Parallel Analog Computation with Charge Coupled Devices

Thesis by

Charles F. Neugebauer

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Applied Physics Department

Pasadena, California

1993

(Defended May 12, 1993)

Acknowledgments

I would like to thank Amnon Yariv for providing an ideal environment for the pursuit of this research as well as unwavering support and encouragement. I am very grateful to Ron Agranat, whose ideas and backing helped start me in this direction. Gert Cauwenberghs and Volnei Pedroni deserve special recognition for their help honing the ideas presented here. Carver Mead and the people in his lab are unique in their openness and enthusiasm and have greatly enriched my time at Caltech.

I am grateful to Barbara Yoon at ARPA and Jim Mulligan of the Army CSW for providing financial support for this work. Thanks to the National Science Foundation and AT&T for graduate fellowship support.

Lastly, I would like to thank Ann Lewis, my family and all my friends for their unceasing support for everything I try to do.

Abstract

Many signal processing and neural network algorithms can be mathematically described in terms of vector matrix multiplication. This thesis introduces two new architectures for computing high-speed vector matrix multiplication using charge coupled devices. These integrated circuits have been designed to accept optical matrix input as well as direct electrical matrix input. In both architectures, the matrix elements are stored as analog charge packets in CCD wells while the vectors are communicated to and from the integrated circuits by electrical means.

The first architecture accomplishes the vector matrix product using a semiparallel computation scheme that requires N clock cycles of the device to complete one vector matrix multiplication where N is the length of the input vector. An analysis of the linearity and charge transfer induced errors is given. The circuit represents an advance over other analog signal processors in density and speed but has serious shortcomings in accuracy, particularly the limited precision of the input vectors.

The second architecture is based on charge injection device (CID) imager arrays and addresses many of the inadequacies of the semiparallel architecture. A fully parallel circuit, the CID has similar density and much higher computation speed and accuracy. A novel digital input method is introduced that extends the input vector precision significantly. In addition, accuracy issues related to charge transfer efficiency are resolved. An analysis of linearity and accuracy is provided showing the advantages of the architecture over previous implementations.

Contents

Chapter 1

1. INTRODUCTION

This thesis discusses the electronic implementation of biologically inspired signal processing algorithms. One feature common to most of these neural computation algorithms is the high degree of interconnection between neurons. Compared to traditional computing structures, neural architectures derive most of their computational abilities from the connections between processors as opposed to the processors themselves. The majority of neural network implementations tend to focus on the interconnect problem and not the neuronal simulation.

It is important to note the physical limitations of trying to map biological networks which have three dimensions at their disposal to form connections onto man-made technologies. The dominant implementation technology for computing devices is silicon, which is essentially restricted to two dimensions for communication. While optical and chip stacking technologies have emerged as solutions to the interconnect density problem, the displacement of conventional flat silicon by these and other esoteric approaches is unlikely in the near future. As such, the circuit designer is left to grapple with the transformation of three-dimensional circuits to two dimensions.

The application of CCD signal processing techniques to parallel computation is addressed in this thesis. Analog CCD signal processing circuits have been described in the past and have the

advantageous features of dense integration, small size, low power, and high speed. Exploiting these features in the hope of creating a compact neural signal processor is the goal of this work. As in most analog hardware of this sort, the choice of information representation is tightly coupled with the resulting circuit implementation. The primary emphasis of this thesis is on the device analysis at the simple connection level and how this pertains to system level performance with the proper choice of information representation.

The thesis is organized as follows:

- Chapter 1 presents an overview of system level considerations which influence the choice of interconnection devices. A cursory description of algorithms is given and various computation schemes and technologies are presented.

- Chapter 2 gives an introduction to CCD device physics and develops a simple formalism for the performance analysis of these CCD circuits relevant to the architectures discussed in this thesis. A new set of practical voltage limits is derived for buried channel devices.

- Chapter 3 deals with the first circuit implementation of a vector matrix multiplier using a semiparallel CCD architecture.

- Chapter 4 describes a fully parallel CID architecture that supersedes the one described in Chapter 3 by virtue of its improved speed, accuracy and density.

- Chapter 5 comments on the relative performance of these architectures versus other implementations.

Due to the limited exposure most readers have to CCD signal processing devices, an introductory text on CCD circuits such as [Séquin et al., 1975] or [Beynon et al., 1980] can be useful for understanding this thesis. Although an introduction to CCD devices is given in Chapter 2, the historical background and device evolution presented in the introductory texts is of significant value as it provides a proper perspective from which to view this work.

Also, the short introduction to neural networks is only intended to illuminate the implementation details necessary for developing the devices. The reader is encouraged to read [Kohonen, 1984][McClelland et al., 1986][Lippmann, 1987] if further background information is required.

1.1 Neural Networks and Adaptive Signal Processing

The development of computational models of biological neurons has been undertaken to understand the principle mechanisms of neural system behavior and to mimic this behavior in artificial machine-based neurons. Various levels of abstraction characterize the major theories' mathematical descriptions of the operation of neurons. Real neurons are naturally very complex, responding to both electrical and chemical stimuli often from thousands of other neurons with diffusion constants, electrical spike generation and propagation and a multitude of interaction pathways complicating the mathematical analysis. While detailed simulation of most of these attributes is possible, the significant computational resources required precludes this approach in cases where many neurons interact. The incorporation of the adaptive mechanisms of neuron behavior to the model, specifically synaptic plasticity and long term potentiation, complicate matters even further. A mathematical abstraction takes place at some level in all simulations, the real question being what level abstraction is necessary to capture the essential behavior without unnecessarily complicating the analysis. While much of the answer to this question depends on the system being studied, a number of useful descriptions of neurons have evolved that provide

adaptive behavior with a minimal amount of computation.

One of the simplest descriptions of the aggregating properties of neurons [Minsky et al., 1969] treats connections (i.e., synapses) as simple linear elements which weight signals. In this model, input signals are multiplied by the connection strength and added together to form the post-synaptic potential. The model lumps excitatory and inhibitory synapses into the same connection by allowing it to be positive or negative. This is the only form of stimulus considered and no synaptic time constants are included. This primitive model can effectively simulate some of the behavior of real neurons, proving that in some cases simple linear synaptic connections are adequate models.

A further abstraction can be made with respect to the electrochemical stimulus of the neurons. A quantity called the excitation is introduced which roughly corresponds to the instantaneous firing rate of a neuron. The effect of pulses can be compressed into a pulse rate, effectively removing the time dimension from the simulation. While this results in significant computational savings, the essential time domain behavior of neurons is lost. The penalties associated with this abstraction vary from system to system and can be significant for coherence sensitive problems such as auditory localization [Mead, 1989]. For a large class of simple problems, however, this abstraction is not detrimental [McClelland et al., 1986].

In [Hopfield, 1982], Hopfield describes a simple distributed neural system that behaves as a content addressable memory. In a later model [Hopfield, 1984] the neuron activity is represented as a single number corresponding to the firing rate. The set of neuron activities was described as a vector while the synaptic strengths were encoded as a matrix. Dynamic interactions were given by a simple differential equation that governed the reaction of each neuron to its input. Implicit in the mathematical description was the connectivity of the neurons expressed as a matrix. The computation involved repeated multiplication of the matrix by the neuron activation vector. In addition, a soft thresholding function related the aggregated input to the neuronal output. This

nonlinearity gives neural networks the ability to perform complex mappings given multiple layers [Minsky et al., 1969][McClelland et al., 1986].

The linear synapse model is the most common description of synaptic activation as it seems to capture enough of the real behavior of neurons to be useful for many problems. In a signal processing sense, the input vector corresponds to an input signal and the matrix corresponds to a linear transform. The information processing capabilities of neural networks are enhanced significantly when the matrix is adaptively modified to improve performance. Such learning behavior in simple linear transforms was first described in [Widrow et al., 1960] and has found widespread use in communications systems. The algorithm was developed by first quantifying the performance of the network into a single number, the global error, then finding the derivative of this error with respect to the matrix elements. The matrix elements are then modified slightly in proportion to the error gradient and the error measurement process is repeated. This incremental learning method is referred to as gradient descent for obvious reasons. While the recipe for gradient descent has been modified extensively for performance reasons, the basic tenet of reducing a global error remains true.

Gradient descent was mathematically motivated by a signal processing need. It is safe to say that biology does not use a simple gradient descent for adaptive behavior. While not biologically motivated, gradient descent is a simple solution to a significant signal processing problem where adaptive systems are needed.

The integration of neural network models and gradient descent occurred in the mid-1980's when back propagation was introduced. Partially motivated by biology in structure and by signal processing in adaptation, back propagation provided a formalism for adapting the weights of a feed-forward nonlinear network [McClelland et al., 1986]. The concept of a multilayer network with multiple stages of neurons feeding into more neurons is diagrammed in Figure 1.1.

Information enters at the top and percolates down. Each layer consists of a vector matrix multiplication and a soft nonlinearity. Such a feed-forward network, when presented patterns at the top as input and a requested target at the bottom as desired output, is able to learn the association or mapping between input and output vectors under certain conditions. The learning process is governed by

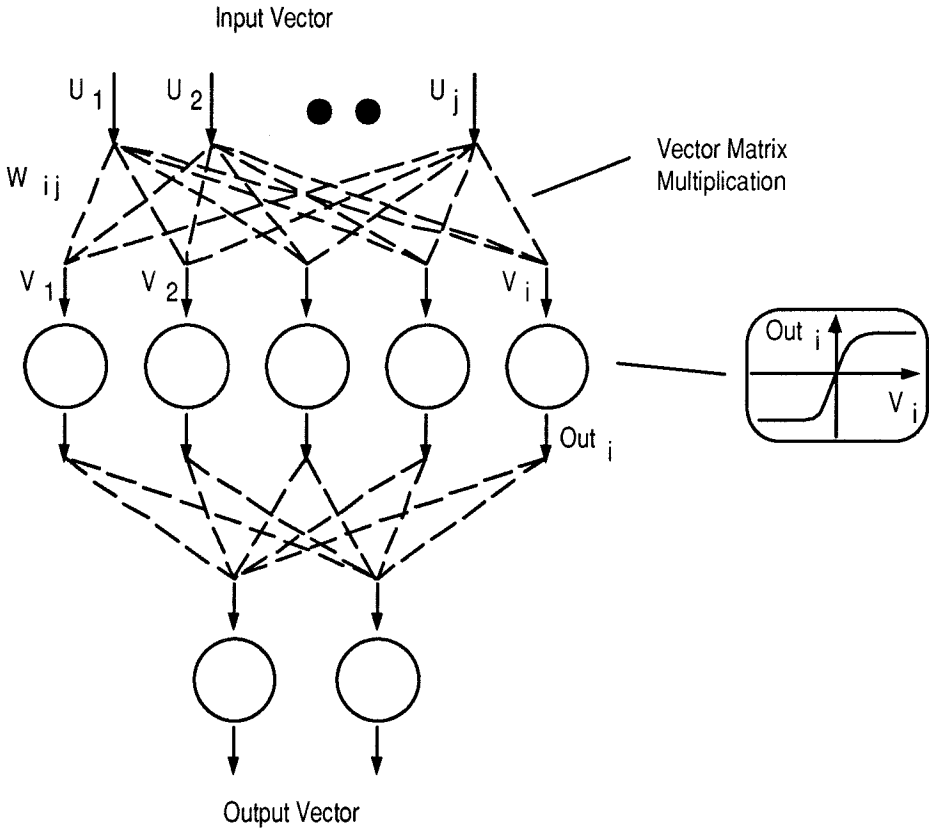

Figure 1.1. A feed forward neural network with a single hidden layer. Information flow is top to bottom. The input vector pattern U_j is linearly weighted and summed by vector-matrix multiplication with the W_{ij} matrix. A sigmoidal function is applied to the aggregated signals to form the output of the hidden layer. This output is similarly weighted, summed and put through the sigmoidal nonlinearity to form the final output vector.

gradient descent, which is not guaranteed to converge for multilayer nonlinear networks and makes

the entire procedure fraught with pitfalls in the areas of choosing network size, learning parameters, etc.. While of limited usefulness in understanding and mimicking neurobiology, the back propagation approach has proven to be useful in applications where the relationship between input and output data is unknown and target outputs are easily generated for teaching purposes.

1.2 Vector Matrix Multiplication

The algorithms described above center around vector matrix multiplication which typically constitutes the bulk of the processor load when such algorithms are simulated on a computer. Obvious advantages exist in implementing the algorithms in special purpose hardware to take advantage of the inherent parallelism of neural calculations, although the general purpose flexibility of a conventional computer is lost. Mapping the connectivity to a matrix is a natural process for two-dimensional silicon technology. The work described in this thesis is targeted at efficient silicon implementations of the vector matrix multiplication.

The vector matrix product is useful in a number of other signal processing applications. Correlations, pattern matching and simple linear transforms such as wavelet and chirp-z transforms are all functionally described as vector matrix multiplication. Hence the functionality of the devices presented extends beyond the field of neural networks.

1.3 Technologies for Parallel Computation

The success of digital computing is due to a number of factors such as its programmability, high accuracy, noise insensitivity and low cost. Although limited in accuracy, analog circuits have held the promise of performing a given computation much more efficiently in terms of area, power and

speed than their digital functional equivalent. Neural network models are intentionally robust with respect to accuracy limitations and offset errors due to their distributed and adaptive behavior, making them ideal candidates for analog implementations. Furthermore, the real world sensor data that most neural systems deal with is often low accuracy and high bandwidth, such as imager data which typically comes from an analog CCD sensor. Thus for certain low accuracy problems, the precision limitations of analog circuitry are far outweighed by the potential gains in speed, density and reduced power consumption. Although analog circuitry will never match digital in terms of accuracy and flexibility, for certain niche markets analog has the upper hand.

The choice of implementation technology relies on the consideration of a number factors. Computational devices require information storage, either short term or long term depending on the application. For the systems mentioned above, the need for accurate long term storage is a primary concern as the matrices are often fixed or adapt slowly. On the other hand, time multiplexing a single chip with a few hundred neurons to take the place of hundreds of thousands of neurons requires that the many different connection matrices be swapped quickly, implying off-chip long term storage.

Many methods of storing information in silicon technology, including DRAMs, EPROMs and CCDs, involve charge storage on a capacitor. Except for floating gate capacitors, these MOS capacitors exhibit leakage currents which tend to degrade the stored information within tens of milliseconds. Floating gate structures, while exhibiting extremely long storage times, are characterized by a number of inconveniences such as slow write times and a limited number of write cycles. The loading limitations on floating gate circuits precludes them from consideration for time multiplexed neural systems.

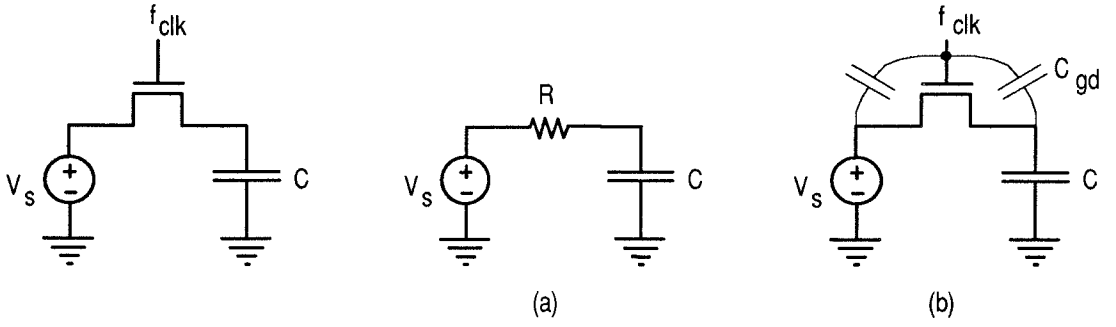

Figure 1.2. Error sources in switched capacitor circuits. The clock signal controls the gate of the MOSFET device which has a finite 'on' resistance shown in (a) and contributes Johnson noise. The parasitic capacitance of the clock signal to the storage capacitor, C_{gd}, is another source of errors.

The comparison of conventional switched capacitor storage [Kub et al., 1990] and CCD circuits is straightforward since they both use capacitors as the storage medium. Voltages in switched capacitor circuits are transferred from place to place through a series of amplifiers, switches and capacitors. In a typical arrangement shown in Figure 1.2, switched capacitor circuits store charge on a capacitor by using a FET as a switch connected to a voltage source, which is most often an on-chip operational amplifier. The two dominant error sources are thermal noise and clock feedthrough, shown schematically in Figures 1.2(a) and 1.2(b). Thermal noise is due to the finite 'on' resistance of the FET during the write procedure which introduces a Johnson noise-induced voltage variation on the storage capacitor given by

$$\overline{V}_n^2 = \frac{kT}{C} \tag{1.1}$$

which is derived in Chapter 2. Note that the FET 'on' resistance does not appear in Equation 1.1. The parasitic capacitive coupling between the drain and gate of the switch transistor causes a voltage variation on the storage capacitor given approximately by

$$\Delta V_c \cong \frac{C_{gd}}{C+C_{gd}}\left(V_S + V_T + \gamma\left[\sqrt{2|\phi_F| + V_S} - \sqrt{2|\phi_F|}\right]\right) \tag{1.2}$$

where V_S is the source voltage, $\gamma \approx 0.7V^{1/2}$ is the bulk threshold parameter and $\phi_F \approx 0.7V$ is the strong inversion surface potential [Allen et al., 1987]. Here it is important to note that the parasitic gate-to-drain capacitor is in general proportional to the size of the FET. An engineering trade-off exists between the speed of the circuit and the amount of clock feedthrough error. If the FET is too small, the circuit will be speed-limited by the RC circuit formed by the FET and the storage capacitor. In practice, however, a minimum sized switch requires a relatively large capacitor to achieve even modest levels of accuracy due to the combination of these two error sources. This large capacitor requirement sets a density limit of switched capacitor storage circuits.

In contrast, CCD circuits move charge between elements as shown in Figure 1.3 which depicts a simple four phase CCD shift register. Clock feedthrough, i.e., a modification of the stored charge due to a transition of one of the clocked electrodes, does not occur due to charge conservation within the CCD. Also, thermal noise associated with creating charge packets is reduced since charge can be loaded optically or with a charge creation circuit designed to minimize this effect. The chip area penalty for low noise charge loading is negligible for large CCD arrays since one large charge creation area can be used to create the charge for many small storage areas. The upshot of these differences between CCD and switched capacitor circuits is that the capacitor storage cells can be made much smaller for CCD circuits (i.e., much higher density) without sacrificing accuracy. In addition, the natural ability of CCDs to form multiplexors, analog shift registers and to be optically loaded is very useful for signal processing systems. For these reasons, CCD technology was chosen.

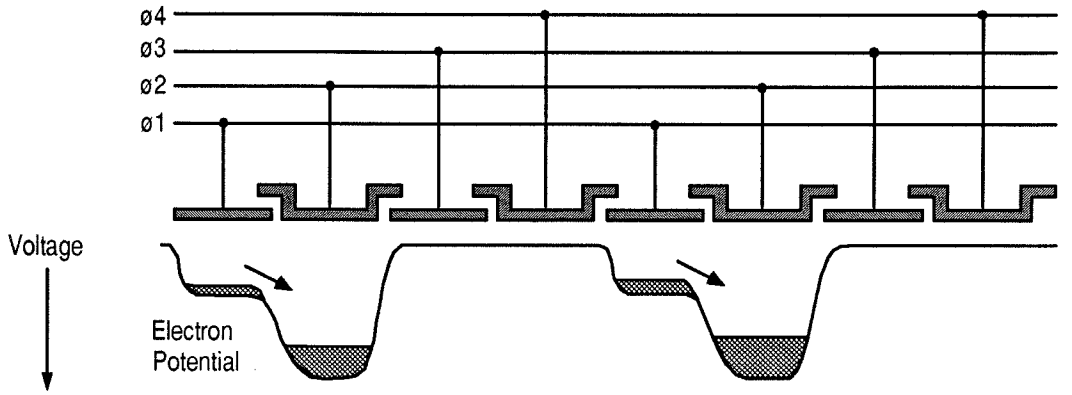

Figure 1.3. CCD shift register side view. Electron potential is an inverted plot of surface voltage of the channel. The surface potential is controlled by the polysilicon gates connected to the four-phase clock. The charge behaves like a fluid, moving to its local potential minimum.

1.4 Architectural Considerations

The essential element of an analog vector matrix multiplier is the matrix element which performs the synaptic weighting function. Each matrix element in the circuits of this thesis stores a given weight, W_{ij}, as a charge packet and performs a multiply-accumulate operation. The input vector element U_j is multiplied by W_{ij} at each matrix element and the product is added to the output vector element, V_i. A typical vector matrix multiplier chip block diagram is shown in Figure 1.4. Given that both input and output vectors are O(N) in size, the matrix contains $O(N^2)$ elements and typically requires the most chip area for large N. Thus it is important to concentrate on minimizing the matrix element size for large matrix designs.

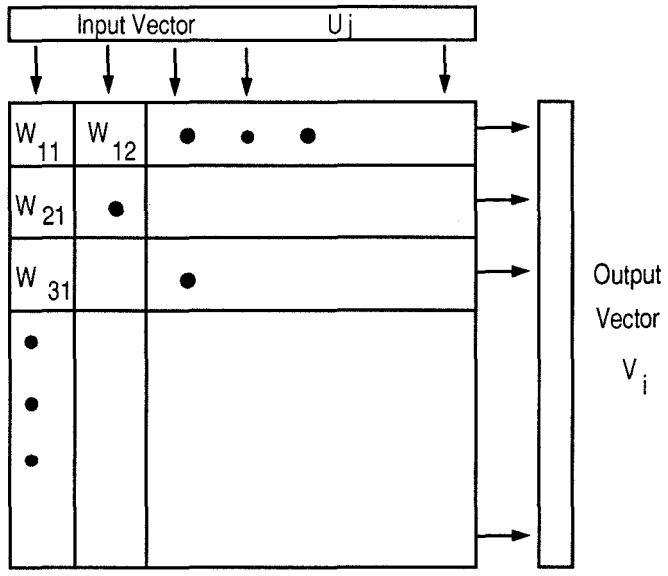

Figure 1.4. Vector Matrix Multiplier (VMM) block diagram. Each matrix element performs a multiply accumulate function to form the output vector. Weights are physically stored as charge packets in the array.

CCD technology has experienced widespread use in the field of electronic imaging where pixel density and analog performance are constant areas of improvement. The high quantum efficiency and low noise of CCD detectors make them ideal for light sensing applications. The stated goal of vector matrix multiplication can exploit this aspect of CCDs as a means of matrix input. Two-dimensional imager arrays can be used as the storage medium for matrix elements. The use of optical input as a parallel matrix communication scheme increases the bandwidth of the system beyond traditional pin limited designs and takes advantage of the parallelism of optics in a simple, implementable way.

The optoelectronic computation scheme is shown in Figure 1.5 where a spatial light modulator creates a two-dimensional pattern of optical intensities encoding the matrix which is projected onto the CCD vector matrix multiplier chip. This concept of hybrid CCD/optical systems is not new -- in [Copeland et al., 1976] for example, an optically encoded filter characteristic is loaded onto an IC for subsequent FIR filtering of an electrically input signal. Although spatial light modulator technology lags significantly behind silicon technology, the advent of sophisticated holographic

crystal storage mechanisms and other exotic technologies may lead to significant system performance enhancements.

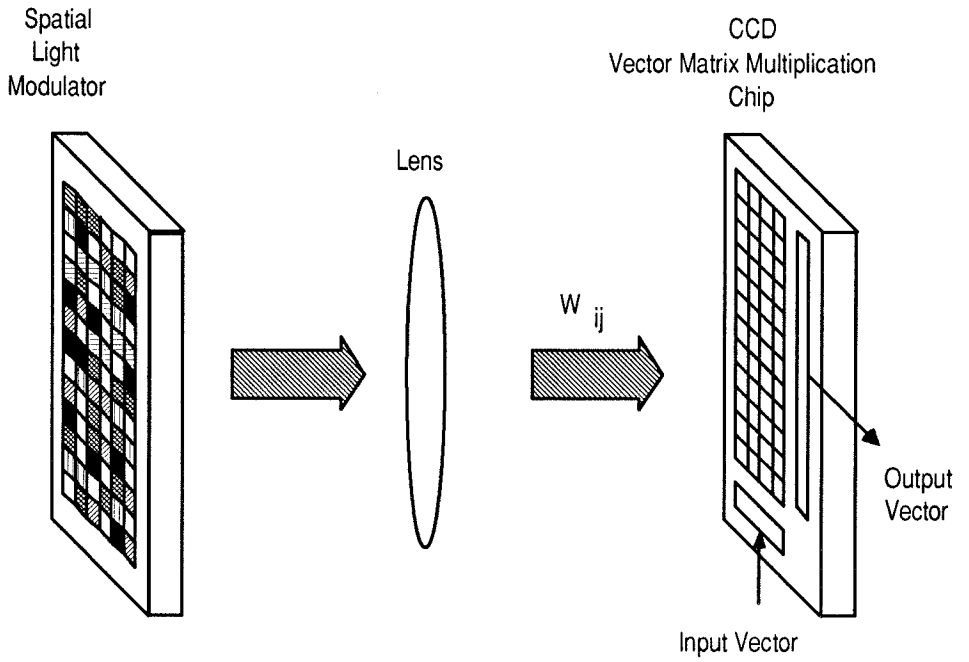

Figure 1.5. Optical matrix input. The matrix elements of the CCD VMM chip are light sensitive allowing matrix input to be performed optically. The pattern of intensities on the spatial light modulator is imaged onto the CCD chip for a brief period of time to accomplish loading.

Another option for creating a matrix of stored charge is an electrically controlled charge generator, described in Chapter 2. The charge generator creates precise charge packets which can then be shifted by CCD registers to the proper locations. This mode of operation makes for very compact systems without the need for a complicated optical arrangement at the expense of complex clocking and a significant reduction in matrix input bandwidth.

As mentioned above, the matrix cell size must be kept as small as possible which, in turn, implies a simple cell. The standard figures of merit for a given implementation are its power-delay product for functional efficiency and its area-delay product for silicon efficiency. As CCDs are inherently

low power and dense, the circuits presented here are expected to fare well by these measures. It is important to avoid comparisons with general purpose digital hardware such as microprocessors. The accuracy and programmability of conventional digital machines limits their efficiency of implementation by these two metrics. Only comparisons with special purpose digital hardware that has been optimized for one specific task with limited accuracy are valid. Such comparisons are presented in Chapter 5.

Given the task of computing a vector matrix multiplication, CCDs are one of many technologies that can be used, each of which has its own advantages and disadvantages. In the following chapter, the limitations of CCD devices in performing analog computation are explored at the device level. In later chapters, the specific architectures are discussed.

1.5 References:

[Allen et al., 1987] P.E. Allen and D.R. Holberg, *CMOS Analog Circuit Design.* New York: Holt, Rinehart and Winston, 1987.

[Beynon et al., 1980] J.D.E. Beynon and D.R. Lamb, *Charge Coupled Devices and Their Applications.* London: McGraw-Hill, 1980.

[Copeland et al., 1976] M.A. Copeland, D. Roy, J.D.E. Beynon and F.Y.K. Dea, "An optical CCD convolver," *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 152-155, 1976.

[Hopfield, 1982] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proceedings of the National Academy of Sciences*, USA, vol. 79, pp. 2554-2558, 1982.

[Hopfield, 1984] J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," in *Proceedings of the National Academy of Sciences, USA*, vol. 81, pp. 3088-3092, 1984.

[Hopfield et al., 1986] J.J. Hopfield ad D.W. Tank, "Computing with neural circuits: a model," *Science*, vol. 233, pp. 625-633, 1986.

[Howes et al., 1979] M.J. Howes and D.V. Morgan, *Charge Coupled Devices and Systems*. London: John Wiley & Sons, 1979.

[Kohonen, 1984] T. Kohonen, *Self-Organization and Associative Memory*. Berlin: Springer-Verlag, 1984.

[Kub et al., 1990] F.J. Kub, K.K. Moon, I.A. Mack and F.M. Long, "Programmable analog vector-matrix multipliers," *IEEE Journal of Solid State Circuits*, vol. SC-25(1), pp. 207-214, 1990.

[Lippmann, 1987] R.P. Lippmann, "An introduction to computing with neural networks," *IEEE ASSP Magazine*, pp. 4-22, 1987.

[McClelland et al., 1986] J.L. McClelland and D.E. Rumelhart, *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986.

[Mead, 1989] C.A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[Minsky et al., 1969] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press, 1969.

[Roth, 1988] M.W. Roth, "Neural-network technology and its applications," in *Johns Hopkins APL Technical Digest*, vol. 9(3), 1988.

