Copyright © 1981, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

CHARGE CIRCUITS FOR ANALOG LSI

by

R. H. McCharles

Memorandum No. UCB/ERL M81/24

March 1981

(cover)

CHARGE CIRCUITS FOR ANALOG LSI

by

Robert H. McCharles

Memorandum No. UCB/ERL M81/24

March 1981

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

CHARGE CIRCUITS FOR ANALOG LSI

Robert H. McCharles

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

ABSTRACT

A charge circuit is a clocked analog circuit which is amenable to realization in large-scale integrated (LSI) form. Analog functions including arithmetic, delay, and many time-dependent functions can be realized with precision approaching 0.1%. Complex charge circuits are best designed from algorithmic representatives of the desired function. Error-correcting algorithms can be applied to improve on the inherent accuracy if desired.

Metal-oxide-semiconductor technology is most suitable for implementation of charge circuits. Experimental results for recursive or cyclic analog-to-digital converter are reported. A cyclic converter with 8-bit inherent accuracy was used with an error-correcting algorithm to achieve 12 bit resolution and accuracy. Conversion time was 200 microseconds. Die area for this converter totaled less than 9 mm^2.

Research sponsored by the National Science Foundation Grants ENG78-11397 and ENG73-04184-A01.

Table of Contents

Chapter I

Chapter II

Chapter III

Chapter I

Charge Circuits

Large scale integration has substantially reduced the cost of digital logic, but it has had less effect on the cost of analog circuits. There are several reasons for this. While digital systems are constructed of easily-integrated standard modules such as gates, memories and arithmetic units, standard analog modules of comparable utility do not exist. While digital circuits do not require adjustment, analog circuits usually require trimming. While digital circuits use a clock to generate time-dependent functions, analog circuits use RC or LC products which are difficult to integrate. These problems have limited the complexity of analog integrated circuits.

These problems are not inherent in analog circuits, but result from the techniques used to integrate them. They can be reduced by using the clocked analog circuits discussed here. We will call these circuits "charge circuits." Before we consider what this term means we will give some examples.

A Charge Circuit Filter

Fig 1 shows a simple charge circuit filter. This filter is the charge circuit equivalent of a single-pole RC filter [1,13]. Fig 1c shows a timing diagram and the circuit response to a step input. The approximate exponential decay is shown by the dashed line, and the equivalent RC circuit is shown in fig 1b.

In this circuit, as in all other circuits in this chapter, we will assume a two-phase nonoverlapping clock. The clock phases will be called PHI1 and PHI2. Positive voltage, switch closure and a logic one will all be equivalent.

The two switches of the filter are connected to the clock phases as shown, so that in each clock cycle a charge of $(V_{in}-V_{out})*C$ is transfered from C_1 to C_2. If the clock period is T then the average current is $(V_{in}-V_{out})*C_2/T$. If T is large compared with the maximum frequency of interest, the combination of S_1, S_2 and C_2 acts like a resistor of value T/C_2.

Despite its simplicity, this charge circuit filter demonstrates significant features of our approach. Before we consider these features, we will consider another simple cirucit.

A Charge Circuit Digital-to-Analog Converter

Fig 2 shows a charge circuit digital-to-analog converter and an example of the circuit operation. To begin conversion the converter is cleared by shorting both S_b and S_c. Switch S_c is connected to PHI2. During the following cycles, the digital input will determine whether S_a or S_b is closed during each PHI1 clock phase, starting with the least significant bit. If this bit is a one, switch S_a is closed; if this bit is a zero, switch S_b is closed. During the PHI2 phases both S_a and S_b are open, and charge is shared between C_1 and C_2. At the end of cycle N+1 an N-bit result has been obtained in C_2. This circuit is sometimes called the charge-equalizing digital-to-analog converter [2], and it is the charge circuit equivalent of the Shannon-Rack decoder [3].

The Charge Circuit

Now that we have seen some examples of "charge circuits," we can define what the term means. Charge circuits are clocked analog circuits which use the clock as a reference for time-dependent functions. Charge circuits use precision matched capacitors to replace matched resistors. The relevant circuit equations for charge circuits are conservation of charge and Kirkoff's voltage law instead of Kirkoff's current and voltage laws. For this reason, among others, computer-aided design programs are often ineffiecient, and they may give erroneous results.

