deteriorating the phase margin. In a simple two stage op amp, Miller compensation attempts to drive the pole at an output beyond the GB, while making the internal pole dominant. This scheme does not completely eliminate the output pole problem, since for large load capacitances, the output pole has tendency to shift back toward the origin resulting in unstable operation [46]. Since the input resistance of a folded cascode stage is very low ( $1/g_m$ ), the folded cascode eliminates the high impedance nodes and thus only one dominant pole exist at the output. In contrast to the two stage Miller compensated op amp, any increase in the load capacitance for the folded cascode increases the

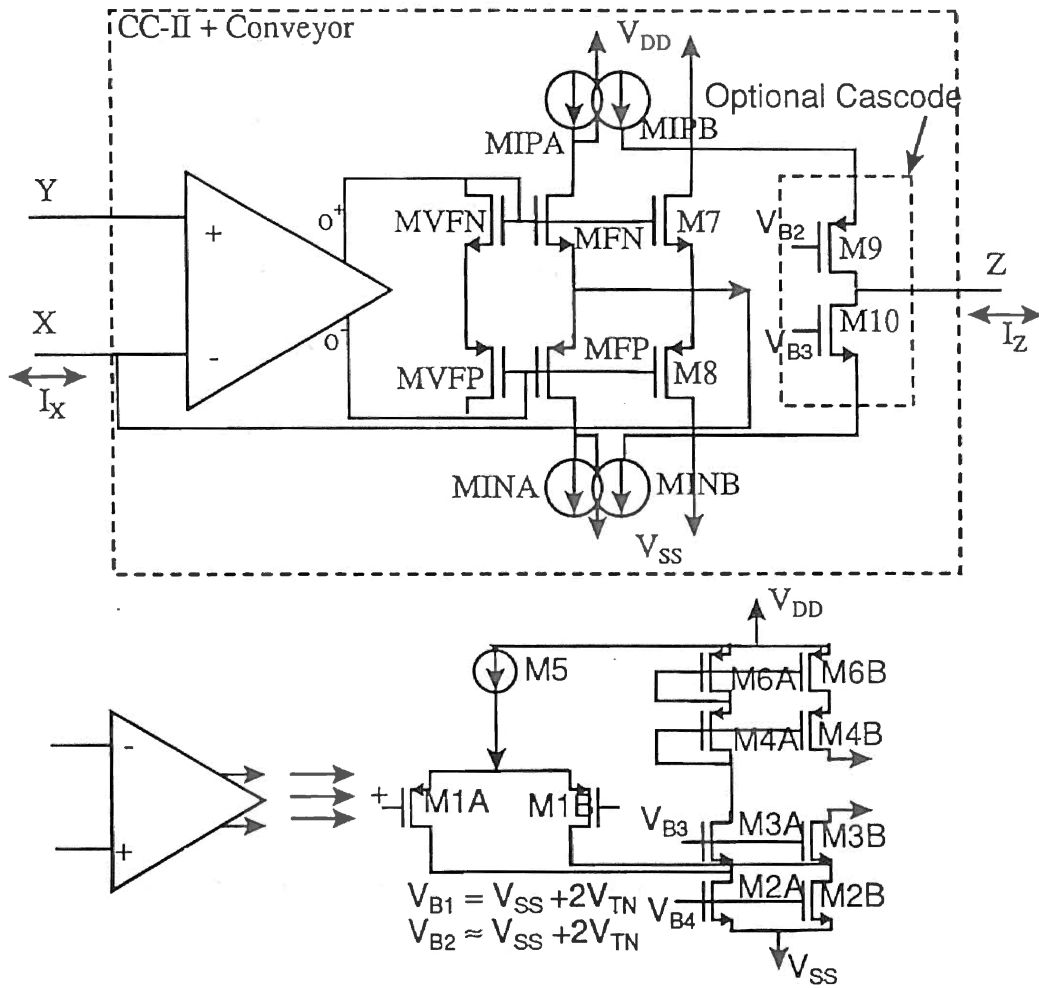


Figure 27. Bi-directional Voltage/Current Conveyor

compensation, resulting in a further increase in the phase margin [46].

In Figure 25, switch  $S_1$  is activated by the forward digital LOT signal and switch  $S_2$  is activated by the backward digital signal from the tie resolver. During the forward cycle, switch  $S_1$  is switched to  $V_{ref}$  while switch  $S_2$  is switched to ground. The voltage controlled voltage source configured CC ensures  $V_{X1}$  equals to  $V_{Y1}$ .  $S_2$  is switched to ground forcing  $X_2$  to the ground reference. If the LOT line is digitally high, then the potential difference between  $X_1$  and  $X_2$  ( $V_{ref}$ ), causes a current  $I_{X1}$  to flowing in the forward direction proportional to the charge on the floating gate and the value of  $V_{ref}$ . The current controlled current source configured CC ensures  $I_{Z2}$  equals to  $I_{X2}$ .  $I_{Z2}$  is then processed further by the winner take all circuit.

The BiVI buffers must be able to supply the total weight current in one column of the weight matrix. To accomplish this, the source/sink transistors,  $M_{IPA}$ ,  $M_{FN}$ ,  $M_{INA}$ , and  $M_{FP}$  must be sized appropriately. The current sources are sized to source currents in the voltage controlled voltage source mode, while the current sinks are sized to sink weighted currents in the current controlled current source mode.

During the backward phase,  $S_2$  is activated by lines from tie resolver switching  $Y_1$  either to the reference voltage or to the ground potential, depending on the state of the corresponding resolver line. A winning state results in  $Y_1$  being switched to  $V_{ref}$ . The voltage controlled voltage source configured CC ensures  $V_{X2}$  equals to  $V_{Y2}$ .  $S_1$  is switched to ground forcing  $X_1$  to be a virtual ground. If a WTA line is a logic high, then the potential difference between  $X_2$  and  $X_1$  causes current  $I_{X2}$  to flowing during the backward phase proportionally to the charge on the floating gate and the value of  $V_{ref}$ . The current controlled current source configured CC ensures  $I_{Z1}$  equals to  $I_{X1}$ . The

resulting  $I_{z1}$  is processed further by the current copier integrator circuit.

### Simulations

Figure 28 shows the transient response and Figure 29 shows the ac response of the CCI circuit. The CCII+ circuit is capable of source/sink 1 Ma of current while slewing a single weight current ( $40 \mu\text{A}$ ) in less than 400 ns into a 1 K ohm load. A small signal bandwidth is greater than 10 Mhz frequency.

The SPICE simulations of the DC transfer characteristics of the CCII+ conveyer are shown in Figure 30. The characteristics are obtained by ramping the input current  $I_x$  from the negative to the positive value. Over the range -2 Ma to 2 Ma, the output current  $I_z$  is a linear function of the input. The CCII+ loses its linearity as the internal transistors  $M_{FN}$  and  $M_{FP}$  begin to fall out of saturation.

### Testing

The DC transfer characteristics are obtained by ramping voltage  $V_Y$  from -2.5 V to 2.5 V. According to equation 55,  $V_X=V_Y$ . The test data indicates that  $V_X$  exactly tracks  $V_Y$ . The resistance connected between terminal X and ground, thus draws current  $I_x$  proportional to  $V_Y$ .  $I_x$  is conveyed to the output as  $I_z$  via CC. The  $I_x$ - $I_z$  transfer curve is shown in Figure 31. The test results are comparable to the simulations except that CC is linear over  $I_x$  range -2 Ma to 1.75 Ma compared to  $\pm 2$  Ma for the simulations.

The small signal transient step response rise and fall times are found to both be 2.5  $\mu\text{s}$ . The transient response times are limited by the parasitic capacitances at nodes X, Y, and Z.

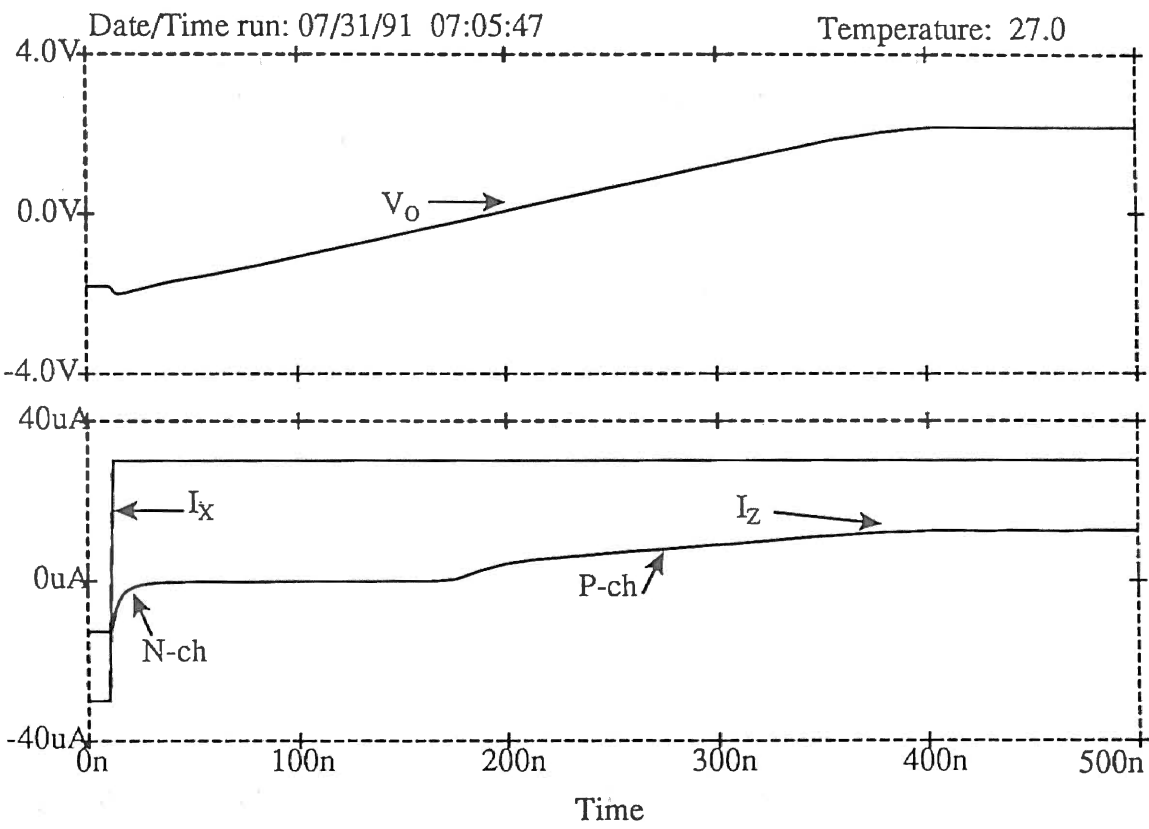


Figure 28. Transient Response of the Current Conveyor

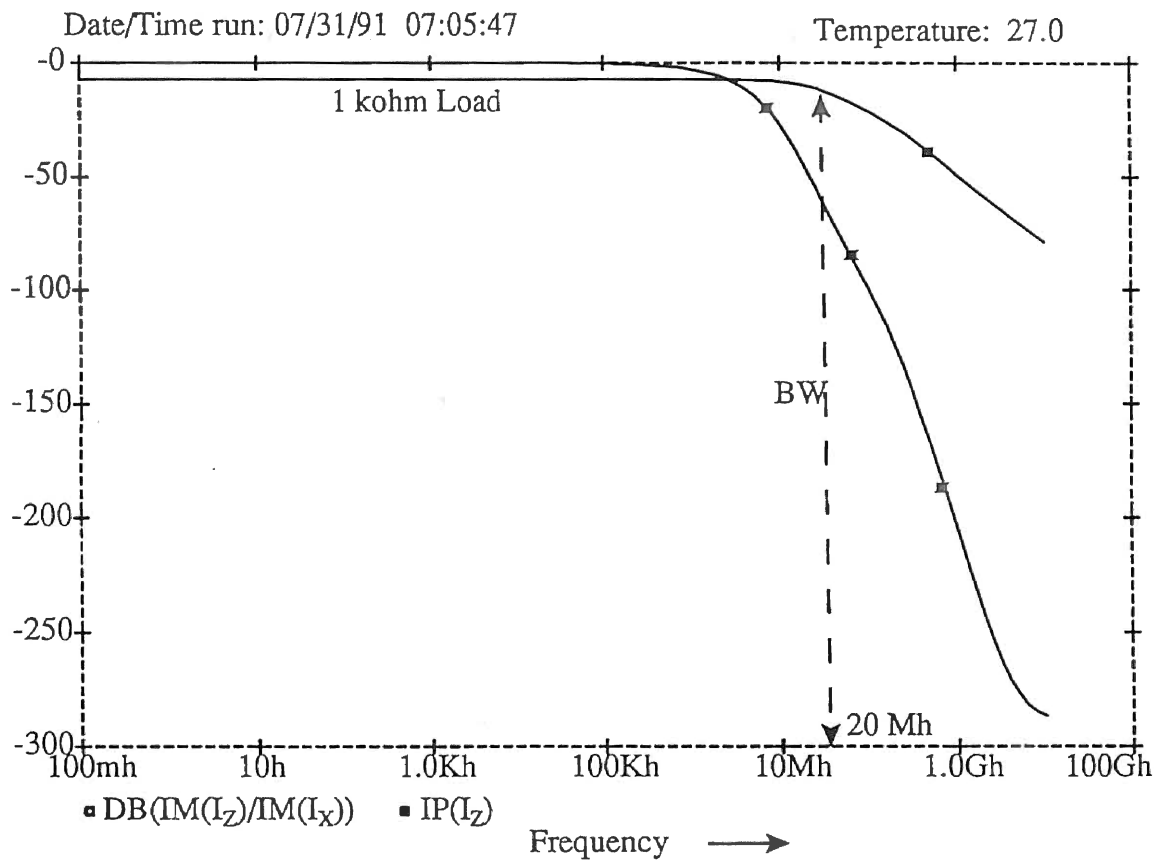


Figure 29. Current Conveyor AC Response

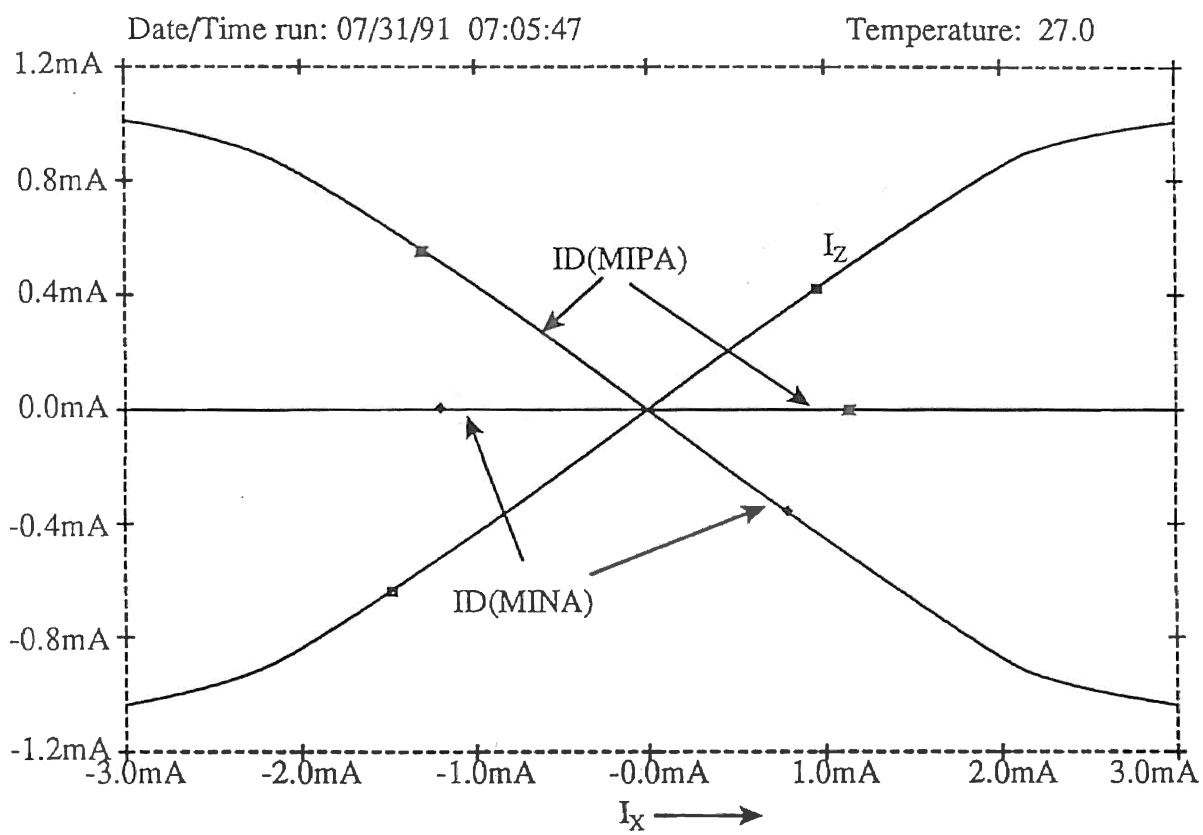


Figure 30. DC Transfer Characteristics of the CC Obtained From Simulations



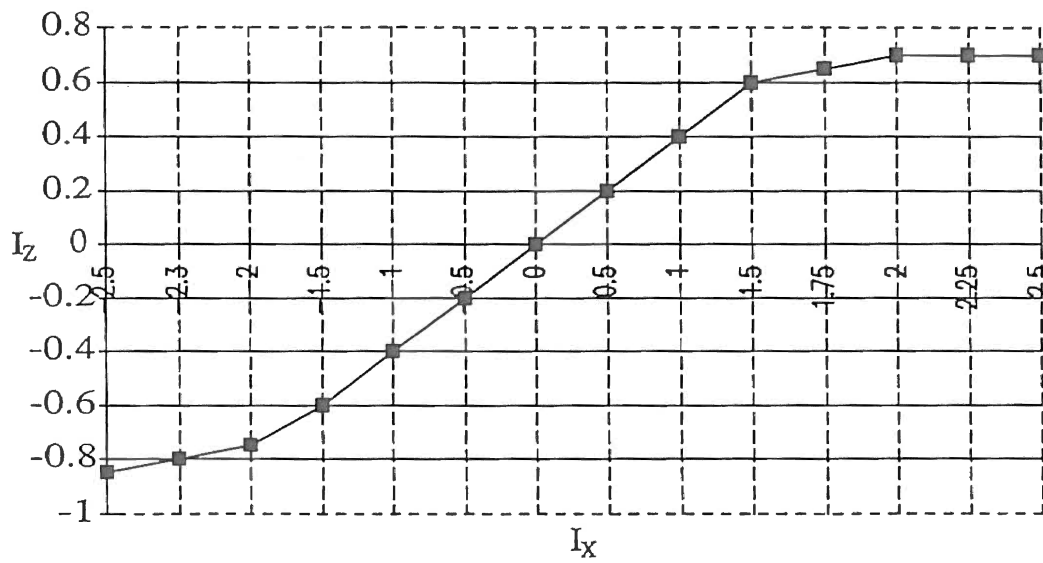


Figure 31. DC Transfer Characteristics of the CC Obtained from Test Results

## Weight Matrix

One of the most onerous requirements facing the designers of the neural networks integrated circuit (NNIC) is the appropriate selection of technology and circuit configuration to produce a memory with suitable characteristics. In general, from the electronic neural networks perspective, a memory element can be characterized by: (1) nature of memory, analog or digital (2) location, on-chip or off-chip (3) volatility, volatile or nonvolatile (4) programming/erasing method, electrical or non-electrical, and (5) the precision in bits. More often than not, the technology selection is restricted by factors such as the cost and availability of a particular process by the commercial vendors. Most of the research reported to date, requires a special processes such as an ultrathin window, nitrite oxide, and textured polysilicon.