[Séquin et al., 1975] C.H. Séquin and M.F. Tompsett, *Charge Transfer Devices*. New York: Academic Press, 1975.

[Tank et al., 1987] D.W. Tank and J.J. Hopfield, "Collective computation in neuron-like circuits," *Scientific American*, vol. 257, pp. 104-114, 1987.

[Widrow et al., 1960] B. Widrow and M.E. Hoff, "Adaptive switching circuits," in *IRE Western Electronic Show and Convention*, Part 4, pp. 96-104, 1960.

Chapter 2

2. CHARGE COUPLED DEVICES

The charge coupled device was conceived in 1970 [Boyle et al., 1970][Amelio et al., 1970] as a replacement for bucket brigade devices that had been used as analog delay lines. A good review of the history of charge transfer devices can be found in [Séquin et al., 1975]. CCDs were later adapted for use as imaging devices, a market they currently dominate. The basic principals of CCDs are presented in condensed form in this chapter with some new results regarding voltage limits, and then a framework for understanding the nonlinear effects inherent in charge manipulation is developed. In the following chapters, significant use of this formalism will be made to explain the limitations of analog CCD computation architectures.

2.1 MOS Capacitor Physics

The basic MOS capacitor is constructed of a polysilicon gate above a p-type substrate separated by silicon dioxide. The well known behavior of this device is best illustrated by the energy band diagrams of Figure 2.1. It is assumed the reader is familiar with this representation of semiconductor devices. A good introduction for the uninitiated can be found in [Sze, 1985]. All semiconductor voltages are given with respect to the conduction band and the gate voltages with

respect to the flat band voltage as in [Howes et al., 1979]. Figure 2.1(a) shows the flat band condition where the gate is grounded. For small positive gate voltages, the area beneath the polysilicon is depleted of majority carriers (holes) resulting in band bending in the depletion region, depicted in Figure 2.1(b), due to the space charge.

At higher gate voltages, minority carriers can accumulate beneath the gate, a condition called inversion, which is shown in Figure 2.1(c). Inversion charge can be generated by a number of sources with the major generation mechanism for a simple isolated MOS capacitor being thermal generation of charge due to traps [Howes et al., 1979].

Dynamically, if the gate is kept at a low voltage and then pulsed to a high voltage, the inversion charge will not have time to accumulate. This situation is referred to as deep depletion and is a pseudo-equilibrium state that exists before thermal inversion charge is accumulated. Shown in Figure 2.1(d), deep depletion is the mode in which the CCD circuits operate. The surface potential is a strong function of the gate voltage and forms a potential well for inversion electrons. The higher the gate voltage, the deeper the well. A MOS capacitor can hold varying amounts of inversion charge in the short time span (less than 100 milliseconds, typically) in which this pseudo-equilibrium approximation is valid. This charge can encode a signal level, as in CCDs, and can be moved by changing the potential of adjacent polysilicon gates as explained below.

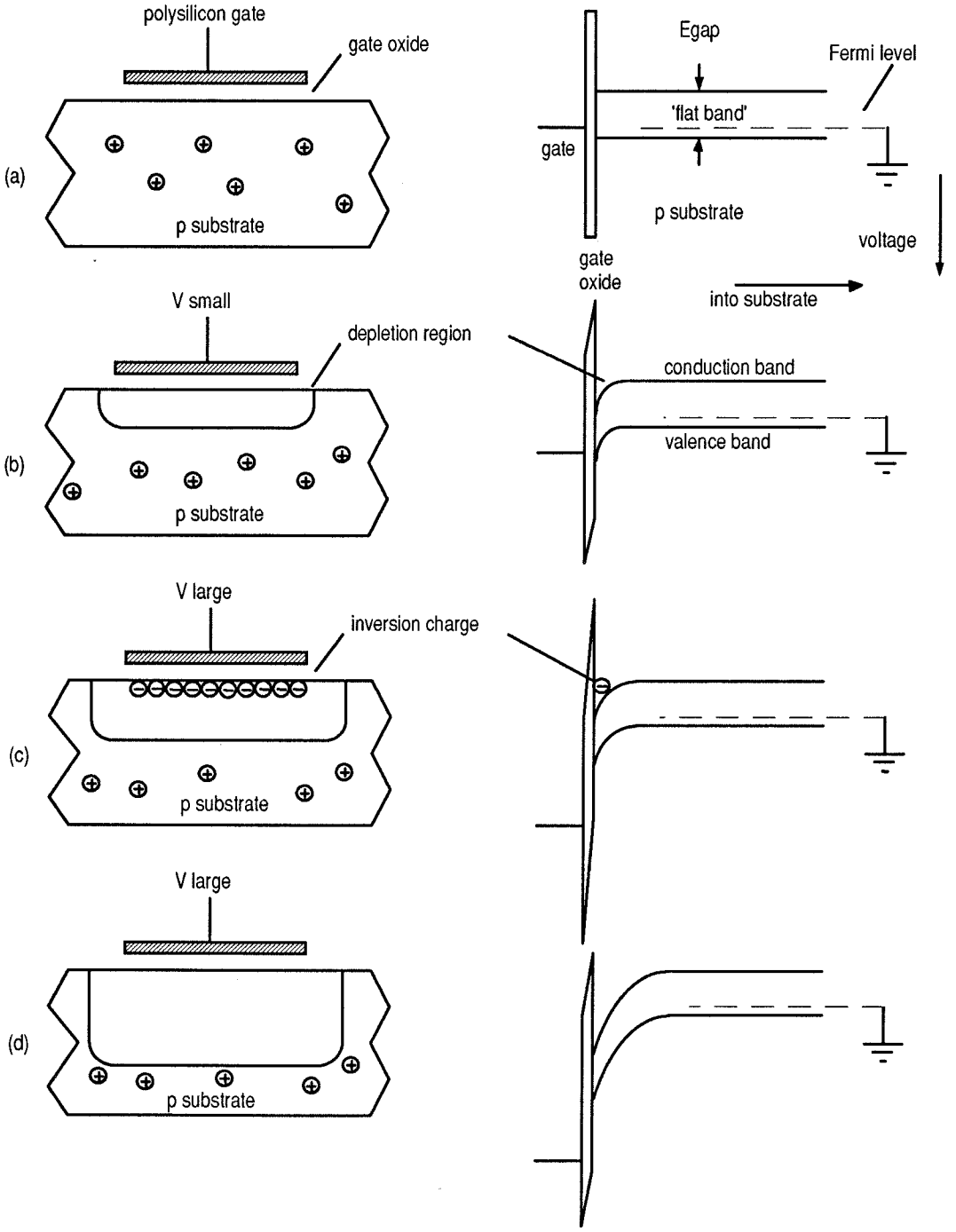

Figure 2.1. MOS capacitor band diagrams. In (a), the substrate is in thermal equilibrium with a uniform hole concentration (majority carriers) determined by the substrate doping density. In (b) the gate voltage is raised, repelling the holes in the substrate immediately beneath the gate to create the depletion region. Raising the gate voltage further (c) attracts electrons to the surface to form an inversion charge layer. In (d), if the gate is pulsed quickly to a high voltage, electrons do not have time to accumulate and a deep depletion is formed.

2.2 The CCD Shift Register

To introduce the operation of CCDs, the example of a shift register is explored. A CCD shift register consists of multiple overlapping polysilicon gates that are connected to time varying potentials, shown in Figure 2.2. The surface potential of the CCD channel is manipulated by the voltages on the polysilicon gates. Charge can be abstracted to behave as a fluid, moving to find the minimum potential (i.e., maximum voltage) level. In this abstraction, the gate voltages set the depth of the potential wells and hence control the relative potential energy of the electrons beneath the gates. If one gate is at a higher voltage with respect to its neighbors, its potential well is deeper and it will attract the electrons from under the neighboring gates. CCD shift register operation occurs when the polysilicon gates are driven with overlapping four-phase clock signals as depicted in Figure 2.2. Charge introduced from the left is moved to the right by the four-phase clock signals which act to repeatedly create a deeper potential well to the right of the charge packet, inducing the charge to move.

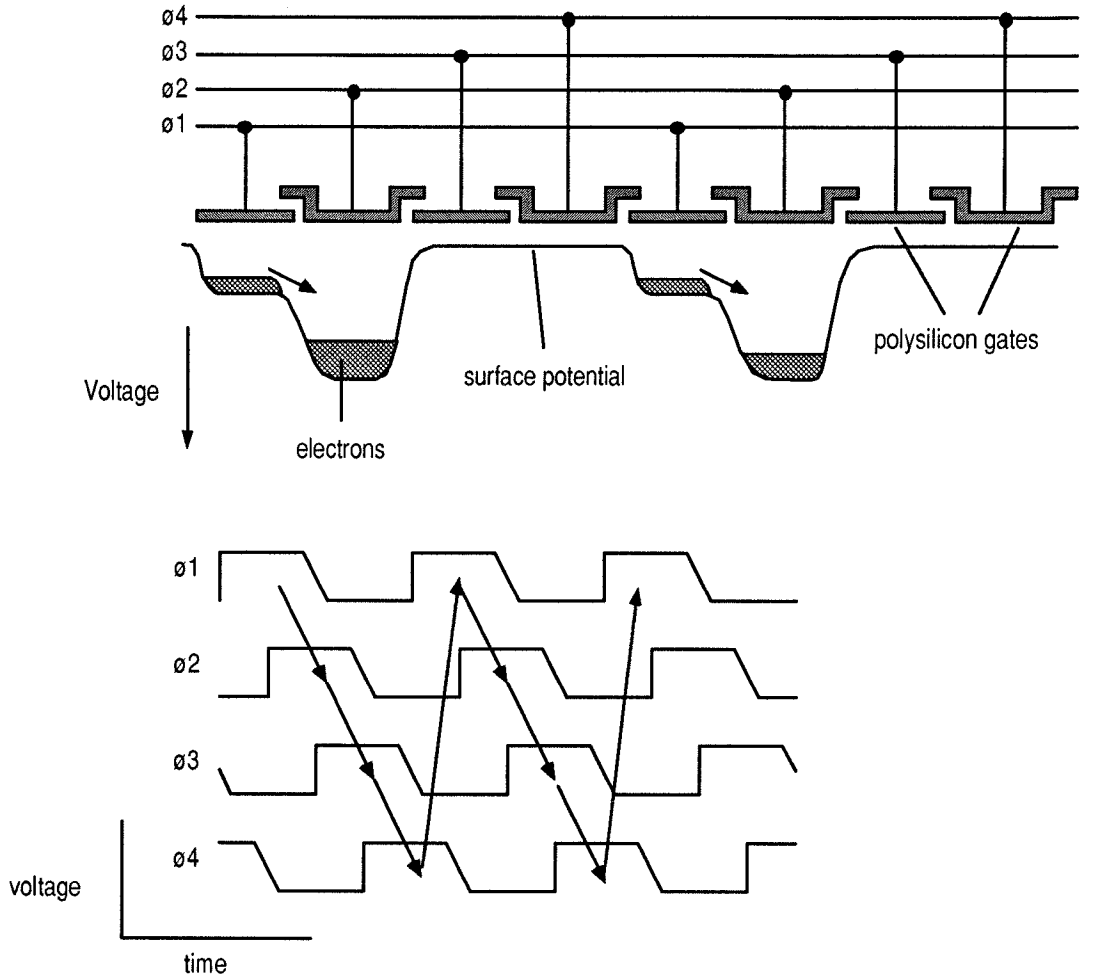

Figure 2.2. Clock voltages used to operate a CCD shift register. The charge is moved left to right by application of the four-phase clock signals to the polysilicon gates. The arrows on the voltage waveforms signify the charge packet following the voltage maximum in the channel.

2.3 Charge Transport in CCDs

Charge transport in CCDs is due to two fundamental processes, drift and diffusion. The condensed material presented here is based on the work of [Carnes et al., 1971] and that found in [Howes et al., 1979]. When the channel potential profile is changed to move charge from one gate to another, the initial movement of charge is primarily due to self-induced drift. Once a significant fraction of the charges have moved, diffusion takes over and empties the source well further. After almost all of the electrons have left the initial gate area, diffusion is no longer a strong driving force and

electric field induced drift completes the transfer. In typical CCDs the transfer of the last few electrons determines the transfer performance. Horizontal electric fields, also known as 'fringing fields', directly beneath a gate sweep out this small remaining charge. The time required to transfer the last few electrons is dependent on the gate length and the fringing field and is approximated by

$$\tau_E \approx \frac{L_{gate}}{E_{fringe}\mu_{e-Si}} \tag{2.1}$$

A first-order approximation to the fringing fields along the channel for a surface channel three-phase device can be found in [Carnes et al., 1971] and is given by

$$E_{fringe} \approx 6.5\frac{t_{ox}V}{L_{gate}^2}\left[\frac{\frac{5x_d}{L_{gate}}}{\frac{5x_d}{L_{gate}}+1}\right]^4 \tag{2.2}$$

where X_D is the depletion depth and V is the voltage difference between gates. Thus transfer time varies as

$$\tau_E \propto \frac{L_{gate}^3}{t_{ox}V} \quad \text{to} \quad \frac{L_{gate}^7}{t_{ox}V}. \tag{2.3}$$

Short gates are thus a prerequisite for high-speed CCD circuits, noting the strong dependence on gate length. A more complete review with numerical simulations can be found in [Mohsen et al., 1973][Mohsen et al., 1975] and more recently in [Bakker, 1991].

2.4 Charge Transfer Efficiency and Dark Current

The simple operation of the shift register is complicated by the effects of incomplete charge transfer [Berglund et al., 1973]. Fabrication process, clock voltages and edge rates all have significant effects on charge transfer. Charge Transfer Efficiency (CTE) is the common performance measure of CCDs and is defined as the fraction of charge that is actually transferred from one gate to an adjacent gate. Some fraction, ε, is left behind and is added to the trailing charge packet in the shift register. More advanced abstractions of CTE beyond this simple linear (i.e., fractional) model have been studied [Séquin et al., 1975], but for this analysis, the simple model will suffice. In the following sections, clock voltage ranges and edge rate limits are examined with the performance criterion of maximizing CTE.

An easily controllable source of transfer inefficiency is fast clock edge rates. The dynamics of the fluid-like charge cause it to equilibrate with a design-dependent time constant as discussed in Section 2.3. Referring to Figure 2.3, if a gate voltage transition occurs quicker than the charge can equilibrate, a small amount of charge can be transferred backwards in a shift register operation [Singh et al., 1974]. This results in a smearing of the charge encoded data residing in the shift register. In practice, gate voltage fall times must be limited to prevent this type of transfer inefficiency. An experimental measurement of transfer efficiency loss is shown in Figure 2.4 where a buried channel shift register was clocked with a variable edge rate clock.

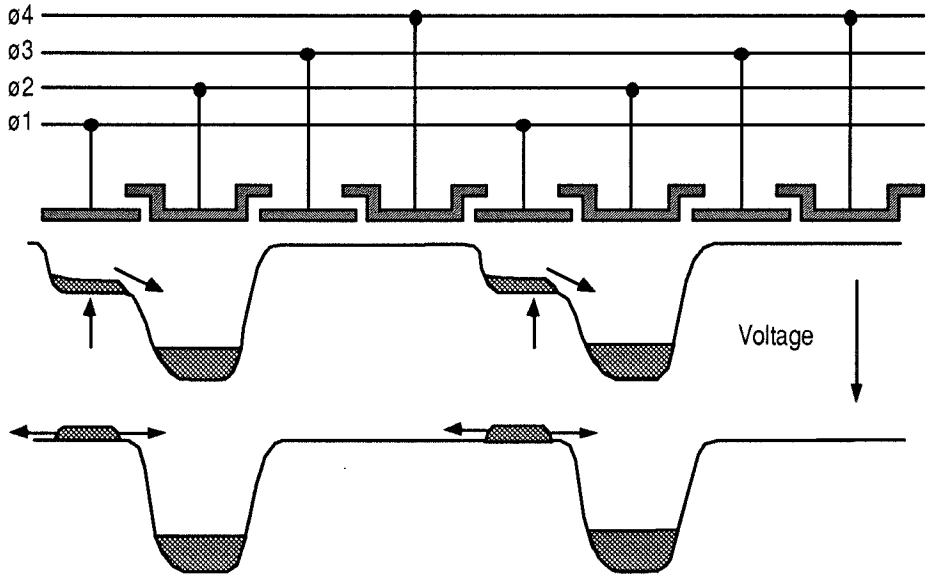

Figure 2.3. Incomplete charge transfer caused by rapid clock voltage fall times. If clock voltages fall too quickly, charge remaining in the collapsing well will not have time to transfer completely, causing the remaining charge to diffuse along the channel and corrupt the stored information.

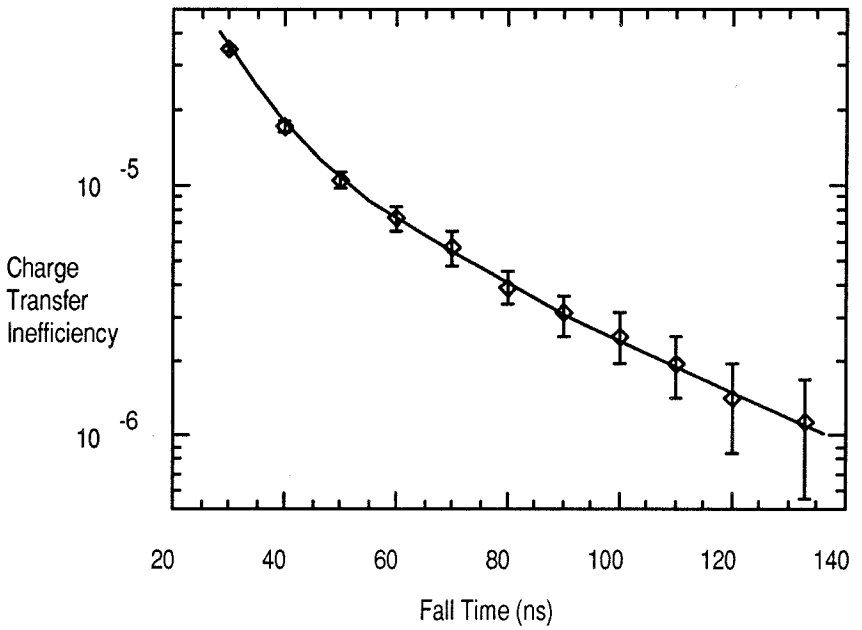

Figure 2.4. Experimental charge transfer inefficiency as a function of fall time. The test was performed on a 724 stage buried channel shift register with 8um channel length per phase at 1MHz clock frequency. The fall time of only one clock phase was varied -- the other three phases were held at a constant rise/fall time of 200ns.

Another less controllable form of charge loss is caused by electron traps, which are very dense at the substrate/oxide interface. If charge is introduced into a previously empty CCD shift register, the traps under each gate will capture a fraction of the electrons from the first charge packet. These electrons are randomly released from the traps at various later times with time constants ranging out to milliseconds. The charge retention behavior of the traps tends to smear out information between charge packets after they have been shifted many times. By encoding the zero signal level at some finite amount of charge [Strain, 1972] (a technique called 'fat zero') some of the traps are kept constantly filled and the overall CTE can be improved.

Errors also arise from thermal generation of charge in the substrate which causes potential wells to slowly fill. The quality of the substrate and fabrication process heavily influences this effect. In CCD cameras, a non-illuminated CCD will integrate this thermal charge and output an image of the thermal generation current, which is commonly called 'dark current'. Due to the fabrication dependent nature of the generation sites, the patterns of dark current seen on wafer scale plots are often large swirls, the remnants of a spin-on fabrication step. Locally, over a few pixels distance, the dark current is relatively uniform except where there are crystal dislocation faults, which cause large dark current spikes [Howes et al., 1979]. In signal processing designs, the effect of dark current can be effectively canceled out to first order by implementing differential charge storage.

Charge levels are easily controlled in signal processing CCD circuits to implement the 'fat zero' technique to improve CTE. Clock edge rates are also easily controlled to improve CTE. However, a certain amount of CTE loss always exists and must be considered when designing large signal processing architectures where many transfers are needed.

2.5 Surface Channel Devices

The simple shift register described in section 2.2 is called a surface channel CCD in which the signal charge resides at the substrate/oxide interface. An optimal clock voltage range can be determined from analysis of the device. With respect to low clock voltages, the intent of a low voltage on the gate is to prevent charge from moving along the channel. The charge barrier along the channel obviously cannot be lower than the intrinsic barrier of the substrate, or else charge will be injected into the substrate, so it makes no sense to lower gate voltages below ground. Furthermore, the surface channel CCD gate voltage must remain above zero to prevent any charge at the oxide interface from being injected into the substrate and to prevent holes from accumulating at the interface, as depicted in Figure 2.5. A 'softer' limit for surface channel devices is to require that the gate voltages should always be above V_T, the inversion threshold of the substrate, to ensure that the depletion region boundary remains at least a fixed minimum distance from the interface [Singh et al., 1976]. A more negative gate voltage during one of the clock phases can eject trapped charge into the substrate or accumulate holes which rapidly recombine with trapped signal electrons, causing net charge loss. Transmission line effects, which result in ringing of long clock electrodes, can cause the gate voltage to momentarily dip below ground and must be considered for fast circuits. It is also important to control the size of CCD gates properly. If a large charge packet is moved under a small gate, charge will spill out into adjacent cells or the substrate. For a positive voltage limit, surface channel devices have an upper gate voltage limit which is determined by the breakdown voltage of the oxide, a limit that is rarely reached in practice.

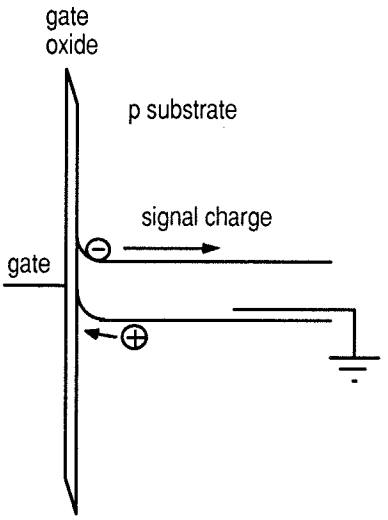

Figure 2.5. Slightly negative clock voltages cause electron injection into the substrate. Holes can also accumulate at the surface, resulting in recombination and charge loss. Fast edge rates and negative clock voltages can result in significant loss of charge to the substrate.

In addition to preventing ringing, slew rate limited clock drivers are essential to proper operation of surface channel CCD circuits. The backflow of charge caused by rapid fall times can significantly impact CTE as discussed in Section 2.3 and in [Singh et al., 1974], especially in devices such as the surface channel CCD where fringing fields are low. The slew limits required for near optimal CTE are the dominant factor in determining the maximum clock rate of the device.

2.6 Buried Channel Devices

A clever fabrication enhancement can significantly improve the CTE and speed of surface channel devices by forcing the charge to travel beneath the surface in a 'buried channel' [Walden et al., 1972][Kim et al., 1972][Esser, 1972]. By introducing a weak n-type implant at the substrate surface beneath the gates, the potential energy bands can be made to curve the opposite direction near the surface, shown in Figure 2.6. The solution to the one-dimensional Poisson equation with the depletion approximation (i.e., uniform space charge) is a parabola with the curvature dependent

on the sign of the space charge. A representative energy band diagram of the structure under operating conditions is shown in Figure 2.6. Note that the potential minimum has been shifted beneath the substrate surface which causes signal charge to travel beneath the surface, effectively isolating the charge packets from the oxide/substrate interface traps. Buried channel CCDs are widely used in commercial imagers since they permit high speed operation and high transfer efficiency. However, buried channel devices often require clock voltages that are outside the normal TTL voltage levels and cannot easily be generated on-chip. The tradeoffs between buried channel and surface channel CCDs are extremely system dependent, since computing architectures can be designed to counter the effects of the poor CTE and low speed of surface channel devices.

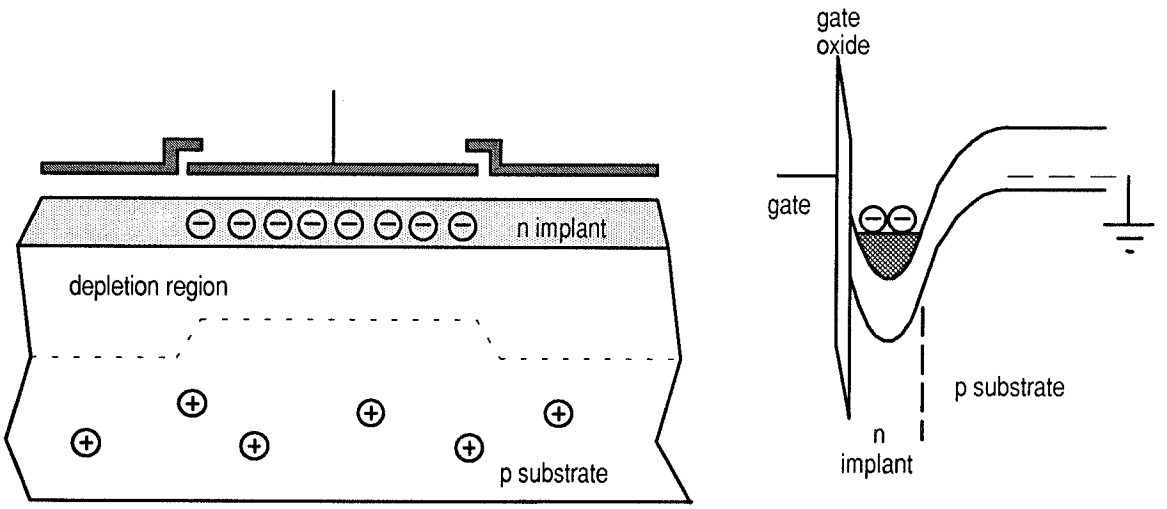

Figure 2.6. The buried channel CCD implant causes the potential minimum to move away from the surface. Charge introduced into the channel resides beneath the surface and encounters fewer traps than surface channel devices which results in improved transfer efficiency. In addition, the higher fringing fields of buried channel devices allow much higher clock frequencies.

Voltage levels required for buried channel devices are somewhat different than those for surface channel CCDs. As shown in Figure 2.7, the introduction of the n-type implant beneath the gates reverses the sign of electric field in the gate oxide. Thus to achieve similar potential minimums, the

buried channel gate voltage must be significantly less than the surface channel gate voltage.

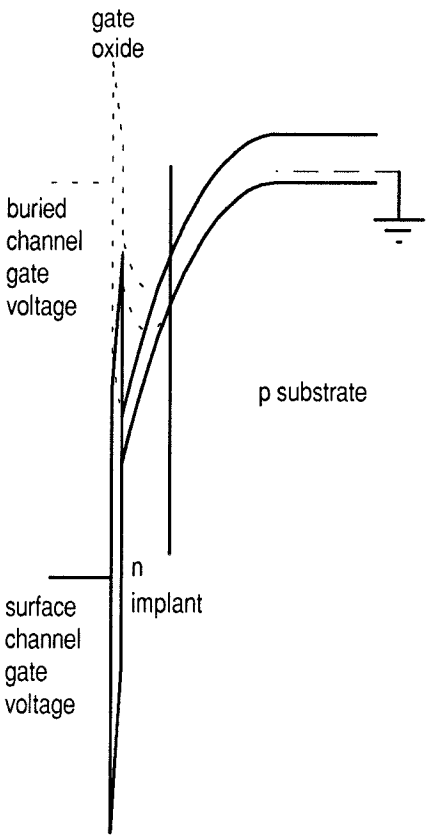

Figure 2.7. Buried channel gate voltages compared to surface channel gate voltages. The n implant reverses the sign of curvature of the bands, resulting in opposite sign field in the oxide. For equivalent depth potential wells, the buried channel voltage is significantly lower than that of the surface channel device.

Buried channel devices have a more complicated set of voltage limits. In practice, the buried channel n-type implant is positively biased through taps at the ends of the CCD channel so that the n-implant is completely depleted. The potential minimum in the semiconductor with no charge can be calculated using the depletion approximation. The equations presented here follow [Beynon et al., 1980]. Assuming a gate voltage V_{gate} and a uniform n-type implant of N_D density and X_D depth with a substrate doping of N_A, the voltage is given by the solution of the one-dimensional Poisson Equation

$$\frac{d^2V}{dx^2} = -\frac{q(N_+ - N_-)}{\varepsilon_{si}} \tag{2.4}$$

which, when integrated twice results in the general solution

$$V = -\frac{q(N_+ - N_-)x^2}{2\varepsilon_{si}} + C_1 x + C_0 \tag{2.5}$$

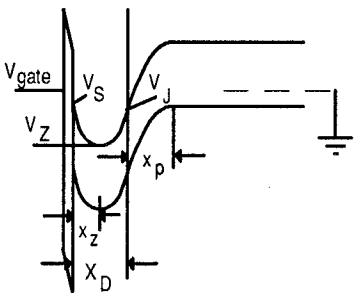

Figure 2.8. Buried channel energy band diagram. The potential minimum of the channel, V_Z, is located at a depth X_Z from the substrate surface.