The motivation for the development of charge cirucits has been the desire to implement analog functions in MOS technologies. These technologies produce high-quality capacitors, but resistors have wide tolerances and values which are often too low to be useful. Capacitors also have advantages in linearity and temperature sensitivity [2]. A deceptively large number of analog functions can be performed with only capacitors and switches, but the majority of practical circuits will also require amplifiers. To see why this is so we will have to examine what circuit functions are required in analog systems.

Analog Circuit Functions

Analog circuits perform only a small number of basic operations. for example, the charge cirucit filter of fig 1 performs multiplication by $C_1/(C_1+C_2)$, adding C_2*V_{in} and delay. The charge cirucit digital-to-analog converter performs the same basic operation but also performs discharging to zero. These operations are controlled digitally by analog switches.

Typical basic operations used to perform some analog functions in both continuous and sampled-data fashion are shown in Table I. Notice that the continuous-time operation of integration is replaced by the discrete-time operations of add and delay, and the continuous-time operation of multiply by 2 can be replaced by a multiply by two and delay repeated N times. This allows us to reduce the number of basic operations that sampled-data analog circuits must perform. It is true in general that the basic operations of add, subtract, multiply by fixed ratios and delay can be combined under digital control to perform any analog function.

Need for Amplifiers

Since the basic operations of add, multiply by fixed ratios and delay can be performed under digital control by bucket-brigade circuits and CCD's it is plausible to adapt such devices for general purpose analog functions without the need for amplifiers. This line of research lead to the charge multiplier [4]. Unfortuately, this reasoning is chimerical. There are three reasons why amplifiers will almost always be required in discrete-time analog systems.

One reason for using amplifiers is that they are needed to perform some of our basic operations. Bucket-brigade circuits and charge multipliers operate by charge conservation, and they have charge gain less than unity. Without gain we cannot implement multiplication by two, nor can we implement subtraction (since this would enable gain greater than one). Multiplication by two simplifies analog-to-digital converters. Multiplication by numbers greater than one is important for most filters. Subtraction is necessary both for recursive filters and for bipolarity in analog-to-digital and digital-to-analog converters. A lack of amplifiers restricts the functions we can perform.

Another reason for using amplifiers in charge circuits is for parasitic suppression. Integrated circuits have parasitic capacitance to the substrate. These parasitic effects are nonlinear and reduce the precision of our circuits. If we use capacitors in amplifier feedback loops the parasitic effects can be reduced by a factor of the loop gain.

Our final reason for using amplifiers is for buffering. The two circuits we showed at the beginning of this chapter do not require amplifiers. Neither do CCD's or bucket-brigade circuits. But to make these circuits functional we must sense the output. If these outputs are analog they will require buffering amplifiers.

The design of MOS analog amplifiers is developing rapidly. While charge circuits can be fabricated in bipolar technologies, they are more attractive in MOS technologies due to requirements for low amplifier input current, many analog switches and large logic swings to drive them. Amplifiers can be fabricated in the simplest digital MOS technolgies [5], but their performance is low. The performance of CMOS [7] and NMOS depletion-load [8] amplifiers is high enough to be used with the circuits discussed here.

The Clocked Analog Inverter

Amplifiers can perfom three functions: gain, parasitic suppression and buffering. These functions can be performed by a single circuit that adds, subtracts and multiplies by fixed ratios under digital control. In many ways this circuit is an analog equivalent of a digital gate.

The clocked analog inverter is shown in fig 3 [8]. The inverter has two inputs, V_a and V_b; two input capacitors, C_1 and C_2; switch S_e and the summing capacitor C_s. Switches S_a, S_b, S_c and S_d convert the continuous inputs V_1 and V_2 into the voltage transitions V_a and V_b. They are not part of the clocked analog inverter.

During the PHI1 clock phase switch Se is closed, shorting the output of the amplifier to its negative input. During the PHI2 clock phase, switch E is opened and any voltage transitions at inputs V_a and V_b will appear inverted and multiplied by C_1/C_s and C_2/C_s at the output. Fig 3b shows timing for the circuit performing its basic repetoire of operations.