Knowledge in the analog artificial neural networks is stored in the form of variable weights. Neural networks adapt themselves by modifying the strength of connecting weights according to the specific learning algorithms. This requires that the weight be easily altered in order to take a wide range of positive values. These weights must allow long term storage and must be locally stored to allow easy and rapid access. Storage of analog weights necessitates analog memories that are (1) truly non-volatility, for long term retention of the stored knowledge, (2) on-chip and rapidly programmable, to expedite the network learning by minimizing read and write times, and (3) application specific yet simple, for ease of fabrication. Strictly speaking, due to factors such as the learning rate in an ENN, discrete programming of true analog memories results in finite resolution, usually specified in bits. The electronic implementations of most widely used networks

including back propagation typically require resolution on the order of 5 bits or greater [47].

The favorable learning features of the GLA model are that the weights require only low precision on the order of three to five bits. The learning in the network comprises of course, unidirectional, and parallel real time weight updates which take place according to a simple Hebb-type co-active based update rule. The inherently slow multi-sampling process at theta rhythm (200 ms) can tolerate long programming times although fast updates are preferred. Due to the coarse learning, retentivity of 3-5 bit over 10 years at room temperature is allowed. Thus, in summary, to implement network learning with a sparse synaptic weights requires coarsely analog, non-volatile, electrically programmable/erasable memory with programming time on the order of 200 ms. Each memory element should be configured with a variable conductance synapse, whose conductance can be modulated by the nonvolatile weight. The sparse weight matrix  $W$  consist of sparsely placed electrically erasable/programmable transistors and randomly arranged in a 4x5 sub-matrix as shown in Figure 32.

In the past, attempts to build neural weights have resulted in simplified non-adaptable or discontinuously adaptable synaptic weights [48]. Some provide a continuous true analog nature, but do not store the weights locally on chip. This limits the computational capability of the NNIC or neural systems because the read and write become input/output limited resulting in very large developmental time. The numerous possibilities to build a memory element can be broadly classified as: digital semiconductor memories, analog semiconductor memories, i.e. CCD, and floating gate analog semiconductor memories.

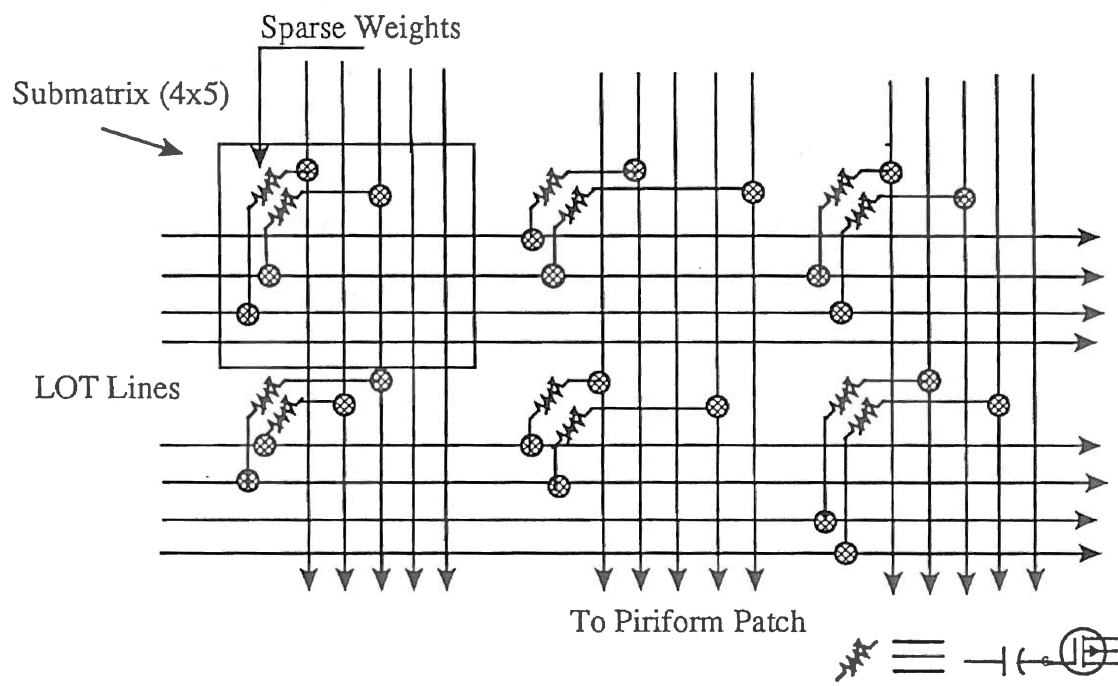


Figure 32. Weight Matrix

Semiconductor memories (e.g. SRAMs) are volatile in nature, that is, data content is lost when power is removed. This problem can be solved by using the fixed programmable memories or the mask programmable read only memories (ROM's), where data content is placed in the memory during the manufacturing process. This makes them non adaptable. Also, these memories require a large manufacturing volume of a particular program to recover the high fabrication cost. Programmable read only memories (PROM's) allow programming prior to use. These memories can be built using either bipolar technology (fusible link) or the MOS technology, e.g., the floating gate avalanche injection MOS (FAMOS) [49]. Bipolar devices are non adaptable because they cannot be erased once programmed. However, FAMOS can be erased by exposing it to ultraviolet rays. Unfortunately, none of the above memories truly satisfies the need of electrically programmable/erasable analog memory.

In digital semiconductor memories, the MOS capacitor holds data which is dynamically refreshed to preserve the data content. Weights can be stored in digital form and then converted into analog form by D/A converters (for example, M-DAC). This technique relies on the fact that the conductance or transconductance of a MOS transistor can be modulated by changing the transistor gate voltage. The transistor is operated in the triode region where non-linearity of the synapse is fairly low. Multiplexing and routing complexities make the parallel updating of weights in such architectures slow and complex. Proper trade off between quantization error and silicon area (RAM memory) is necessary. Along the similar lines, another technique is suggested by Y. Tsividis and S. Satyanarayana [50] where analog voltages are stored at the gate capacitance of the synaptic MOS transistor itself. They suggest canceling the inherent

non-linearity of a transistor by using complementary input voltages through the matched weighing transistor, or by passing the same voltages through the complementary weighing transistors: the n-channel and the p-channel. Learning takes place by addressing the proper capacitors and charging them according to a specified learning algorithm. Once the weights are settled (RC time constant), the capacitors are periodically accessed for reading, charging, and refreshing. This scheme suffers from a relatively short retentivity resulting in decreased accuracy. As a result, the network becomes "absent minded", forgetting information shortly after learning.

Floating-gate analog semiconductor memories have been proposed and studied by a number of researchers [51] as a suitable analog medium for the long-term storage of the weights. They serve the dual purpose of providing local on-chip weight storage on the floating gate of synaptic transistor. The transistor intern can be used as the variable synapse. The strength of the synaptic weight depends upon the stored charge on the floating gate. This type of memory element exhibits long term retention because no discharge path is available since the gate is surrounded by the dielectric material  $\text{SiO}_2$ .

The charge transport mechanism used by floating gate memories can broadly be classified as the avalanche injection of electrons [52], and Fowler-Nordheim tunneling of electrons [53]. Some use a combination of these two. There are four basic categories of avalanche injection [52]. In the avalanche injection of electrons, high energy electrons are generated within a substrate, to surmount the  $\text{SiO}_2$  barrier and to be injected onto the conductive floating silicon gate. While in the Fowler-Nordheim tunneling of electrons, a high voltage is placed across a thin oxide, typically a window across the floating gate. This impart sufficient energy to the electrons within the substrate to tunnel through the

SiO<sub>2</sub> barrier. However, the process by which the stored charge may be altered is highly nonlinear, sensitive to geometric and processing parameters, and can require high programming voltages (greater than 5 V). In general, it is a function of the applied electric field intensity, programming duration, and back emf. It is difficult to conceive precise modification of analog weights without feedback control. The most obvious solution is the use of coarse weights. A few researchers have proposed modification of established algorithms by using very coarse quantization weight updates [54].

One well known solution for adaptable weight is the metal nitride oxide-semiconductor (MNOS) technology [48]. A MNOS device has a variable threshold which can be electrically changed by a tunneling charge into an interfacial layer in the gate dielectric. By reversing the polarizing field, the charge can be tunneled out of the interfacial layer, thus making the device electrically writable/erasable. The MNOS fabrication process is complicated because the control of the silicon nitride-tunneling gate oxide is difficult. With some modifications to the FAMOS structure, it is possible to have an electrically programmable/erasable non-volatile memory. The most recent development is the dual injector floating gate MOS (DIFMOS). In the DIFMOS, like the FAMOS, data are stored on the floating gate which is charged by the avalanche injection of electrons. But unlike the FAMOS, erasure is achieved by the avalanche injection of holes. However, hole injection is an order of magnitude slower than electron injection [49].

The following sections briefly review the widely used floating gate semiconductor technologies.

### Floating Gate Avalanche Injection MOS Memory

The concept of an insulated gate field effect transistor with a floating gate as a nonvolatile memory element was first advanced by Khang and Sze [55]. The operation of the proposed structure is based on the charge transport from the silicon substrate across a thin insulator layer ( $\approx 50 \text{ \AA}$ ) to a floating metal electrode which is covered by a second insulator and the upper metal gate. The charge is stored in the floating metal gate in response to the applied voltage between the upper metal and the substrate. The formation of the metal gate over a very thin dielectric layer is the major obstacle in the practical realization of the proposed structure. The similar concept is involved in the MNOS structure in which the floating metal gate is replaced by a layer of traps of the silicon nitride. MNOS technology will be discussed in detail in the next section.

Nicollian et al. [56] reported that the high electron current densities can be achieved in the MOS capacitors by avalanche injection from the P type substrate at considerably lower current density than the hole injection from the N type substrate. The FAMOS structure uses this principle to avoid the basic drawback of Khang and Sze's structure. FAMOS combines the floating gate concept with an avalanche injection of electrons to yield a nonvolatile memory element [57,58].

The cross section of a FAMOS structure is shown in Figure 33. It is essentially a p-channel device in which no electrical contact is made with the silicon gate. The floating gate is formed by depositing a polysilicon layer over  $1000 \text{ \AA}$  or thinner gate oxide. Gate is isolated from the top by a  $1 \text{ \mu m}$  thick oxide. Initially all the terminals are at a common ground potential and the floating gate is at neutral. Also consider that a



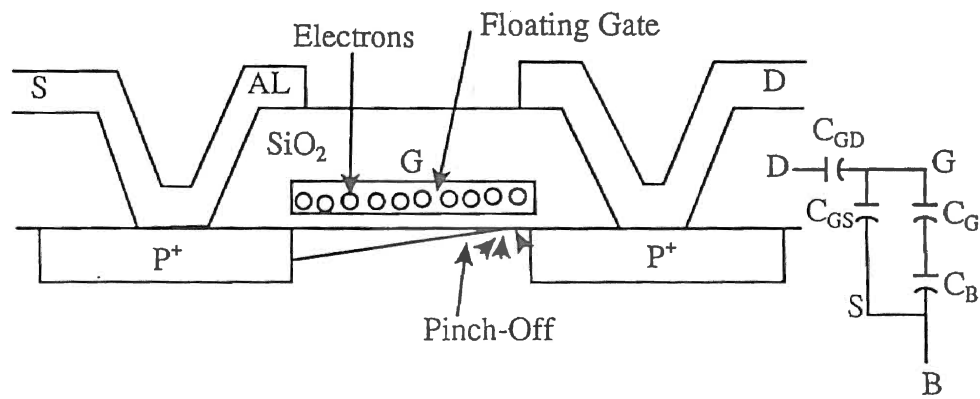


Figure 33. Cross-section of the FAMOS Structure

negative drain to the source voltage is applied. As the voltage increases, a positive drop appears across the overlap region between the floating gate and the P+ drain region. This drop tries to invert the heavily doped drain region. As a result, depletion takes place at the drain end near the SiO<sub>2</sub> interface. Eventually, the electric field induced in the surface depletion region reaches a point at which avalanche multiplication occurs. The generated high energy electrons acquire sufficient energy to surmount the SiO<sub>2</sub> barrier and to be swept towards the conductive floating silicon gate. This charge is responsible for the inversion layer underneath the Si-SiO<sub>2</sub> boundary. The amount of the charge transferred to the floating gate is a function of the amplitude and the duration of the applied p-n junction potential. The amount of the transferred charge can be determined by measuring drain to source conduction. The accumulation of charge changes the threshold of the MOS structure [57]. The change in the threshold voltage is given by:

$$\Delta V_T = \frac{-(Q_G - Q_{G(0)})}{C_o} \quad (56)$$

where  $Q_G$  is the final stored charge,  $Q_{G(0)}$  is the initial charge (if any), and  $C_o$  is the oxide capacitance. In general, the threshold voltage is given by:

$$V_T = V_{FB} + 2\phi_F - \frac{Q_B}{C_o} - \frac{Q_{G(0)}}{C_o} \quad (57)$$

where,  $V_{FB} = \phi_{SS} - \frac{Q_{SS}}{C_o}$

where  $V_{FB}$  is the flat-band voltage,  $\phi_{SS}$  is the polysilicon work function,  $Q_{SS}$  is the fixed charge at the Si-SiO<sub>2</sub> interface,  $\phi_F$  is the Fermi potential, and  $Q_B$  is the charge within the substrate.

The  $I_D$ - $V_{DS}$  characteristics of a charged and uncharged FAMOS device reveal that the device conducts even when there is no charge on the floating gate. This is due to the capacitive feedthrough voltage from drain to gate. The feedback voltage is given by:

$$V_{GF} = V_D \frac{C_{GD}}{C_{GD} + C_{GS} + C_{GB}} \quad (58)$$

where  $C_{GB}$  is the series combination of  $C_G$  and  $C_B$ . The  $I_D$ - $V_{DS}$  characteristics of the FAMOS device and the ordinary MOS device with its gate voltage equivalent to the amount of charge transferred on the floating gate of the FAMOS, are not the same. This is mainly due to the capacitive feedback to the floating gate. The amount of the feedback voltage depends on the value of the drain voltage. The variation in the feedback factor ( $\delta V_{GF} / \delta V_{DS}$ ) stems from the variation of the inter-electrode capacitance as a function of the drain voltage. When  $V_G > V_{DS} - V_T$  (triode region), the inversion layer extends from source to drain. Thus,  $C_G$  is splitted between drain and source equally. This increases the numerator of equation 58, thus increasing the feedback factor. At higher values of drain voltage (saturation), due to the pinch off of the channel,  $C_G$  is diverted to source. This decreases the numerator of equation 58, thus reducing the feedback factor.

Charge accumulation in a FAMOS is identical to that of a MOS whose gate is kept floating. A MOS transistor with a gate oxide thickness of 1000 Å takes approximately 80 V across drain to source before any appreciable gate current can be observed. In the same structure, avalanche-junction breakdown can occur at 30 V. Had this gate been floating, the avalanche injection would have resulted in the transfer of an equivalent amount of charge to the gate. This charge, divided by the oxide capacitance, gives the change in the threshold voltage. The amount of charge transferred is a function of the

applied junction voltage, programming duration, and the charge stored on the floating gate.

A stored charge of  $4 \times 10^6$  electron/cm<sup>2</sup> results in an electric field intensity of approximately  $2 \times 10^6$  V/cm across the thermal oxide [57]. If the polysilicon-SiO<sub>2</sub> barrier is assumed to be 3.2 eV, then the discharge current due to oxide leakage will be of the order of  $10^{-40}$  amp/cm<sup>2</sup> at 300° C. Retentivity plots at different initial charge and temperatures reveal drastic initial decay and thereafter a logarithmic decay [57]. Initially, negatively charged electrons counterbalance the positive charge accumulated at the Si-SiO<sub>2</sub> interface (due to the high dielectric field in the oxide created by the floating gate at elevated temperatures). The logarithmic retentivity is due to leakage through oxide.

Since the gate is surrounded by a dielectric, it is not accessible. Thus FAMOS is not electrically erasable. Due to a lack of evidence of substantial hole conduction through the oxide, the possibility of neutralizing electrons by the injection of holes from the substrate is doubtful. But Tarui et al. [59] have reported that hole injection is possible. With a slight modification of the basic FAMOS structure, electrical erasure is theoretically possible. In the modified FAMOS, like Khang and Sze's structure, the top gate is added to facilitate electrical programming/erasure. The device is held at a high positive voltage and programmed similar to the FAMOS structure. Erasure takes place with the top gate at ground or negative potential to favor hole injection into the floating gate. Classically, the device is restored to its neutral condition by exposing it to ultraviolet or X-ray radiation. Rays with suitable wavelength excite electrons to overcome the oxide barrier of approximately 4.3 eV. Erasure by X-ray radiation involves the generation of a hole electron pair in the oxide.