When there is no charge in the channel, the buried channel n-implant is completely depleted and its maximum voltage can be calculated as a function of the gate voltage. Using the distances and voltages labeled in Figure 2.8 and starting from the right (i.e., the substrate), the potential at the metallurgical junction, V_J, can be written as

$$V_J = \frac{qN_A X_p^2}{2\varepsilon_{Si}}. \tag{2.6}$$

The electric field between the n-implant and p-substrate must be continuous, which requires

$$\frac{qN_A X_p}{\varepsilon_{Si}} = E_J = \frac{qN_D(X_D - X_Z)}{\varepsilon_{Si}}. \tag{2.7}$$

The potential minimum of the channel (i.e., maximum voltage, V_Z) is given by

$$V_Z = \frac{qN_A X_p^2}{2\varepsilon_{Si}} + \frac{qN_D(X_D - X_Z)^2}{2\varepsilon_{Si}}. \tag{2.8}$$

Continuing left, the voltage at the oxide/substrate interface is given by

$$V_S = V_Z - \frac{qN_D X_Z^2}{2\varepsilon_{Si}}. \tag{2.9}$$

Setting the displacement fields equal across the interface results in

$$qN_D X_Z = D_S = \frac{(V_S - V_{Gate})}{t_{ox}} \varepsilon_{SiO_2}. \tag{2.10}$$

These equations can be manipulated to solve for the maximum channel voltage, V_Z [Séquin et al., 1975][Beynon et al., 1980].

$$V_Z = \frac{q\varepsilon_{Si} N_A (N_A + N_D)}{2N_D}\left[\sqrt{\beta^2 + \frac{2}{N_A \varepsilon_{Si}}\left(\frac{V_{Gate}}{q} + N_D X_D\left(\beta - \frac{X_D}{2\varepsilon_{Si}}\right)\right)} - \beta\right]^2 \tag{2.11}$$

where $\beta = \frac{t_{ox}}{\varepsilon_{SiO_2}} + \frac{X_D}{\varepsilon_{Si}}$.

Channel voltage thus depends nonlinearly on the gate voltage, the non linearity being most severe when the gate voltage is at its most negative. This function is plotted in Figure 2.9 along with a straight-line reference for the fabrication process listed in Table 2.1. The lower limit of channel voltage is the 'built-in' potential of the channel-substrate diode, V_{bi}.

N_A	$1x10^{15}$
N_D	$3.5x10^{16}$
t_{ox}	450Å
ε_{Si}	$9.74x10^{-13}$
ε_{SiO_2}	$2.66x10^{-13}$
V_{bi}	0.67V
X_D	0.3um

Table 2.1. Process parameters of the 2um buried channel CMOS process used in this thesis. The effective uniform channel doping N_D is calculated from measurements of the depletion mode FET formed by the channel.

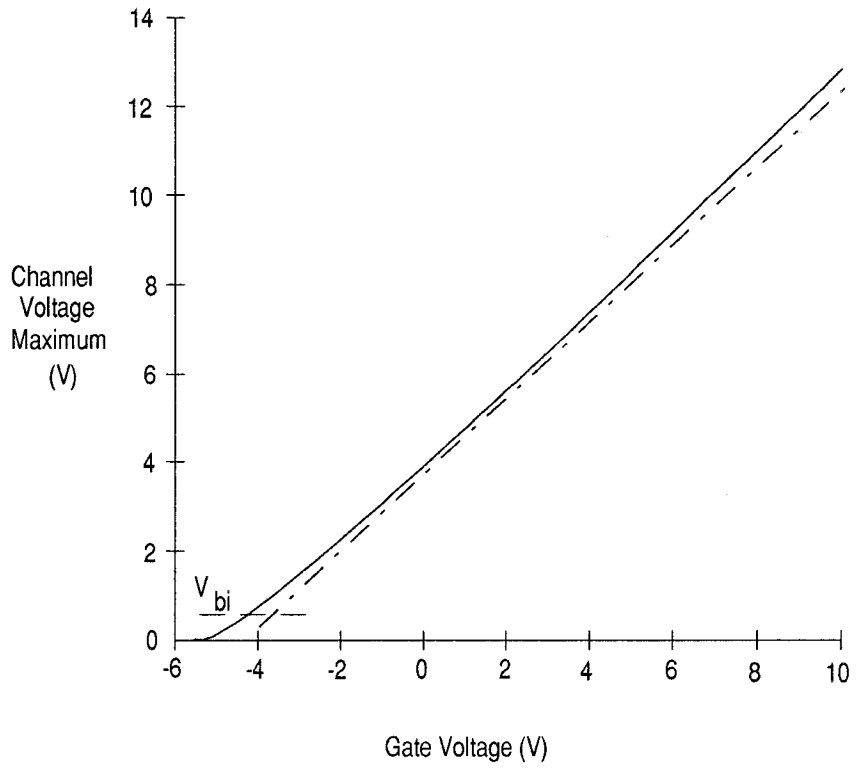

Figure 2.9. Plot of Equation 2.11 showing the nonlinearity of the channel potential as a function of gate voltage along with a straight-line reference. The channel voltage cannot go below V_{bi}, since this would forward bias the implant/substrate diode and cause hole accumulation.

To determine the operating conditions for CCDs, the analysis found in [Séquin et al.,

1975][Beynon et al., 1980] is extended here. The voltage limits determined in what follows are not generally known and are of practical importance. The minimum gate voltage for a buried channel device occurs when the channel voltage maximum, V_Z, equals the built in voltage, V_{bi}, of the implant-substrate diode, which is given by [Sze, 1985]

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right). \tag{2.12}$$

The gate voltage with $V_Z = V_{bi}$ as shown in Figure 2.10 is given by

$$V_{gate}^{(min)} = V_{bi} - \frac{qN_D}{2\varepsilon_{Si}} X_z^2 - \frac{qN_D t_{ox}}{\varepsilon_{SiO_2}} X_z, \tag{2.13}$$

where X_Z is the potential minimum distance from the surface for the potential minimum of V_{bi}

$$X_z = X_D - \sqrt{\frac{2\varepsilon_{Si} V_{bi}}{qN_D\left(1+\frac{N_D}{N_A}\right)}} \tag{2.14}$$

which for the fabrication process listed in Table 2.1 results in

$$V_{gate}^{(min)} \approx -4\text{v}. \tag{2.15}$$

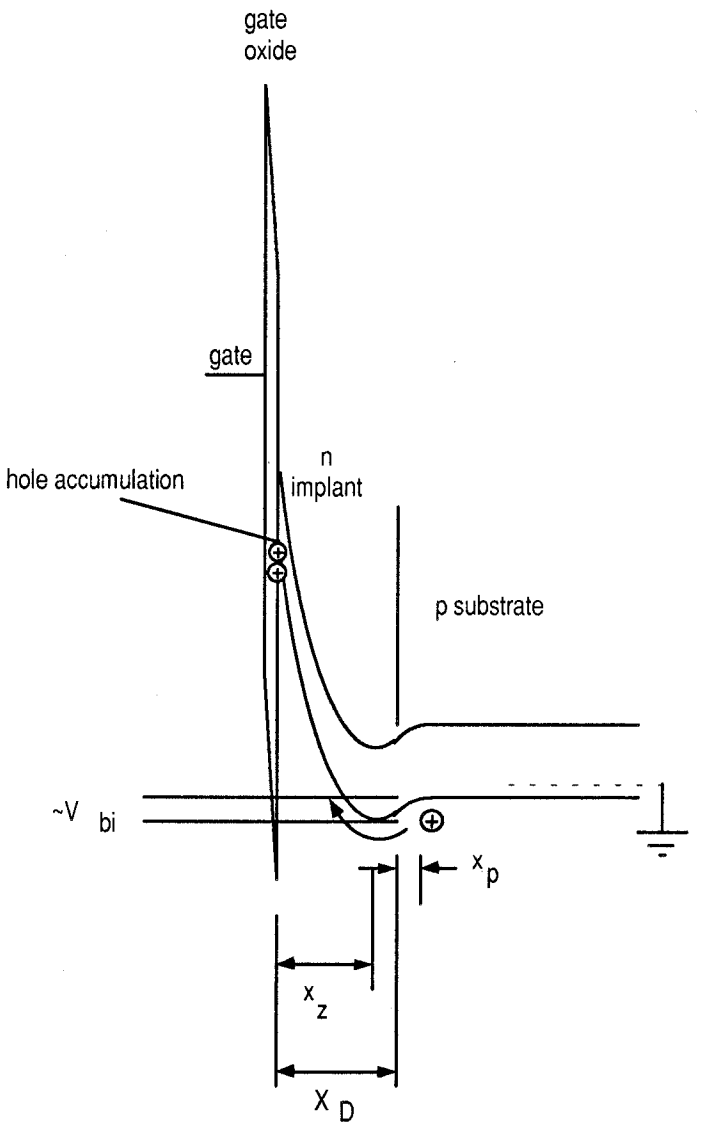

Figure 2.10. Past the minimum gate voltage, holes can jump the barrier and accumulate at the surface. Electrons in the well are injected into the substrate. More negative voltages are completely compensated by hole accumulation at the surface and only result in increased electric field across the oxide.

For more negative gate voltages, any charge left under the gate will be pushed into the substrate via the forward biased diode and holes are allowed to accumulate at the surface, as shown in Figure 2.10. As with the surface channel CCD, the minimum gate potential is intended to act as a barrier to the flow of charge along the channel. The minimum gate voltage described by Equation 2.13 presents the same barrier as the substrate, hence lower gate voltages are not beneficial. A two-

dimensional Poisson simulation of the process of Table 2.1 is shown in Figure 2.11 with two gates, one at a positive voltage and the other at the voltage given by Equation 2.15. Further reducing the gate voltage would cause holes to accumulate at the interface, as shown in Figure 2.10, which would compensate for any additional negative gate voltage changes. Thus lower gate voltages would not modify the potential well diagram of Figure 2.11 and would only cause hole accumulation at the surface, which can adversely impact performance.

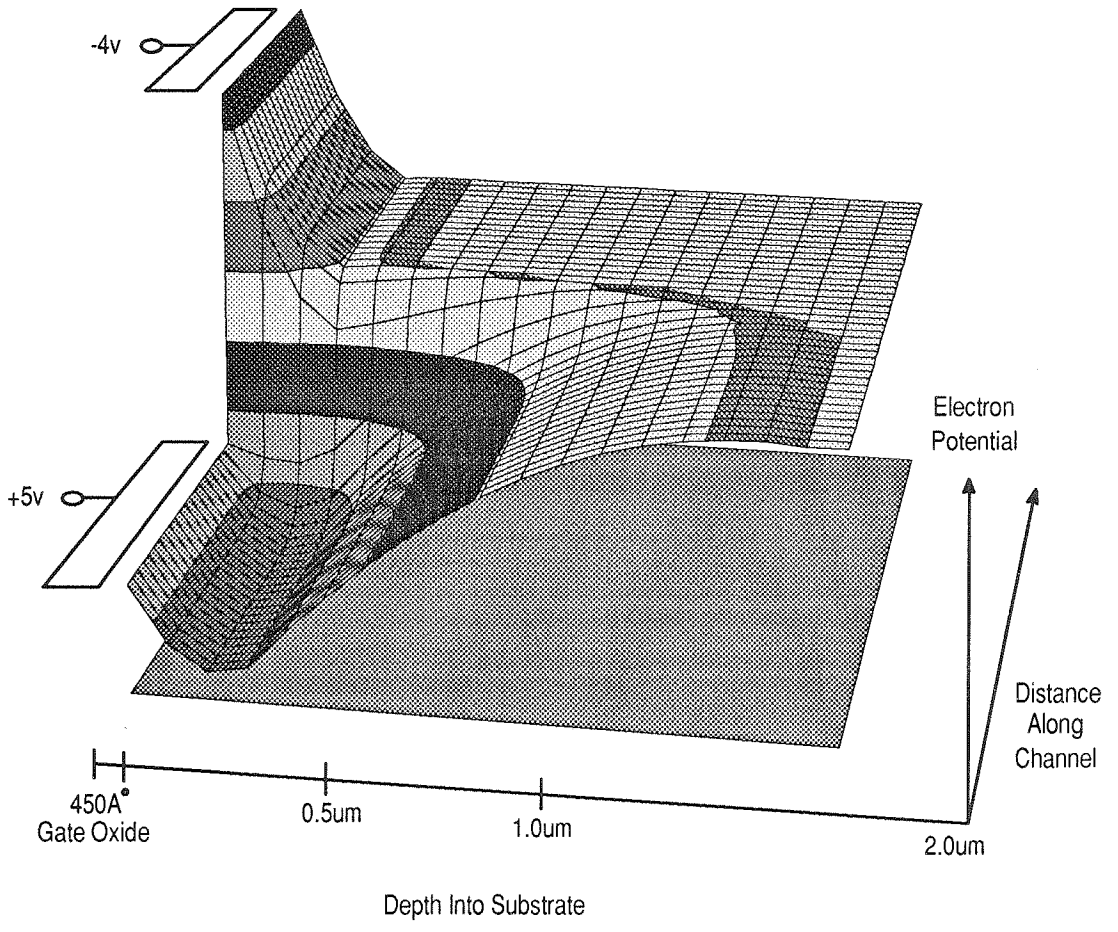

Figure 2.11. Simulated electron potential well diagram of the process listed in Table 2.1 with +5 and -4v gates and no charge. The -4v gate should not be lowered further since hole accumulation at the surface would compensate for any additional negative gate voltage excursion.

This negative gate voltage limit is not found in the literature and has added to the misconception

that CCDs are unpredictable and that finding proper operating voltages is a special art.

In order to elucidate the maximum allowable clock voltages, the behavior of the device with a charge packet needs to be analyzed. When charge is introduced into a buried channel device, a portion of the n implant returns to its undepleted neutral state. Electrons in buried channel devices are majority carriers and have a charge density equal to the doping density of the n-implant. The charge packet has a physical volume and depth from the surface determined by the doping profile of the implant, the gate voltage and the number of electrons in the charge packet. This is in sharp contrast to surface channel devices in which electrons are minority carriers and aggregate in a thin sheet of charge at the oxide/substrate interface with much higher charge density. A diagram showing a buried channel with some charge is shown in Figure 12. The neutral region (i.e., flat band) in the middle of the implant is the signal charge. Its width, X_C, determines the amount of charge in the packet, namely

$$Q = qX_C AN_D \tag{2.16}$$

where A is the area of the gate. The potential of the channel holding charge Q and with gate voltage V_{gate} can be easily calculated using a similar procedure to that used above and the nomenclature of Figure 2.11.

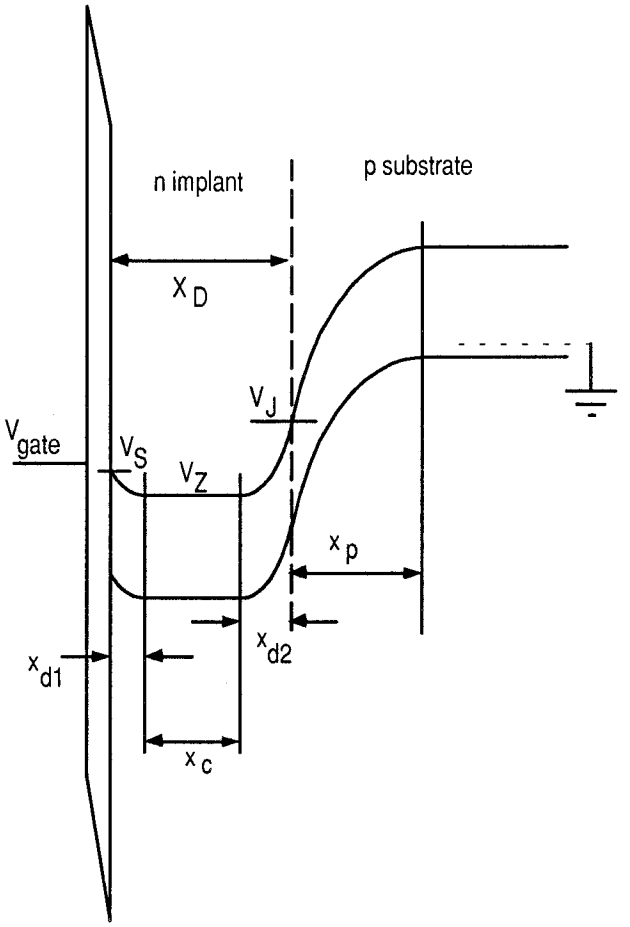

Figure 2.12. Buried channel band diagram with charge. The charge packet forms a neutral region in the n implant that has a charge density equal to that of the doping density. This packet occupies physical volume in contrast to surface channel devices in which charge resides in a thin sheet at the oxide/substrate interface.

The sum of the depleted and neutral region thicknesses in the implant are constrained by

$$X_{d1} + X_{d2} + X_c = X_D. \tag{2.17}$$

The surface voltage can be calculated from these quantities using the depletion approximation, giving

$$V_s = V_J + V_Z - \frac{qN_D X_{d1}^2}{2\varepsilon_{Si}} = \frac{q}{2\varepsilon_{si}}\left(N_A X_p^2 + N_D\left(X_{d2}^2 - X_{d1}^2\right)\right). \tag{2.18}$$

Using the electric field at the substrate-oxide interface, we can calculate the gate voltage given the

thickness of the oxide and its permittivity as before

$$V_{gate} = V_s - \frac{t_{ox} q N_D}{\varepsilon_{SiO_2}} X_{d1} \quad (2.19)$$

The fact that the electric field is zero inside the charge packet requires

$$N_D X_{d2} = N_A X_p \quad (2.20)$$

which gives the gate voltage as a function of the signal charge

$$V_{gate} = \frac{qN_D}{2\varepsilon_{si}} \left[\left(1 + \frac{N_D}{N_A} \right) \left(X_D - \frac{Q_{sig}}{qAN_D} - X_{d1} \right)^2 - X_{d2}^2 \right] - \frac{t_{ox} q N_D X_{d1}}{\varepsilon_{SiO_2}} \quad (2.21)$$

Equations 2.16-2.21 parallel those found in one of the basic texts on CCDs [Séquin et al., 1975][Howes et al., 1979][Beynon et al., 1980]. The limitations of upper clock voltages is briefly touched upon in [Gunsagar et al., 1973] and [Howes et al., 1979], but in inadequate detail. In what follows, a detailed analysis of the limitation of upper clock voltages is developed from equations 2.16-2.21. With respect to maximum gate voltages, the common wisdom is that the more positive the gate voltages, the larger the maximum signal charge packet [Esser, 1972][Beynon et al., 1980]. A voltage limit does exist, however, and it is not simply determined by oxide breakdown. To start, a plot of the maximum charge capacity as a function of voltage is shown in Figure 2.13, which has two different phenomena responsible for its shape which are explained below.

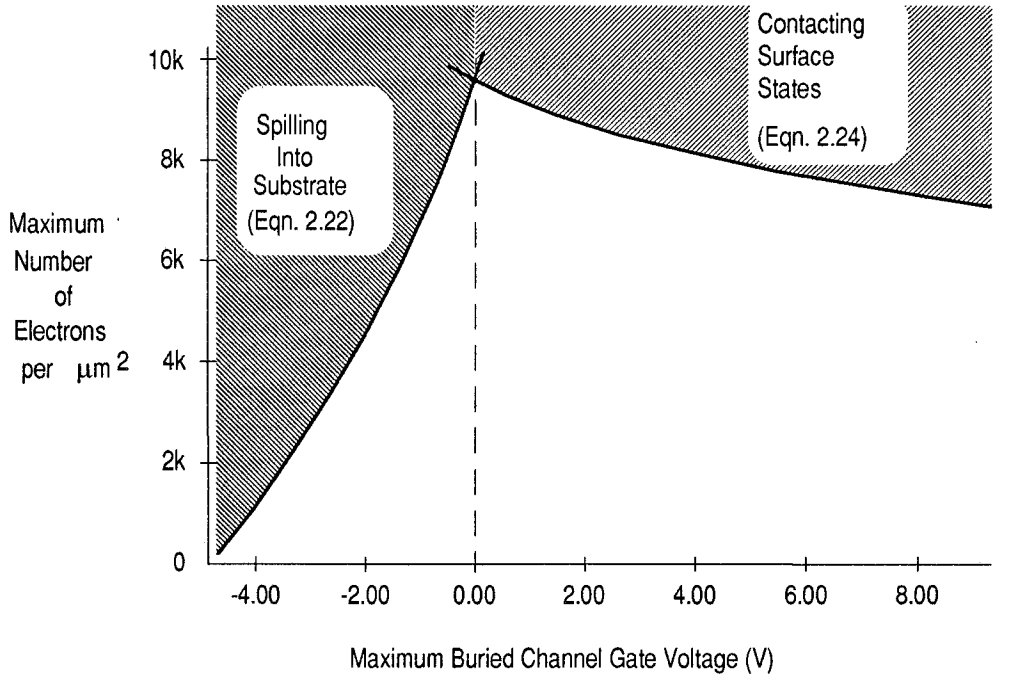

Figure 2.13. Theoretical maximum charge packet size as a function of positive gate voltage swing. The gate voltage is with respect to the flat band gate voltage of Figure 2.1(a). For gate voltages less than the flat band voltage, charge added to the channel will spill into the substrate. For higher gate voltages, contact with surface states limits the charge packet size.

Starting from the left of Figure 2.13 where the gate voltage does not swing very high, the channel charge capacity is limited by charge spilling into the substrate. This occurs when the channel potential with charge Q reaches the built in diode voltage, V_{bi}. Adding more charge forward biases the substrate diode and leads to charge loss. This limit restricts the channel charge capacity. Given that the gate voltage is defined as zero in the flat band condition of Figure 2.1(a) and using Equations 2.16-2.21, the maximum charge capacity as a function of gate voltage is derived

$$Q_{max} = AN_D\left(X_D - X_{bi} - \frac{t_{ox}\varepsilon_{Si}}{\varepsilon_{SiO_2}}\left(\sqrt{1-\frac{2\varepsilon_{SiO_2}^2 V_{gate}}{qN_D\varepsilon_{Si}t_{ox}^2}}-1\right)\right) \qquad V_{gate} < 0 \qquad (2.22)$$

where the built-in depletion width to the left of the metallurgical junction, X_{bi}, is given by

$$X_{bi} = \sqrt{\frac{2\varepsilon_{Si} V_{bi}}{qN_D\left(1+\frac{N_D}{N_A}\right)}}. \tag{2.23}$$

Note that when the gate voltage is at zero, the channel is completely neutral except for the intrinsic diode depletion. Greater gate voltages invoke the next limiting situation which occurs when signal charge introduced into the buried channel comes in contact with the interface traps, which causes a loss of transfer efficiency. As shown in Figure 2.14, for larger amounts of signal charge, the boundary of the upper depletion region moves closer to the surface. The maximum gate potential is directly related to the maximum charge capacity if the interface traps are to be avoided. From Equations 2.16-2.21 the maximum charge packet that avoids interface traps is given by

$$Q_{max} = AN_D\left(X_D - \sqrt{\frac{2\varepsilon_{Si}\left(V_{gate}+V_{bi}\right)}{qN_D\left(1+\frac{N_D}{N_A}\right)}}\right) \quad , \quad V_{gate} > 0 \ . \tag{2.24}$$

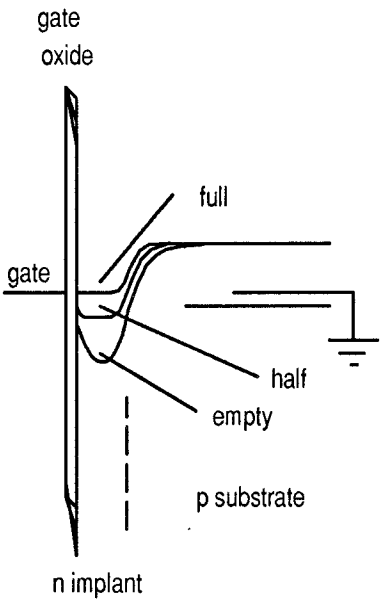

Figure 2.14. Buried channel potential diagram. As the channel fills up, the charge can come in contact with surface states which degrade CTE to that of a surface channel device.

To see the effects on charge transfer efficiency, an experiment was performed on a buried channel shift register. A near maximum size charge packet with near optimal CTE was introduced into the shift register and the maximum voltage of the four-phase clock was varied. The result, shown in Figure 2.15 demonstrates the charge packet coming in contact with interface states.

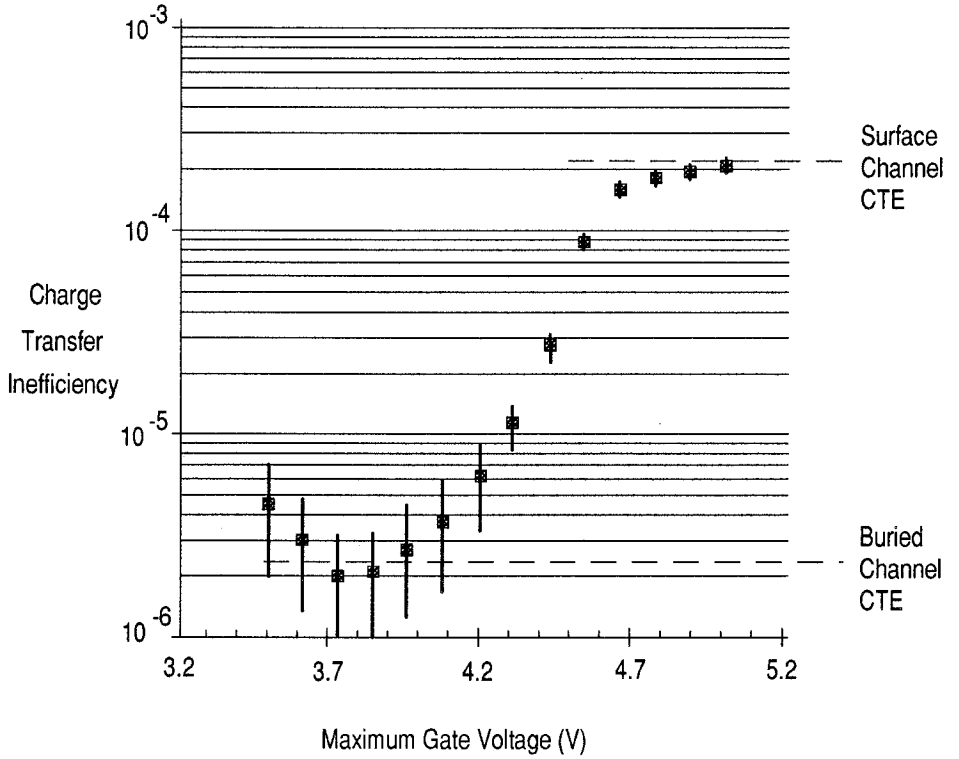

Figure 2.15. Experimental charge transfer loss as a function of maximum gate voltage showing effect of surface states on buried channel performance. A constant size charge packet was input into a buried channel CCD shift register while the maximum voltages of the four central clock lines were varied. Above ~4.2v for this particular size charge packet, the charge comes in contact with surface states which degrade the performance to that of a surface channel device.

As can be seen, higher gate voltages reduce the charge handling capability of the buried channel CCD, a result which is not found in the literature. However, higher gate voltages also increase the fringing fields in Equation 2.2 and hence speed up operation. For imagers where charge packet size is rarely near maximum, the upper gate voltage limit is not restrictive in practice. In signal processing devices, however, the charge packets are of user controllable size and these limits must be understood.

The restricted voltage swing of the buried channel devices typically requires clock drivers which are able to pulse gate voltages lower than the substrate voltage. Unfortunately, this often prevents the inclusion of on-chip clock generation due to the nature of the fabrication process.