Compatible Delay Element

During each clock cycle a clocked analog inverter has two transitions. The first transition has a negative sense, and the second transition has a positive sense and a half-cycle of delay. If the output of an inverter goes to another inverter clocked with the opposite phase, a full cycle of delay is produced. This gives the analog shift register shown in fig 4.

A timing example for the analog shift register is shown in fig 4b. Four inverters are connected to make a total of two clock cycles of delay. Both inverting and noninverting buffered outputs are available for connection to other circuits. Multiple weighted inputs can be added to any stage by ading extra input capacitors. This one circuit, the clocked analog inverter, performs add, subtract, multiply by fixed ratios and delay under digital control. Any sampled data analog function can be performed by clocked inverters and digital controlling logic.

Algorithms, Replication and Iteration

Discrete-time analog functions can be implemented using a two-step design process. First we specify a procedure, or algorithm, which performs the analog functions using our basic operations. This algorithm is the abstract model for the circuit operation. Second we identify each instance of performing a basic operation with the physical circuit element which performs it and the time at which it is to be performed. Circuits designed in this way will be called algorithmic circuits.

There are several methods available for specifying algorithms. We can use a language developed for that purpose, we can use flow charts [9] or we can use signal flow diagrams [10]. The last method is a natural means of describing algorithms for circuits. Signal flow diagrams for a digital-to-analog converter, an analog-to-digital converter and a second-order recursive filter section are shown in figs 5a, 5b and 5c. The N-bit digital-to-analog converter consists of N stages, each of which takes the output from the previous stage, multiplies it by one-half and adds or subtracts the one-half of the reference. The output from the last stage is an N-bit bipolar representation of the digital signal used to control the add/subtract funciton. The N-bit analog-to-digital converter consists of N stages and N comparators for determining the signs of the N outputs. Each stage multiples its input by two and adds or subtracts the reference depending on the sign of that input. The comparator outputs form an N-bit digital representation of the bipolar analog input to the first stage. The second order filter consists of two weighted sum-and-delay stages. By varying the values of r and THETA, the poles of the filter can be placed anywhere in the z-plane; in most cases r will be slightly less than one and THETA will be close to zero. These three algorithms were chosen for their simplicity. In practice, error properties and hardware complexity must be considered.

There are two simple ways of identifying our basic operations with the circuit parts which perform them. These ways will be called replication and iteration. Replication is a direct implementation of the signal flow by many separate circuits. Iteration implements the algorithm by repeated use of the same physical circuit. Replication and iteration can be combined to trade circuit complexity against speed. Replication is sometimes called "iteration in space" or "pipelining" while iteration is sometimes called "iteration in time." [11]

Replication

Figs 6, 7 and 8 show the algorithms of figs. 5a, 5b and 5c implemented by replication. The signal flow digrams of fig 5 are composed of sum-and-delay stages that are each realized by two clocked analog inverters. The N-bit digital-to-analog converter and the N-bit analog-to-digital converter are each realized with N of these two-inverter stages. The second order filter is realized with two of the two-inverter stages plus one extra inverter to produce the negative coefficient $-R_2$. Circuits designed by replication represent the extreme of maximizing speed by sacrificing chip area.

Analog Storage

Since second order filters have two state variables, analog storage is required to implement them by iteration. Analog storage is also required to implement more complex analog-to-digital and digital-to-analog conversion algorithms than those shown here.

Three methods of analog storage are shown in fig 9. Fig 9a shows a method of storage which uses switched input capacitors. This is the method used by McCreary [12] to store the input to his analog-to-digital converter. The method of switched summing capacitors was used by Young [8] to allow a single inverter to perform summation in his filter. Both of these methods have the advantage of requiring only one extra capacitor and one extra switch per analog word, and these methods are immune to amplifier offset. The method shown in fig 9b is also immune to low-frequency noise in the amplifier. The method shown in fig 9c is a buffered sample-and-hold circuit. This requires two switches and one amplifier for each word and does not provide offset immunity. The advantages of this storage method are that it provides a latched and buffered output, and that it has a gain of one independant of capacitor size. We will use this last method in our examples, and we will show how offset immunity can be provided by modifying the algorithm.