An interesting problem occurs when the device is in the read mode. Generally, the memory cell is read by sensing the drain current. This is done by applying low negative voltage, around -15 V, to the drain of the FAMOS. This raises the possibility of whether an uncharged memory cell can be slowly charged by repeatedly selecting it in the read mode, which is of-course undesirable. Empirical experiments demonstrate that such parasitic charging does not present a potential programming problem in memory cell operation [57].

### Metal Nitrite Oxide Silicon Memory

The process limitation in the formation of a metal layer over a very thin dielectric in Khang and Sze's structure, has led to the invention of the MNOS structure. It is typically used as a digital memory element in EEPROM. The structure is the same as the modified FAMOS, except that in lieu of a metal gate, a nitride layer is laid on the thin oxide. The top gate is made of polysilicon. For an n-channel MNOS device, a high gate voltage causes electrons to be injected from the substrate to the insulating silicon nitride layer. The injection uses the modified Fowler-Nordheim tunneling and other mechanisms [60,61]. The oxide thickness must be less than 50 Å. Trapped electrons in the dielectric nitrite layer result in a positive shift of the threshold. During electrical erasure, high negative gate voltages repel or drive electrons from the nitrite trap layer to the substrate. The threshold window (minimum and maximum amplitude in the threshold swings) is limited by the number of write/erase cycles. The degradation in swing is caused by a creation of surface states and surface charges due to the high field across the oxide layer applied during the first few programming/erasing pulses. The increased number of states

results in a loss of stored charge from the oxide-nitride surface, short term retention of weights and a reduction in threshold swing. The retention time in MNOS devices ranges from one to ten years, depending on the permittivity of nitride silicon [48].

### Dual Injector Floating Gate MOS Memory

From a neural networks integrated circuits perspective, one of the problems in the discussed memories, is the learning time. Any basic weighing memory cell operates in two modes: read and write. In the discussed memories, reading and programming can not be done simultaneously since terminals are common for read and write operations. In order to achieve both operations simultaneously, separate read and write terminals are necessary. The DIFMOS is a four terminal device [49]. Two of the terminals, the drain and source, function as a built-in electrometer for measuring the charge stored on the floating gate. The other two electrodes belong to the electron and hole injector diodes. The DIFMOS structure is shown in Figure 34. When reverse biased into avalanche breakdown, these injectors inject electrons and holes into the floating gate. Both injectors are excited by negative current sources. As programming proceeds, the charge on floating gate retards further accumulation of electrons due to back emf but encourages the injection of positively charged holes. The level of the drain current indicates the state of the device.

The DIFMOS basically consists of a sensing transistor, a floating gate, an electron injector, and a hole injector. The bootstrap capacitor functions as a part of the hole injector by providing a favorable electric field for hole injection. Because greater current densities can be achieved with lower electric fields from the majority injection, DIFMOS

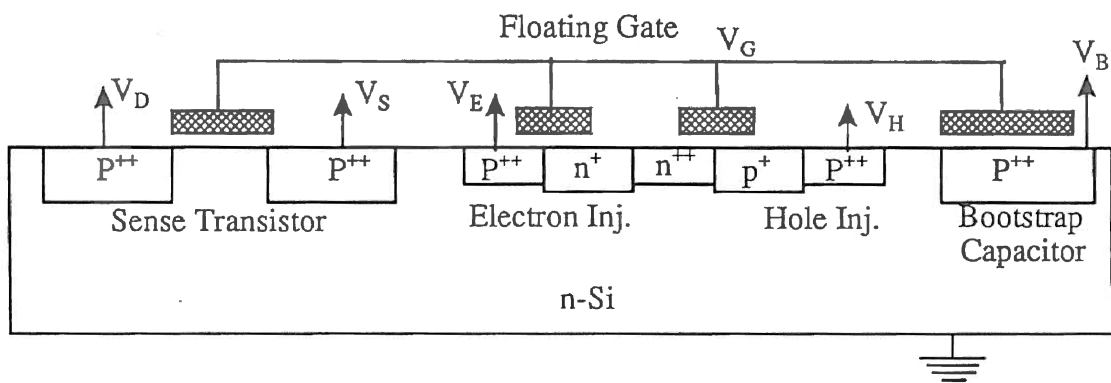


Figure 34. DIFMOS Structure

uses p<sup>+</sup>p<sup>-</sup> junctions for its electron injection and p-n<sup>+</sup> junctions for hole injectors.

Ideally, the hole injector should discharge the floating gate to the cutoff voltage. But it is not capable of discharging the floating gate below the threshold  $V_T$  of the hole injector. This problem can be overcome by using a bootstrap capacitor. The capacitor is formed between a floating gate and the p diffusion. During normal operation,  $V_B$  is held at the substrate potential. During the erase operation, sufficiently negative voltage is applied to  $V_B$  which capacitively couples a voltage to the floating gate equivalent to the minimum discharge threshold voltage. Erasing action only occurs when both, the bootstrap capacitor and the hole injector are operated simultaneously. The capacitively coupled voltage is given by:

$$\delta V_{GB} = \frac{C_B}{C} \delta V_B \quad (59)$$

where  $C_B$  is the bootstrap capacitor,  $C$  is the total floating gate capacitance,  $C_B/C$  is referred to as coupling ratio. The minimum bootstrap voltage required is given by:

$$V_{Bmin} = \frac{C}{C_B} V_{thx} \quad (60)$$

where  $V_T$  is the threshold voltage.

Performance measurements were reported by M. Gosney [48]. Programming pulses of 500  $\mu$ A and 50  $\mu$ s duration were used for the write and erase operations. The bootstrap voltage pulse was -40 V for 100 ms. The bootstrap voltage is applied just before the hole injector avalanche is turned on. After avalanche is over, the bootstrap is removed. Timing of the bootstrap and avalanche is not critical, but both must be present for the erase operation. The device suffers from write/erase time limitations which are



several orders of magnitude slower than read time. Therefore, the DIFMOS will generally be limited to read-mostly applications. The device suffers from the trapping of holes and electrons in the oxide as all others memories do. As traps are filled, the charging and discharging times become longer. For a given voltage configuration, the decay in gate voltage is approximately linear with the logarithm of the number of write/erase cycles. Endurance (life) is a function of the cumulative trap charge. Trapped charges reduces the gate voltage window. At room temperature, retention is measured at 0.06 percent/decade, while at elevated temperatures of 80° C, it is approximately 1 percent per decade [48].

From a fabrication perspective, the DIFMOS and the CMOS are nearly equal in process complexity. The FAMOS is much simpler but has no electrical erasure ability. The process comparison among the PMOS, FAMOS, CMOS, DIFMOS, and MNOS, is reported by Gosney [48].

#### Floating Gate Analog Memory in Standard CMOS Process

The memories discussed above require a special fabrication process such as ultrathin window, nitrite trap oxide, or a conventional textured polysilicon. These processes are not yet matured. Usually, these special processes are expensive and simply not available in many design environments, especially universities. In order to fulfill the need of an analog neural network designers for programmable memories, existing standard CMOS process without modifications must be able to provide a solution to realize floating gate memories. Recently several such implementations have been reported [62,63].

Based on the limitations discussed above, we propose that the sparse weight matrix  $W$  to be implemented in a standard CMOS process. This memory takes advantage of the mask geometries to cause the field-enhanced Fowler-Nordheim tunneling of the electrons from a substrate through a standard gate oxide of thickness 40 nm at relatively low programming voltages. Unlike the existing methods for the tunneling of electrons through a thick oxide by field enhancement, this method does not require a special process for textured-surface polysilicon, nor does it require an ultrathin gate oxide. Instead, the mask geometric factors induced by the physical shape of the gate are used to enhance the electric field strength at the  $\text{SiO}_2$  interface. The following section discusses this weighing memory in detail.

### Memory Structure

The test structure designed to understand the charge transport mechanism in the floating gate memory is fabricated in the two micron, p-well, double poly, double metal CMOS process with a gate-oxide thickness of 40 nm. The electrical equivalent schematic of the layout is shown in Figure 35.

There are four basic test cells. Each test cell consists of the following: (1) a current injector  $C_{inj}$ , for injecting and removing electrons to and from the floating gate, (2) a PMOS sense transistor  $M$ , for sensing charge on the floating gate, and finally (3) a bootstrap capacitor  $C_B$ , to allow external control and programming of the floating gate voltages without actually having an electrical connection between the programming gate and the floating gate. All four cells are identical except for their injector structures and the value of their bootstrap capacitor. The different injector shapes are deliberately

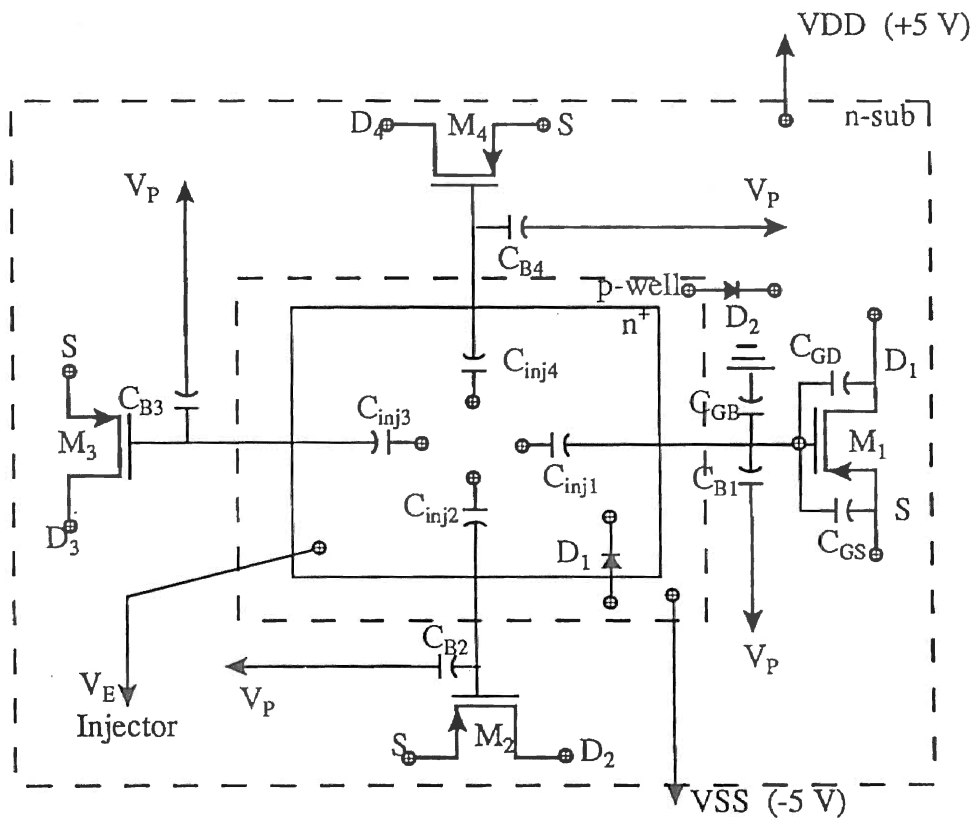


Figure 35. Electrical Equivalent Schematic of the Layout

chosen to assist in determining the effect of the injector structure on the tunneled charge. The  $C_B$  is sized to approximately maintain a constant  $C_B/C_{inj}$  ratio among all the test cells. The injector structure details are summarized in table I.

All four cells with the different injector structure are intended to be programmed or erased simultaneously in order to compare the geometrically dependent behavior of the charge injection at various points during programming. This arrangement removes effects that are present due to the variation in amplitude and duration of the programming pulses as well as the variation due to different drain to source voltage of the sense transistors. These effects are present if the devices are tested separately. The sources of all of the identical sense transistors are connected together and the same drain voltage is simultaneously impressed across them. This ensures the equal drain to source voltage across each memory cell and thus removes the effect of channel length modulation on the drain current.

In the injector structure, a self alignment process results in a lateral diffusion of the n+ region under the floating gate by a lateral diffusion factor WD. A floating polysilicon gate, ends up with its peripheral edge and corners over the n+ diffusion. Theoretically, the electric field due to the floating gate voltage is concentrated locally at the corners and may be along the peripheral edge. The exact field distribution density of the electric field is complex and believed to be a function of the geometry of the injector. Experimental results indicate that a field enhancement factor of 2 to 4 can be obtained [62]. In order to experimentally predict I-V curves, different combinations of corners and periphery as given in table 1 have been selected. We theorize that injector area does not play an important role in the tunneling process since the electric field generated by the

TABLE I  
INJECTOR STRUCTURES

Cell #	Corners		Injector Perimeter $\mu\text{m}$	Injector Area $\mu\text{m}^2$	Bootstrap Area $\mu\text{m}^2$	$C_B / C_{inj}$
	Internal	External				
1	4	2	30	54	900	9.92
2	6	4	24	14	225	9.57
3	10	8	40	38	625	9.79
4	4	2	18	22	361	9.77

$$C_B = 0.5 \text{ ffd}/\mu\text{m}^2$$

$$C_{inj} = 0.84 \text{ ffd}/\mu\text{m}^2$$

programming voltages is not sufficiently high to cause a significant amount of tunneling from the p well to the floating gate. However, the tunneling may be present along the edges of the injector.

The bootstrap capacitor is formed between poly-1 and poly-2. Poly 1 serves as a floating gate as well as the lower plate of the capacitor, while poly-2 acts as the upper plate of the capacitor. Thus, poly-1 is floated, i.e., electrically isolated from all the nodes. The poly-1 to poly-2 oxide thickness is 50 nm. Since the floating gate is surrounded by insulating SiO<sub>2</sub> from all the sides, charge leakage will be insignificant.

During the programming and erasing, the voltage difference between the floating gate and the n+ diffusion is responsible for the Fowler Nordheim tunneling of the electrons. The bootstrap capacitor is necessary to control and isolate the floating gate voltage. Figure 35 also shows the different parasitic capacitances associated with a memory cell. The percentage of the programming voltage that appears on the floating gate depends on the capacitive coupling ratio  $\alpha$ . This ratio is given by:

$$\alpha = \frac{C_B}{C_B + C_{GS} + C_{GB} + C_{GD} + C_{inj}} \quad (61)$$

where  $C_B$  is the bootstrap capacitor between the floating gate and the control gate across the poly-1 to poly-2 oxide,  $C_{GS}$  is the floating gate to source capacitance,  $C_{GB}$  is the capacitance between the floating gate and the bulk,  $C_{GD}$  is the capacitance between the floating gate and the drain, and  $C_{inj}$  is the injector capacitance across the gate oxide.

The voltage responsible for the tunneling is thus given by:

$$V_{tun} = \alpha V_P \quad (62)$$

Clearly for the given tunneling voltages, tighter coupling minimizes the required programming voltages  $V_p$ . For this reason, the bootstrap capacitor should be at least one order of magnitude larger than the sum of  $C_{GS}$ ,  $C_{GB}$ ,  $C_{GD}$  and  $C_{inj}$ . Taking into account the circuit area, proper trade offs between the size of the bootstrap capacitor and programming voltage have to be made. Using typical assumptions, the approximate bootstrap coupling ratio for all the four cells in this case is 10/11.

The bootstrap capacitor  $C_B$  formed between poly-1 and poly-2 does not impose a significant limitation on the highest value of the programming voltage. The diffusion-poly capacitor, on the other hands, would have limited the maximum signal peak to approximately  $\pm 14$  V (for the orbit process) to save the device from avalanche breakdown either between the diffusion and the well or between the well and the substrate. However, for the chosen process, the per unit capacitance formed between the diffusion and the poly is more area efficient than that formed between the two polys.

The sense transistor and the charge on its floating gate represent the synapse in the weight matrix  $W$  and the value of the weight respectively. As the network learns, the strength of the synapse increases. This is the electrical equivalent of dumping more charge on the floating gate, i.e., programming. Programming modulates the electrical conductivity of the synapse (P-MOS) device. Thus during programming, the electrical conductivity of the synapse is expected to increase. The P-sense transistor was specifically chosen to achieve this operation. During programming, the floating gate acquires electrons. Trapped electrons develop a negative potential on the floating gate

of the P-MOS sense transistor. The floating gate voltage tends to become more negative as programming proceeds. Therefore, the drain current through the device increases, i.e., conductivity increases. This would not have been possible with a N-MOS because conductivity of NMOS decreases with the decrease in gate voltage. To avoid the problem associated with the N-MOS as a sense transistor, the synapses would initially have to be driven to the cutoff region by programming. Then by removing electrons in the erasing mode and superimposing fixed bias voltage on the controlling gate, weights would be loaded. Another reason for using the P-MOS sense transistor is to avoid an erroneous change in the gate voltage due to the generation of hot electrons near the floating gate. N-MOS transistors operating at higher values of  $V_{DS}$  are more prone to such effects [62].

#### Field Enhanced Fowler-Nordheim Tunneling

A simplified explanation of the Fowler-Nordheim tunneling is as follows [62]. There exist an energy barrier of approximately 3.2 eV that prevents the escape of electrons from the substrate to the  $\text{SiO}_2$ . At room temperature, the kinetic energy of the electrons allows them to tunnel through an oxide barrier whose thickness is approximately 5 nm. If the favorable electric field (generated due to external potential within this 5 nm range near the oxide silicon interface) is less than 3.2 eV, then the electrons are pulled back into silicon. However, if the external field strength in this region is greater than 3.2 eV, a percentage of the total electrons continue to travel in the direction of the external field and thus a small current flows from the Si surface. Increasing the electric field increases the electron flow and thus the electron current. Keeping these numbers in mind, it takes approximately 25 V to tunnel electron across a thickness of 40 nm. This voltage should



be well below the gate-oxide breakdown voltage, which is about 28 V for MOSIS process.

According to this theory, the electric field within 5 nm of the SiO<sub>2</sub> interface plays an important role in the tunneling process. The electron emitting surface can be structured in order to increase the local electric field at the SiO<sub>2</sub> interface, thus allowing electron currents to be induced at much lower external voltages. Commercial EEPROMs use the same concept by deliberately introducing spikes or other non uniformities, such as surface textures, on the Si-SiO<sub>2</sub> interface. Enhancement of the electric field in such cases is reported to be by factor 4 to 5. In the present case, instead of special processing such as textured polysilicon, the lithographic features have been used to enhance the local field intensity. The field enhancement factor obtained by lithographic features (2 to 4) [62] is less than the field enhancement factor obtained by the textured polysilicon injector (4 to 5) [62].