2.7 Charge Input

A number of structures exist for the creation of a charge packet based on a voltage level. One circuit in particular will be examined here as it has the best noise performance and is used almost exclusively in the field for CCD charge input. The circuit consists of an input diode and two gates which set the size of the charge packet, as shown in Figure 2.16, and is called a *potential equilibration* circuit or more commonly a 'fill-and-spill' circuit [Carnes et al., 1973][Tompsett et al., 1973][Tompsett, 1975]. The operation of the device occurs in two phases. In the fill phase, the diode supplies electrons up to a certain potential level and fills the structure above the potentials of the DC and Signal gates, as in Figure 2.16(a). The second phase occurs when the diode potential is dropped and charge spills back out of the region, shown in Figure 2.16(b). The charge left underneath the signal gate is proportional to the channel potential level difference between the Signal and DC gates. Once the charge packet is created, it can be shifted to the right using a CCD shift register.

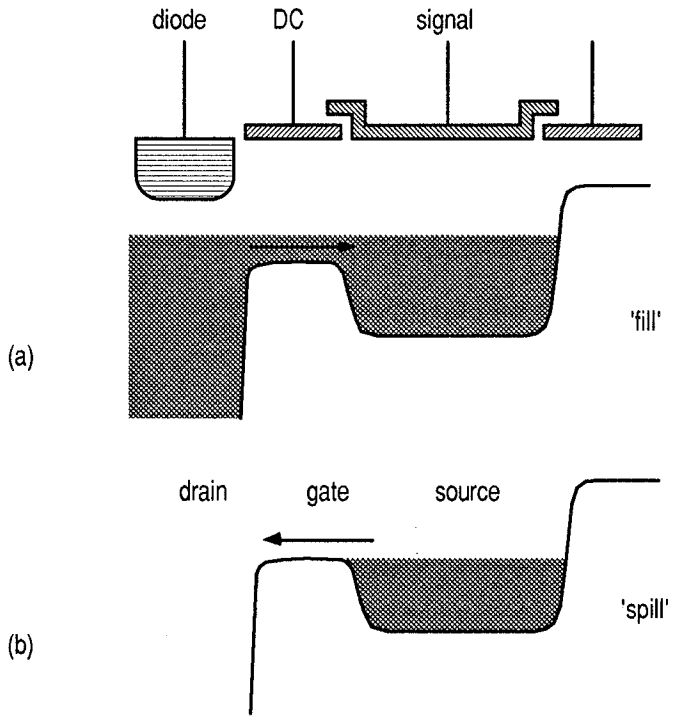

Figure 2.16. Potential equilibration charge input structure. During the 'fill' phase, the input diode is pulse to a low voltage to inject electrons into the channel. The diode is then returned to a high voltage, causing excess electrons to flow out of the channel, leaving a charge packet proportional to the difference of the DC and signal gates. The operation can also be thought of in terms of an n-channel MOSFET device where the drain is the input diode, the DC CCD gate acts as the gate and the substrate beneath the signal gate acts as the source. The 'fill' phase is an 'on' transistor with mobile charge in the channel. When the diode (drain) is pulled high, the substrate beneath the signal gate (source) goes as high as the gate will let it until the transistor shuts 'off' due to lack of gate-source voltage.

This device exhibits the same noise qualities of the switched capacitor storage device mentioned in Chapter 1 and is fully examined in [Mohsen et al., 1975]. The DC gate acts as the FET and the Signal gate as the storage capacitor. The noise analysis of the fill-and-spill circuit is the same as that for a Johnson noise calculation of a RC circuit. When the circuit is in the spill phase, the finite channel resistance underneath the DC gate exhibits Johnson noise which is given by

$$\overline{V}_r^2 df = 4kTR_{fet}df \tag{2.25}$$

where $\overline{V}_r^2$ is the mean squared thermal voltage of the equivalent resistor in the bandwidth *df*. The equivalent circuit is shown in Figure 2.17.

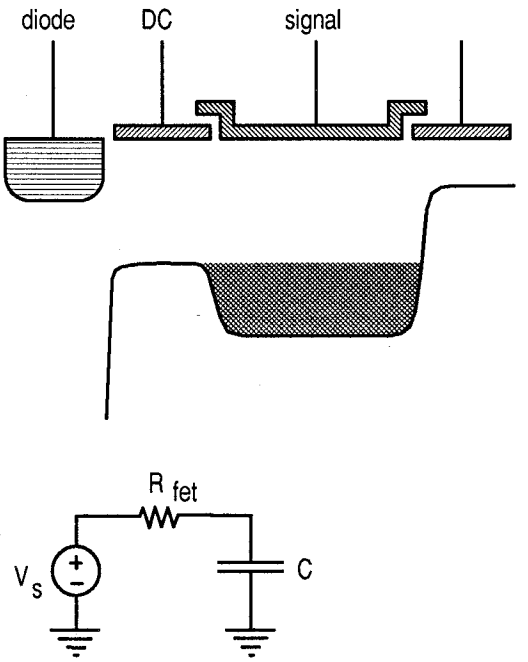

Figure 2.17. Equivalent circuit of the 'fill and spill' circuit for noise analysis. The finite 'on' resistance of the FET channel underneath the DC gate contributes a Johnson noise component to the forming charge packet.

Passed through the low-pass filter of Figure 2.17 results in a noise voltage on the capacitor in band *df* of

$$\overline{V}_c^2 df = \frac{4kTR_{fet}}{1+\omega^2 C^2 R_{fet}^2} df \tag{2.26}$$

The integral of the response of the RC circuit over all frequencies results in the well known noise voltage

$$\overline{V}_n^2 = \int_0^\infty \overline{V}_c^2 df = \frac{kT}{C} \tag{2.27}$$

Note that the resistance of the channel does not influence the noise of the circuit. Another source

of noise is the phenomena of charge splitting, where charge left under the DC gate is split, with the noise level determined by how much remains under the Signal gate.

The linearity of the fill-and-spill circuit, i.e., voltage to charge linearity, can be calculated to first order by finding the Q_{sig} that results from a given voltage difference on the two gates of the circuit of Figure 2.16. For surface channel devices, the charge at the interface can be modeled as a thin sheet of charge, Q_{sig}. The amount of charge left behind in the spill phase adjusts the surface potential under the signal gate until it equals that of the DC gate. For different signal voltages, the surface voltage remains the same -- only the amount of charge changes. The charge is thus linearly proportional to the signal gate voltage

$$Q_{sig} = \frac{A\varepsilon_{SiO_2}}{t_{ox}}\left(V_{gate}^{sig} - V_{gate}^{dc}\right), \tag{2.28}$$

the proportionality constant being the characteristic capacitance of the oxide. Note that the nonlinearities of the MOS capacitor are not present, simply due to the fact that the depletion regions under the DC and Signal gates are the same depth after the spill phase regardless of signal charge size. A plot is given in Figure 2.18 using the process parameters listed in Table 2.1. Nonlinearities occur in practice due to dynamic fluctuations and channel widening in the two dimensions not covered in this analysis which limit the performance to approximately 50-60dB linearity [Mohsen et al., 1975].

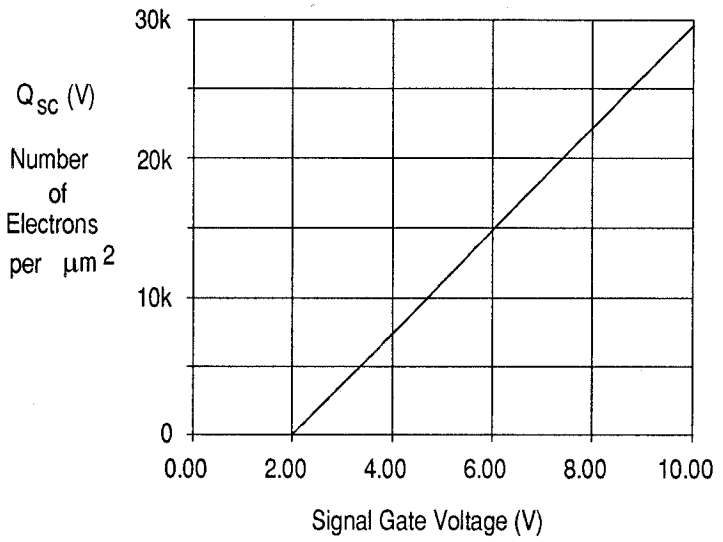

Figure 2.18. Simulated linear charge generation with the surface channel fill-and-spill circuit. The DC gate is held at 2V. The proportionality between charge and voltage is simply the oxide capacitance ($C_{ox} = t_{ox} / \varepsilon_{SiO_2}$). Second-order effects such as dynamic fluctuations and channel widening limit linearity to < 60dB in practice.

For the buried channel device, the calculations require more algebra and yield

$$Q_{sig} = qAN_D\left[X_D - \sqrt{\frac{2\varepsilon_{Si}V_Z^{dc}}{qN_D\left(1+\frac{N_D}{N_A}\right)}} + \frac{t_{ox}\varepsilon_{Si}}{\varepsilon_{SiO_2}} - \varepsilon_{Si}\sqrt{\left(\frac{t_{ox}}{\varepsilon_{SiO_2}}\right)^2 - \frac{2\left(V_{gate}^{sig} - V_Z^{dc}\right)}{qN_D\varepsilon_{Si}}}\right] \tag{2.29}$$

where V_Z^{dc} is the potential minimum of Equation 2.11 for the DC gate. A family of curves is plotted in Figure 2.19 for the process listed in Table 2.1. Due to the inherent linearity of surface channel devices, they are preferentially used for charge input. It is common in signal processing devices to use surface channel input structures with buried channel transfers [Wen, 1976]. The conversion functions shown in Figures 2.18 and 2.19 are labeled $Q_{sc}(V)$ and $Q_{bc}(V)$ for surface channel and buried channel, respectively, for use in later chapters.

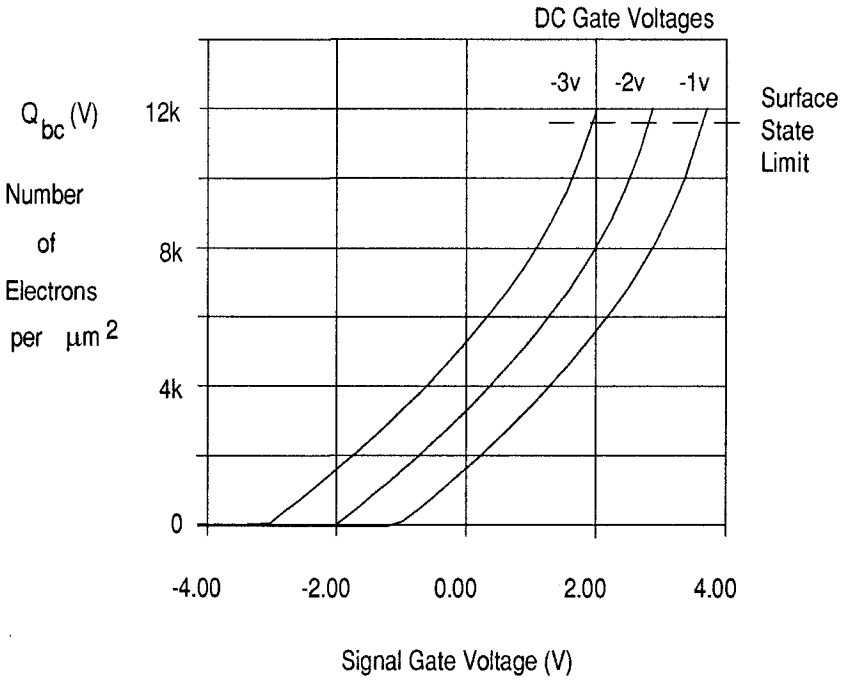

Figure 2.19. Simulated charge generation linearity for a buried channel fill-and-spill circuit is shown for various DC gate voltages for the process of Table 2.1. The nonlinearity is due to the depletion depth changes which significantly modify the total capacitance seen by the charge.

2.8 Charge Output

CCD output circuits convert a charge packet to a voltage, usually by transferring the charge packet of interest onto a capacitor, which experiences a voltage change. This capacitor is typically either a floating diffusion or a floating gate [Engeler et al., 1970][Séquin et al., 1975]. Diffusion sensing destroys the charge packet being sensed, making it less attractive for signal processing where utilization of the same charge for many computations is required. For the purposes of this thesis, floating gate sensing is examined exclusively as it provides the nondestructive sensing required for the signal processing devices discussed in later chapters. The equivalent circuit of Figure 2.20 is used to analyze the charge-to-voltage conversion.

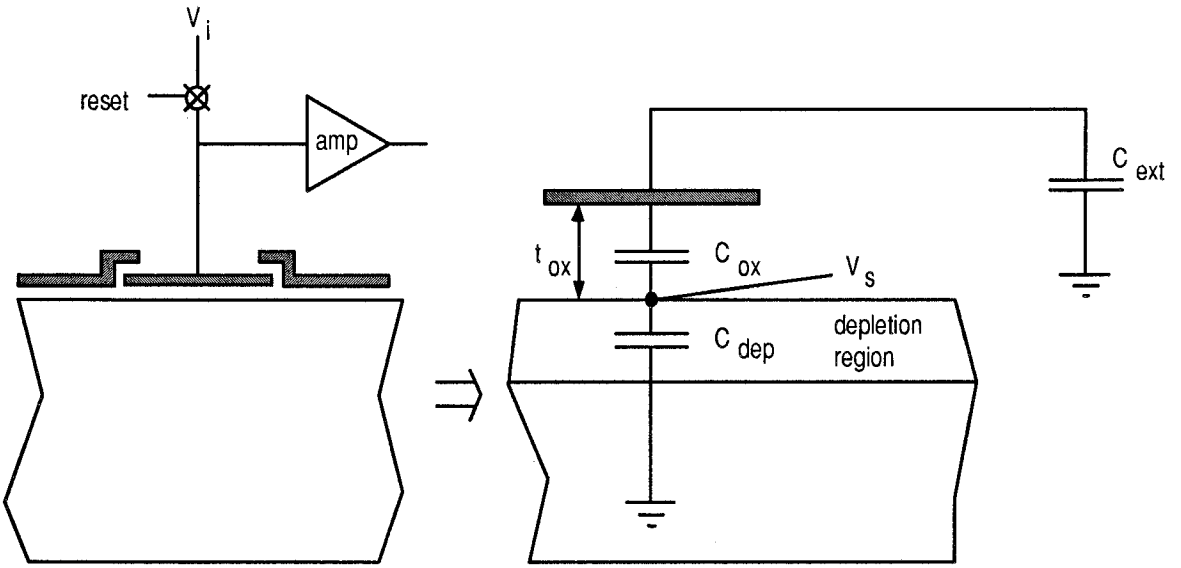

Figure 2.20. In the floating gate charge sensing output circuit the sense gate is reset to a known voltage then the charge is moved beneath it. The charge sees a capacitance shown in the model to the right and produces a surface voltage change. The surface potential change is seen at the output through the capacitive divider formed by the oxide and external capacitances.

In operation, the sense gate is reset to a particular voltage, V_i, with no charge beneath it and then allowed to float. When the charge is transferred underneath the sense electrode by the action of the neighboring gates, the sense electrode experiences a voltage change. C_{ext} is the parasitic capacitance of the sensing gate which includes all sidewall and amplifier input capacitance. Because the depletion boundary changes with the amount of charge present, the capacitance to the substrate introduces a nonlinear component into the charge-to-voltage conversion. In the surface channel device of Figure 2.20, the capacitance the signal charge sees [Séquin et al., 1975] is given by

$$C_{tot}(V_S) = \frac{C_{ox}C_{ext}}{C_{ox}+C_{ext}} + C_{dep}(V_S) \tag{2.30}$$

The depletion layer capacitance depends on the surface potential,

$$V_S = \frac{qN_A}{2\varepsilon_{Si}} X_p^2 \tag{2.31}$$

which inserted into

$$C_{dep}(V_S) = \frac{A\varepsilon_{Si}}{X_p} \tag{2.32}$$

results in

$$C_{dep}(V_S) = A\sqrt{\frac{qN_A\varepsilon_{Si}}{2V_S}} \tag{2.33}$$

Adding *dQ* charge into the channel causes a voltage change via the relationship

$$dQ = C_{tot}(V_S)dV_S \tag{2.34}$$

Integrating from the surface potential to the final surface potential

$$Q_{sig} = \int_{V_i}^{V_f} \left(\frac{C_{ox}C_{ext}}{C_{ox}+C_{ext}} + A\sqrt{\frac{qN_A\varepsilon_{Si}}{2V}} \right) dV \tag{2.35}$$

results in the following expression for the change in surface potential when solved for the channel potential change ΔV_S.

$$\Delta V_S = V_f - V_i \cong V_i \left[\frac{Q_{sig}}{\beta V_i + \frac{\alpha\sqrt{V_i}}{2}} + \frac{Q_{sig}^2}{8\alpha^2 V_i \left[\frac{\beta\sqrt{V_i}}{\alpha} + \frac{1}{2} \right]^3} + \ldots \right] \tag{2.36}$$

$$\alpha = A\sqrt{2qN_A\varepsilon_{Si}} \qquad \beta = \frac{C_{ox}C_{ext}}{C_{ox}+C_{ext}}$$

The output voltage that enters the amplifier is related to the channel potential change of Equation 2.36 through the capacitive divider

$$\Delta V_{out} = \Delta V_S \frac{C_{ox}}{C_{ox} + C_{ext}} \quad . \tag{2.37}$$

The relationship between charge and output voltage is graphed in Figure 2.21 for a reset voltage of 5v along with a straight line reference. In later chapters this function for surface channel devices is referred to as $V_{sc}(Q)$.

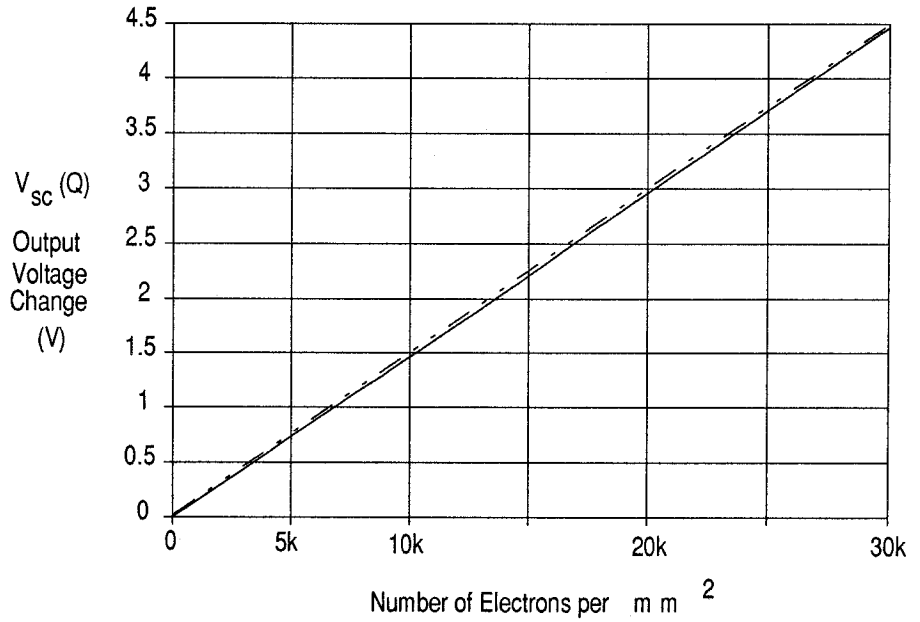

Figure 2.21. Simulated voltage output linearity of surface channel floating gate sense circuit for the process of Table 2.1 which is reset to an initial voltage of 5v. The parasitic capacitance C_{ext} is 100fF. The nonlinear depletion capacitance limits accuracy to ~ 40dB.

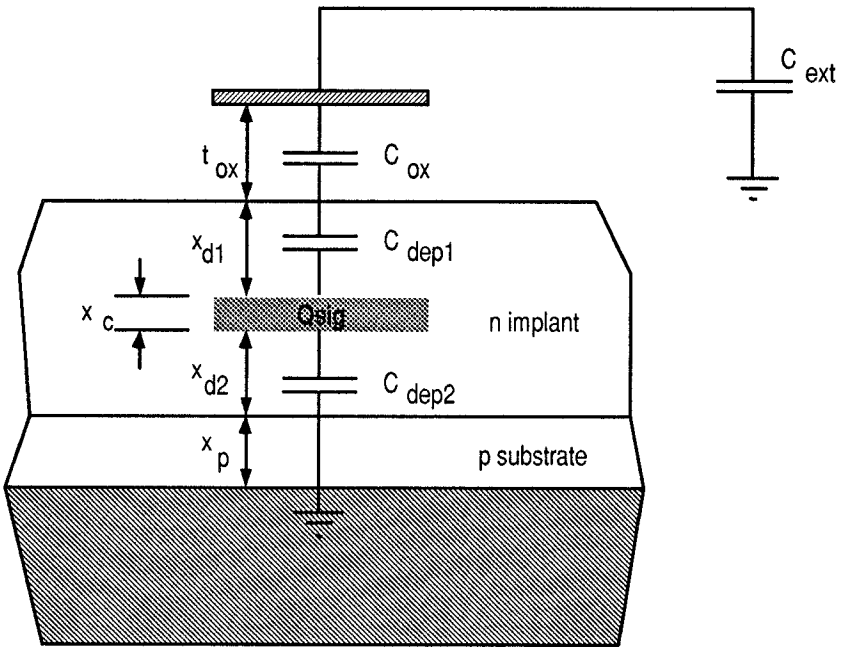

Figure 2.22. Buried channel sensing has to contend with two depletion depths, one above and the other below the charge packet. The additional nonlinear capacitors depend strongly on reset voltage and add significantly more nonlinearity compared to surface channel devices.

For the buried channel device, the analysis follows a similar path. Complications arise because the signal charge is spatially distributed in the channel which leads to additional capacitance variations. The model used in this analysis of the buried channel output circuit is shown in Figure 2.22. Because the potential within the charge packet is uniform, it acts as a single node whose capacitance to ground is given by

$$C_{tot} = \frac{1}{\frac{1}{C_{dep1}} + \frac{1}{C_{ox}} + \frac{1}{C_{ext}}} + C_{dep2} \tag{2.38}$$

The depletion layer thicknesses are given by

$$X_{d2} + X_p = \sqrt{\frac{2}{q}\left(\frac{1}{N_D} + \frac{1}{N_A}\right)\varepsilon_{Si} V_{chan}} \tag{2.39}$$

and

$$X_{d1} = X_D - X_{d2} + \frac{Q_{sig}}{qAN_D} \tag{2.40}$$

so that the capacitance of the signal charge region as a function of channel voltage and charge is

$$C_{tot} = \frac{1}{\frac{1}{C_{ext}} + \frac{t_{ox}}{A\varepsilon_{SiO_2}} + \frac{X_D - \sqrt{\frac{2\varepsilon_{Si}V_{chan}}{qN_D(1+N_D/N_A)}} + \frac{Q_{sig}}{qAN_D}}{A\varepsilon_{Si}}} + A\sqrt{\frac{q\varepsilon_{Si}}{2V_{chan}\left(\frac{1}{N_D}+\frac{1}{N_A}\right)}}. \tag{2.41}$$

The numerical solution to the differential equation governing the charge to voltage relationship

$$\frac{dQ}{dV_{chan}} = C_{tot}(Q, V_{chan}) \tag{2.42}$$

results in the channel potential change, ΔV_{chan} due to the added charge. The voltage seen at the output is through the capacitive divider

$$\Delta V_{out} = \Delta V_{chan} \frac{1}{1 + C_{ext}\left(\frac{1}{C_{depl}} + \frac{1}{C_{ox}}\right)}. \tag{2.43}$$

and is shown in Figure 2.23 for various initial sense gate voltages given the process of Table 2.1. This buried channel charge to voltage function is labeled $V_{bc}(Q)$ for future use. As can be seen,

the linearity of the buried channel is significantly less than that of the surface channel CCD. The cause is simply the increased role of the voltage dependent depletion layer capacitances in C_{tot}. Also, the voltage change for the maximum size charge packet is significantly less than that of the equivalent surface channel device of Figure 2.21.

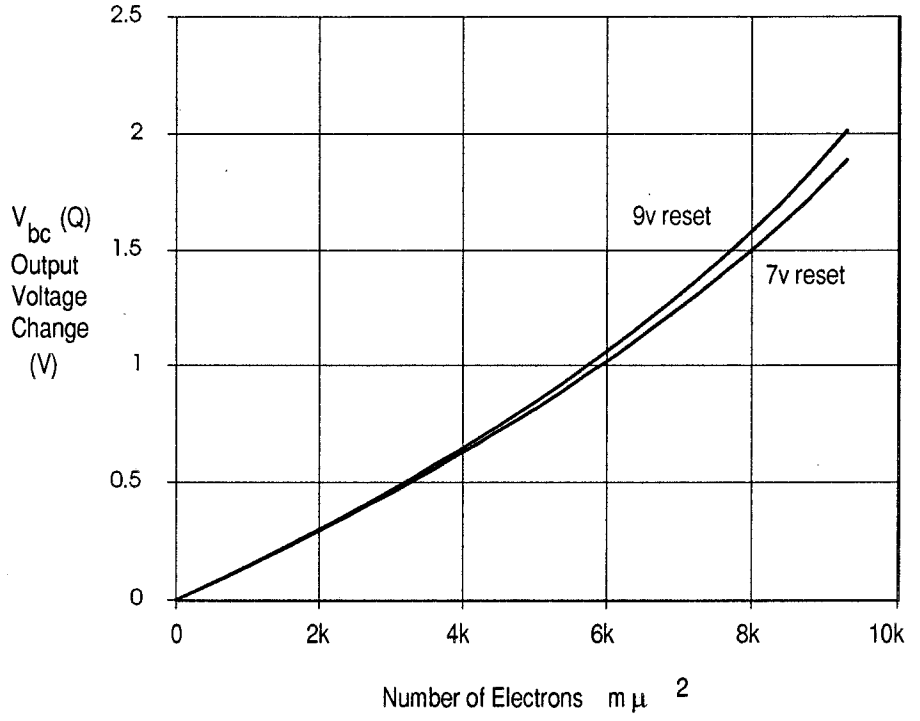

Figure 2.23. A numerical simulation of the buried channel sensor output voltage as a function of charge for two different reset voltages for the process listed in Table 2.1 and an external capacitance of 100fF. The nonlinear capacitance of the channel varies with reset voltage resulting in different Q-V transform curves.

It is important to choose operating conditions, such as reset voltage and limited charge packet sizes, to optimize linearity. In the architectures discussed in the following chapters, the operating points have other constraints and such optimizations are not always possible.

A complete voltage-to-charge and back to voltage conversion adds significant nonlinearities to any

signal processing system. Figure 2.24 shows the simulated and measured input and output curves for a buried channel CCD shift register. By choosing different operating points for the charge input and output circuits, the shape of the response can be modified somewhat. The nonlinearity shown in Figure 2.24 illustrates the points that the buried channel devices are not inherently linear and that the plethora of user defined operating parameters have significant effects on linearity.

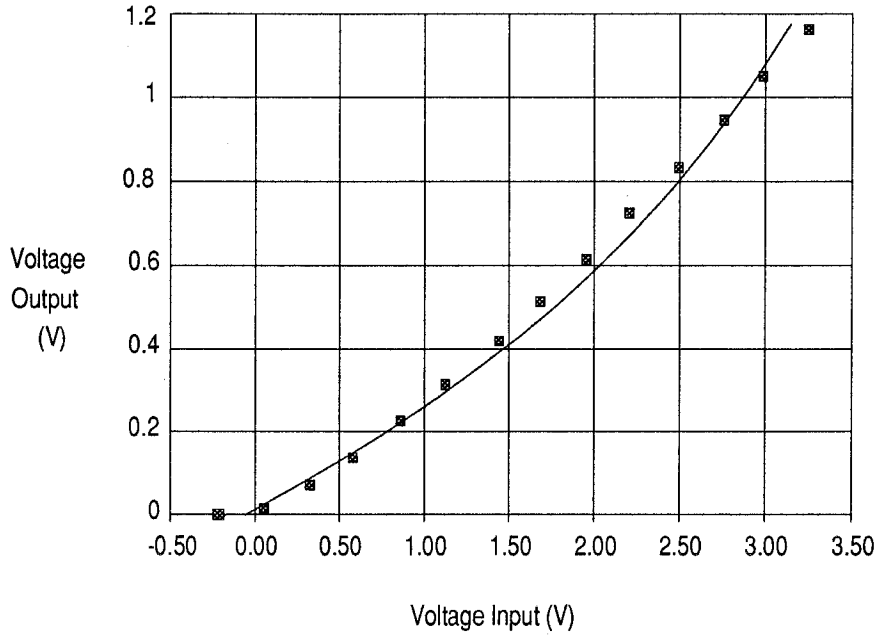

Figure 2.24. Experimental voltage-charge-voltage conversion of a buried channel CCD. Experimental data taken at 1MHz shift rate with an input DC gate of -0.3v, a reset voltage of 9v and an estimated parasitic capacitance of 100fF.