Iteration

By combining one sum-and-delay stage with analog storage and digital logic, any analog function can be performed. Figs 10, 11 and 12 show the same algorithms implemented by iteration.

A sample timing sequence is shown in fig 10b for the digital-to-analog converter. In the first clock cycle switch S_c is closed during the PHI1 phase. This places a zero in the delay stage. During the remaining N cycles, switch S_c is connected to PHI2, and the phasing of switches zero or a one. Input to the D/A converter is least-significant bit first.

A sample timing sequence for the analog-to-digital converter is shown in fig 11b. During the first cycle, switch S_d is closed during the PHI1 phase to zero the delay path. During the remaining N phases switch S_d is connected to PHI2. Switch S_a is closed during the PHI1 phase of the second cyle. S_a is opened and S_b closed during the PHI2 clock phase. This places a change in voltage of $-V_{in}$ at the inverter input between the trailing edges of PHI1 and PHI2. During the PHI2 phase of the second cycle, the comparator will show the sign of the input. This is the most-significant bit. During the remaining N cycles, the closing of switches S_b and S_c is phased to either add or subtract the reference to reduce the magnitude of the feedback signal. The sequence of comparator outputs gives a digital representation of the analog signal, most-significant bit first.

The timing diagram of fig 12 shows the iterated filter responding to an impulse. The second stage of delay is introduced by alternately storing the single-stage outputs in two storage registers. Subtraction is performed by switching the input capacitor from the input voltage to alternate storage registers on the transition from PHI1 to PHI2. The voltage transition at the sum-and-delay stage input is the difference between the input voltage and the output delayed by two clock cycles.

Analog Arithmetic Unit

The sum-and-delay stage can be used to construct an analog arithmetic unit capable of executing the same basic instructions as a digital arithmetic unit. Fig 13 shows how this can be done. The feedback capacitor can be switched between values of 0, C/2, C and 2C under digital control. This allows the feedback multiplier to take values of 0, 1/2, 1 and 2. These correspond to digital operations of clear, arithmetic right shift, store, and arithmetic left shift. These operations can be performed simultaneously with an add or subtract of an input. The resulting combinations of operations, together with mnemonic abbreviations are shown in Table II. Fig 13b is a sample timing diagram showing the register performing clear and add (CLA), multiply by one-half and add (RSA), subtract (SUB), and clear (CLR).

Variable Coefficients

We have so far assumed that we are free to choose the capacitor values in our designs. This is true if the design is a custom integrated circuit, but this is not true if we are designing or using standard parts. Analog-to-digital and digital-to-analog converters can be constructed by using gains of 1/2 and 2, but filters and other functions require multiplication by numbers which are not so standard. Variable coefficients can be realized by performing a multiplying D/A conversion using the analog quantity to be muliplied as a reference. This can be done by replication or by iteration.

Strict replication requires N two-inverter stages for an N-bit coefficient, but these can be merged into a single inverter with capacitor array as shown in fig 14. The output of the inverter can be switched to any value between $-2*V_{in}$ and $2*V_{in}$ in a single clock cycle. This array can be used with the delay element or the analog arithmetic unit to implement variable coefficients. McCreary has fabricated ten-bit arrays of this type with high yield [12].

The method of iteration allows us to use a storage register and a single sum-and-delay stage to perform the same mutiplication in N cycles. This is done by storing the input we wish to multiply in the register and then using the register as a reference in performing a digital-to-analog conversion as in fig 10. By combining the analog arithmetic register of fig 13 with three storage registers we can implement the second order filter at a cost of N+1 clock cycles per N-bit coefficient.

Error Correction

An important feature of algorithmic circuits is that we can invent algorithms that are insensitive to arithmetic errors. One way of doing this is to discover an algorithm which estimates the error made by another algorithm. These two algorithms can then be combined to create an algorithm which has higher accuracy. This is an alternative to having external adjustments or trimming on-chip components. We will give a simple example here. A more sophisticated example will be given in chapter III.