The theorized area of a gate that is influenced by sufficient field strength is very small (probably only corners). Thus programming and erasing are extremely slow. However, this is not critical for the implementation of the plasticity in the electronic olfactory system.

### Programming

The test setup is shown in Figure 36. The setup is configured to measure the threshold voltage of the sense transistors  $M_{1-4}$  before and after every programming attempt. Programming results in the tunneling of the electrons onto the floating gate, which according to equation 56 produces the negative shift in the threshold voltage of the

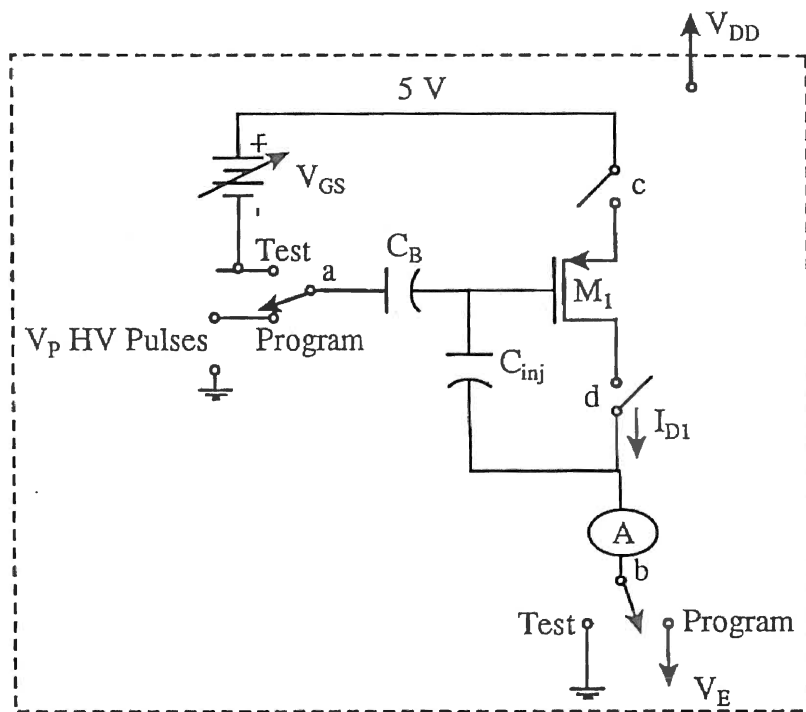


Figure 36. Test Setup for Testing Weighting Cell

sense transistors. The shift in the threshold voltage is used to confirm the presence of tunneling phenomena.

To measure the threshold voltage, switches a and b are switched over to the test mode while c and d are closed. The plot of square root of the drain currents versus  $V_{GS}$  of the un-programmed devices is shown in Figure 37. The threshold voltage for the cells is found to be approximately -0.8 V.

To program the memory cells, switches a and b are switched over to the program mode while c and d were left open. With  $V_E$  set at 0 V, -5 V, and -10 V, programming pulses ( $V_P$ ) of amplitude ranging from 5 V to 16 V with the step of 1 V were applied. The same programming voltages was applied across all of the cells. Thus, any difference in electron current flowing onto the floating gate could be attributed to the differences in the injector structures. The duration of pulses was varied from 2 ms to 40 ms in the steps of 5 ms. The rise and fall times of  $V_P$  were controlled, since it determines the peak capacitive current that flows through the injector. A sufficient rise time [62] of the pulse was used to prevent sharp capacitive current pulses that can result in gate oxide breakdown. After every programming attempt, the threshold voltage of the sense transistors was measured by switching the devices in test mode to observe any shift in the threshold voltage due to the programming. Over numerous such attempts, no significant shift in the threshold voltage was observed. However, a significant shift in the threshold voltage was observed in the last set with  $V_E$  at -10 V and with  $V_P$  pulse amplitude of 16 V. But, in this case,  $V_{DD}$  was left floating instead of connected to the power supply. The resulting shift in the threshold voltages is shown in Figure 38.

Figure 38 together with table 1, does not clearly reveal the possible relationship

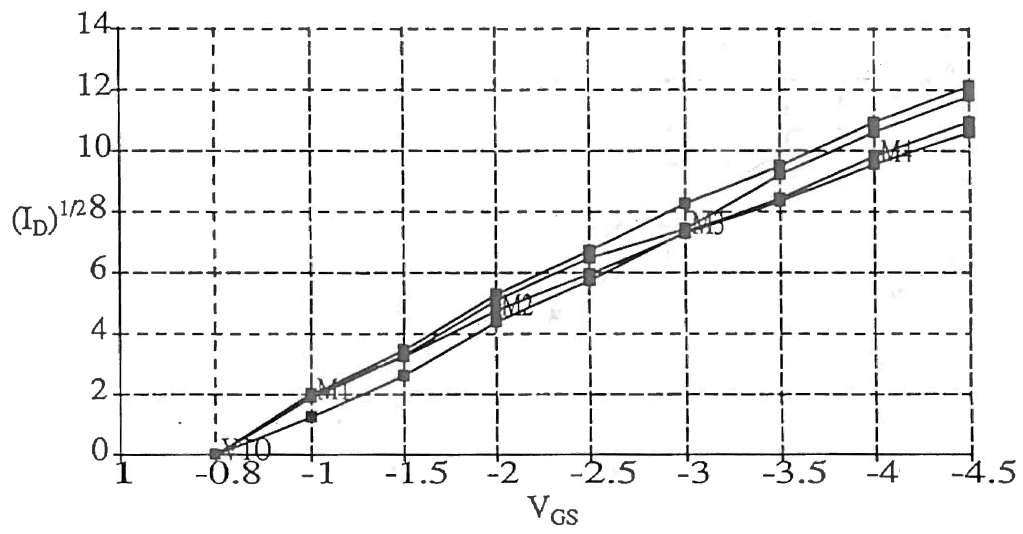


Figure 37. Threshold Voltages of Un-Programmed Devices

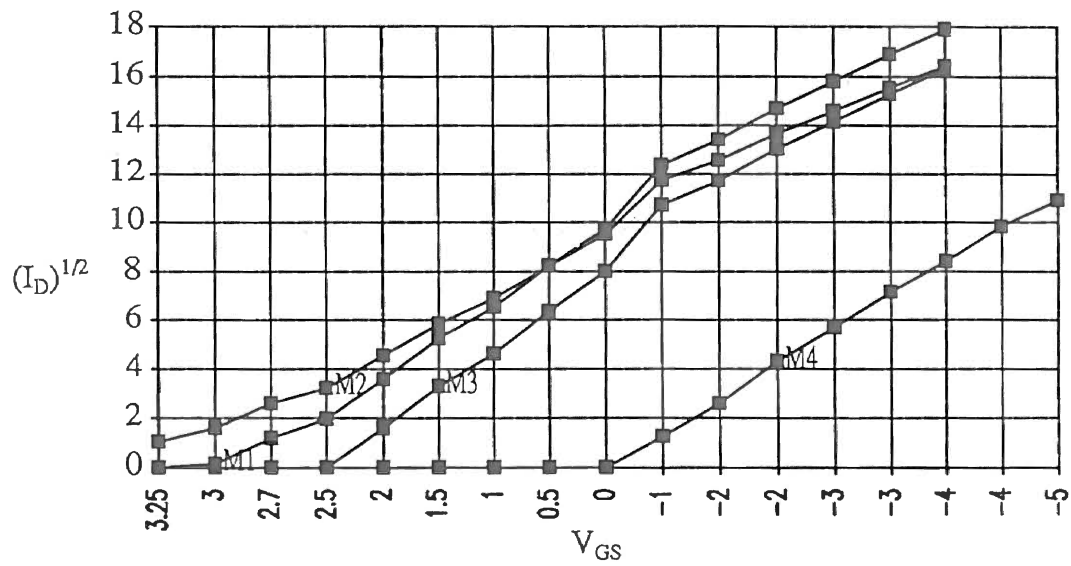


Figure 38. Threshold Voltages of Programmed Devices

between injector structure and programming level. For bootstrap ratio of 10/11, the tunneling voltage of 27 V is comparatively higher than reported by L. R. Carley (18 V to 19 V) [62]. Note that for the same programming voltages, no tunneling was observed when  $V_{DD}$  was used. This raises a question as to whether powering of  $V_{DD}$  (see Figure 35) adds an extra capacitance to the floating gate thereby reducing the effective bootstrap coupling ratio. The decrease in the bootstrap ratio leads to higher programming voltages. The validity of the above statement has not yet been verified.

Retentivity plots taken at room temp (26° C) after 3 and 130 hours are shown in Figure 39, and 40 respectively. The comparison study of Figures 38 and 39 demonstrate excellent short term retentivity. However, the comparison study of Figure 37 and 40 demonstrate that cell does not possess long term retentivity.

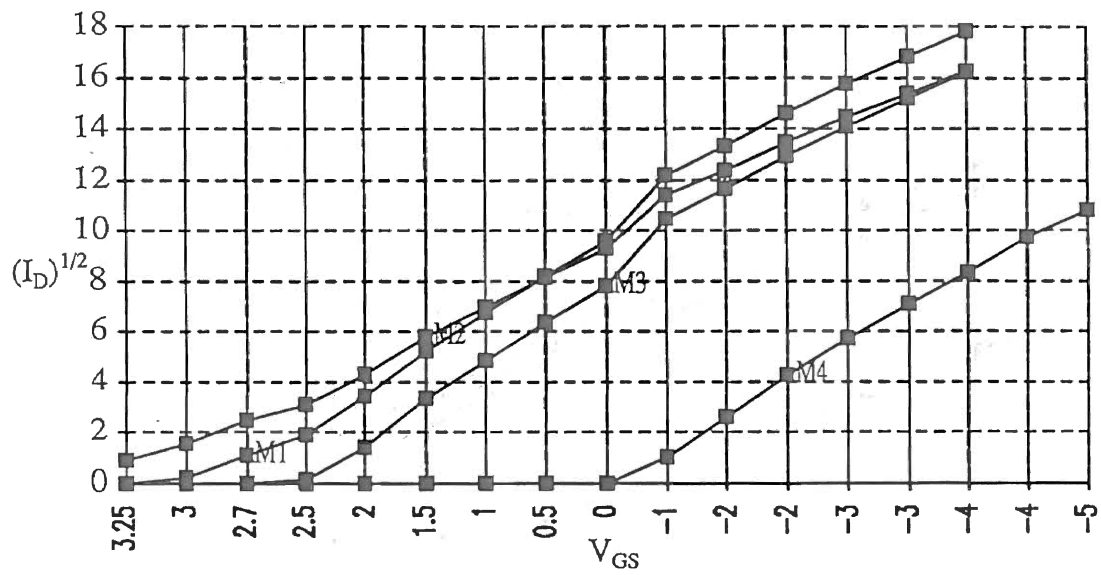


Figure 39. Threshold Voltage Retention After 3 Hours

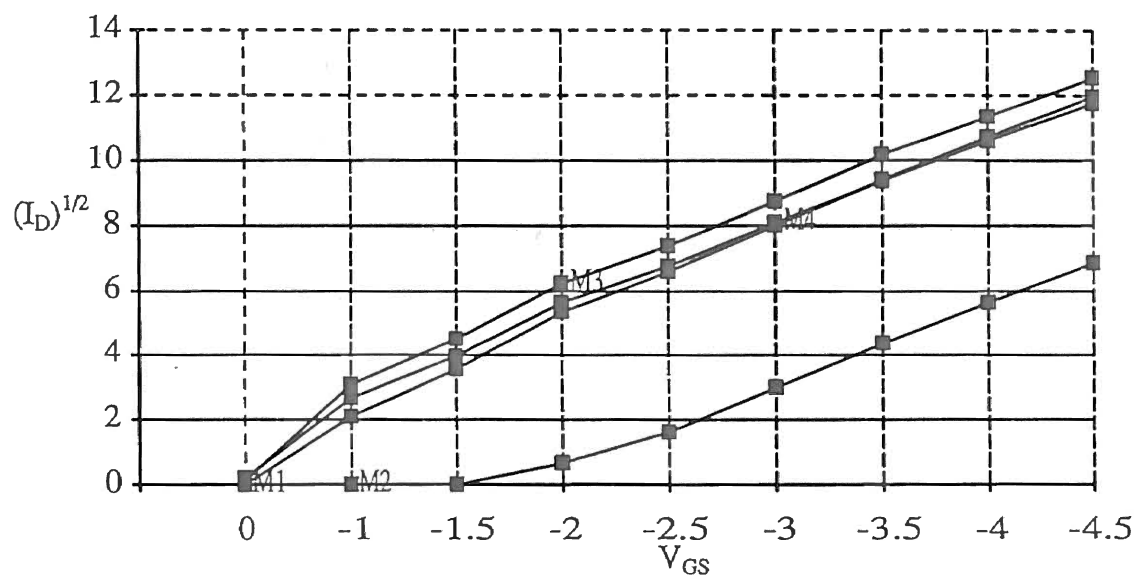


Figure 40. Threshold Voltage Retention After 130 Hours



### Winner Take All

Winner take all (WTA) competition of the piriform cortex is accommodated in the  $p$  identical piriform patches. One ( $k=1$ ) such WTA piriform patch within a PC is shown in Figure 41. The patch consists of  $h$  identical piriform cells connected in parallel. Within a patch, the cells share a common comparison node  $C$ . Node  $C$  serves as a strong local inhibitory feedback, similar to the circuit designed by Lazarro et al. [64]. However, the circuit is designed for improved sensitivity. Each piriform cell receives input current,  $P_{i1}^*$  ( $1 \leq i \leq h$ ), from the piriform BiVI buffer. A single piriform cell is shown by the dotted box. Gate of  $M_1$  is the node where comparison takes place. This node is common with other cells.  $M_7$  is a cascode device provided to minimize the current mirroring error in  $M_1$  which is present due to the channel length modulation.  $M_3$  provides leakage current that is present on the common gate.  $M_5$  provides source for the shortfall in the mirrored current.

The circuit is reset at the beginning of each sniff by pre-charging the common mirroring node  $C$  to  $V_{SS}$  by the switch  $M_9$  which is actually distributed in each of the piriform cell. During the winner take all competition,  $M_9$  is shut off and the circuit is allowed to seek a stable equilibrium. Depending on the time constant at node  $C$ , the common gate voltage starts rising due to the incoming currents and finally settles to the voltage corresponding to the highest value of input current  $P_{i1max}^*$ . Since this voltage is common to all  $h$  piriform cells, the highest input current gets mirrored in the rest of the  $h-1$  cells by the  $M_1$  transistor in the other cells. At this stage, all comparing transistors ( $M_1$ 's) attempt to sink to the maximum input current. Thus in all but the cell with highest

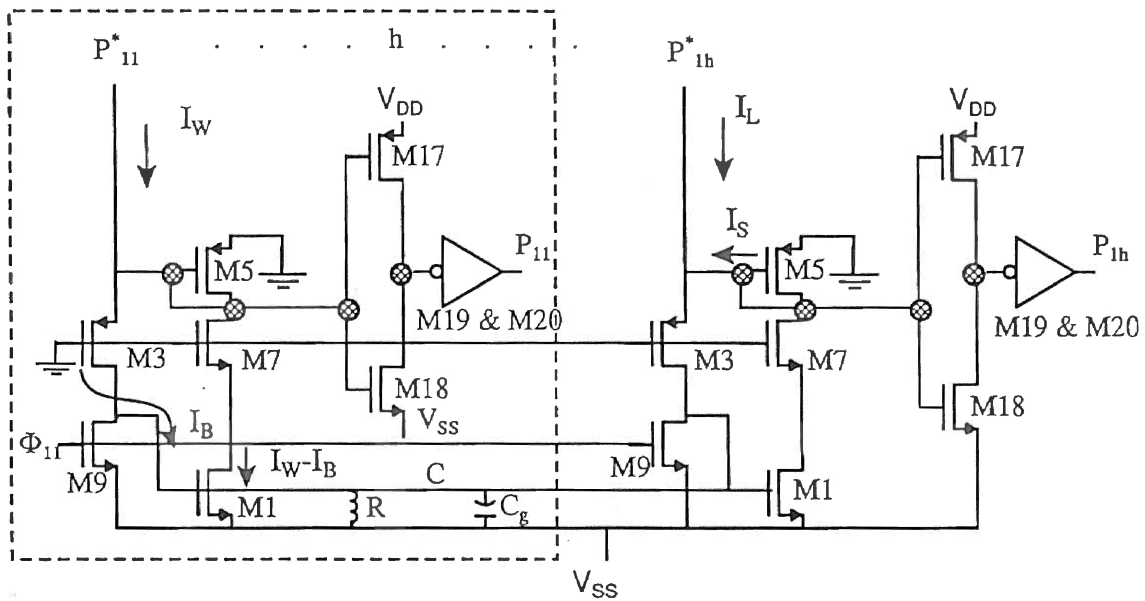


Figure 41. Winner Take All Circuit

input current, sinking currents exceed the input currents  $P_{1i}^*$ . The shortfall, the difference between the maximum current and the corresponding input current in cells ( $P_{1\max}^* - P_{1i}^*$ ), is supplied by the diode connected transistor  $M_5$  connected at the cell input. The differential current results in drop across  $M_5$ . The drop biases  $M_3$  to conduct and  $M_5$  to shut off in the branch associated with  $P_{1\max}^*$ , making the corresponding  $M_1$  a main controlling device while all other  $(h-1)$   $M_1$ 's mirroring devices. At this transition, the voltage at the input drops from a threshold above ground to a threshold below ground at all of the input nodes, except the branch with the highest current since the shortfall in that branch is zero. The resulting change in the diode voltage ( $2 V_T$  approximately) is amplified by the inverter  $M_{17,18}$  and level shifted by inverter  $M_{19,20}$ . Thus, the maximum current results in the logic high at the output of the piriform cell signifying the winner, while all other piriform cells remain low signifying the losers.