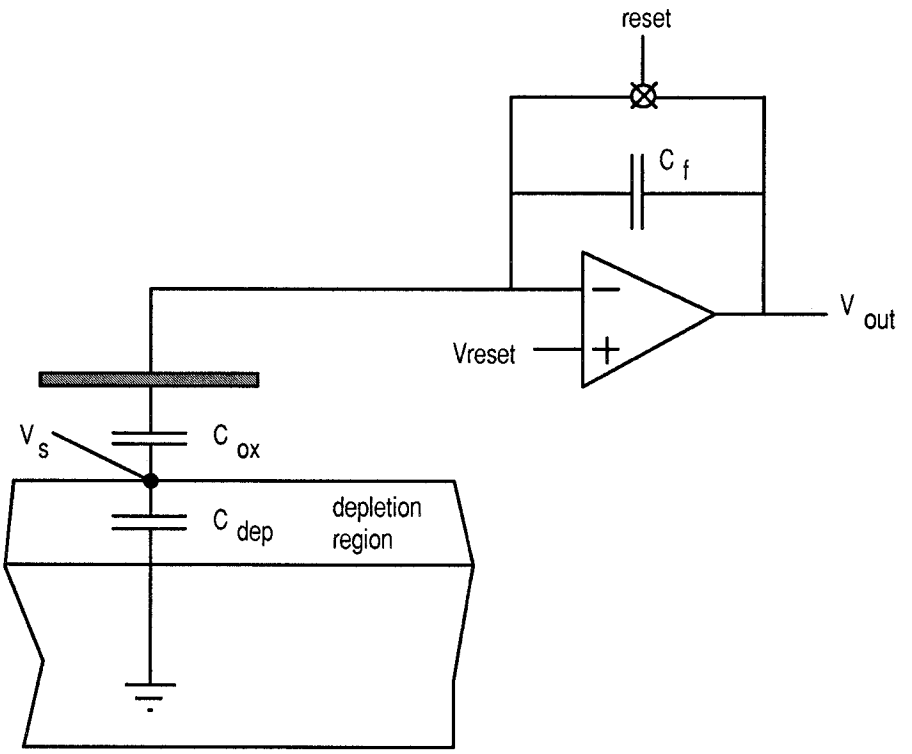

Figure 2.25. Charge feedback sense amplifier. The polysilicon sense gate is held at a virtual ground by the amplifier feedback loop which increases the sensitivity. Additionally, nonlinear parasitic capacitance is canceled and the gain of the sense amp can be specified by selecting the proper C_f independent of polysilicon gate size.

Using a different sensing arrangement like the one shown in Figure 2.25 can increase the sensitivity of the circuit by forcing the gate to remain at a constant potential via the feedback action of the sensing amplifier and feedback capacitor [Beynon et al., 1980][Miida et al., 1991]. The previous analysis can be used by letting $C_{ext} \rightarrow \infty$ and using the following relationship for surface channel devices

$$\Delta V_{out} = \Delta V_S \frac{C_{ox}}{C_f} \tag{2.44}$$

and for buried channel

$$\Delta V_{out} = \Delta V_{chan} \frac{C_{ox} C_{depl}}{C_f \left(C_{ox} + C_{depl} \right)}. \tag{2.45}$$

In contrast to Equations 2.37 and 2.43, these equations can produce voltage gain for small C_f. Linearity is improved since the linear oxide capacitance has effectively been increased by holding the gate at a virtual ground. For future use, the feedback sensing Q-V transforms for surface and buried channel devices are referred to as $V_{fsc}(Q)$ and $V_{fbc}(Q)$ respectively.

The procedure of transforming voltage to charge back to voltage presents a linearity problem that limits the overall accuracy of CCD signal processors. One way to reduce the nonlinearities is to pre-distort the input voltage to create a charge packet which gives linear output voltage response. A feedback charge input system as described in [Maclennan et al., 1975][Hense et al., 1976] can produce good linearity at the expense of slow active elements in the feedback loop. The implementation of such a system is described in Section 4.4.

2.9 References:

[Amelio et al., 1970] G.F. Amelio, M.F. Tompsett and G.E. Smith, "Experimental verification of the charge-coupled device concept," *Bell System Technical Journal*, vol. 49, pp. 593-600, 1970.

[Bakker, 1991] J.G.C. Bakker, "Simple analytical expressions for the fringing field and fringing-field-induced transfer time in charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-38(5), pp. 1152-1161, 1991.

[Berglund et al., 1973] C.N. Berglund and K.K. Thornber, "Incomplete transfer in charge-transfer devices," *IEEE Journal of Solid State Circuits*, vol. SC-8(2), pp. 108-116, 1973.

[Beynon et al., 1980] J.D.E. Beynon and D.R. Lamb, *Charge Coupled Devices and Their Applications*. London: McGraw-Hill, 1980.

[Boyle et al., 1970] W.S. Boyle and G.E. Smith, "Charge-coupled semiconductor devices," *Bell System Technical Journal*, vol. 49, pp. 587-593, 1970.

[Carnes et al., 1971] J.E. Carnes, W.F. Kosonocky and E.G. Ramberg, "Drift aiding fringing fields in charge-coupled devices," *IEEE Journal of Solid State Circuits*, vol. SC-6, pp. 322-326, 1971.

[Carnes et al., 1972] J.E. Carnes, W.F. Kosonocky and E.G. Ramberg, "Free charge transfer in charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-19(6), pp. 798-808, 1972.

[Carnes et al., 1973] J.E Carnes, W.F. Kosonocky and P.A. Levine, "Measurements of noise in charge-coupled devices," *RCA Review*, vol. 34, pp. 553-565, 1973.

[Engeler et al., 1970] W.E. Engeler, J.J. Tiemann and R.D. Baertsch, "Surface charge transport in silicon," *Applied Physics Letters*, vol. 17, pp. 469-472, 1970.

[Esser, 1972] L.J.M. Esser, "Peristaltic charge-coupled device: a new type of charge-transfer device," *Electronics Letters*, vol. 8, pp. 620-621, 1972.

[Gunsagar et al., 1973] K.C. Gunsagar, C.K. Kim and J.D. Phillips, "Performance and operation of buried channel charge-coupled devices," in *Technical Digest of the International Electron Devices Meeting*, p. 21, 1973.

[Hense et al., 1976] K.R. Hense and T.W. Collins, "Linear charge-coupled device signal-processing techniques," *IEEE Transactions on Electron Devices*, vol. ED-23(2), pp. 265-270, 1976.

[Howes et al., 1979] M.J. Howes and D.V. Morgan, *Charge Coupled Devices and Systems*. London: John Wiley & Sons, 1979.

[Kim et al., 1972] C.K. Kim, J.M. Early and G.F. Amelio, "Buried-channel charge-coupled devices," *NEREM, Record of Technical Papers*, Boston, Part I, pp. 161-

164, 1972.

[Lee et al., 1972] H.S. Lee and L.G. Heller, "Charge control method of charge-coupled device transfer analysis," *IEEE Transactions on Electron Devices*, vol. ED-19(12), pp. 1270-1279, 1972.

[Maclennan et al., 1975] D.J. Maclennan and J. Mavor, "Novel technique for the linearization of charge-coupled devices," *Electronic Letters*, vol. 11, pp. 222-223, 1975.

[Maclennan et al., 1975] D.J. Maclennan and J. Mavor, "Linearization of the charge-coupled device transfer function," in *Proceedings of the International Conference on CCDs*, San Diego, pp. 291-294, 1975.

[Miida et al., 1991] T. Miida, Y. Hasegawa, T. Hagiwara and H. Ohshiba, "A CCD video delay line with charge-integrating amplifier," *IEEE Journal of Solid State Circuits*, vol. SC-26(12), pp. 1915-1919, 1991.

[Mohsen et al., 1973] A.M. Mohsen, R. Bower, T.C. McGill and T. Zimmermann, "Overlapping gate buried channel charge-coupled devices," *Electronics Letters*, vol. 9, pp. 396-397, 1973.

[Mohsen et al., 1973] A.M. Mohsen, T.C. McGill and C.A. Mead, "Charge transfer in overlapping gate charge-coupled devices," *IEEE Journal of Solid State Circuits*, vol. SC-8, pp. 191-207, 1973.

[Mohsen et al., 1975] A.M. Mohsen, M.F. Tompsett and C.H. Séquin, "Noise measurements in charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-22(5), pp. 209-218, 1975.

[Séquin et al., 1975] C.H. Séquin and M.F. Tompsett, *Charge Transfer Devices*. New York: Academic Press, 1975.

[Séquin et al., 1975] C.H. Séquin and A.M. Mohsen, "Linearity of electrical charge injection into charge coupled devices," *IEEE Journal of Solid State Circuits*, vol. SC-10(2), pp. 81-92, 1975.

[Singh et al., 1976] M.P. Singh, S.D. Brotherton, P.C.T Roberts and D.R. Lamb, "Influence of clocking waveform on charge transfer in three-phase CCDs," in *Proceedings of the International Conference on Technology and Applications of CCDs*, Edinburgh, p. 39, 1974.

[Singh et al., 1976] M.P. Singh and S.D. Brotherton, "Influence of clocking waveform on charge transfer in three-phase CCDs," *Solid State Electronics*, vol. 19, pp. 279-287, 1976.

[Singh et al., 1976] M.P. Singh and D.R. Lamb, "Theoretical calculations of surface-state loss in three-phase charge-coupled devices," *Journal of Physics D: Applied Physics*, vol. 9, p. 37, 1976.

[Strain, 1972] R.J. Strain, "Properties of an idealized traveling-wave charge-coupled device," *IEEE Transactions on Electron Devices*, vol. ED-19(10), pp. 1119-1130, 1972.

[Sze, 1985] S.M. Sze, *Semiconductor Devices, Physics and Technology*. New York: John Wiley & Sons, 1985.

[Tompsett, 1975] M.F. Tompsett, "Surface potential equilibration method of setting charge in charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-22(6), pp. 305-309, 1975.

[Tompsett et al., 1973] M.F. Tompsett and E.J. Zimany, "Use of charge-coupled devices for delaying analog signals," *IEEE Journal of Solid State Circuits*, vol. SC-8, pp. 151-157, 1973.

[Walden et al., 1972] R.H. Walden, R.H. Krambeck, R.J. Strain, J. McKenna, N.L. Schryer and G.E. Smith, "The buried channel charge-coupled device," *Bell System Technical Journal*, vol. 51, pp. 1635-1640, 1972.

[Wen, 1976] D.D. Wen, "A CCD video delay line," in *IEEE International Solid State Circuits Conference Digest of Technical Papers*, Philadelphia, pp. 204-205, 1976.

Chapter 3

3. SEMIPARALLEL CCD PROCESSOR

In this chapter, a semiparallel CCD vector matrix multiplier is described. The basis for this work is an architecture first described in [Agranat et al., 1987] in which the semiparallel CCD architecture was disclosed. Before describing the semiparallel CCD VMM in detail, a brief account of the history of CCDs in signal processing is given. Good reviews can be found in either [Séquin et al., 1975] or [Beynon et al., 1980].

3.1 Historical Background

Performing analog computation with CCDs has been discussed from the inception of CCDs due to their natural discrete time storage capability. A good review of CCDs as signal processors can be found in [Barbe et al., 1978]. CCDs make excellent analog shift registers which are an important component of many types of signal processing systems such as Finite Impulse Response (FIR) filters. A number of signal processing architectures have been tried, the most common being variants of the transversal filter [Collins et al., 1972][Baertsch et al., 1976]. In Figure 3.1 a simplified transversal filter is shown, adapted from [Séquin et al., 1975].

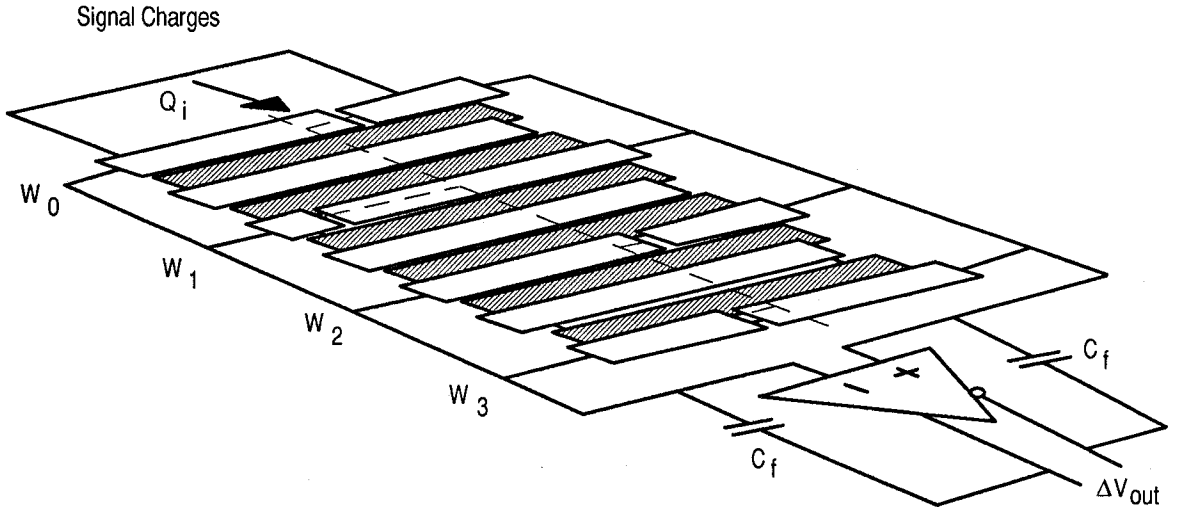

Figure 3.1. A transversal filter adapted from [Séquin et al., 1975]. The signal charge is sensed by split gates where the physical location of the split encodes the filter weight. The signals are aggregated via charge summing on the differential lines and a voltage output is produced which is the convolution of the charge signal and the fixed weights.

The charge domain signal is transferred down the CCD channel which has floating transversally split gates as one of the clock electrodes. These split gates are attached to a differential charge sensing amplifier. When the charge is transferred under the sensing gates, the output voltage is the convolution of the signal charge Q_i and the fixed weights determined by the position of the electrode splits, W_i.

$$\Delta V_{out} = \frac{1}{C_f} \sum_{i=0}^{N-1} Q_{N-i} W_i . \tag{3.1}$$

Bandpass filters with up to 500 taps have been built [Broderson et al., 1976] and tested. Programmable versions of the CCD transversal filter have also been constructed, usually by adding a MOS based multiplier to the shift register, using the CCD only as a delay line [Mavor et al., 1978].

Additional signal processing architectures using simple functional building blocks [Joseph et al., 1984][Vogelsong et al., 1985][Fossum, 1987] like the fill and spill input circuit, the sense amplifiers of Section 2.8 and charge splitting techniques [Bencuya et al., 1984] such as that shown in Figure 3.2 have been used with success for specific applications. Most devices suffer from lack of programmability and limited accuracy, which limits their applicability.

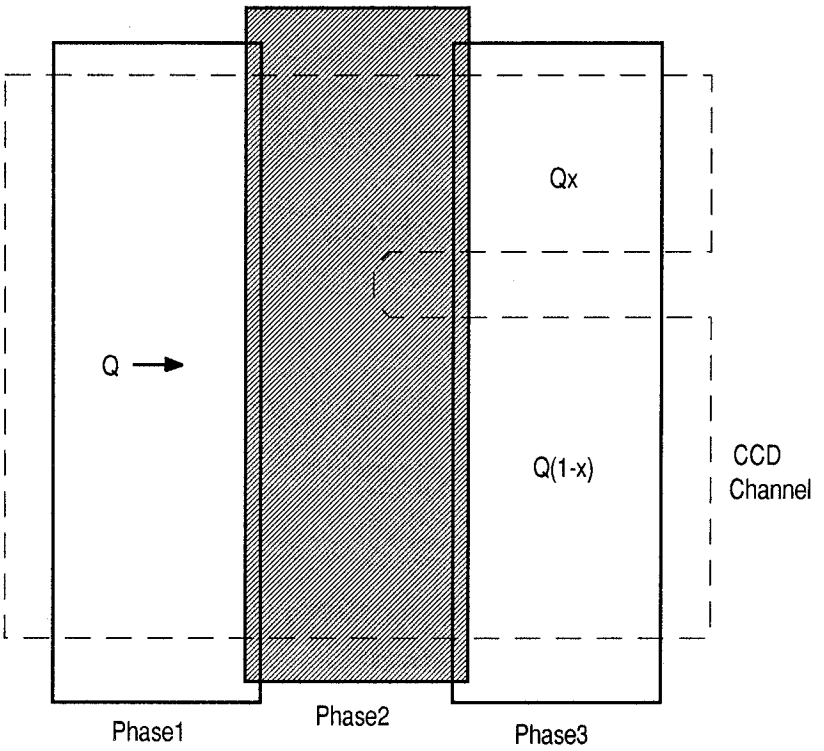

Figure 3.2. A split CCD channel can be used to accurately partition charge according to the relative size of the channel. For an original amount of charge Q entering the device from the left, Qx is available out the top and Q(1-x) out the bottom. The accuracy depends mainly on the size of the structure in a similar way that transistor or capacitor matching depends on size.

The main use of CCD technology today is in imaging devices due to its high sensitivity, controllable offsets, low noise, manufacturability, and natural shift register implementation. The ability to transfer the small charge packets created in low light situations to an extremely small sense capacitor in an amplifier with good offset control is especially noteworthy. Special amplifiers make it possible to now sense discrete electrons at room temperature and with dynamic

ranges of almost 100dB [Matsunaga et al. 1991]. However, as a number of recent articles on CCD imagers for high definition television illustrate, specialized fabrication is required to extract such high levels of performance.

The thrust of this research has been to explore the applicability of standard process technology to parallel analog signal processing, more specifically to vector matrix multiplication. Without resorting to specialized technologies, the designer must make a cognizant effort to design fabrication tolerant architectures that can still retain accuracy even though fabricated in a digital process. The first architecture discussed here required a specialized CCD fabrication where, due to the cost, only one chip run was done. The outcome of this phase of the project was illuminating in two ways, the first being that for this type of research multiple chip runs are required to get devices to meet performance expectations, especially in an unknown process. Secondly, we learned that the choice of architecture should be determined by the available fabrication processes. There is little point in developing an architecture that inherently is difficult and expensive to implement. For the first architecture, however, we had access to a CCD imager fab line and developed the following circuit.

3.2 System Description

The semiparallel CCD vector matrix multiplier incorporates a large matrix storage region that holds the analog weight matrix, W_{ij}, as charge packets in CCD shift registers. N^2 charges are stored in N circulating ring registers which each contain one row of W_{ij}, as shown in the left portion of Figure 3.3. The rings have a common set of four-phase clocks which move charge clockwise. Given N clock cycles, the data in the rings completely revolve once. A compact storage cell was realized for the ring storage cells. The fabricated cells measured 14μm by 40μm for the storage of one analog matrix element. The matrix is loaded from the left by serial input and

rotation of the ring shift register through a set of fill and spill circuits (not shown). The data storage structure is similar to CCD architectures that were used in the 1970's as digital memories [Beynon et al., 1980].

The vector matrix multiplication process is accomplished by adding a column of sense/multiply circuits to one side of the charge storage area, which is shown in Figure 3.3. The sense/multiply circuit non-destructively measures one charge in the ring shift register and multiplies it by a common column line encoding U_j. The nondestructive sensing is done with a floating gate charge sensor as described in Section 2.8. The result of this multiplication, a charge packet, is dumped into one of N large accumulator regions, one for each row.

The sequence of events for a vector matrix multiplication is the following. First, the accumulators are cleared and the matrix loaded with the desired analog matrix of charge. Second, the array is clocked and the data rotated. For the first clock cycle, U_1 is applied to the vector input line and multiplies the first column of W_{i1}. The products are stored in the accumulator regions. During the second clock cycle, the ring registers are rotated one element and U_2 is applied to the vector input line. Thus the second column of W_{ij} is multiplied by the proper element of U_j. Going through N clock cycles completes the vector matrix multiplication where the column of charges in the accumulators represent the output vector. At this point, the data has completely revolved once and the accumulators can be read (and cleared). Recently, other semiparallel architectures have been published that utilize digital weight storage and use multiplying D-A converters as the multiplication elements. [Chiang, 1990][Chiang, 1991].

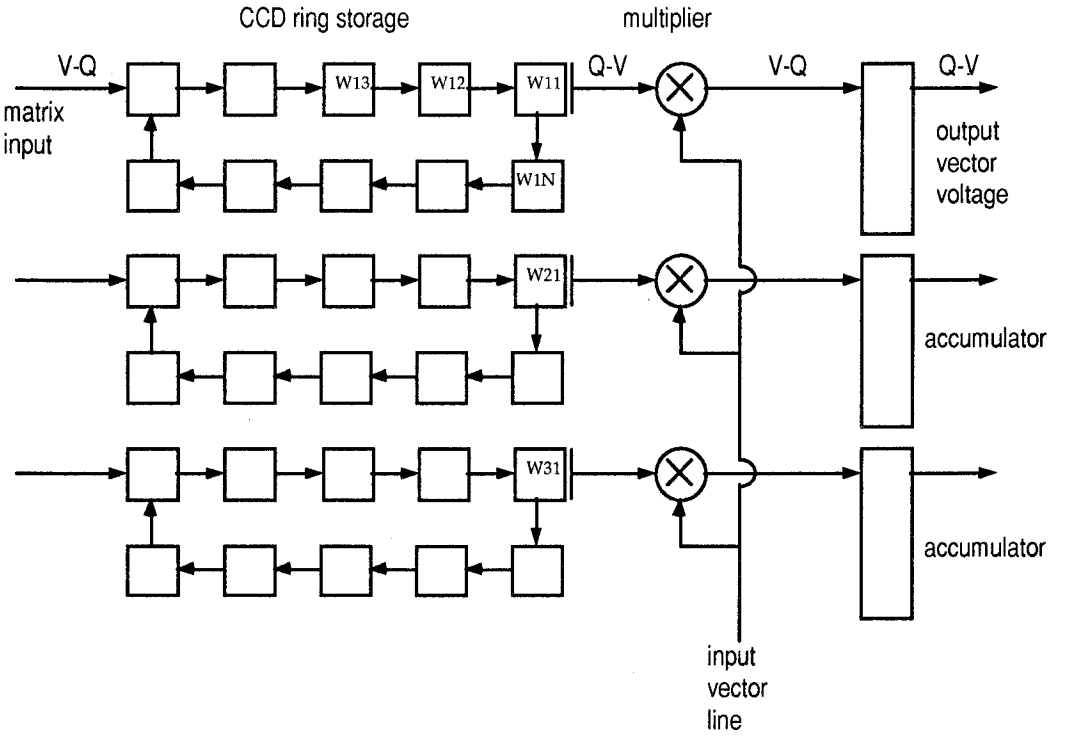

Figure 3.3. The semiparallel CCD architecture. Matrix elements are charges residing in the ring registers which are loaded from the left by fill and spill circuits (not shown). Each column of the matrix is multiplied by a single common input vector line on which the input vector elements are sequentially placed. The products are summed as charges in the accumulators to the right and are read out as a voltage. Note that the progression from matrix input voltage to output voltage involves four nonlinear Q-V transforms.

3.3 Analysis of Charge Storage Capabilities and Linearity

The novel ring storage element has a counter-propagating CCD element with the electrode arrangement of Figure 3.2. The upper channel moves left to right whereas the lower channel moves right to left, achieved simply by swapping phases.

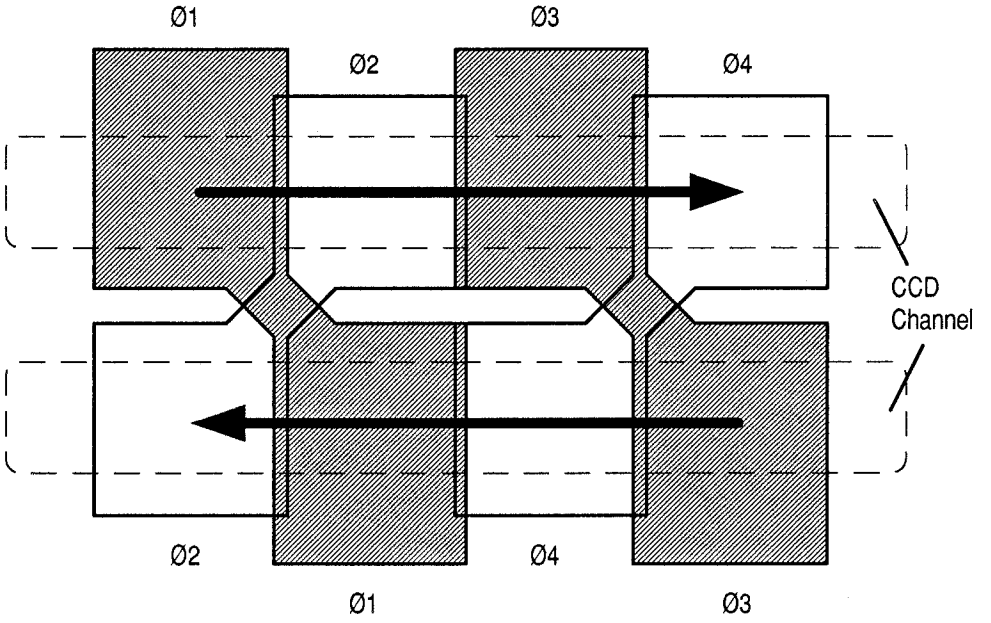

Figure 3.4. The CCD ring register matrix storage cell achieves counter propagating charges by simply swapping two clock lines. The cell stores two charges and the fabricated dimensions were 28um by 40um in a 2um CCD process.

The CCD ring cells, due to their compact layout, are an area efficient method of information storage. Non ideal CCD properties, however, limit the applicability of this storage. Note that the matrix charge is continually rotating in the ring registers during normal operation. Charge transfer inefficiency causes the charge packets to become smeared together. The charge transfer inefficiency tells how much of a given charge packet is left behind. The loss after N transfers is given by

$$(1-\varepsilon)^N \approx 1-N\varepsilon \tag{3.2}$$

where ε is the transfer inefficiency. For a matrix element to retain n-bit accuracy, the transfer loss must be less than 1/2 LSB. Combined with the fact that one semiparallel vector matrix multiplication of the CCD requires 4N transfers given a four-phase clock as shown in Figure 3.4, the number of vector matrix multiplications that can be completed before transfer loss becomes a problem is

$$\#\ \text{of Operations} \leq \frac{1}{4N\varepsilon 2^{n+1}}. \tag{3.3}$$

For a good buried channel CCD process, the efficiency can reach 0.999999, in which case the number of vector matrix multiplications possible before refresh with 4-bit accuracy and N=256 is ~ 32. The matrix thus requires reloading every 32 computation cycles to maintain four-bit accuracy. As for other CCD nonidealities such as dark current, the frequent reloading needed to account for transfer loss causes their effects to be negligible. For surface channel devices, the number of operations is less than one, precluding the use of surface channel circuits for this architecture. It was for this reason that the special buried channel fabrication process was used.

In arriving at the vector-matrix product, there are a number of voltage to charge and charge to voltage conversions. The matrix elements are input electronically and undergo a V-Q conversion. The multiplier converts a sensed voltage multiplied by the U_j line to a charge which is summed and read out as a voltage. In the simplest case, the U_j line encodes a binary number which acts as a gating signal to the charge packet created in the multiplier. This multiplication of a binary U_j and analog W_{ij} has the nonlinearity of the voltage-to-charge conversion. The nonlinearity up to the point of the multiplication is thus $Q_{bc}(V_{bc}(Q_{bc}(V_{ij})))$. These charges are summed together in the accumulator then read out through an additional charge-to-voltage conversion yielding

$$V_i = V_{bc}\left[\sum_j U_j Q_{bc}(V_{bc}(Q_{bc}(V_{ij})))\right]. \tag{3.4}$$

The error introduced into the output can be quantified. Using the most linear portions of the curves from Figures 2.19 and 2.23, the RMS error of Equation 3.4 is ~10 %, or roughly 3 bits. Note that some compensation can be afforded by predistorting the matrix input voltage with a feedback fill-and-spill circuit to account for some of the nonlinearities. However, because the summation appears before the final nonlinearity in Equation 3.4, i.e., the charge-to-voltage conversion of the

output circuit, the predistortion cannot remove all nonlinearities.

The detailed circuit diagram is shown in Figure 3.5. In this imager process, CMOS transistors were not available, resulting in only simple source follower amplifiers and one-sided reset gates, all constructed of the buried channel depletion mode FETs.