Fig 15 shows an analog arithmetic register connected with one storage register and three possible inputs: V_{in}, analog ground and the analog storage register. To sample the input we could perform a clear and add (CLA) and then store the result in the storage register. However both the analog arithmetic register and the analog storage register have offset errors which will be added to the result. These errors can be eliminated by performing instead the sequence of operations shown in fig 15b. This sequence first samples and holds a zero in the storage register and then samples and holds the difference between the input and the stored zero (which has offset error). In this way we can eliminate the need for an external trimming adjustment.

Conclusion

Charge circuits minimize some problems which have limited the complexity of analog integrated circuits. Matching of RC products is eliminated by using a clock as a time reference. Any analog function can be implemented using analog storage and clocked analog inverters. Algorithms used to define analog circuit functions can be implemented by iteration or by replication to minimize chip area or to maximize speed. Self-correcting algorithms can be found to eliminate the need for trimming. In many ways charge circuits are more similar to digital circuits than to conventional analog circuits, and programmable charge circuits can be constructed to perform general-purpose signal-processing functions.

References for Chapter I

[1] Fried, D. L.,

Analog Sample-Data Filters. IEEE Journal of Solid-State Circuits, Vol SC-7, No. 4, August 1972,pp 302-304.

[2] Suarez, R. E., Gray, P. R. and Hodges, D. A.

All MOS Charge Redistribution Analog-to Digital Conversion Techniques. IEEE journal of Solid-State Circuits, vol. SC-10, No. 6, pp 379-385, December 1975.

[3] Schmidt, H.

Electronic Analog/Digital Conversions. New York: Van Nostrand-Reinholt, 1970, pp 188-191.

[4] Albarran, J. F. and Hodges, D. A.

A Charge-Transfer Multiplying Digital-to-Analog Converter. IEEE Journal of Solid-State Circuits, vol SC-11, No. 6, pp. 772-778

[5] Tsividis, Y. P. and Gray, P. R.

An Integrated NMOS Operational Amplifier with Internal Compensation IEEE journal of Solid-State Circuits, vol SC-11, no. 6, pp. 748-753, December 1976.

[6] McCharles, R. H., Saletore, V. A., Black, W. C., and Hodges, D. A.
An Algorithmic Analog-to-Digital Converter ISSCC Digest of Technical Papers, February 1977, pp. 96-97.

[7] Senderowicz, D., Hodges, D. A. and Gray, P. R.
High-Performance NMOS Operational Amplifier. IEEE Journal of Solid-State Circuits, vol. SC-13 no. 6, December 1978, pp. 760-766.

[8] Young, I. A., Hodges, D. A. and Gray, P. R.
Analog NMOS Sampled-Data Recursive Filter. ISSCC Digest of Technical Papers, February 1977, pp 156-157.

[9] Oppenheim, A. V., and Schafer, R. W.
Digital Signal Processing. New Jersey: Prentice-Hall, 1975, pp 137 ff.

[10] Blakeslee, T. R.
Digital Design with Standard MSI and LSI. New York: 1975, Wiley Interscience, pp. 172-175.

[11] Ibid. pp. 233 ff.

[12] McCreary, J. L. and Gray, P. R.
All-MOS Charge Redistribution Analog-to-Digital Conversion Techniques -- Part I. IEEE journal of

Solid-State Circuits, vol SC-10, no. 6, pp 371-379, December 1975

[13] Hosticka, B. J., Brodersen, R. W. and Gray, P. R. MOS Sampled-Data Recursive Filters Using State Variable Techniques. IEEE Journal of Solid-State Circuits, december 1977, pp 600-608.