In Figure 41, if  $I_w$  is the winner's input current,  $I_L$  is looser's input current,  $I_{Dw}$  is the drain current via  $M_1$  of winner,  $I_{DL}$  is the drain current via  $M_1$  of loser,  $I_s$  is drain current via  $M_5$ , and  $I_b$  is the leakage current on node C, then

$$I_w = I_B + I_{Dw} \quad (63)$$

and

$$I_L = I_{DL} - I_s \quad (64)$$

The common node C attends a gate voltage of

$$V_{GS} = \sqrt{\frac{2(I_w - I_B)}{\beta}} + V_T \quad (65)$$

Theoretically, current mirroring should result in  $I_{DL}$  is equal to  $I_{Dw}$ . However, due

to the beta and threshold mismatch, and channel length modulation associated with the  $M_1$ 's,  $I_{DL}$  is equal to  $I_{DW} \pm \Delta I$ , where  $\Delta I$  is given by:

$$\begin{aligned} \Delta I &= \Delta \beta_1 (V_{GSI} - V_{TI}) \pm \beta_1 \Delta V_{TI} \pm \lambda \Delta V_{DSI} \beta_1 \\ &\approx \Delta \beta_1 (V_{GSI} - V_{TI}) \pm \beta_1 \Delta V_{TI} \end{aligned} \quad (66)$$

Subtracting equation 64 from equation 63 results in

$$I_S = I_W - I_L - I_B \pm \Delta I \quad (67)$$

The  $I_S$  is responsible for exhibition of WTA competition. To be able to resolve the winner and the loser,  $I_S$  should be greater than the resolution capacity of the WTA circuit.

The circuit has limited resolution, due to the mirroring error associated with  $M_1$ .

### Simulations

The SPICE simulations are shown in Figures 42 and 43. Figure 42 demonstrates the ability to resolve the winner between inputs, which differ in amplitude by  $1 \mu A$  at low levels of input currents while Figure 43 demonstrates its inability to resolve the same differential at high levels of input currents. The settling time is a complex function of both the magnitude of all the currents and the differential between the winning and losing currents. In general, the settling time  $\tau_s$  is the inverse function of the differential. The worst case time is derived from a pair of closely matched low amplitude input currents. With all identical losers, it is found to be typically  $1 \mu s$ .

In Figure 43, the loser becomes high, even if the inputs have a differential of  $1 \mu A$ , whereby the circuit fails to resolve a winner. However, since the settling time is

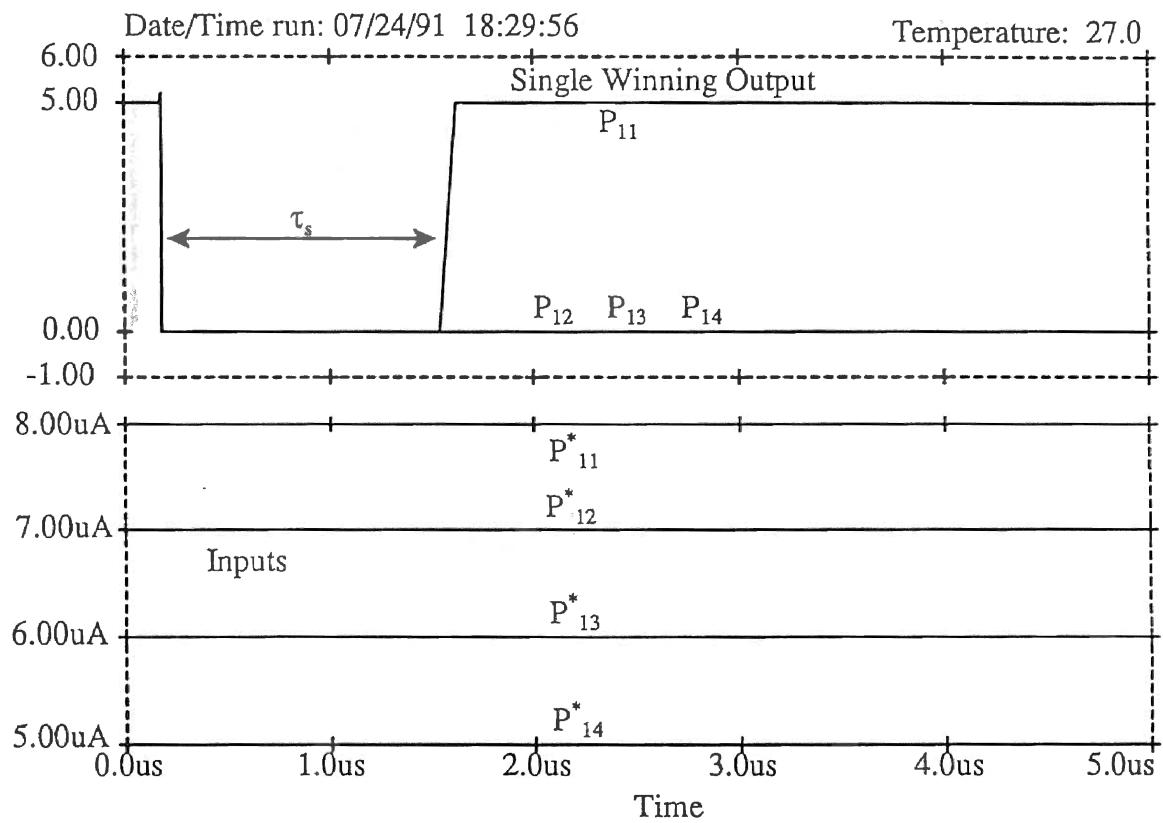


Figure 42. The Demonstration of Resolving WTA Inputs

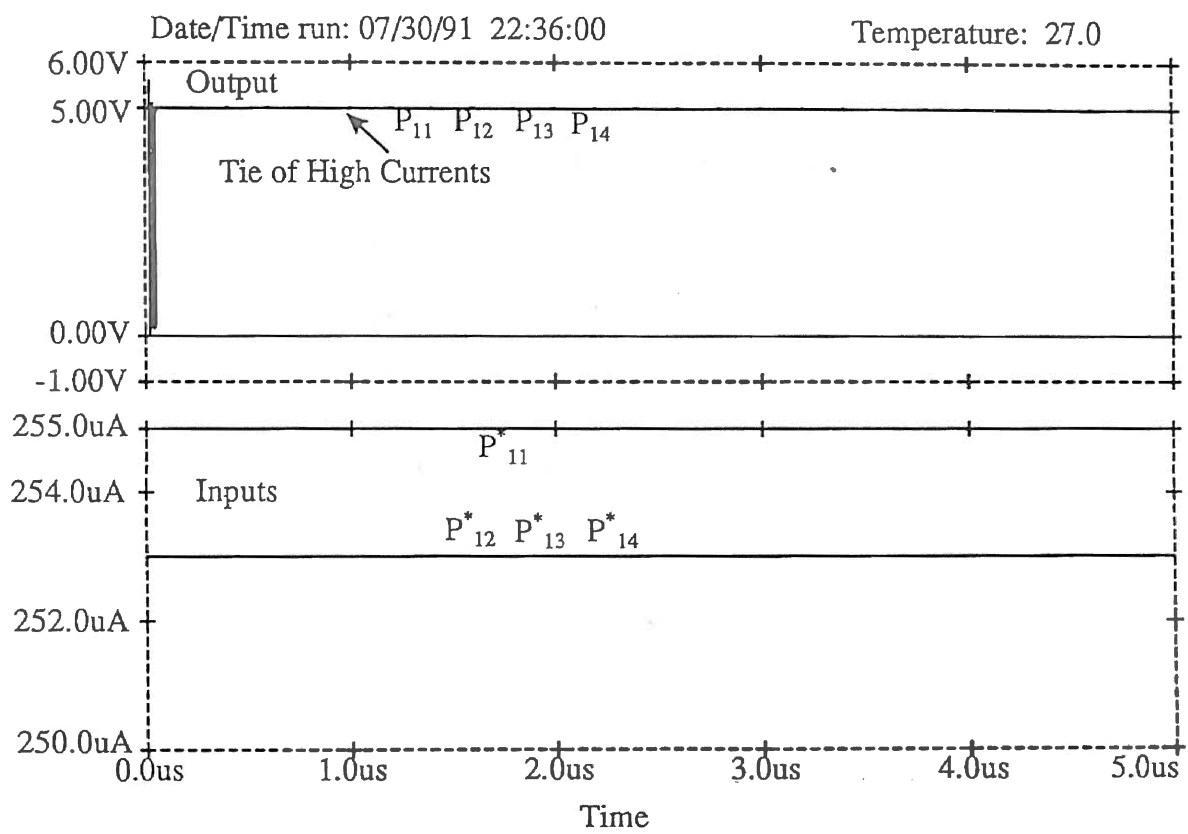


Figure 43. Ties in the WTA Outputs

the inverse function of the differential, the settling time of the true winner is always less than the other winners. This fact may be used in the future applications to an advantage in separating the true winner from many winners.

Simulations have been performed on as many as 250 WTA cells operating in parallel within a single piriform patch. It is observed that number of active WTA cells has a limited influence on the timing performance of the circuit.

### Testing

Due to the pin limitation, only four WTA cells were fabricated. The cell inputs  $P^*_{11}$ ,  $P^*_{12}$ ,  $P^*_{13}$ , and  $P^*_{14}$  are supplied through the high resolution current sinks. While testing,  $P^*_{12}$ ,  $P^*_{13}$ , and  $P^*_{14}$  are grouped together and supplied by a common current sink. The function generator is used to reset the circuit, thus when  $\phi_{12}$  is pulled to a logic low, the circuit is allowed to seek the stable equilibrium.

Experiments are carried out keeping in mind the effect of the mean value of the input currents, and the differential current between the winner and the loser, on the settling time of the circuit. Figure 44 shows the settling time as a function of the input current level with the difference current ( $5 \mu\text{A}$ ) as a constant parameter. For the same differential, any increase in the mean level of the input current beyond the shown current range results in failure to resolve the inputs. It can be seen that with an increase in the current level, the settling time of the winner  $\tau_w$  increases while the settling time of loser  $\tau_L$  decreases. For the present circuit geometries, the current level for the minimum settling time is found to be approximately  $40 \mu\text{A}$ .

Figure 45 shows the effect of the current difference when current level is set as a

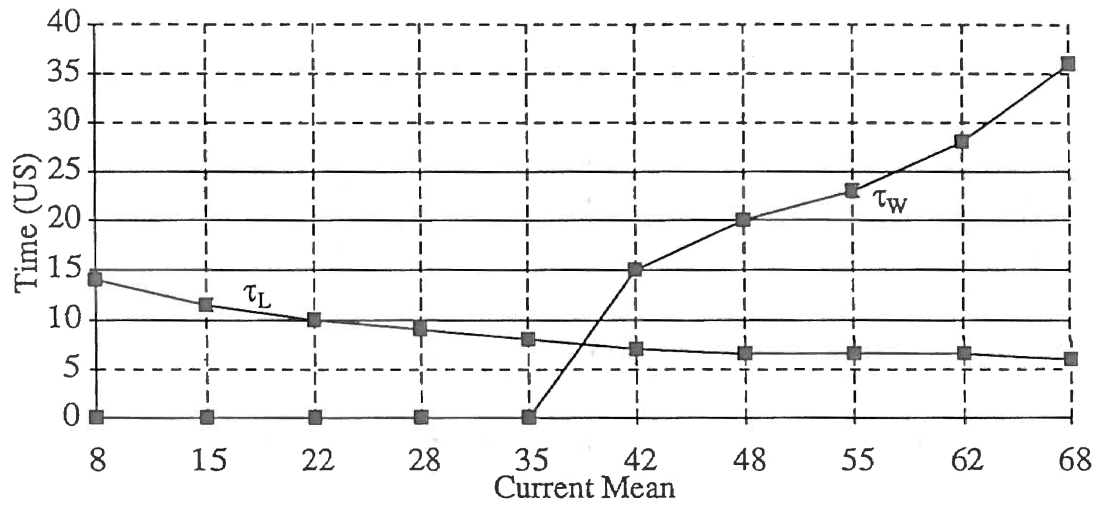


Figure 44. Effect of Mean on the Settling Time



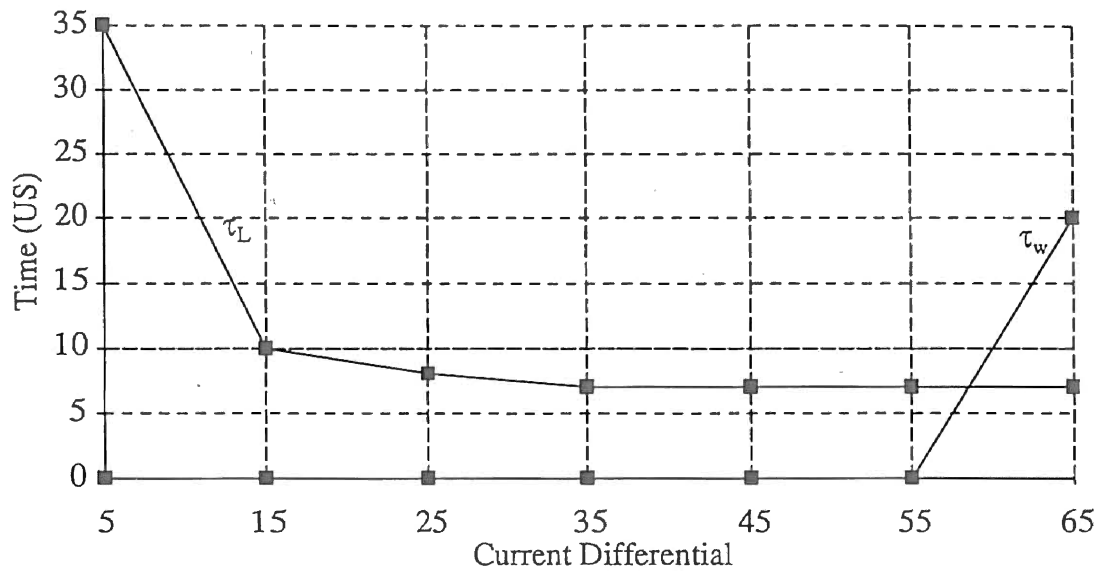


Figure 45. Effect of Difference Current on the Settling Time

constant parameter. As expected, the graph reveals the discrepancy between winner and losing currents.

Testing results taken at a low level of the input currents (near  $5 \mu\text{A}$ ) show that the circuit is capable of resolving difference currents as small as  $2 \mu\text{A}$ . As the current level goes higher, the resolution decreases. Testing results taken at current mean equal to  $70 \mu\text{A}$  shows that the circuit is capable of resolving difference current of about  $5 \mu\text{A}$ .

### Tie Resolver

The short coming, i.e., the finite resolutions of the WTA circuit was discussed in the pervious section. To ensure only one winner in a single piriform patch, a resolver circuit is required to post-process the WTA circuit output.

The tie resolver element is shown in the Figure 46. This element digitally resolves the ties among the winners. In the circuit, inputs and outputs are defined as follows:  $L$  is the learn,  $TI_i$  is the control input,  $TO_i$  is the control output,  $P_{kl}$  is the unresolved input, and  $PW_{kl}$  is the resolved output. The 1 bit resolver is formed by connecting 1 resolver cells in a chained fashion, where  $TI_i$  is propagated across the entire input vector  $P_{kl}$  from left to right. The control output of the preceding resolver element forms the control input to the next element. That is,  $TO_{(i)}=TI_{(i+1)}$ , except with  $TI_1$ .

The truth table II for the resolving logic function states that, with learn high, the high  $TI_i$  is propagated from left to right until it encounters the first winner, making  $PW_{kl}$  of the corresponding element high and negating its control output  $TO_i$ . For  $PW_{kl}$  to be high, both  $TI_i$  and  $P_{kl}$  must be high. This ensures that if there is more than one winner,

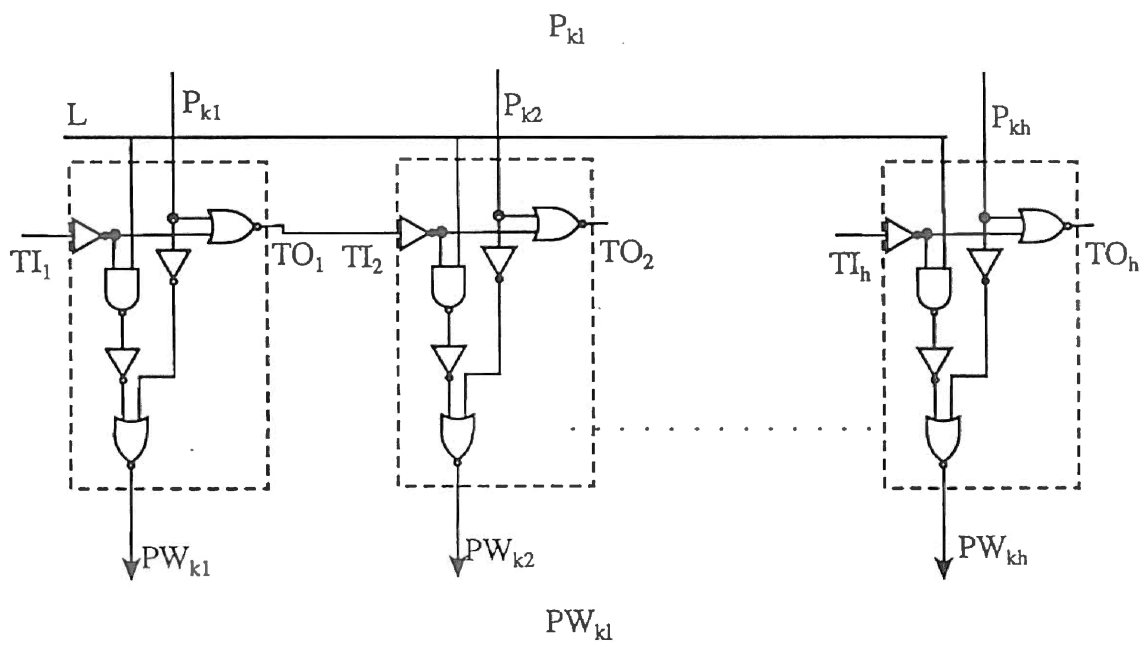


Figure 46. Tie Resolver

TABLE II  
TRUTH TABLE

L	$TI_1$	$P_{kl}$	$PW_{kl}$	$TO_1$
1	1	0	0	1
1	1	1	1	0
1	0	X	0	0
0	X	0	0	X
0	X	1	1	X

only the one with the lowest settling time corresponding to the highest input, will be transferred to the output. Since  $TI_i$  is propagated from left to right, the left most winner is selected and declared as the final winner.