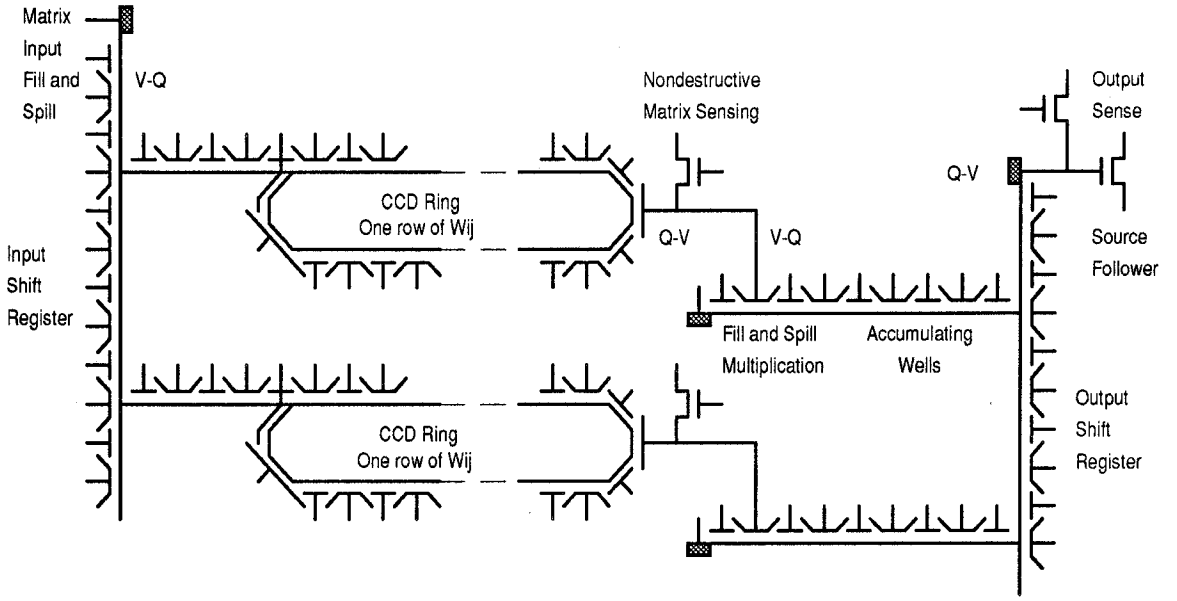

Figure 3.5. Four nonlinear charge/voltage conversions are performed on between matrix input to signal output. The clock signals required to operate this device were non trivial. The multiplication in this device was performed by simply gating the fill and spill circuit connected to the nondestructive matrix sensing. The resulting analog/binary product was then summed and read out through the mux on the left.

3.4 Experimental Results

A large CCD chip containing 65,536 storage elements and 256 multipliers was fabricated and tested and was functional on the first and only run [Agranat et al., 1990]. The relevant chip data is listed in Table 3.1 and a chip photograph is shown in Figure 3.6. This chip had ~40 externally driven clock signals with ~80 voltage levels to adjust. Getting the timing and voltage levels

adjusted properly was a time consuming process and limited the amount of experimental data gathered. Furthermore, the limited test equipment available at the time was incapable of analog matrix input. Only binary vectors and matrices were possible at the time although both analog matrix and analog vector functionalities were included on the chip. A layout error limited the device to operating only a few of its 16 outputs at any one time. However, this problem did not compromise the tests on linearity and speed.

Technology	2um CCD (Ford Aerospace/Loral) Double poly/double metal
Transfer Efficiency	> 0.999999
Number of Matrix Elements	65,536
Size	9mm x 13mm
Accuracy	Matrix Elements - 4 bits Vector Elements - 1 bit Output - 3 bits
Clock Frequency	1.5 MHz (limit of test equip.)
Multiply-accumulates/sec	$4x10^8$ (limit of test equip.)

Table 3.1. The architecture was successfully fabricated and tested in one iteration using a CCD imager fabrication line. The test station generating the clock signals was incapable of supporting higher clock rates.

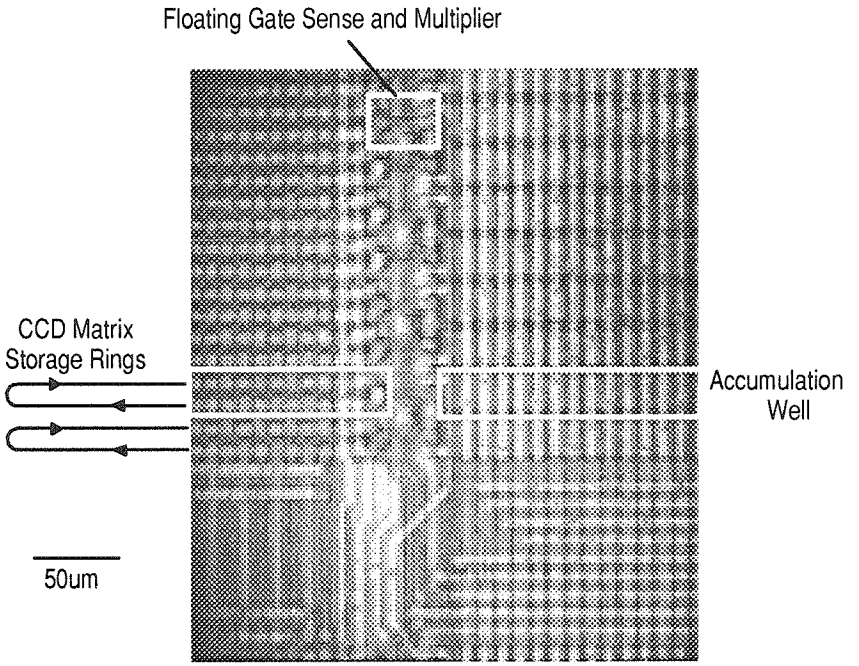

Figure 3.6. A portion of the chip showing the edge of the ring storage area, a multiplier and the large accumulator well. The cell pitch was 28um vertically in the 2um CCD process.

A crude linearity test was performed in which the matrix contained a number of columns filled with charge and the vector elements were left on. The output is expected to vary linearly with the number of columns that are 'on'. A similar test was performed by filling the matrix entirely with charge and varying the number of vector elements in U_j that were 'on'. Once again, the output should vary linearly with the number of 'on' vector elements. The results of these tests are shown in Figure 3.7. Due to the difficult adjustment procedure, voltage ranges were not optimized for linearity, exacerbating the nonlinearity error in Equation 3.4. The measured RMS error of these transfer curves was around 15%, or roughly 3 bits.

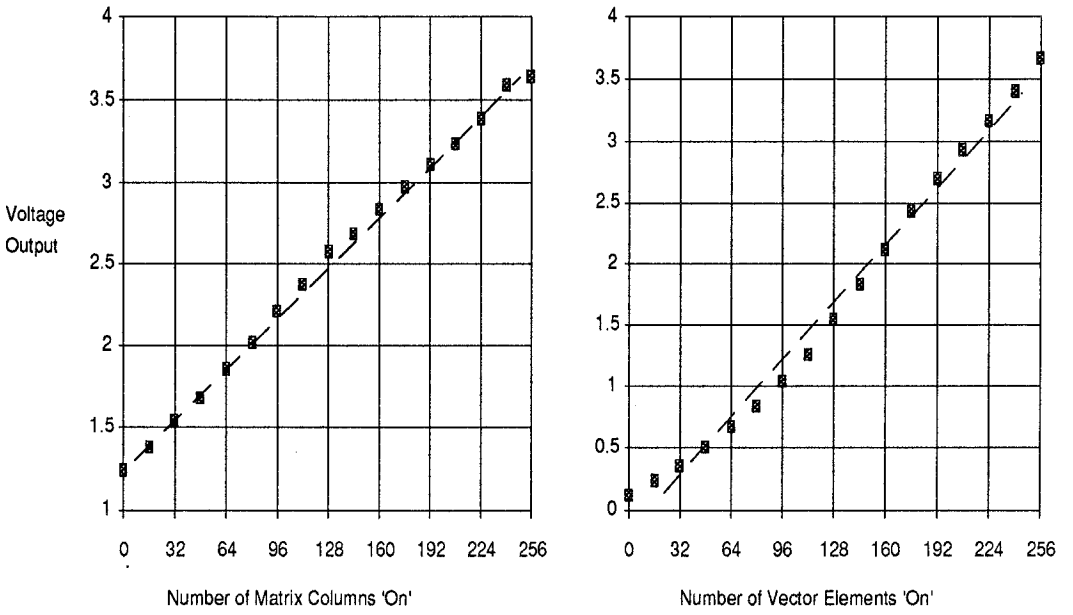

Figure 3.7. A fully 'on' vector was multiplied by a matrix with a variable number of 'on' columns, giving the result on the left. A similar test was performed for a fully 'on' matrix and a variable number of input vector elements being 'on' giving the result on the right. The limited test equipment and inherent nonlinearities of buried channel devices resulted in the poor (15% RMS error) linearity measured here.

A second, easier to implement test was performed by loading the matrix with half the columns turned 'on'. This matrix was multiplied by a vector that was also half 'on'. During the first computation cycle, the 'on' halves of the matrix and input vector overlap completely, giving maximum output. Prior to the next vector matrix multiplication, the matrix is rotated slightly by clocking the ring storage registers a few extra times. During the next computation, the 'on' halves of the matrix and input vector are slightly out of phase, resulting in diminished output. After many repetitions of this calculation-plus-extra-data-rotation cycle, the matrix and input vector will be completely out of phase, thus giving zero output. The output over many cycles will look like a triangle wave, as the matrix and vector come in and out of phase. The output of this experiment is shown in Figure 3.8.

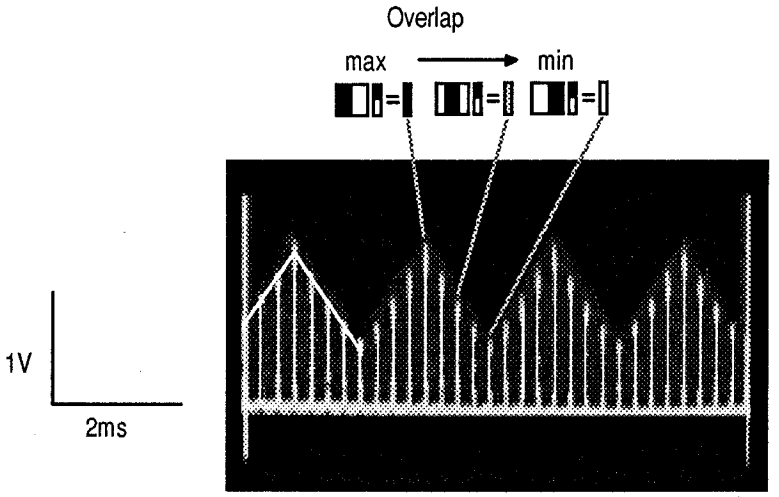

Figure 3.8. The vector and matrix are both half 'on' and in phase for the first computation. After each computation, the ring registers are given a few extra clock cycles. Repeated many times, the 'on' halves of the matrix and vector come in and out of phase which results in a triangle wave output over many computation cycles. Limited transfer efficiency tends to reduce the sharpness of the peaks after many transfers.

3.5 Limitations of the Semiparallel Architecture

The semiparallel CCD processor performs only N multiply-accumulate operations per clock cycle. The bulk of the chip area is used as storage space for the W_{ij}. The same function could be accomplished with a physically smaller digital memory which are much denser than CCD circuits. Also, due to the transfer loss, it is not even quality storage space, as it requires frequent reloading. In addition, the multiple nonlinearities associated with its operation, as evidenced by Equation 3.3 and Figure 3.7, limit the circuit's applicability to problems where low accuracy can be tolerated. More linear surface channel technology cannot be used due to its low transfer efficiency. Furthermore, the multiplier is limited in its analog accuracy to only a few bits. Altogether, these limitations are significant deterrents to using this architecture for system implementations. The device did serve a purpose, however, in that it pointed out the technology limitations and gave us insight on how to overcome them.

It is important to note that in the semiparallel architecture of [Chiang, 1991] the memory refresh problem has been resolved at the expense of area by using binary CCD storage with on-chip refresh. In addition, the multiplier in [Chiang, 1991] is a multiplying D-A converter which has good resolution but also requires significant area.

3.6 References:

[Agranat et al., 1987] A. Agranat and A. Yariv, "Semiparallel microelectronic implementation of neural network models," *Electronic Letters*, vol. 23, pp. 580-581, 1987.

[Agranat et al., 1990] A.J. Agranat, C.F. Neugebauer, R.D. Nelson and A. Yariv, "The CCD neural processor: a neural integrated circuit with 65,536 programmable analog synapses," *IEEE Transactions on Circuits and Systems*, vol. 37, pp. 1073-1075, 1990.

[Baertsch et al., 1976] R.D. Baertsch, W.E. Engeler, H.S. Goldberg, C.M. Puckette and J.J. Tiemann, "The design and operation of practical charge-transfer transversal filters," *IEEE Transactions on Electron Devices*, vol. ED-23(2), pp. 133-141, 1976.

[Barbe et al., 1978] D.F. Barbe, W.D. Baker and K.L. Davis, "Signal processing with charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-25(2), pp. 108-125, 1978.

[Bencuya et al., 1984] S.S. Bencuya and A.J. Steckl, "Charge packet splitting in charge domain devices," *IEEE Transactions on Electron Devices*, vol. ED-31(10), pp. 1494-1501, 1984.

[Beynon et al., 1980] J.D.E. Beynon and D.R. Lamb, *Charge Coupled Devices and Their Applications*. London: McGraw-Hill, 1980.

[Broderson et al., 1976] R.W. Broderson, C.R. Hewes and D.D. Buss, "A 500-stage CCD transversal filter for spectral analysis," *IEEE Transactions on Electron Devices*, vol. ED-23(2), pp. 143-151, 1976.

[Buss et al., 1973] D.D. Buss, D.R. Collins, W.H. Bailey and C.R. Reeves, "Transversal filtering using charge-transfer devices," *IEEE Journal of Solid State Circuits*, vol. SC-8, pp. 138-146, 1973.

[Chiang, 1990] A.M. Chiang, "A CCD programmable signal processor," *IEEE Journal of Solid State Circuits*, vol. SC-25(6), pp. 1510-1517, 1990.

[Chiang, 1991] A.M. Chiang, "A CCD programmable image processor and its neural network applications," *IEEE Journal of Solid State Circuits*, vol. SC-26(12), pp. 1894-1901, 1991.

[Chiang, 1987] A.M. Chiang, "A video-rate CCD two-dimensional cosine transform processor," in *Visual Communications and Image Processing II*, T.R. Hsing, ed., Proceedings of the SPIE, vol. 845, pp. 2-5, 1987.

[Collins et al., 1972] D.R. Collins, W.H. Bailey, W.M. Gosney and D.D. Buss, "Charge-coupled device analogue matched filters," *Electronics Letters*, vol. 8, pp. 328-329, 1972.

[Fossum, 1987] E.R. Fossum, "Charge-coupled computing for focal plane image preprocessing," *Optical Engineering*, vol. 26(9), pp. 916-922, 1987.

[Hartmann et al., 1973] C.S. Hartmann, L.T. Claiborne, D.D. Buss and E.J. Staples, "Programmable transversal filters using surface wave devices, charge-transfer devices and conventional digital approaches," in *Proceedings of the International Specialist Seminar on Component Performance and System Applications of Acoustic Surface Wave Devices*, pp. 102-114, Aviemore, 1973.

[Joseph et al., 1984] J.D. Joseph, P.C.T. Roberts, J.A. Hoschette, B.R. Hanzel and J.C. Schwanebeck, "A CCD-based parallel analog processor," in *State-of-the-Art Imaging Arrays and Their Applications*, K.N. Prettyjohns, ad., Proceedings of the SPIE, vol. 501, pp. 238-241, 1984.

[Matsunaga et al., 1991] Y. Matsunaga, H. Yamashita, S. Manabe and N. Harada, "A high-sensitivity MOS photo-transistor for area image sensor," *IEEE Transactions on Electron Devices*, vol. ED-38(5), pp. 1044-1047, 1991.

[Matsunaga et al., 1991] Y. Matsunaga and S. Ohsawa, "A 1/3-in interline transfer CCD image sensor with a negative-feedback-type charge detector," *IEEE Journal of Solid State Circuits*, vol. SC-26(12), pp. 1902-1906, 1991.

[Mavor et al., 1978] J. Mavor and P.B. Denyer, "Design and development of CCD programmable transversal filters," *IEE Journal of Electronic Circuits and Systems*, vol. 2, pp. 1-8, 1978.

[Séquin et al., 1975] C.H. Séquin and M.F. Tompsett, *Charge Transfer Devices*. New York: Academic Press, 1975.

[Volgelsong et al., 1985] T.L. Vogelsong and J.J Tiemann, "Charge domain integrated circuits for signal processing," *IEEE Journal of Solid State Circuits*, vol. SC-20(2), pp. 562-570, 1985.

Chapter 4

4. PARALLEL CHARGE INJECTION DEVICE PROCESSOR

With the knowledge gained in testing the semiparallel CCD processor, a new architecture was developed in which many of the nonidealities discussed in the previous chapter were overcome. The ease of fabrication and testing played a major role in choice of architecture, which is discussed below.

4.1 Background

An outgrowth of the semiparallel work, the Charge Injection Device (CID) processor is a fully parallel charge domain processor that avoids most of the pitfalls encountered in the previous chapter. The name 'CID' comes from the imager type the processor closely resembles [Arnold et al., 1971][Michon et al., 1973][Michon et al., 1974][Burke et al., 1976]. A good review can be

found in [Michon et al., 1980]. Invented as an imaging device to compete against CCDs, CIDs have failed to become mainstream products for a simple physical consequence of their architecture. A good comparison of the different types of imagers can be found in [Barbe, 1976]. A simplified CID imager is shown in Figure 4.1 using a readout technique described in [Michon et al., 1975]. The imaging area consists of a grid of CID cells, each of which has connections to horizontal and vertical lines. The single CID cell, shown in Figure 4.1(a), is initially empty. Illumination causes the charge well to fill as in 4.1(a). For readout, a sense amp is connected to the desired column line and the charge is moved beneath the row lines, shown in 4.1(b). One row line is then pulsed to a low voltage, causing the charge for that particular row to transfer to the column lines, shown in 4.1(c). The sense amp sees this charge and creates a voltage output.

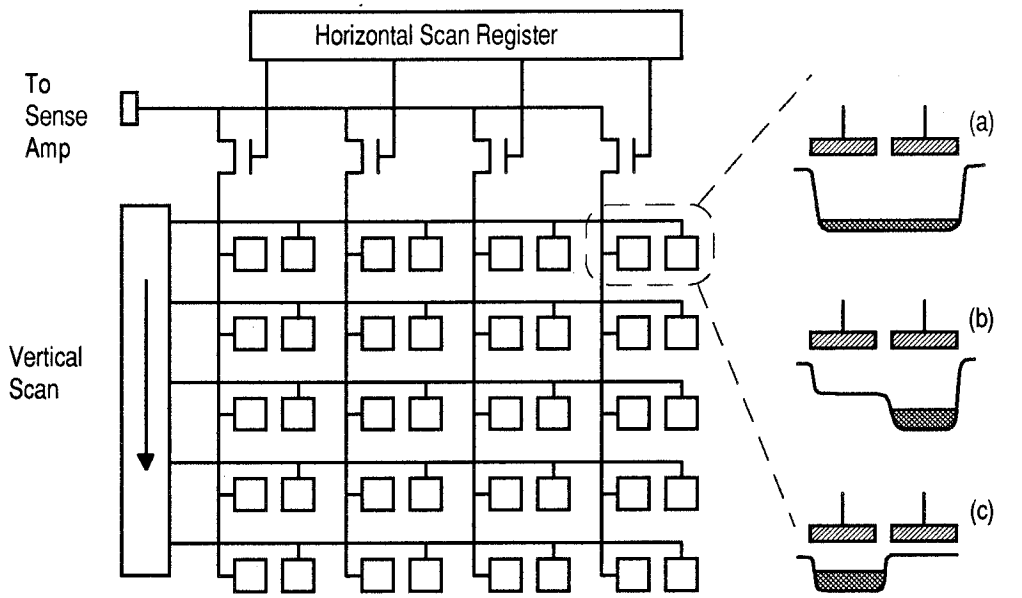

Figure 4.1. A general CID architecture, adapted from [Séquin et al., 1975]. Each pixel has two gates which integrate photo-generated electrons (a). For readout, all charge in the matrix is moved to the row gate side of each cell (b). The column line of interest is selected and the row line of interest is pulsed low, transferring the charge to the column lines (c). The sense amp outputs a voltage proportional to the charge at the selected pixel. The pixel charges are cleared by pulsing all row and column lines negative, injecting the charge into the substrate.

The reason CID imagers have failed to gain mainstream application is that the sensing capacitance,

i.e., the column capacitance, is very large compared to the sensing capacitance possible with conventional shift register based CCD imagers. In CCD imagers, the charge packet can be maneuvered onto a tiny sense capacitance, giving a large output voltage for small charge. Recent imagers have achieved single electron sensitivity at room temperature [Matsunaga et al., 1991] In CIDs, however, the charges remain in the array and are restricted to be sensed by a amplifier at the end of a long sense line. Furthermore, in shift register based CCD imagers, only one highly sensitive amplifier needs to be built and any offsets it has are simply a DC correction to the output. In CID imagers, however, each column has its own switches and parasitic capacitances which result in significantly more fixed pattern noise and much less sensitivity [Grafinger et al., c. 1982]. Despite being an elegant implementation that is easier to fabricate, the CID as an imager has significant shortcomings.

As a signal processing device, the CID can be transformed into something very useful. Charge levels are not restricted by low light levels and can be kept at the capacity limit of the device. The CID operation described above does not require very good transfer efficiency, making it a candidate for surface channel implementations. Additionally the simple clocking, compared to the semiparallel circuit, is much more amenable to testing and later system integration.

4.2 System Description

As in the semiparallel device, the matrix is stored as a set of charges. The novel architecture of the CID processor consists of a large array of simple cells, each of which contains two electrodes and stores one matrix element [Agranat et al., 1988]. One of the electrodes is connected vertically and the other horizontally. The simplified block diagram is shown in Figure 4.2. Additional gates used to load the device, which are held at DC potentials during computation, are not shown. In its most basic configuration, the circuit computes the product of a binary input vector, U_j, and the analog

matrix of charge. The computation done at each CID cell is a multiply-accumulate operation in which the charge, Q_{ij}, is multiplied by a binary input vector element, U_j, encoded on the column line. This product is summed with other products in the same row along the row line to form the output vector, V_i. Binary multiplication at each matrix cell is equivalent to adding or not adding the charge Q_{ij} to the sum formed on the row line.

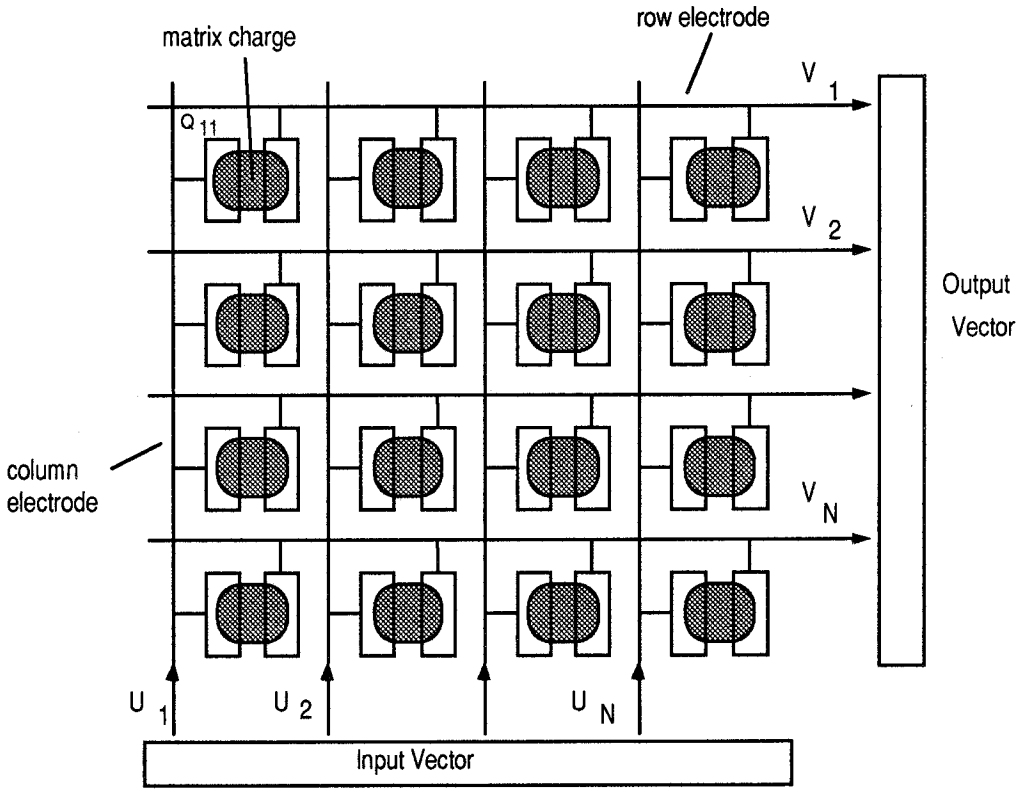

Figure 4.2. The CID vector matrix multiplier architecture borrows heavily from the imager except that the column scanner has been replaced by an input vector register. Sense amps along each row line measure the charge that is moved by action of the input vector.

The matrix cell operation is shown in Figure 4.3, which shows one row of the matrix. Three gates are shown, the middle being held at a constant voltage throughout operation. The device is reset to the state of Figure 4.3(a) in which all charges are beneath the column lines.

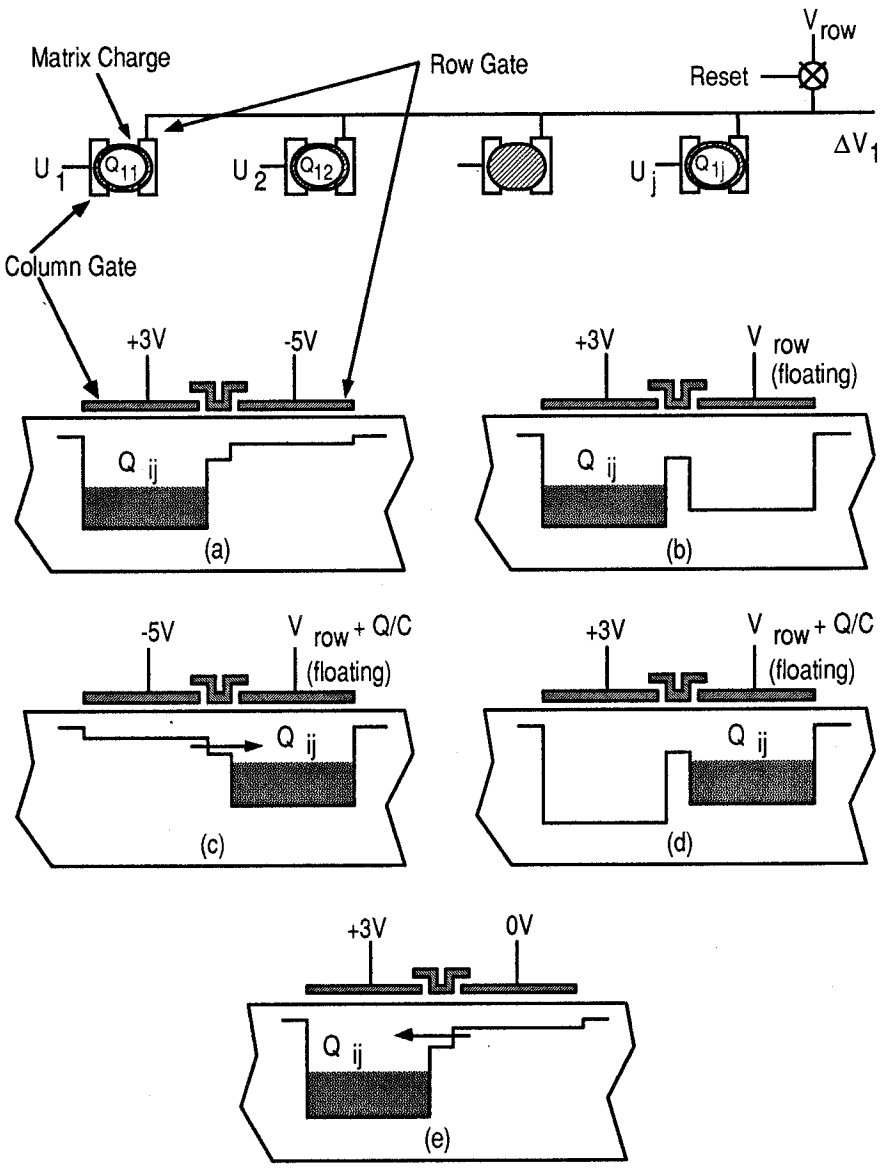

Figure 4.3. The CID cell in operation. Charge initially under the column lines (a) is transferred (c) depending on the corresponding vector element's binary state causing a voltage change in the row line (d). Many charges are summed along the row simultaneously. The charge is returned (e) to get ready for another computation.