The Karnaugh map of the tie resolver logic results in

$$\begin{aligned} TO_i &= \overline{(\overline{TI_i} + P_k)} \\ PW_k &= \overline{P_k + (L \cdot TI_i)} \end{aligned} \quad (68)$$

The hardware implementation of the above Boolean equations is shown in Figure 46. The standard CMOS gates have been used from a standard library to fabricate the resolver circuit. The layout is done by using the VLSI tool LAGER. Logical simulations are carried out by the built in circuit simulator IRSIM.

### Testing

The testing results agreed with the Boolean equation 68. However, with the 5 V supply voltage, the high logic level on the  $TO_i$  is found to be only 1.6 V. We attribute this fault to the possible defect in the mask since the standard LAGER cells were used to build the circuit. The rise and fall times are found to be 0.56  $\mu$ s and 0.5  $\mu$ s respectively.

### Dynamic Current Copier Integrator

The current copier integrator (CCI) provides collateral feedback inhibition from the active piriform patches to glomeruli. Winning piriform neurons are applied to the  $W^T$  matrix generating feedback currents. Feedback currents are sampled, stored, and integrated in the CCI. During the backward phase and at the end of each minor cycle,

inhibition is applied to the glomerulus. This inhibition persists to be used during the forward phase of the next cycle. During successive cycles, all of the inhibition currents that are generated in the backward phase are sampled and summed with previously stored inhibition. In this way, according to GLA olfactory model, as the multi-sampling proceeds, the cumulative inhibition up to that cycle is applied to the glomerulus to inhibit the stronger components in the input vector. This allows weaker components to become comparatively significant thus taking an active part in the overall clustering process. The CCI is a dynamic, yet discrete analog memory element to compute and store the accumulation of the sampled feedback analog currents. The circuit is based on dynamic current copier principle. The following text describes the electronic implementation of the CCI.

### Background

The standard current mirror is the most widely used block in analog integrated circuits. The current mirror concept was originally applied in bipolar technology. It is now extensively used in the CMOS process to duplicate, multiply or divide the currents. Current error due to the threshold mismatch and  $1/f$  flicker noise is the most significant limitation of the standard MOS current mirrors, when used in a high precision analog circuits. In spite of the various circuit design techniques reported, these errors typically could not be reduced below 1% [65]. The dynamic current copier, also referred to by many other names such as a current copier, current self calibrating circuit, and dynamic current mirror, etc., is a recent innovation. They completely overcome the limitations of the standard current mirrors, and moves achievable precision to tighter limits [65]. The

circuit is essentially a sample and hold cell that supplies current by storing a voltage at the gate of a MOS transistor through which current flows. Current copiers can replace the standard current mirrors to achieve multiple copies of a reference current with an accuracy of several PPM as compare to the typical one percent accuracy in standard current mirrors. This advantage led to the invention of the dynamic current copying techniques.

Since the gate of the MOS device has practically infinite input impedance, it can be used to store the information on the gate capacitor for a short time period, i.e., for a few ms. Figure 47 shows the basic N-copier cell. To copy the current  $I_0$  into the cell, switches  $S_1$  and  $S_2$  are closed (sample phase). The capacitor is charged to the gate voltage required by the transistor to achieve the drain current  $I_0$ . If  $M_1$  is in saturation, the gate voltage is given by:

$$V_{GS} = \sqrt{\frac{2I_0}{\beta_1}} + V_{TI} \quad (69)$$

The capacitor  $C_1$  will be charged to a voltage  $V_{GS}$ . The switches may then be opened ( $S_1$  must be opened before  $S_2$  to avoid the discharge of  $C_1$  via  $M_1$ ). Ideally, the cell is capable of sinking  $I_0$  when connected to the a load via  $S_3$  (hold phase). Several cells can sequentially be loaded from the same source. Note that a P-copier cell can be obtained by replacing the N-MOS transistor with its equivalent P-MOS transistor, and by reversing the supply polarities and the direction of currents. In such a case, the cell sources  $I_0$  when connected to the load. The cells need not be accurately matched with respect to the transistor dimensions or the capacitor values since the current copying operation in each case results in the appropriate transistor gate voltage being stored on the gate capacitor

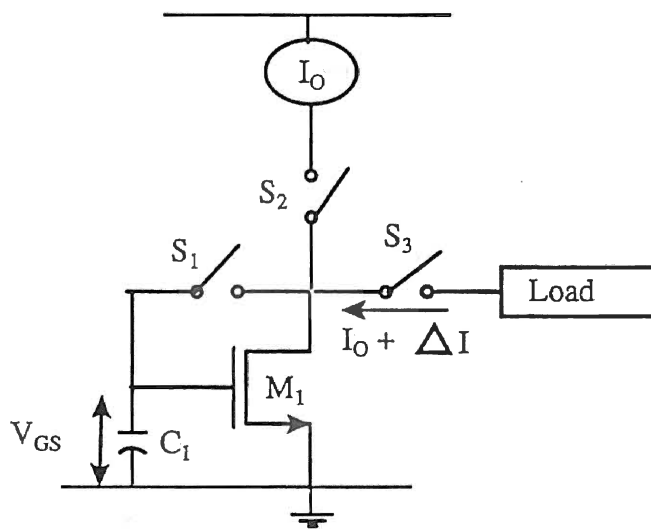


Figure 47. Basic Current Copier



of the selected transistor. Since the same transistor is used for sampling and holding, beta and threshold mismatches are completely eliminated. However, inevitable circuit flaws result in an error current causing the  $I_O$  retrieved from a cell in hold phase different from  $I_O$  of sampled phase. This error current is denoted by  $\Delta I$ . The mechanisms of the original-to-copy error include: (1) switch charge feedthrough, limiting the initial accuracy of the current sample, (2) channel length modulation, producing a change in the retrieved current as the voltage  $V_{DS}$  changes (as with standard current sources), (3) junction leakage associated with  $S_1$ , causing a steady discharge of the storage capacitor, (4) channel charge injection associated with switch  $S_1$ , causing a change in  $V_{GS}$  when  $S_1$  is opened, and other flaws in the circuit.

### The Operating Principle of the Current Copier Integrator

In Figure 47, integer multiplication of  $I_O$  by variable  $n$  can be achieved by making  $n$  copies of  $I_O$ . These copies can be added together through a common load. This would require  $n$  identical current copier cells, whereby after adding them together would give a load current of  $n \times I_O$ . However, serial discrete integration of  $I_O$  ( $\Sigma$ ) can be obtained by using a pair of complementary (N & P) current copying cells connected in a circular fashion where during any instance, one of them acts as a temporary memory. Figure 48 shows such a current copier integrator. The N-cell acts as a temporary memory while the P-cell acts as the sampler and summer. The circuit operates in two phases requiring two non-overlapping switching clocks of the same frequency,  $\phi_{21}$  and  $\phi_{22}$ . During phase 1,  $S_1$  is closed and  $S_2$  is opened while during phase 2,  $S_2$  is closed and  $S_1$  is opened. During phase 1,  $S_1$  is closed on phase  $\phi_{21}$  and the steady state input current  $I_i^*$  is sampled into

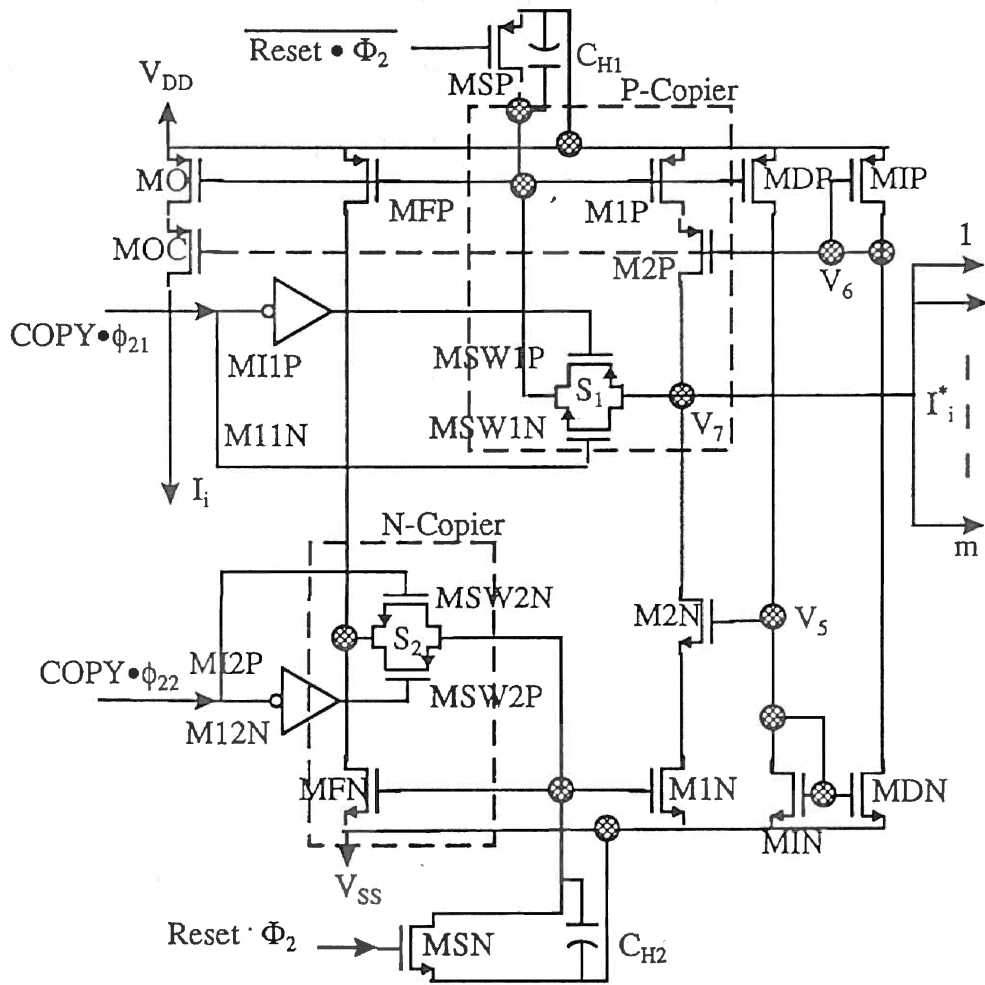


Figure 48. Current Copier Integrator

the P-cell. Capacitor  $C_{H1}$  is charged to the gate voltage  $V_{CH1}$  corresponding to the drain current  $I_i^*$  that is flowing through  $M_{1P}$ . During phase 2,  $V_{CH1}$  is transferred and memorized on  $C_{H2}$  by closing switch  $S_2$  on  $\phi_{22}$ . This completes one cycle. Note that  $S_1$  should be opened before closer of  $S_2$ , and vice a versa to avoid the improper operation of the circuit. At this stage, transistor  $M_{1N}$  is capable of sinking exactly  $I_i^*$ . During the next cycle when  $S_1$  is closed again,  $C_{H1}$  is charged to the gate voltage corresponding to the drain current  $2I_i^*$  ( $I_i^*$  from input plus  $I_i^*$  from  $M_{1N}$ ) that is flowing through  $M_{1P}$ . For a steady state input current, over  $n$  cycles, a total of  $n \times I_i^*$  current flows through  $M_{1P}$ , which when mirrored by  $M_O$  is available as an output current  $I_i$ . However, if the input is a time varying analog signal  $I_i^*(t)$ , then the output over  $n$  cycles is given by:

$$I_i(n) = \sum^n I_i^*(n) ; \quad n=0,1,2,\dots,J \quad (70)$$

where

$$I_i^*(n) = I_i^*(nT) \quad (71)$$

The parameter  $T$  is the time period of the switching frequency  $\phi_2$ .  $I_i^*(t)$  is assumed constant during sampling. From the above equation, the output is clearly a discrete integration of the time varying input current. The initialization of the integrator is essential in order to restart the inhibition for different sets of inputs. The minimum dimension switches,  $M_{SP}$  and  $M_{SN}$  are used to reset the gate voltage or hold capacitors. The circuit is initialized by resetting  $V_{CH1}$  and  $V_{CH2}$  to zero on RESET  $\phi_2$ . To reduce the error due to the channel length modulation, cascode devices  $M_{2P}$  and  $M_{2N}$  are added. Dynamic biasing of these cascodes gives improved cascoding. This is achieved by using

additional dynamic biasing circuitry consisting of  $M_{DP}$ ,  $M_{IP}$ ,  $M_{IN}$  and  $M_{DN}$ . Cascode  $M_{OC}$  serves the same purpose of reducing error due to channel length modulation. Switches  $S_1$  and  $S_2$  are made of transmission gates to cancel the effects of the feed through and channel charge injection. The precision MOS capacitors  $C_{H1}$  and  $C_{H2}$  are realized between poly-1 and poly-2.

### Circuit Design

This section addresses the CCI design. It addresses the limitations imposed on maximum integration level, the maximum switching frequency, and the minimum switching frequency.

#### Upper Integration Limit

Equation 70 to be accurate within 5%, the transistors  $M_{FN}$ ,  $M_{IN}$ ,  $M_{IP}$ , and  $M_{FP}$  must stay in saturation over the entire dynamic range. As the integration progresses, for the unidirectional input current, the current (integrand) in the circuit rises. For the selected geometries, let  $I_{max}$  be the maximum attainable current that can be delivered without any of the transistors slipping out of the saturation region. Thus  $I_{max}$  determines the upper limit on integration. For any current above  $I_{max}$ , the circuit loses its accuracy as either one or all of the transistors fall into the triode region. This forms the design criteria for  $I_{max}$ . Assuming  $\beta_{M1P} = \beta_{M2P} = \beta_{M2N} = \beta_{M1N} = \beta$  and assuming all corresponding transistors in the P-copier leg operating in saturation, the maximum current that can be pushed through P-copier leg is given by:

$$I_{\max} = \frac{(V_{DD} - V_{SS})^2 \beta}{32} \quad (72)$$

The sampled current in the sampler has to be exactly transferred to the hold cell. This requires  $\beta_{\text{MFP}}$  equals to  $\beta_{\text{MIP}}$ , and  $\beta_{\text{MFN}}$  equals to  $\beta_{\text{MIN}}$ . Finally, to mirror the integrand in the circuit to the output with the unity ratio requires  $\beta_{\text{MO}}$  equals to  $\beta_{\text{MIP}}$ .

The bias voltages  $V_5$  and  $V_6$  of the cascode transistors should be maintained as low as possible to maximize the full-scale current range. The dynamic cascoding is essential for optimized cascoding effects at all integrated levels of the input current. To achieve this, the bias voltages  $V_5$  and  $V_6$  must be a function of the present level of the current in the circuit.  $M_{\text{DP}}$  is a 1:1 biasing current mirror that copies the present current level and feeds it into the biasing circuitry. Considering the worst case that occurs when the current level is  $I_{\max}$ , the biasing voltages needed for proper functioning of the circuit are given by:

$$V_5 = 2 \sqrt{\frac{2 I_{\max}}{\beta}} + V_{\text{TM2N}} + V_{\text{SS}} \quad (73)$$

Similarly,

$$V_6 = -2 \sqrt{\frac{2 I_{\max}}{\beta}} + V_{\text{TM2P}} + V_{\text{DD}} \quad (74)$$

$M_{\text{IN}}$  is an active resistor used to dynamically bias  $M_{2\text{N}}$ . The bias voltage  $V_5$  at the current level  $I_{\max}$  requires the geometry of  $M_{\text{IN}}$  to be:

$$\beta_{\text{MIN}} = \frac{2 I_{\max}}{(V_5 - V_T)^2} \quad (75)$$

Similarly, if  $\beta_{\text{MIN}}$  equals to  $\beta_{\text{MDN}}$ , then the geometry of  $M_{\text{IP}}$  is given by:

$$\beta_{MIP} = \frac{2I_{\max}}{(V_6 - V_T)^2} \quad (76)$$

### Maximum Switching Frequency

In order to calculate bandwidth of the CCI circuit, it is essential to know that how fast circuit can be run without adding excessive error in the integration. The transient response of the CCI is limited by the settling time of RC network formed by the switching elements, sample and hold capacitors.

As the integration progresses, the accumulative sum of the sampled current (integrand) in the circuit continuously changes its value. During both phases, it is essential to update the value of the last stored voltages  $V_{CH1}$  and  $V_{CH2}$  to a new voltage corresponding to the latest sum. This stores the integrand up to that point in the copier cell and allows the variation to be followed by the output current  $I_o$ . If switch  $S_1$  during phase 1 remains closed for duration  $t_1$ , then correct updating is only possible if the time duration  $t_1$  is longer than the settling time of the sample and the hold formed by  $M_{1P}$ ,  $S_1$  and  $C_{HI}$ . Assuming small perturbations, the settling behavior of this circuit can be examined by means of the P-cell of Figure 49. Opening the loop between capacitor  $C_{HI}$  and the gate, the open loop transfer function is [65] given by:

$$G(s) = \frac{V(s)}{V_G(s)} = -\frac{1}{s\tau_1(1+s\tau_2)} \quad (77)$$

where,

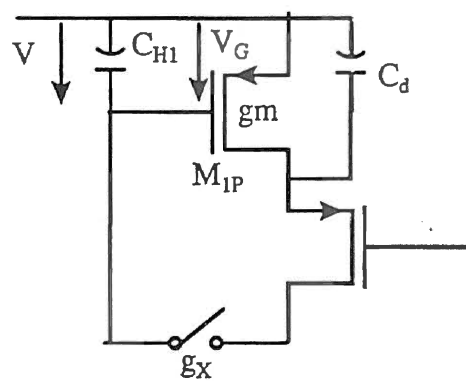


Figure 49. The P-Cell

$$\tau_1 = \frac{C+C_d}{g_m} \quad (78)$$

$$\tau_2 = \frac{C+C_d}{g_x(C+C_d)}$$

The  $g_m$  and  $g_x$  are the transconductance of  $M_1$  and switch respectively. The two poles of the closed loop circuit are the roots of equation 1-  $G(s) = 0$

$$s_{1,2} = -\frac{1}{2\tau_2} \pm \sqrt{\frac{1}{4\tau_2^2} - \frac{1}{\tau_1\tau_2}} \quad (79)$$

For  $4\tau_2 > \tau_1$ , the response is a damped oscillations with an envelope time constant of  $2\tau_2$ .