During this phase, the reset switch is 'on' so that the row line is precharged to a fixed potential. The reset switch is then turned 'off' so that the row line is floating, shown in Figure 4.3(b). The computation occurs by presenting the binary input vector to the column lines. If the vector element

is 'on', then the column line is pulsed to a negative voltage as in Figure 4.3(c)-(d), which transfers the matrix charge in that column. If the vector element is 'off', the column line remains at its positive potential. Since the row line is floating, it experiences a voltage change approximately proportional to the total amount of charge moved beneath it by the action of the column lines. Thus the voltage change in the row lines is given approximately by

$$\Delta V_i = \frac{\sum_{j=0}^{N-1} U_j Q_{ij}}{C_{row}} \tag{4.1}$$

where C_{row} is the capacitance of the row electrode. Thus in one clock cycle, the product of a binary vector and an analog matrix of charge is available in the form of row voltage changes. The voltage on the row lines is sampled in Figure 4.3(d). The charge can then be returned to the column side during the reset phase by pulsing the row lines to a negative potential, shown in Figure 4.3(e), and subsequently reused for another computation.

4.3 Loading the Matrix

As described above, the CID processor can be optically loaded with charge. For systems where this is not practical or the matrix bandwidth requirements are low, electrical loading can be done. Additional electrodes are required for electrical loading the matrix of charge [Neugebauer et al., 1990]. The actual constructed cell is shown in Figure 4.4 in both the loading and computation phases of operation. Each cell has four electrodes. If a standard four-phase clock is applied to these electrodes, they will act as one element of a shift register, moving charge from left to right as in Figure 4.4(a). The device is loaded using this shift register mode of the CID cell as shown in block diagram form in Figure 4.5. Matrix input is accomplished through a single pin by adding a vertical CCD shift register. The matrix is loaded by shifting in an entire column of the matrix into

the vertical shift register, then transferring this column of data into the array using the clocking shown in Figure 4.4(a). Once the matrix is loaded, two of the electrodes at each cell become dormant and are held at fixed potentials. The other two electrodes are used as the row and column lines of the CID device, as shown in Figure 4.4(b).

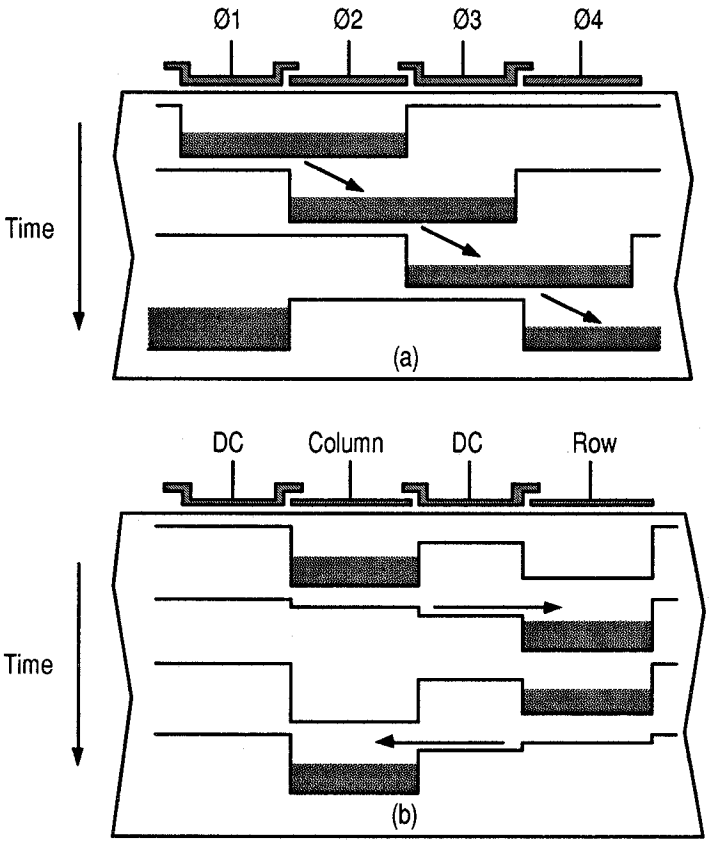

Figure 4.4. The CID cell has four electrodes, all of which are used for a loading cycles in which each CID cell acts as one stage in a shift register (a). In computation mode (b), two of the electrodes are kept at a constant potential and provide isolation between cells and reduce row/column line parasitic capacitance.

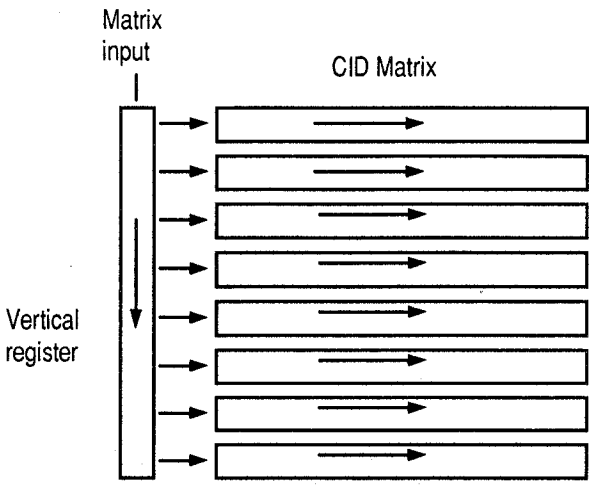

Figure 4.5. A vertically oriented shift register is added for loading. Charges are created from a single fill and spill circuit and shifted into the column which, when full, is transferred across the matrix.

The electrode between the row and column gates serves an important function other than matrix loading. During the computation phase shown in Figure 4.3(c)-(d), the column gate is returned to its original potential level prior to the row voltages, i.e., VMM outputs, being sampled. The constant potential gate between the row and column gates prevents the charge from flowing back beneath the column gate. Note that the 'before' and 'after' voltages of all the column lines are the same. Stray capacitance between row and column lines would lead to an error in the output voltage if the column lines changed in potential. By allowing the column lines to return to their original voltage levels prior to sampling, the only difference between Figure 4.3(b) and 4.3(d) is the charge movement, hence the stray capacitance between row and column lines does not influence the VMM outputs.

4.4 Improving Linearity

The row capacitance changes with its voltage, causing a nonlinearity in the summation. This nonlinearity involves the interdependence of the charge packets and the row voltage, which is a

function of the charge packets and the input vector. Each charge has a variable effect on the row voltage. One way to remove this nonlinearity is to use the feedback amplifier circuit discussed in section 2.8 and shown in Figure 4.6 [Anagnostopoulos, 1978][Neugebauer et al., 1992].

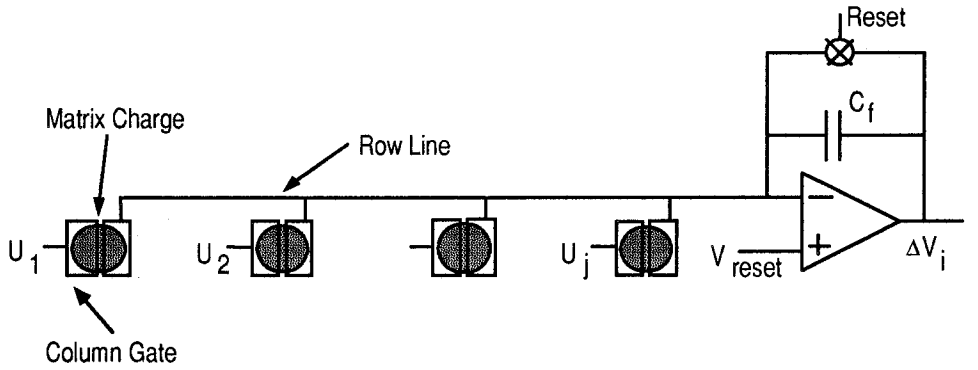

Figure 4.6. Extending the capacitive sense amplifier of Section 2.8 to multiple sense gates, as in the transversal filter, the effect of the nonlinear row capacitance is removed. Wider output voltage ranges can be achieved through proper choice of the feedback capacitor than with the sense configuration of Figure 4.3.

The action of the amplifier holds the row voltage at a virtual ground, eliminating the linearity associated with row voltage changes. The summation done by the row line can be abstracted to be an inverting amplifier with many inputs, as shown in Figure 4.7. The potential change of the channel caused by the movement of charges is added through the capacitance between the channel and the row line.

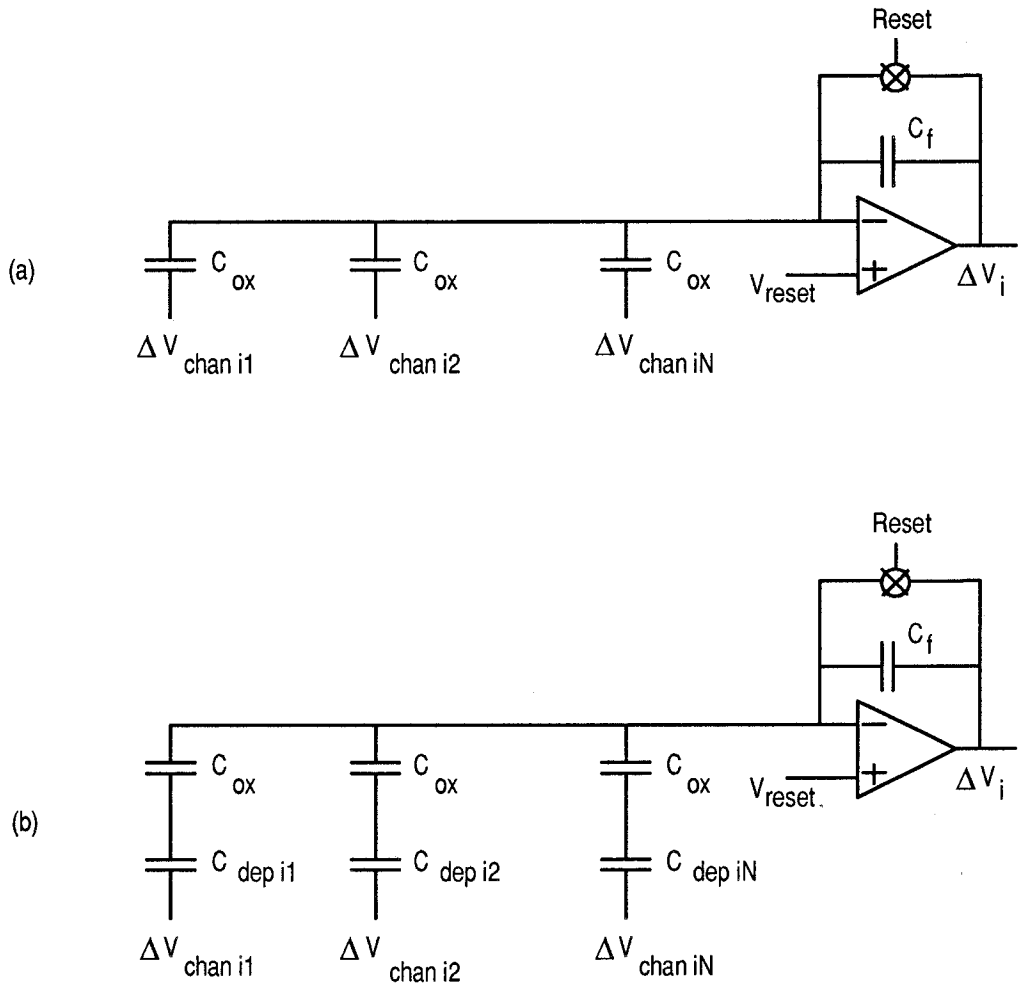

Figure 4.7. The feedback sense amp capacitive model includes a depletion depth dependent term for the buried channel model (b) whereas the surface channel model (a) has no such nonlinear elements. If the matrix charges are loaded so that they result in a linear channel potential change, the surface channel CID VMM of Figure 4.6 is linear.

For surface channel devices, depicted in Figure 4.7(a), the capacitance between the channel and the row line is just C_{ox}, which is linear. For the buried channel device of Figure 4.7(b), however, the channel to row line capacitance depends on the depletion depth of the channel. Given that the charges are input electrically, for the surface channel device the output of the binary vector/analog matrix multiplication is

$$\Delta V_i = \sum_{j=0}^{N-1} V_{fsc}(U_j Q_{sc}(V_{ij}^{(in)})) = \sum_{j=0}^{N-1} U_j V_{fsc}(Q_{sc}(V_{ij}^{(in)})) \tag{4.2}$$

using the charge to voltage characteristic for the feedback sense amp, $V_{fsc}(Q)$. The buried channel equivalent to Equation 4.2 using the transfer characteristic of Figure 2.21 is

$$\Delta V_i = \sum_{j=0}^{N-1} V_{fbc}(U_j Q_{bc}(V_{ij}^{(in)})) = \sum_{j=0}^{N-1} U_j V_{fbc}(Q_{bc}(V_{ij}^{(in)})). \tag{4.3}$$

The binary U_j term can be brought outside the nonlinearity because both $V_{fsc}(Q)$ and $V_{fbc}(Q)$ have the property that zero charge gives zero output voltage change.

Note that in both cases, the summation occurs without additional nonlinearities, allowing the use of compensation that was not possible with the semiparallel CCD processor of Chapter 3. The matrix input voltage can be pre-scaled by the use of a feedback linearization circuit to make the device immune to the CCD nonlinearities inherent in the Q-V and V-Q conversions. A feedback linearization circuit is shown in Figure 4.8, adapted from [Beynon et al., 1980].

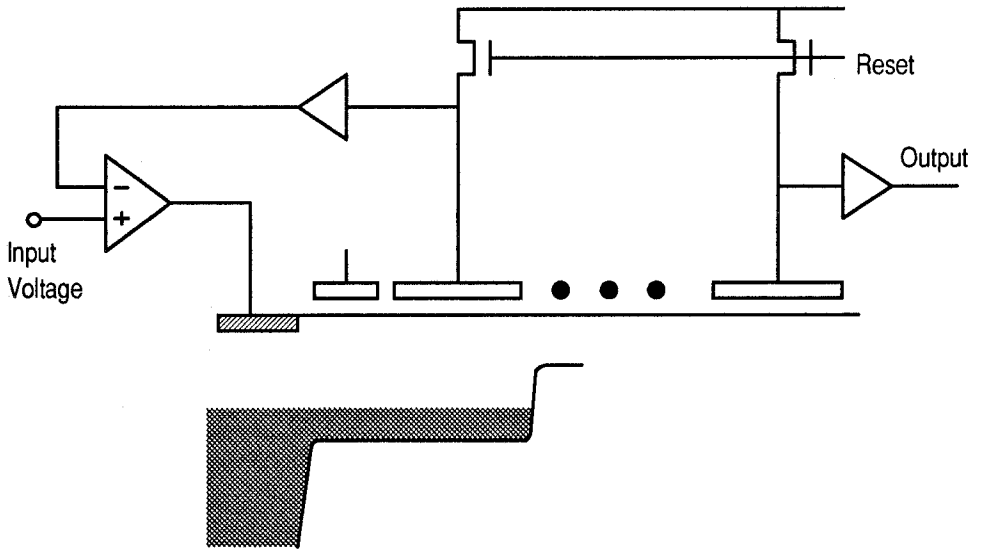

Figure 4.8. A feedback charge linearization circuit duplicates the output sense layout as close as possible in the fill-and-spill circuit. The sensed voltage on the second gate controls an opamp which adjusts the input diode level so that the charge under the second gate results in a voltage change equal to the input voltage.

In subsequent sections of this chapter, the following definitions are used

$$V_{ij} = V_{fsc}(Q_{sc}(V_{ij}^{(in)})) \quad \text{or} \quad V_{ij} = V_{fbc}(Q_{bc}(V_{ij}^{(in)})) \tag{4.4}$$

which account for the V-Q-V conversion done during the computation for either surface or buried channel devices. When feedback linearization is performed, the transform is simple a scaling factor.

4.5 Extending Input Vector Precision

Many applications such as image processing require multilevel input capability. This can easily be implemented using the architecture described in Section 4.2 in a 'bit-serial' mode [Neugebauer et al., 1990]. The operation of the device is identical to the description given above except that processing n-bit input precision requires n cycles of the device. Digital shift registers are added to each column input line which drive the column lines with successively more significant bits of the input vector, shown in Figure 4.9.

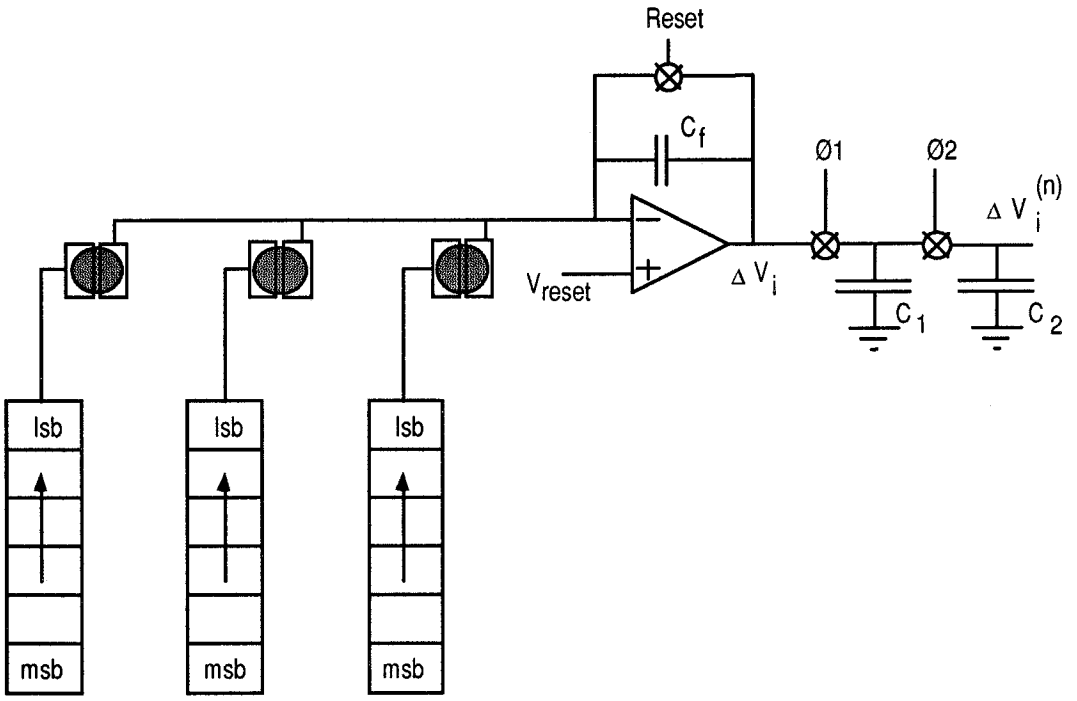

Figure 4.9. Adding a shift register to each input line and a divide-by-2 switched capacitor circuit to each output enables multivalued input vectors instead of only binary input. The operation is 'bit-serial' where the least significant bit (lsb) is first multiplied, resulting in a set of analog voltages at the output. The output vector is divided in half by the switched capacitors and the next most significant bit is multiplied. The analog result is added to the previous result, having twice the weight. By repeatedly summing and dividing, each bit receives the proper power of two weighting. An n-bit conversion takes n clock cycles, a direct speed/accuracy tradeoff.

Using the notation $U_j^{(n-1)}$ to represent the binary vector formed by taking the n^{th} bits of all the digital input elements, the first binary vector/analog matrix multiplication done by the circuit results in row voltage changes that are stored on capacitors at the end of each row, C_1 and allowed to share charge with another set of equally sized capacitors, C_2. This has the effect of dividing the row voltage changes in half, i.e.,

$$\Delta V_i^{(0)} = \frac{1}{2}\sum_{j=0}^{N-1} U_j^{(0)} V_{ij} \tag{4.5}$$

The next most significant binary input vector, $U_j^{(1)}$, is then presented and creates another set of row voltage changes which are stored on C_1 and shared with C_2 which gives

$$\Delta V_i^{(1)} = \frac{1}{2}\sum_{j=0}^{N-1} U_j^{(1)} V_{ij} + \frac{1}{4}\sum_{j=0}^{N-1} U_j^{(0)} V_{ij} \tag{4.6}$$

The is repeated n times, effectively weighting each successive bit's data by the proper power of two factor giving a total output voltage of

$$\Delta V_i^{(n-1)} = \sum_{j=0}^{N-1} V_{ij}\left(\sum_{k=0}^{n-1} 2^{k-n} U_j^{(k)}\right) = \sum_{j=0}^{N-1} V_{ij} D_j \tag{4.7}$$

after n clock cycles where D_j represents the multilevel digital words of the input vector. In this manner, multivalued input of n-bit precision can be processed where n is only limited by the analog accuracy of the components. If 4-bit input precision is required, the device is simply clocked four times. Since the power of two weighting is divisive, the most significant bit is always given the same weighting regardless of input word length.

4.6 Charge Storage Capabilities

In contrast to the semiparallel circuit of Chapter 3, the CID processor does not suffer from charge transfer efficiency problems. During sequential computations, the charge is repeatedly transferred locally between row and column gates of each matrix element. Incomplete charge transfer does not degrade performance since any charge left under the column gates in Figure 4.2(d) is picked up again in the next computation cycle when the charge packet is transferred back beneath the column gate. Only dark current accumulation presents a problem, requiring periodic refresh every fifty milliseconds or so at room temperature for moderate accuracy. When required, this effect can be easily countered by having a differential CID cell which stores two charges whose difference encodes the matrix element. Besides reducing common mode noise and providing four quadrant multiplication, such a cell can reduce the effects of dark current generation tremendously. Dark

current is fairly uniform over the spatial scale of the CID matrix cells, causing both charges of a differential matrix element to increase at roughly the same rate and canceling the dark current to first order. The CID storage represents a significant improvement over the storage employed by the semiparallel CCD circuit of Chapter 3. Clock speed can be much higher than the semiparallel CCD since poor CTE can be tolerated. Furthermore, due to the insensitivity to charge transfer losses, the device can be fabricated in the more linear and more accessible surface channel technology.

4.7 Input and Output Compatibility Issues

The extension of the input precision by using the device in a bit serial mode allows the input vector to be in digital form. This is advantageous from the system point of view where a digital interface eases the integration of an analog processing chip with standard digital memories and microcontrollers. The output, however, is in analog form. Making multilayer neural network systems as shown in Figure 1.1 requires an analog-to-digital conversion at the output of the CID chip. A number of circuits have been designed to accomplish this goal using both CCD and switched capacitor circuitry. Another alternative is to employ a competitive analog network, i.e., a 'winner-take-all' circuit, immediately after the vector matrix multiplication which activates only the output element with the greatest value. This type of network functions as a classifier, telling which row of the matrix most closely matches the input vector using a correlation measure. The classification network has the added benefit of having only one output, namely the address of the winner, instead of N outputs if the entire vector had to be communicated off chip. This reduction in output bandwidth can reduce power dissipation and ease pinout requirements significantly. This is in keeping with the popular view of neural networks as computational structures which take many inputs and reduce them to a few relevant quantities, reducing output bandwidth while retaining the important information.

4.8 Experimental Results

A number of circuits have been fabricated to test different variations of these ideas. Both buried channel and surface channel devices have been employed using a 2μm CCD/CMOS process with varying degrees of success. The largest CID processor built has a 128 by 128 matrix of buried channel charge storage elements, digital input and analog output and has been reported in [Neugebauer et al., 1992]. A set relevant parameters of this processor and the fabrication process are given in Table 4.1 and a chip photograph is shown in Figure 4.10. Other test versions, while not quite as large, were used to test the functionality of the surface channel devices and the capacitive feedback amplifiers.

Technology	2um CCD (Orbit Semiconductor/MOSIS) Double poly/double metal
Transfer Efficiency	0.999998
Number of Matrix Elements	16,384
Size	4.6mm x 6.7mm
Accuracy	Matrix Elements - 4 bits Vector Elements - 4 bits Output - 3 bits
Clock Frequency	4 MHz
Multiply-accumulates/sec	1.6×10^{10}

Table 4.1. A buried channel CID chip was fabricated early in this research containing 16,384 multipliers operating at 4MHz.

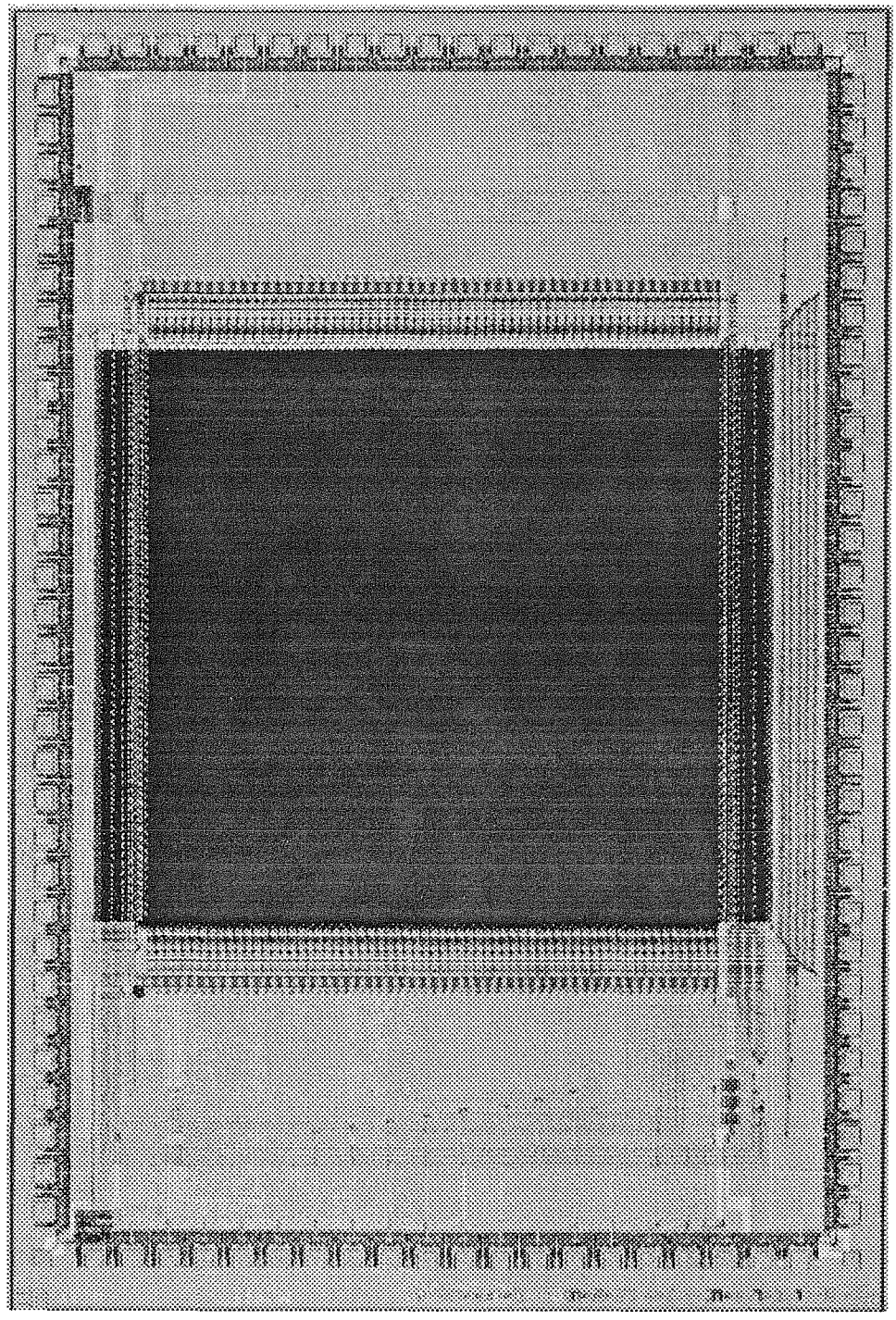

Figure 4.10. Chip photo of the IC of Table 4.1. It contains 16,384 CID cells implemented in 2um CCD/CMOS. The CID array measures 3.2mm x 3.2mm and was operated at 4MHz bit rate.