For  $\tau_1 \gg 4\tau_2$ , it settles exponentially with the time constant  $\tau_1$ . The global settling time constant may be reasonably approximated by:

$$\tau_s = \tau_1 + 2\tau_2 \quad (80)$$

$\tau_s$  must be 5 to 7 times (depending on the desired accuracy) smaller than  $t_1$  to ensure that equilibrium is reached. Applying similar treatment during phase 2 for the N-cell, results in  $t_2$ . Therefore, the maximum switching frequency is:

$$\phi_{2_{\max}} \propto \frac{1}{t_1 + t_2} \quad (81)$$

These conditions place an upper limit on the operational frequency, and upper limit on values of  $C_{H1}$  and  $C_{H2}$ . They also place a lower limit on  $g_x$  and  $g_m$ .

### Minimum Switching Frequency

During normal operation,  $S_1$  is opened followed by the closer of  $S_2$  and vice a versa. The time intervals,  $t_{12}$  and  $t_{21}$  between these two instances determine the minimum



possible switching frequency. During these intervals, the circuit is idle since neither of the capacitors are connected to their respective drains through switches since both of the switches are off. Thus, the gate voltages  $V_{CH1}$  and  $V_{CH2}$  float to their pre-charged voltages. The voltages stored on the MOS capacitors at the gates of  $M_{FN}$  and  $M_{1P}$  are affected by the leakage currents that is flowing from the gate. The peak to peak variation caused by the leakage current is given by:

$$\Delta V_{PP} = I_{leak} \frac{t_{12}}{C_{HI}} \quad (82)$$

Variation in the gate voltage produces a variation in the drain current

$$\Delta I = \Delta V_{PP} gm \quad (83)$$

The leakage current is present due to the reverse biased diode current associated with transmission gates. Longer  $t_{12}$  and  $t_{21}$  result in larger drain current errors. These relations impose the upper limit on the  $t_{12}$  and  $t_{21}$  for a given tolerance in drain current error.

These times may be referred to as circuit idle time. The idle time is given by:

$$t_{12} = \frac{\Delta I C_{HI}}{gm I_{leak}} \quad (84)$$

Note that  $t_{21}$  equals to  $t_{12}$ . They set the minimum allowable operational frequency at:

$$\phi_{2min} = \frac{1}{t_1 + t_2 + t_{12} + t_{21}} \quad (85)$$

Switching frequencies below  $\phi_{2min}$  introduce unacceptable errors in the output current. Increasing  $C_{HI}$  and  $C_{H2}$  results in a lower operational frequency but increases the settling time. Hence, a proper trade off has to be made.

## Mechanisms of Errors

This section addresses the errors that are present in the output current of CCI circuit. These errors are result of charge injection, switch feedthrough, channel length modulation, and leakage current. The error due to the leakage current was discussed previously.

### Charge Injection

A significant limitation to the precision of the current copiers is due to the realization of the various switches by means of transistors. To close the switch, the switching transistor is made conductive by mobile carriers that are attracted into the channel by the gate voltage. For charge equilibrium, the total charge of the mobile carriers in the channel must be equal to the total charge stored on the gate. The charge stored on the gate in strong inversion is given by:

$$q = (W L COX')_{S1} (V_{GS} - V_T)_{S1} \quad (86)$$

When the switch is opened, these carriers are released from the channel in order to block the transistor. The channel charge flows into the source and drain. Thus in the N-copier when switch  $S_1$  opens, a fraction  $\delta q$  of  $q$  is dumped on the capacitor  $C_{H2}$ . The factor  $\delta$  determines the amount of charge that is dumped on the source of the MOS transistor. In some literature, it is specified to be 0.5 [66]. This causes gate voltage error given by

$$\begin{aligned} \Delta V &= \frac{\delta q}{C_{H2}} \\ &= \delta \frac{(W L COX')_{S1} (V_{GS} - V_T)_{S1}}{C_{H2}} \end{aligned} \quad (87)$$

in the stored voltage  $V_{CH2}$ . This voltage error in turn creates a relative error in the output

current of the copier as

$$\begin{aligned} \frac{\Delta I_{DMFN}}{I_{DMFN}} &= \frac{g_{m_{MFN}} \Delta V}{I_{DMFN}} \\ &= \frac{(W L COX')_{SI} (V_{GS} - V_T)_{SI}}{C_{H2} (V_{GS} - V_T)_{MFN}} \end{aligned} \quad (88)$$

$\Delta V$  can be decreased by making gate oxide capacitance of the switch a small percentage of the  $C_{H2}$  where one limit is given by the area of the  $C_{H2}$ . It can also be decreased by reducing the total charge  $q$  in the channel which intern reduces the fraction  $\Delta q$  that flows onto  $C_{H2}$ . This can be achieved by minimizing the gate area  $W \times L$  and/or by controlling the gate voltages of the switch. The percentage error also tends to be low at higher values of  $V_{CH2}$ . A similar treatment applies to the P-copier cell for determination of the error due to the charge injection.

### Switch Feedthrough

Switch feedthrough contribution is due to the clock voltage that is coupled to the gate via  $C_{GS}$ . The clock voltages is partially transferred to the gate via the capacitive network. The transferred voltage is given by,

$$\Delta V = -\frac{C_{GS}}{C_{GS} + C_{HI}} (V_g + V_T - V_{SS}) \quad (89)$$

where  $V_g$  is the gate voltage of switch transistor and  $C_{GS}$  is gate to drain capacitance of the transmission gate. The change in the gate voltage multiplied by the transconductance reflects an error in the drain current.

### Cascode Configurations

Consider the structure illustrated in Figure 48 without the cascode devices  $M_{2P}$  and  $M_{2N}$ . For any cycle, during phase 1 and 2, currents are sampled on  $C_{H1}$  and  $C_{H2}$ , respectively. While during the remainder of the cycle time, the copier hold these sampled currents on their gates. Considering the P-cell, let  $V_{7S}$  and  $V_{7H}$  be the voltages attended by node 7 during the sample phase and the hold phase, respectively. The  $V_7$  must return to the value  $V_{7S}$  equals to  $V_{CH1}$  during sample phase 1. During the hold phase,  $S_1$  is open and  $V_7$  jumps to the voltage  $V_{7H}$ , imposed by the relative impedances of  $M_{1N}$ ,  $M_{1P}$ , and the input current sink. Since  $V_{7S}$  is not equal to  $V_{7H}$ , the difference in drain voltages during the two phases produces additional contributions to the inaccuracy of the integration.

The first contribution is due to the channel length modulation producing change in the drain current as the drain to source voltage changes. Mathematically, this can be represented in terms of the effective output conductance  $g_o$ , where  $g_o$  is the combined transconductance of cascoded M1P and M2P. Thus, the relative error in the output current of the copier can be written as

$$\frac{\Delta I_D}{I_D} = \frac{g_o(V_{7H} - V_{7S})}{I_D} \quad (90)$$

The second contribution is due to the drain voltage transferred to the gate via  $C_{GD}$ . The difference in the two voltages is partially transferred to the gate via the capacitive network as

$$\Delta V = \frac{C_{GD}}{C_{HI} + C_{GD}} (V_{D_{sample}} - V_{D_{hold}}) \quad (91)$$

where  $V_{D_{sample}}$  and  $V_{D_{hold}}$  are the drain voltages attended by  $M_{1P}$  in sample and hold phases respectively. The change in the gate voltage multiplied by the transconductance reflects an error in the drain current.

### Simulations

The transient simulation of the CCI is shown in Figure 50. With a 30  $\mu A$  steady state input current applied, integration over 5 cycles is observed.  $\phi_{21}$  and  $\phi_{22}$  are the switching clocks. The output current is sampled via  $R_O$ .

Initially, the circuit is reset to the initial conditions. The first output sample is found to be approximately 35  $\mu A$ . A successive increase in step size of the output current is attributed to the cumulative integration of an error term that is present due to previously described factors specifically channel length modulation effects and channel charge injection. Over 5 cycles, the error is found to be 33%. For the designed geometries, the circuit saturates above 300  $\mu A$ .

This simulation demonstrates maximum switching frequencies in excess of 10 MHz.

### Testing

The test set up consisted of two variable duty cycle non-overlapping clocks  $\phi_{21}$  and  $\phi_{22}$ , derived from the pulse generator and applied to CCI. The auxiliary bread boarded circuit which was driven by  $\phi_{21}$  was used to generate complementary reset pulses after every 8 clock cycles. Thus, throughout the testing, integration is performed over 8 clock

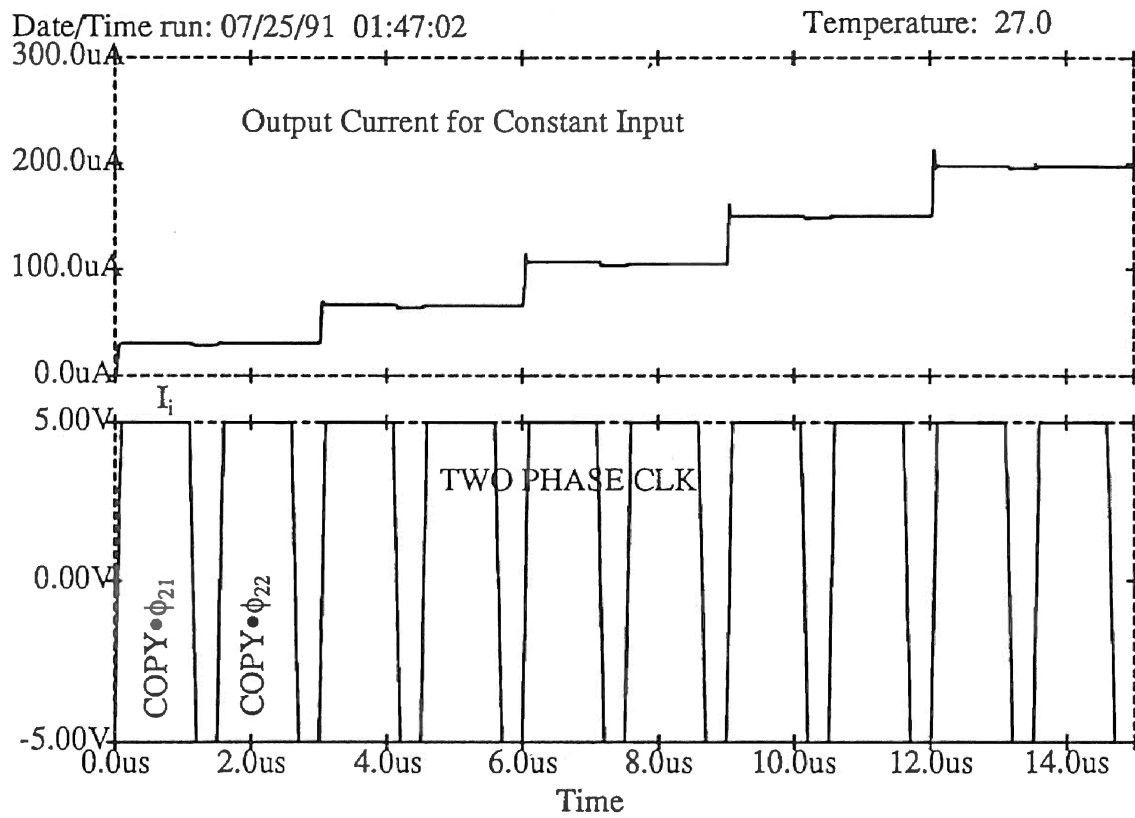


Figure 50. Transient Operation of the CCI

cycles by periodically resetting the  $V_{CH1}$  and  $V_{CH2}$  with reset pulses. The auxiliary circuit was bread boarded to produce a precision current sink  $I_i^*$ , where current was controlled in steps. The output current  $I_i$  was sampled across a precision 10 K $\Omega$  sampling resistor.

The performance was observed and recorded under two conditions: with, and without the external gate capacitances added to the internal MOS capacitances  $C_{H1}$  and  $C_{H2}$ . With an external capacitance of 200 pF each is added to both  $C_{H1}$  and  $C_{H2}$ , the clock speed is set at the low value (1 KHz). In this case for  $I_i^*$  equal to 20  $\mu$ A, during the first clock cycle,  $I_i$  is found to be 20  $\mu$ A which is in exact agreement with the integrator theory. But, during the second clock cycle,  $I_i$  raised to 70  $\mu$ A instead of the theoretical 40  $\mu$ A value, leading to a 75% error in the integrand. During the subsequent cycles, the output current is observed to be increasingly deviating from its expected theoretical values. Clearly, this is due to the cumulative integration of an error term, which is being added during every cycle along with the information signal. Thus, as the integration progresses, a larger error accumulates leading to a substantial error term during the later part of the prolonged integration cycles (200%). Hence, in the subsequent designs the following factors should be considered: (1) an error compensation scheme to the basic circuitry of Figure 48, in order to cancel the error in the integrand before it is processed further, and (2) the additional cascodes to MFN and MFP.

The circuit conditions without the external capacitance added to the circuit are identical to the earlier case except that  $I_i^*$  is set at 5  $\mu$ A. In this case, a step in the output current due to channel charge injection and/or feedthrough is observed. It occurs when  $S_1$  and  $S_2$  are opened. From equation 87, the channel charge injection error is a function of the ratio of switch oxide capacitance and hold capacitance. It has been suggested

previously that such an error can be reduced by making switch oxide capacitance a small percentage of  $C_{H1}$  and  $C_{H2}$ . Therefore, during the previous experiment, the off-chip 200 pF capacitances were added to the internal MOS capacitances  $CH_1$  and  $CH_2$ . With the external capacitances removed,  $\Delta V$  due to the channel charge injection or switch feedthrough is comparatively high, resulting in a false step in the output current in every cycle. The resultant error in the output current due to the channel charge injection or switch feedthrough was found to be as high as 40%.

### Summary

From the above discussion, it is clear that dynamic current copying techniques are potentially superior to the normal current mirroring techniques, due to the complete elimination of threshold mismatch errors and potential removal of flicker noise. However, proper selection of switching frequencies, incorporation of proper compensating schemes, and proper circuit design techniques cannot be overlooked when attempting to minimize errors. Appropriate device geometries, switches, and gate capacitances are important factors in the design of the CCI. Increasing gate capacitance  $C_{H1}$  and  $C_{H2}$  improves accuracy but lowers the operating frequencies. Therefore proper trade off between space and accuracy is required. In summary, charge injection, switch feedthrough and other errors due to a variation in drain voltage, are the main sources of errors which occur during the integration. Complementary clocks which are suggested for driving complementary transmission switches to cancel the effect of switch feedthrough appear to be of little value. Therefore, future designs will make use of dummy switches in conjunction with single channel transistor switches.



## CHAPTER IV

### CONCLUSIONS AND FUTURE PROSPECTS

Analog circuits are often criticized for their functionality when compared to their digital counterparts. Usually, analog integrated circuits designed by even the most experienced designers require multiple attempts to achieve desired results. Our experience in this regard is the other way around. Seven of nine blocks showed satisfactory DC behavior, due to the utmost care in design, simulation, and layout. However, the electronic olfaction topic is still open to many improvements, both in the olfactory model and in the refinement of electronic building blocks. The following suggestions provide the future scope that will help in realizing the system level integration of the GLA olfactory model.

The GLA olfactory model described in chapter II is most definitely biologically inspired, but the basic idea in the minds of the original investigators initially may not have been its hardware implementation. In other words, the model may need additional simplifications that favor a simple electronic implementation while retaining the model's essential clustering properties. The ongoing simulation efforts of the simplified model by our group at Oklahoma State University and some researchers elsewhere [47] will hopefully lead to further simplified but computationally efficient model in the near future. In spite of the number of favorable features that make the GLA model suitable for direct implementation, we foresee some problems that may be encountered during

implementation. The system level integration of the olfactory model will require additional knowledge of specific model parameters values (g, m, p, h). The primary task of selecting the best set of implementation strategies for an olfactory architecture is a rather difficult issue since olfaction is poorly understood. Extensive computer simulations will be required to analyze the effect of various model parameters such as number of glomerulus and mitral patches etc. on the clustering properties. This will assist in selecting the most optimal parameters thus providing efficient use of the silicon area. These parameters will have a direct impact on the transistor level design.

Two dimensional connectivity may form a bottleneck. In this regard, techniques like multiplexing, and inherent sparse and spatially local interconnect or shared wires will help to reduce routing complexity.

The problems of communication, weight representation, and learning will also be of particular importance. To achieve effective communication on the memory front, local storage of the weight in close proximity of the multiplier hardware is the preferred solution. The task of weight updates is complex since it involves issues related to high voltage non-linear programming, learning algorithm, weight storage, on/off chip learning etc. In other words local optimization will dominate design and will remain a key focus in any olfactory system design.

From an electronic perspective, the future prospects for electronic olfaction are unlimited. Adhering to the sequence as it is presented in chapter III, the multiplier testing results closely match with simulation and theoretical results. However, the present multiplier circuit is area consuming. The possibility of an alternative area efficient single quadrant multiplier needs to be investigated. The mitral patch circuit needs thorough

analysis. The idea of incorporating the sigmoidal function within the mitral patch by arranging thresholds in a nonlinear fashion certainly deserves some attention. The offset circuit associated with the multiplier needs special attention. The area efficient way must be found to realize an internal high leakage resistance. MOSFET operating in the subthreshold region should be investigated for this purpose.

In electronic neural networks, the problem of realizing a trainable analog medium is current subject of high interest. Floating gate memories provide the best answer to electrically programmable/erasable non-volatile semiconductor memories. Out of the numerous possibilities, the concept of standard CMOS floating gate memory, based on the field enhancement due to mask geometries, is relatively new and poorly understood. These memories may not ever be suitable due to their heavy dependence on the manufacturing process. Precise control of the weight needs extensive experimentation to mathematical model and understand and the programming and erasing behaviors. This will assist in uncovering the basic physical principals hidden behind the field enhancement due to mask geometries and the retention of charge.

System level integration will require a suitable programming scheme. An algorithm has to be devised to convert the inherently complex and non-linear programming into a relatively simplified and hopefully linearized learning algorithm.