The matrix cell in most of the tested chips was 24μm by 24μm. A few later versions used a more aggressive interpretation of the 2μm CMOS design rules to squeeze this to 18μm by 20μm. For

the largest CID chip, the matrix contained 16,384 multiply-accumulate analog memory elements in a square 3.2mm on a side. The row and column lines were connected with aluminum to attain higher speeds. Later versions had metal straps for all four electrodes in each cell.

Many circuit variations were tested, including buried and surface channel devices. The most accurate was the surface channel feedback sensing device. The extended voltage range, 0v to 5v clocking and good linearity were clear benefits of this architecture. The equivalent circuit of a tested 64 cell surface channel CID processor is shown in Figure 4.11.

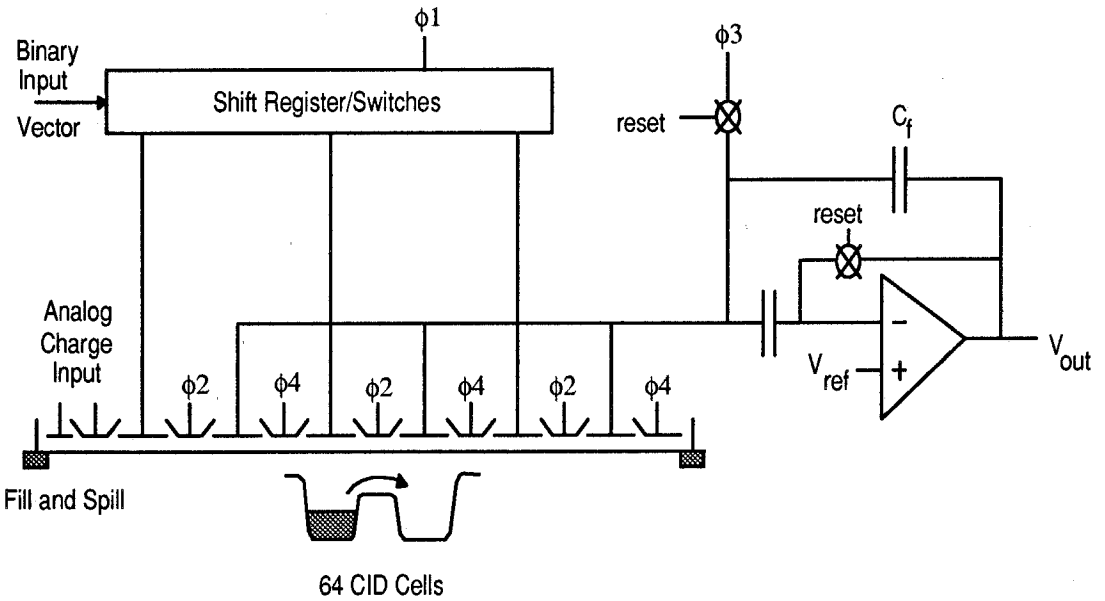

Figure 4.11. A 64 CID cell processor fabricated in a 2um p-well process was designed with a series capacitor coupling the sense line to the operational amplifier. The coupling capacitance holds a fixed dc bias during sensing that extends the output range of the amplifier considerably. The CID cell measured 18um x 20um.

The experimental setup is shown in Figure 4.12. The input vectors and matrix charges were controlled by the computer. The output data was sampled by a digital oscilloscope. Not shown is the extensive programmable clock generation circuitry. A number of tests, described below, were used to evaluate the relative performance of the different technologies and sensing methods. The data from these tests for the top performing surface channel device is given in Table 4.2.

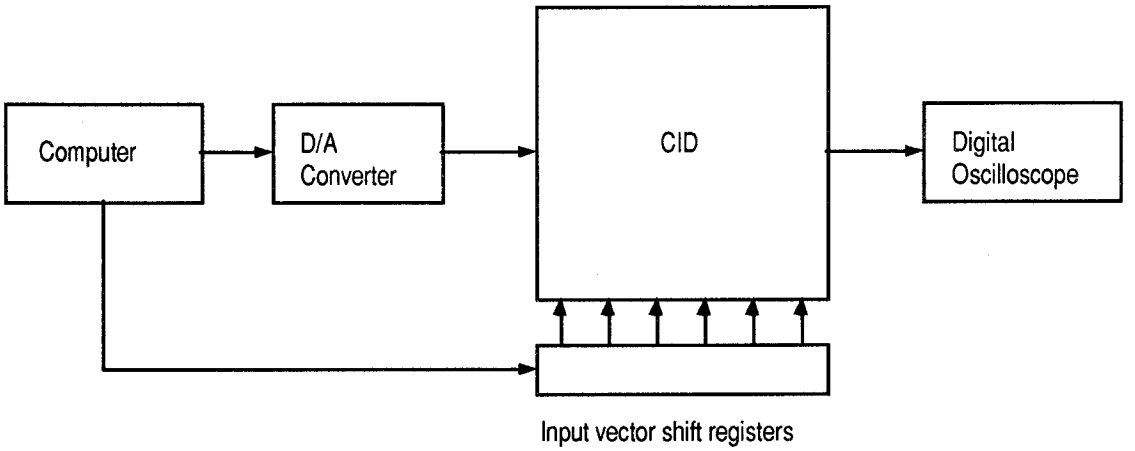

Figure 4.12. The experimental setup allows matrix charges and input vectors to be programmed from the computer which also controls all clock waveforms.

- Input & Output Voltage Range - The input/output voltage characteristic was measured for the device with the binary input vector completely 'on'. Choosing the most linear section of the transfer function gives the input and output voltage ranges along with a linearity measurement.

- Noise Floor - The output noise as a percentage of the full scale output swing represents the limitations of the device due to clock coupling, power supply noise, ground bounce, etc..

- Linearity - A calculation of the total RMS error of the transform shown in Figure 4.13 gives the linearity of the device as a transform engine.

Technology	2um CMOS (Orbit Semiconductor/MOSIS) Double poly/double metal
Transfer Efficiency	0.9998
Transform Length	64
Cell Size	18um x 20um
Accuracy	Matrix Elements - 6 bits Vector Elements - 6 bits Output - 6 bits
Noise Level	7 bits
RMS Error	<1.0% (limit of test)
Clock Frequency	1 MHz
Voltages	Power and Clocks between 0v and 5v Input Range - 2v Output Range - 1.5v

Table 4.2. The surface channel devices have much lower transfer efficiency but much better linearity than the buried channel devices of Table 4.1. The clock frequency was limited by digital shift register load rates, a curable problem.

As can be seen by Table 4.2, the surface channel feedback amplifier device offers much better accuracy compared to the buried channel device of Table 4.1. In addition, the surface channel device allows the future migration of all the clock driver circuits onto the same chip by virtue of its strictly positive clock voltage range.

A transform experiment was performed on the surface channel capacitive feedback device. Two binary signals with the same spatial frequency were convolved together using the circuit of Figure 4.11. A Walsh transform basis vector was convolved repeatedly with a signal constructed two out of phase copies of the same Walsh basis vector. The output results are plotted in Figure 4.13 along with a normalized computer generated plot of the numerically correct output. The measured RMS error was less than 1%, which is at the sensitivity limit of the experimental setup.

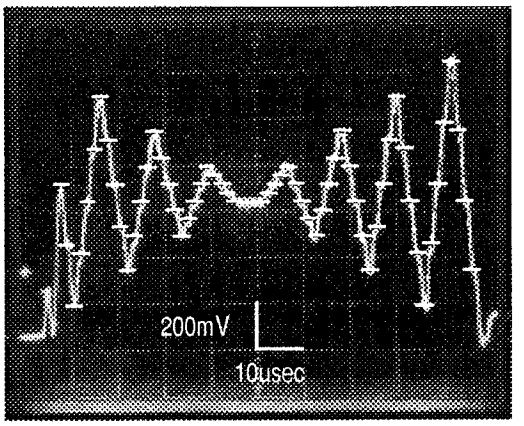

Figure 4.13. Two vectors with the same frequency were convolved, one shifted as a function of time. A half cycle phase shift in one of the vectors results in total cancellation in the middle. The computer simulated data points are also shown. The measured RMS error is <1.0%, limited by test equipment.

4.9 Power Dissipation

The total energy dissipated in performing a single computation can be calculated and used to compare the efficiency of this architecture to other dissimilar technologies, such as digital CMOS, optical, and even biology in Chapter 5. The majority of the power is dissipated by the column driver lines which move the charge from column to row gate. The energy required to execute the different phases of the computation depicted in Figure 4.3 is given approximately by

$$\Delta E = 2CV_{clock}^2 \tag{4.8}$$

where V_{clock} is the voltage swing of the row and column clocks. Using the voltages and parameters of the tested devices, with the average charge being 1/2 the full scale value setting the depletion capacitance depth and input vector bits being 'on' half the time, this energy is, on average, roughly 0.25 picojoule for the surface channel devices. This is the energy required to process one bit of the input vector -- a full 8-bit multiply-accumulate operation requires ~2pJ.

4.10 Comparison with Semiparallel Processor

The parallel CID processor has many advantages over the semiparallel CCD processor. The ability to compute the vector matrix multiplication in a single clock cycle in addition to using less power, providing more accuracy, eliminating the effects of transfer loss, and being more compact are clearly in the CID's favor. From a system point of view, the flexible digital input precision afforded by the CID and the capability of the CID to have all clock waveforms generated on-chip make the CID easier to integrate at the system level.

4.11 Future Directions

The CID processor as described above multiplies a positive digital vector by a positive matrix. Some applications require four quadrant multiplication with both vectors and matrix elements being positive and negative. Such a scheme is easily implemented, as alluded to before, by using differential charge storage for each matrix element. The equivalent schematic is shown in Figure 4.14 with a differential output amplifier. To achieve positive and negative input, each binary vector formed by taking the n^{th} bit of the input vector digital words is presented in its normal and complemented form. These two presentations are subtracted from each other by the operation of switches during ϕ_3 in Figure 4.14. The presentation of the normal and complement of each binary input vector allows the input vector to have positive and negative values. This implementation has the added benefits of canceling out dark current effects, diminishing the susceptibility to common mode noise due to power supply noise, ground bounce and column line voltage variations as well as compensating for amplifier offsets.

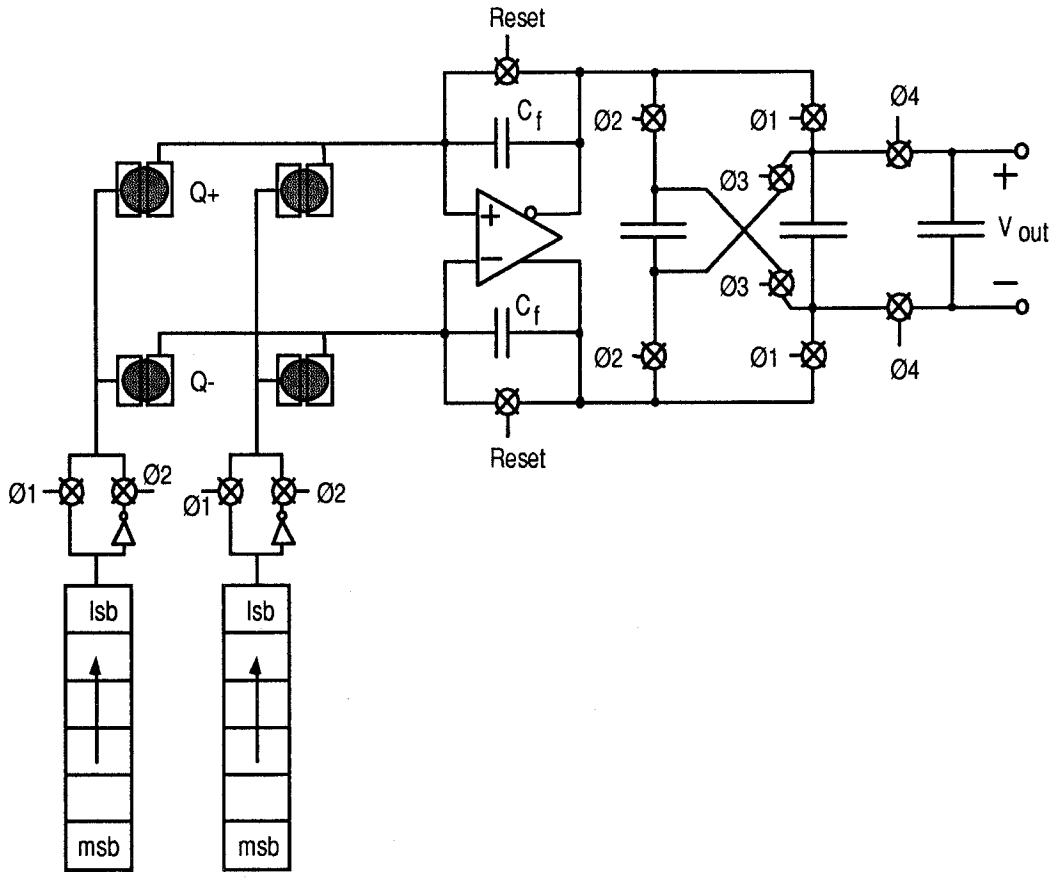

Figure 4.14. The differential CID stores two matrix charges per cell which are multiplied twice by each bit to achieve four quadrant multiplication. Accuracy and linearity are improved significantly with this scheme.

Many directions exist for exploring the potential of these devices in a system context. The integration of the clock generator circuitry onto the same chip as the processor is possible with the CID processor if surface channel technology is used. Including a 'winner-take-all' circuit at the output forms a classifier chip with numerous applications. Integrating the CID processor on the same chip as a CCD imager for simple edge enhancement and pattern recognition primitives is another potential direction. By scaling the devices as in [Yau, 1976], finer line lithography processes can be utilized and much higher CID densities achieved.

4.12 References:

[Agranat et al., 1987] A.J. Agranat and A. Yariv, "Semiparallel microelectronic implementation of neural network models," *Electronic Letters*, vol. 23, pp. 580-581, 1987.

[Agranat et al., 1988] A.J. Agranat, C.F. Neugebauer and A. Yariv, "Parallel optoelectronic realization of neural network models using CID technology," *Applied Optics*, vol. 27, pp. 4354-4355, 1988.

[Anagnostopoulos, 1978] C. Anagnostopoulos, "Signal readout in CID image sensors," *IEEE Transactions on Electron Devices*, vol. ED-25(2), pp. 85-89, 1978.

[Arnold et al., 1971] E. Arnold, M.H. Crowell, R.D. Geyer and D.P. Mathur, "Video signals and switching transients in capacitor-photodiode and capacitor-phototransistor image sensors," *IEEE Transactions on Electron Devices*, vol. ED-18, pp. 1003-1010, 1971.

[Barbe, 1976] D.F. Barbe, "Charge-coupled device and charge-injection device imaging," *IEEE Transactions on Electron Devices*, Vol. ED-23(2), pp. 177-182, 1976.

[Beynon et al., 1980] J. Beynon and D. Lamb, *Charge Coupled Devices and Their Applications*, McGraw-Hill, London, 1980.

[Brown et al., 1978] D.M. Brown, M. Ghezzo and P.L. Sargent, "High density CID imagers," *IEEE Transactions on Electron Devices*, vol. ED-25(2), pp. 79-84, 1978.

[Burke et al., 1976] K.K. Burke and G.J. Michon, "Charge-injection imaging: operating techniques and performance characteristics," *IEEE Transactions on Electron Devices*, vol. ED-23(2), pp. 189-195, 1976.

[Grafinger et al., c. 1982] A.B. Grafinger and G.J. Michon, "Review of charge injection device (CID) technology," *General Electric Optoelectronic Systems Operation*

Application Note, General Electric Co., Syracuse, NY, c. 1982.

[Matsunaga et al., 1991] Y. Matsunaga, H. Yamashita and S. Ohsawa, "A highly sensitive on-chip charge detector for CCD area image sensor," *IEEE Journal of Solid State Circuits*, vol. SC-26(4), pp. 652-656, 1991.

[Matsunaga et al., 1991] Y. Matsunaga, H. Yamashita, S. Manabe and N. Harada, "A high-sensitivity MOS photo-transistor for area image sensor," *IEEE Transactions on Electron Devices*, vol. ED-38(5), pp. 1044-1047, 1991.

[Michon et al., 1973] G.J. Michon and H.K. Burke, "Charge injection imaging," in *International Solid State Circuits Conference Digest of Technical Papers*, Philadelphia, pp. 138-139, 1973.

[Michon et al., 1974] G.J. Michon and H.K. Burke, "Operational characteristics of CID imager," in *International Solid State Circuits Conference Digest of Technical Papers*, Philadelphia, pp. 26-27, 1974.

[Michon et al., 1975] G.J. Michon and H.K. Burke, "Recent developments in CID imaging," in *Symposium on CCD Technology for Scientific Imaging Applications*, Pasadena, CA, 1975.

[Michon et al., 1980] G.J. Michon and H.K. Burke, "CID image sensing," in *Topics in Applied Physics, Volume 38: Charge-Coupled Devices*, 1980.

[Neugebauer et al., 1990] C.F. Neugebauer, A. Agranat and A. Yariv, "A charge domain bit serial vector-matrix multiplier and method thereof," U.S. Patent Application #07-552,772, filed 1990.

[Neugebauer et al., 1992] C.F. Neugebauer and A. Yariv, "A parallel analog CCD/CMOS signal processor," in *Advances in Neural Information Processing Systems*, Eds. J.E. Moody, S.J. Hansen and R.P. Lippmann, Morgan-Kaufmann, San Mateo, pp. 748-755, 1992.

[Séquin et al., 1975] C. Séquin and M Tompsett, *Charge Transfer Devices*, Academic Press, New York, 1975.

[Yau, 1976] L.D. Yau, "Scaling of surface-channel charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. ED-23(2), pp. 282-287, 1976.

Chapter 5

5. COMPARISONS

5.1 Analog vs. Digital

Many different technologies are capable of performing a generic computation such as a vector matrix multiplication. An interesting comparison can be made between digital and analog silicon technology, since they both employ the same technological base. Silicon process technology has been driven by the memory industry, which for the past twenty or so years has enjoyed an unparalleled development schedule of quadrupling memory density every three years. As both analog and digital designs can take advantage of the finer lithography and faster devices, comparisons tend to center around silicon usage efficiency and power efficiency. For the same given computation and fabrication process, the comparison of area-delay and power-delay products for analog and digital implementations is of interest. Depending on the nature of the problem and process, either digital or analog can be the winner. From an industrial point of view, characteristics such as time-to-market, manufacturability and availability cause digital to dominate, but for this discussion only performance criterion are examined.

The types of problems for which analog computation is appropriate can be segregated from those for which digital implementations are more appropriate. To begin with, the general purpose computer will never be analog. One prevalent characteristic of digital designs is the flexibility afforded by their serial, instruction oriented operation. Modifying functionality by reprogramming is a great strength of digital processors in that the manufacturer can make a chip whose ultimate use is not predetermined at fabrication time. Most analog chips, on the other hand, are inflexible by comparison. Analog computation chips are typically designed to perform a specific computation which is completely understood at design time. An operational amplifier cannot be reprogrammed to act as a capacitor.

Another immediately apparent difference is the ability of digital designs to be of arbitrary accuracy. Adding more accuracy, i.e., more bits, to a digital processor is a straightforward scaling issue. Squeezing a few more dB out of an analog chip, in contrast, is much more time consuming. The only types of problems in which analog computations are appropriate are low resolution tasks with accuracy on the order of 8 bits or less. Otherwise, the area and power requirements of extra precision typically negate any advantages analog may have over digital.

Lest this comparison look too bleak for analog, there is a large body of problems which are ideally suited for analog electronics. Any real world sensory data sources, i.e., audio, video, etc., are analog signals of limited precision. Sensory computation problems are typified by a large amount of highly correlated data of low precision where the task of recognition or processing usually involves extracting the correlations of interest. While the flexibility of digital processors make them applicable to any computation problem, digital technology is not a good match for the data precision, bandwidth requirements and biologically inspired solutions to the problem of sensory processing. It is in this realm that analog computers can live up to their potential. A certain amount of the 'chicken and the egg' phenomena is present, however, in that the types of

computational problems people work on have been influenced by the predominantly digital means of solution. The success of massively parallel analog computation for sensory processing depends on the field attaining a critical mass where algorithms, applications and technologies are developed in parallel.

A simple generalization can be made regarding digital and analog computing. With the applications restricted to low accuracy, highly parallel tasks where a special purpose, non-programmable processor can work, analog can often implement the same function as digital in a fraction of the area and consume significantly less power. By exploiting the device physics to perform a computation, analog devices have a significant advantage over digital implementations. This fundamental advantage stems from the fact that digital logic has only two signal levels, whereas analog devices utilize the whole dynamic range of the devices. By using each transistor and wire to handle more information, the density of analog is increased and the power reduced. The benefits are not without cost, since digital designers do not have to worry about clock lines crossing signal lines or other such noise issues.

5.2 Other Technologies for Analog Computation

Many technologies have been used to perform parallel analog computation, such as switched capacitor CMOS, EEPROM, and volume holography, to name a few. All technologies have their unique characteristics which can be detrimental or beneficial, depending on the computational problem at hand. A large ongoing effort aimed at exploiting silicon technology for analog computation has produced a number of interesting implementations of neural inspired functions, such as vector matrix multiplication. The work at Intel [Intel, 1991] on the ETANN chip, a vector matrix multiplier with sigmoidal nonlinearity using EPROM analog storage is a good example of the processors being built.

More often than not, these devices are built outside the scope of a particular application, an attempt at a 'general purpose' analog processor. To accurately gauge the computational power of these devices, they must be viewed in a system context to illuminate all communication, memory, and flexibility limitations. However, pursuing system architectures and algorithms requires more resources than just device work, a fact that has limited application oriented research to resource rich environments, typically in industry. The work at AT&T [Boser et al., 1991] is significant in that a particular problem, namely digit recognition for the Postal Service, has been selected and requisite algorithms were identified before silicon was built. This enabled the chip designers to pay more attention to system level integration aspects such as bandwidth, pinouts, etc. that are often ignored or given short treatment in the general purpose designs. This approach has also been followed by a number of other companies and institutions to varying degrees of success.

5.3 Comparison of the CID Processor with Other Technologies

While making a comparison outside the context of a particular system application is of limited usefulness, it can illuminate the relative merits of the various technologies. For the given computation of vector matrix multiplication, the base unit of computation is the multiply-accumulate operation. The two relevant comparison quantities are energy per operation (i.e., power-delay product) and silicon area per operation (i.e., area-delay product). Table 5.1 lists a number of different computational structures, from digital processors to specialized analog devices such as the ETANN chip and the CID processor. Tables like 5.1 are outdated before anyone reads them -- however, the table clearly shows where the strengths of the CID processor and analog technology in general with respect to more traditional computational methods. The power dissipation numbers are for the whole chip, including output driver power. For specialized hardware implementations of low resolution parallel processing, analog has a clear advantage.

The conventional wisdom that efficiency depends on how well the computation is matched to the architecture and implementation certainly holds in this case. It is important from an applications point of view to pursue computations where low power, low cost, highly parallel processing is needed.

	Technology	Accuracy (bits) (A = Analog, D = Digital)			Synapses	Clock (MHz)	Power (W)	Power*Delay (pico Joules)	Area*Delay (usec*um^2)	Other
Analog		Matrix	Input	Output						
[Boser et al., 1991]	0.9um CMOS	6 A	3 D	3 D	4,096	5	0.5	25	1,000	Programmable topology, cap. storage, digital I/O
[Intel et al., 1991]	1.5um EEPROM/ CMOS	6 A	6 A	6 A	10,240	0.33	1.5	440	8,000	Nonvolatile weight storage, prog. topology, analog I/O
[Chiang, 1991]	2um CCD/CMOS	8 D	8 A	8 A	980	10	2	200	3,000	Digital weights, analog I/O
[Neugebauer et al., 1992]	2um CCD/CMOS	6 A	4 D	3 A	16,384	1	0.1	6	1,000	Analog weights, digital input, analog output
[Arima et al., 1992]	0.8um CMOS	5 A	1 D	2 D	40,000	1	4	100	5,000	Boltzmann machine
Digital										
[Uchimura et al., 1992]	0.8um CMOS	8 D	8 D	8 D	832	16	0.3	25	11,000	Low power digital with algorithmic benefits
[Uramoto et al., 1992]	0.8um CMOS	12 D	9 D	9 D	6	100	3	5,000	35,000	Discrete cosine transform engine
[Newell et al., 1991]	1.6um CMOS	8 D	8 D	8 D	16	40	1	1,500	150,000	Digital neural processor, mostly memory area.

Table 5.1. A comparison of silicon based parallel processors.

5.4 References:

[Arima et al., 1992] Y. Arima, M. Murasaki, T. Yamada, A. Maeda and H. Shinohara, "A refreshable analog VLSI neural network chip with 400 neurons and 40k synapses," *IEEE Journal of Solid State Circuits*, vol. SC-27(12), pp. 1854-1861, 1992.

[Boser et al., 1991] B.E. Boser, E. Sackinger, J. Bromley, Y. Le Cun and L.D. Jackel, "An analog neural network processor with programmable topology," *IEEE Journal of Solid State Circuits*, vol. 26(12), pp. 2017-2025, 1991.

[Chiang, 1991] A.M. Chiang, "A CCD programmable image processor and its neural network applications," *IEEE Journal of Solid State Circuits*, vol. SC-26(12), pp. 1894-1901, 1991

[Intel, 1991] Intel Corporation, "80170NX electrically trainable analog neural network," *Product Specification*, 1991.

[Kub et al., 1990] F.J. Kub, K.K. Moon, I.A. Mack and F.M. Long, "Programmable analog vector-matrix multipliers," *IEEE Journal of Solid State Circuits*, vol. SC-25(1), pp. 207-214, 1990.

[Neugebauer et al., 1992] C.F. Neugebauer and A. Yariv, "A parallel analog CCD/CMOS signal processor," in *Advances in Neural Information Processing Systems*, Eds. J.E. Moody, S.J. Hansen and R.P. Lippmann, Morgan-Kaufmann, San Mateo, pp. 748-755, 1992.

[Newell et al., 1992] M. Newell and J. Rasure, "A VLSI system for real-time linear operations and transforms," *IEEE Transactions on Signal Processing*, vol. 39(8), pp. 1914-1917.

[Uchimura et al., 1992] K. Uchimura, O. Saito and Y. Amemiya, "A high-speed digital neural network chip with low-power chain-reaction architecture," *IEEE Journal of Solid State Circuits*, vol. SC-27(12), pp. 1862-1867, 1992.

[Uramoto et al., 1992] S. Uramoto, Y. Inoue, A. Takabatake, J. Takeda, Y. Yamashita, H. Terane and M. Yoshimoto, "A 100-MHz 2-D discrete cosine transform core processor," *IEEE Journal of Solid State Circuits*, vol. SC-27(4), pp. 492-499, 1992.

Chapter 6

6. SUMMARY

This thesis introduces a number of innovations tied to the CID processor, which grew out of work on the semiparallel CCD processor. A number of U.S. patent applications have been filed as a result of this work [Agranat et al., 1991][Neugebauer et al., 1990][Yariv et al., 1991]. Using standard process technology, the CID processor implements an extremely low power and high speed vector matrix multiplication. A novel bit serial method of achieving flexible digital input without much additional circuitry has been demonstrated for the first time. The bit serial technique is generally applicable to most vector matrix multiplication technologies where it can extend dynamic range and offer significant system integration benefits by affording digital input. A 128x128 vector matrix multiplication chip has been fabricated and tested, along with numerous other circuits which explore the effects of different CCD technologies and circuit configurations. A formalism for analyzing the effects of the nonlinearities associated with CCD circuits has been developed and applied to the CID device to eliminate errors.

In general, the field of parallel analog computation has significant applicability towards problems associated with the processing of real world sensory data. The CID processor's power and area efficiency significantly extends the state of the art in parallel analog processing and will enable the

exploration of new algorithms and architectures.

6.1 References:

[Agranat et al., 1991] A.J. Agranat, C.F. Neugebauer and A. Yariv, "Parallel optoelectronic neural network processors," U.S. Patent #5,008,833, 1991.

[Neugebauer et al., 1990] C.F. Neugebauer, A. Agranat and A. Yariv, "A charge domain bit serial vector-matrix multiplier and method thereof," U.S. Patent Application #07-552,772, filed 1990.

[Yariv et al., 1991] A. Yariv, C.F. Neugebauer and A.J. Agranat, "Non-destructive charge domain multiplier and process thereof," U.S. Patent #5,054,040, 1991.