The on-chip generation of high voltage poses a real challenge. However, the tunneling physics and high voltage pulse generation are two separate issues and initially should be handled separately for conceptual testing and understanding, and then should be combined together. Other issues relating to the weight matrix are cell layout, placement, and signal routing. Cell layout will have a direct impact on both the silicon

area as well as on the cell performance. Significant expertise is needed to arrive at the optimal design. A suitable signal routing scheme is required since the weight matrix is expected to be dominated by routing wires. In this regard, high voltage concerns such as field threshold, reverse breakdown etc. need special attention.

The testing of the WTA circuit reveals a limited operating range (0-70  $\mu\text{A}$ ). Device geometries have to be pushed to achieve a higher dynamic range. Further, a creative on-chip testing structure must be developed to measure the bandwidth. The CCI circuit has to be modified [65] to incorporate the error compensation scheme, improved dynamic cascoding, and the dummy switches. This will bring down the errors in the output current.

Finally, another milestone of this research, the system level integration of the olfactory model on a single substrate will require a serious effort. Each factor (simplification of the model, suitable programming scheme, on-chip high voltage generation, weight cell characterization etc.) by itself can be significant enough to be another thesis. By no means does the author imply that the above list of problems is complete. But as we dwell into the area, hopefully we will come up with many more opportunities for improvements.

## REFERENCES

1. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," in Parallel Distributed Processing, Explorations in the Microstructure of Cognition, D. E. Rumelhart & J. L. McClelland, Eds. MIT Press, Cambridge, MA, Vol. 1, pp. 318-362, 1986.
2. S. Grossberg, "Nonlinear Neural Networks: Principles, Mechanisms, and Architectures" Neural Networks, Vol. I, pp. 17-61, 1988.
3. R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, pp. 4-22, April 1987.
4. R. Sharda, and R. Patil, "Neural Networks as Forecasting Experts: An Empirical Test", International Joint Conference on the Neural Networks, Washington, D.C., Vol. II, pp. 491-494, 1990.
5. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computation Abilities," Proceedings of the American Academy of Sciences, Vol. 79, pp. 2554-2558, 1982.
6. M. Holler, S. Tam, H. Castro, and R. Benson, " An Electrically Trainable Artificial Neural Network (ETANN) with 10240 Floating Gate Synapses," in Proceedings, International Joint Conferences on the Neural Networks, Washington, DC, Vol. II, pp. 177-182, 1989.
7. B. Furman, and A. Abidi, "CMOS Analog IC Implementing the Back Propagation Algorithm," Neural Networks, Vol. 1, Sup. 1, pp. 381, 1988.
8. L. D. Jackel, H. P. Graf, and R. E. Howard," Electronic Neural Network Chips," Applied Optics, Vol. 26, pp. 5077-5080, 1987.
9. P. Mueller, J. Van der Spiegel, D. Blackman, T. Chiu, T. Clare, J. Dao, C. Donham, T. Hsieh, and M. Loinaz, " A General Purpose Analog Neurocomputer," Proceedings of the International Joint Conference on Neural Networks, Vol. II, pp. 191-196, 1989.
10. A. F. Murray, and A. Smith, " Asynchronous VLSI Neural Networks Using Pulse-Stream Arithmetic," IEEE Journal of Solid-State Circuits, Vol. 23, pp. 688-697, 1988.

11. D. B. Schwartz, R. E. Howard, and W. E. Hubbard, "A Programmable Analog Neural Network Chip," IEEE Journal of Solid-State Circuits, Vol. 24, pp. 313-319, 1989.
12. G. Lynch, and R. Granger, "Simulation and Analysis of a Simple Cortical Network", Psychology of Learning and Motivation, Vol. 22, pp. 1-87, 1988.
13. D. O. Hebb, The Organization of Behavior, Wiley, New York, 1949.
14. G. Lynch, Synapses, Circuits, and the Beginnings of Memory, MIT Press, Cambridge, MA, 1986.
15. D. Hammerstrom, and E. Means, "System Design for a Second Generation Neurocomputer," In Proceedings, International Conference on Neural Networks, Washington, Vol. II, pp. 80-83, 1990.
16. J. Ambros-Ingerson, R. Granger, and G. Lynch, "Simulation of Paleocortex Performs Hierarchical Clustering," Science, Vol. 247, pp. 1344-1348, 1990.
17. J. Ambros-Ingerson, Computational Properties and Behavioral Expression of Cortex-Peripheral Interactions Suggested by a Model of the Olfactory Bulb and Cortex, Ph.D. Dissertation, University of California, Irvine, 1990.
18. G. Lynch, and R. Granger, "Simulation and Analysis of a Simple Cortical Network," Psychology of Learning and Motivation, Vol. 23, pp. 205-241, 1989.
19. P. K. Sipmsom, "A Survey of Artificial Neural Systems,"
20. P. D. Wasserman, Neural Computing Theory and Practice
21. T. Kohonon, Self-Organization and Associative Memory, 2nd Edition, Springer-Verlag, Berlin, 1988.
22. D. Specht, "Probabilistic Neural Networks," Neural Networks, Vol. 3, pp. 109-118, 1990.
23. G. E. Hinton, and T. J. Sejnowski, "Learning and Relearning in Boltzmann Machines," in Parallel Distributed Processing, Explorations in the Microstructure of Cognition, D. E. Rumelhart & J. L. McClelland, Eds. MIT Press, Cambridge, MA, Vol. 1, pp. 282-317, 1986.
24. J. Daugman, "Networks for Image Analysis: Motion and Texture," in proceedings, International Joint Conference on Neural Networks, Washington DC, Vol. I, pp. 189-194, 1989.

25. N. Suga, "Cortical Computational Maps for Auditory Imaging," Neural Networks, Vol. 3, pp. 3-22, 1990.
26. Y. Tao and W. J. Freeman, "Model of Biological Pattern Recognition with Spatially Chaotic Dynamics," Neural Networks, Vol. 3, pp. 153-170, 1990.
27. J. P. Wagon, A Proposal for An Analog CMOS Median Filter System Based On Neural Network Architectural Principles, Masters Thesis, Oklahoma State University, Stillwater, 1988.
28. J. Bailey, D. Hammerstrom, J. Mates, and M. Rudnick, "Silicon Association Cortex. In S. F. Zornetzer," J. L. Davis, and C. Lau, editors, An Introduction to Neural and Electronic Networks, Academic Press, August 1989.
29. C. A. Mead, Analog VLSI and Neural Systems, Addison-Wesley, MA, 1989.
30. C. A. Mead, "Neuromorphic Electronic Systems," Proceedings of the IEEE, Vol. 78, pp. 1629-1636, 1990.
31. S. P. DeWeerth, and C. A. Mead, "An Analog VLSI Model of Adaptation in the Vestibulo-Ocular Reflex," in Advances in Neural Information Processing Systems 2, D. Touretzky, Ed. Morgan-Kaufmann, San Mateo, CA, pp. 742-749, 1990.
32. C. A. Mead, X. Arreguit, and J. Lazzaro, "Analog VLSI Model of Binaural Hearing," IEEE Transactions on Neural Networks, Vol. 2, pp. 230-236, 1991.
33. C. A. Mead, and Mahowald, "A Silicon Model of Early Visual Processing," Neural Networks, Vol. 1, pp. 91-97, 1988.
34. J. G Taylor, "A Silicon Model of Vertebrate Retinal Processing," Neural Networks, Vol. 3, pp. 171-178, 1990.
35. A. Moore, J. Allman, and R. M. Goodman, "A Real-Time Neural System for Color Constancy," IEEE Transactions on Neural Networks, Vol. 2, pp. 237-247, 1991.
36. H. C. Card, and W. R. Moore, "Silicon Models of Associative Learning In Aplysia," Neural Networks, Vol. 3, pp. 333-346, 1990.
37. R. Braham, and J. O. Hamblen, "The Design of a Neural Network with a Biologically Motivated Architecture," IEEE Transactions on Neural Networks, Vol. 1, No. 3, pp. 251-262, 1990.

38. C. Toumazou, J. Lidgley, and D. Haigh "Introduction," Ch. 1 in Analogue IC Design: The Current-Mode Approach, C. Toumazou, F. J. Lidgley, and D. G. Haigh, Eds., Peregrinus, London, 1990.
39. B. Gilbert, "Current-Mode Circuits From A Translinear Viewpoint: A Tutorial," Ch. 2 in Analogue IC Design: The Current-Mode Approach, C. Toumazou, F. J. Lidgley, and D. G. Haigh, Eds., Peregrinus, London, 1990.
40. S. T. Dupuie, and M. Ismail, "High Frequency CMOS Transconductors," Ch. 5, in Analogue IC Design: The Current-Mode Approach, C. Toumazou, F. J. Lidgley, and D. G. Haigh, Eds., Peregrinus, London, 1990.
41. K. Bult, and H. Wallinga, "A Class of Analog CMOS Circuits Based on the Squire-Law Characteristics of an MOS Transistor in Saturation," IEEE J. Solid-State Circuits, Vol. SC-22, pp. 357-365, June 1987.
42. S. B. Patil, and C. G. Hutchens, "A Novel Squashing Function for Electronic Implementation of Neural Networks," 5th Oklahoma Symposium of Artificial Intelligence, 1991.
43. Y. Tsvividis, Operation and Modeling of MOS Transistor,
44. P. E. Allen, and D. R. Holberg, "Two Stage Comparators," Ch. 7, in CMOS Analog Circuit Design, HRW Inc., 1987.
45. A. S. Sedra, and G. W. Roberts, "Current Conveyor Theory and Practice," Ch. 3 in Analogue IC Design: The Current-Mode Approach, C. Toumazou, F. J. Lidgley, and D. G. Haigh, Eds., Peregrinus, London, 1990.
46. VLSI Design Techniques For Analog and Digital Circuits
47. P. A. Shoemaker, C. G. Hutchens and, S. B. Patil, "A Hierarchical Clustering Network Based on a Model of Olfactory Processing," Submitted, 1992.
48. T. H. Borgstrom, M. Ismail, and S. B. Bibyk, "Programmable Current-Mode Neural Network for Implementation in Analogue MOS VLSI," IEE Proceedings, Vol. 137, pt. G, No. 2, pp. 175-183, April 1990.
49. W. M. Gosney, "DIFMOS-A Floating-Gate Electrically Erasable Nonvolatile Semiconductor Memory Technology," IEEE Transactions on Electron Devices, Vol. Ed. 24, No. 5, pp. 594-599, May 1977.
50. Y. Tsvividis, and S. Satyanarayana, "Analog Circuits for Variable-Synapse Electronic Neural Networks," Electronics Letters, Vol. 23, No. 24, pp. 1313-1314, November 1987.



51. R. L. Shimabukuro, and P. A. Shoemaker, "Circuitry for Artificial Neural Networks with Nonvolatile Analog Memories," Proceedings, IEEE International Symposium on Circuits and Systems, pp. 1217-1220, 1989.
52. C. Bulucea, " Avalanche Injection into the Oxide In Silicon Gate Controlled Devices-I. Theory," Solid State Electronics, Vol. 18, pp. 363-374, 1975.
53. S. M. Sze, Physics of Semiconductor Devices, Wiley, Newyork, 1981.
54. P. A. Shoemaker, M. J. Carlin, and R. L. Shimabukuro, "Back-propagation Learning with Trinary Quantization of Weight Updates," Neural Networks, Vol. 4, pp. 231-241, 1991.
55. D. Khang and S. M. Sze, Bell System Tech., J. 46, 1288, 1967.
56. E. H. Nicollian, A. Goetzberger, and C. N. Berglund, Applied Physics Letters 15, pp. 174, 1969.
57. D. Frohman-Bentchkowsky, " Memory Behavior In a Floating-Gate Avalanche-Injection MOS (FAMOS) Structure," Applied Physics Letters, Vol. 18, Number 8, pp. 332-334, April 1971.
58. D. Frohman-Bentchkowsky, "FAMOS - A New Semiconductor Charge Storage Device," Solid state Electronics, Vol. 17, pp. 517-529, 1974.
59. Y. Tarui, Y. Hayashi, and K. Nagoi, "Electrically Reprogrammable Non-volatile Semiconductor Memories," IEEE J., SC-7, pp. 369-375, 1972.
60. T. G. Carlstedt, and Svensson C. M., " MNOS Memory Transistor In Simple Memory Arrays," IEEE J., SC-7, pp. 382-388, 1972.
61. R. A. Williams, and M. M. E. Begueala, "The Effect of Electrical Conduction of  $\text{Si}_3\text{N}_4$  On the Discharge of MNOS Memory Transistor," IEEE Transaction, ED-25, 8, pp. 1019-1022, 1978.
62. L. R. Carley, " Trimming Analog Circuits Using Floating-Gate Analog MOS Memory" Circuits, Vol. 24, No. 6, pp. 1569-1575, December 1989.
63. B. W. Lee, B. J. Sheu, and H. Yang, " Analog Floating-Gate Synapses for General-Purpose VLSI Neural Network Computation," IEEE Transaction on Circuits and Systems, Vol. 38, No. 6, June 1991.
64. J. Lazzaro, S. Ryckebusch, M.A. Mahowald, and C. A. Mead, "Winner-Take-All Networks of  $O(N)$  Complexity", California Institute of Technology Technical Report Caltech-CS-TR-21-88, 1989.

65. E. A. Vittoz, and G. Wegmann, "Dynamic Current Mirrors," Ch. 7 in Analogue IC Design: The Current-Mode Approach, C. Toumazou, F. J. Lidgley, and D. G. Haigh, Eds., Peregrinus, London, 1990.
66. S. J. Daubert, and D. Vallancourt, "Operation and Analysis of Current Copier Circuits", IEE Proceedings, Vol. 137, Pt. G., No. 2, pp. 109-115, April 1990.
67. C. Hutchens, A. Hill, and S. B. Patil "Simulation of an Olfactory Neural Paradigm Suitable for Electronic Clustering," 5th Oklahoma Symposium on Artificial Intelligence, Nov. 1991.
68. J. L. Wyatt, D. L. Standley, and W. Yang, "The MIT Vision Chip Project: Analog VLSI Systems for FAST Image Acquisition and Early Vision Processing," Proceedings of the IEEE, International Conference on Robotics and Automation, pp. 1330-1335, April 1991.
69. M. A. Sivilotti, M. A. Mahowald, and C. A. Mead, "Real-Time Visual Computation Using Analog CMOS Processing Arrays," 1987.
70. K. Goser, U. Hilleringmann, U. Rueckert, and K. Schumacher," VLSI Technologies for Artificial Neural Networks," IEEE Micro, pp. 28-44, December 1989.
71. J. P. Sage, K. Thompson, and R. S. Withers, " An Artificial Neural Network Integrated Circuit Based on MNOS/CCD Principles," American Institute of Physics, pp.381-385, 1986.
72. A. J. Agranat, C. F. Neugebauer, and A. Yariv,"A CCD Based Neural Network Integrated Circuit With 64K Analog Programmable Synapses, IJCNN , pp. II-552-555,
73. Y. P. Tsividis, and D. Anastassiou, " Switched- Capacitor Neural Networks," Electronics Letters, Vol. 23, No. 18, pp. 958-959, August 1987.
74. M. Stanford Tomlison Jr., D. J. Walker, and M. A. Sivilotti," A Digital Neural Network Architecture for VLSI," IJCNN, pp. II 545-550.
75. A. F. Murray, and Anthony V. W. Smith, " Asynchronous VLSI Neural Networks Using Pulse-Stream Arithmetic" IEEE, pp. 688-697, 1988.
76. W. Wike, D. Van den, and T. Miller III, "The VLSI Implementation of STONN, " IJCNN, pp. II-593- 598.
77. S. Satyanarayana, and Y. Tsividis," Analog Neural Networks with Distributed Neurons,"Electronics Letters, Vol. 25, No. 5, pp. 302-303, March 1989.

78. J. Alspector and R. B. Allen, "A Neuromorphic VLSI Learning System," Bell Communication Research, pp. 314-345.
79. H. P. Graf, and P. de Vegvar, "A CMOS Implementation of a Neural Network Model," AT & T Bell Laboratories, Holmdel.
80. D. Hammerstrom, "A VLSI Architecture for High-Performance, Low-Cost, Onchip Learning," Adaptive Solutions Inc., Beaverton, Oregon, pp. II-537-544, February 28, 1990.
81. Y. Hirai, K. Kamada, M. Yamada, and M. Ooyama, "A Digital Neuro-Chip With Unlimited Connectability for Large Scale Neural Networks," Institute of Information Sciences and Electronics, University of Tsukuba, Japan, pp. II-163-169.
82. P. W. Hollis, "Artificial Neural Network Using MOS Analog Multipliers," IEEE Journal of Solid-State Circuits, Vol. 25, No. 3, pp. 849-855, June 1990.
83. N. I. Khachab, and M. Ismile, "MOS Multiplier/Divider Cell For Analogue VLSI," Electronics Letters, Vol. 25, No. 2, pp. 1550-1553, November 1989.
84. B. Hochet, "Multivalued MOS Memory for Variable-Synapse Neural Networks," Electronics Letters, Vol. 25, No. 10, pp. 669-670, May 1989.
85. J. Alspector, R. B. Allen, V. Hu, and S. Satyanarayana, "Stochastic Learning Networks and Their Implementations," In D. Z. Anderson (Ed.), Proceedings of IEEE Conference on Neural Information Processing Systems-Natural and Synthetic, pp. 9-21, 1988.

VITA

Sanjay B. Patil

Candidate for the Degree of

Master of Science

Thesis: VLSI IMPLEMENTATION OF OLFACTORY CORTEX MODEL

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Maharashtra, India, August 28, 1966, son of Dr. Bhagawan Patil and Mrs. Vatsala Patil.

Educational: Graduated from Shri Shivaji Secondary School, Navapur, India, 1981; received a Diploma in Electrical Engineering from Government Polytechnic, Yeotmal in July 1984; received Bachelor of Electrical Engineering from College of Engineering Poona, in July 1987; completed requirements for the Master of Science degree at Oklahoma State University in May, 1993.

Professional Experience: Research Assistant (1991-Present), Dept. of Electrical Engg., OSU; Teaching Assistant (FALL-1991), Dept. of Electrical Engg., OSU; Research Assistant (1990-1991), College of Business Administration, OSU.

Design Executive, Switchgears (1988-89), Siemens Ltd., Bombay, India.

Production Engineer, Switch Boards (1987-88), Siemens Ltd., Bombay, India.