TABLE I

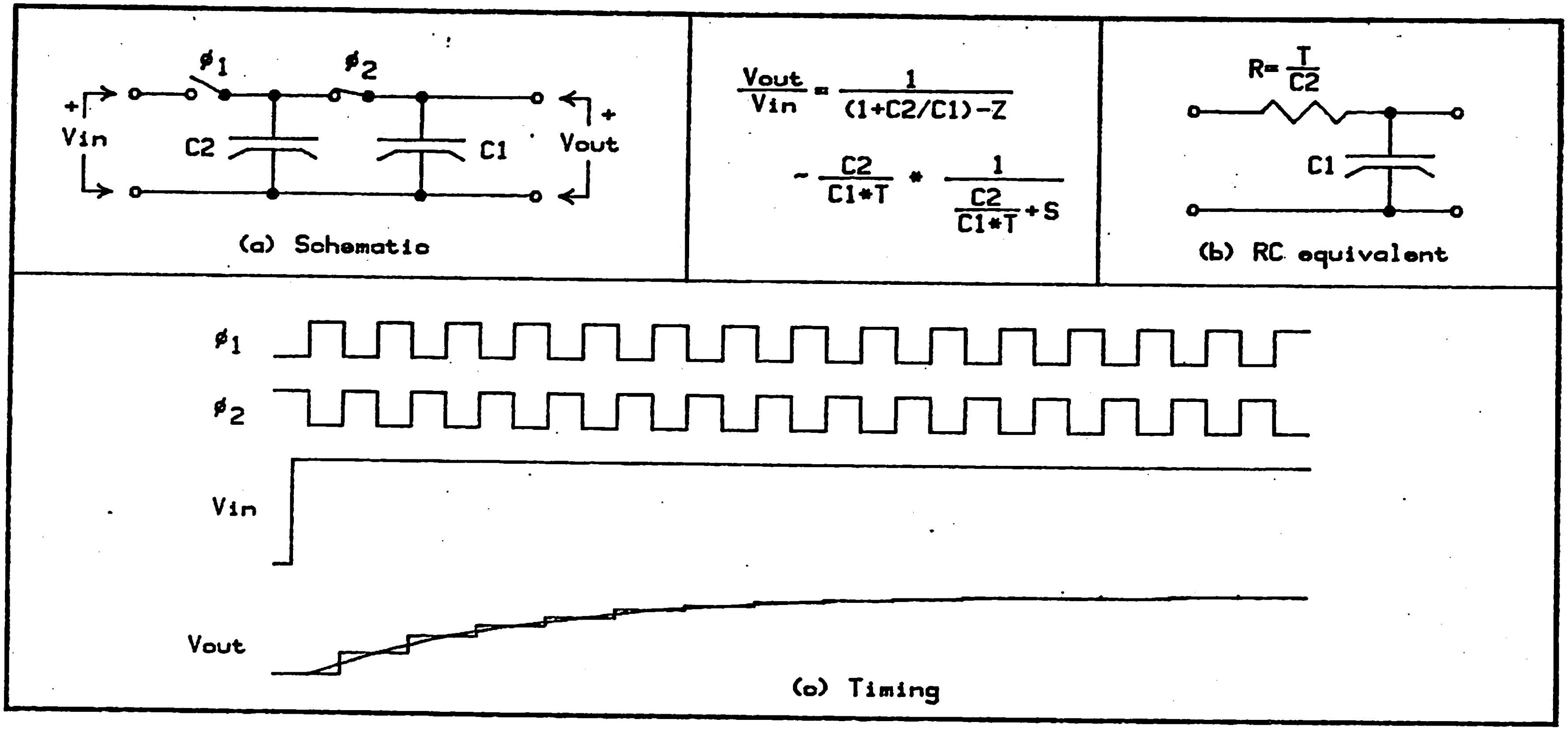
Continuous and Discrete-Time Operations

Function	Continuous-time Operation	Discrete-time Operation
Filtering	Integrate (1/S) Fixed ratio multiply Add Subtract	Delay (Z^{-1}) and add Fixed ratio multiply Add Subtract
A/D Convert	Multiply by 2^{-N} Add Subtract Compare	Multiply by 1/2 Delay Add Subtract Compare
D/A Convert	Multiply by 2^{-N} Add Subtract	Multiply by 1/2 Delay Add Subtract

TABLE II

Analog Arithmetic Unit Operations

Mnemonic	Description
ADD	Add
CLR	Clear
CLA	Clear and add
CLS	Clear and subtract
NOP	Delay
LSA	Multiply by two and add
LSH	Multiply by two
LSS	Multiply by two and subtract
RSA	Divide by two and add
RSH	Divide by two
RSS	Divide by two and subtract
SUB	Subtract



$$\frac{Vout}{Vin} = \frac{1}{(1+C2/C1)-Z}$$

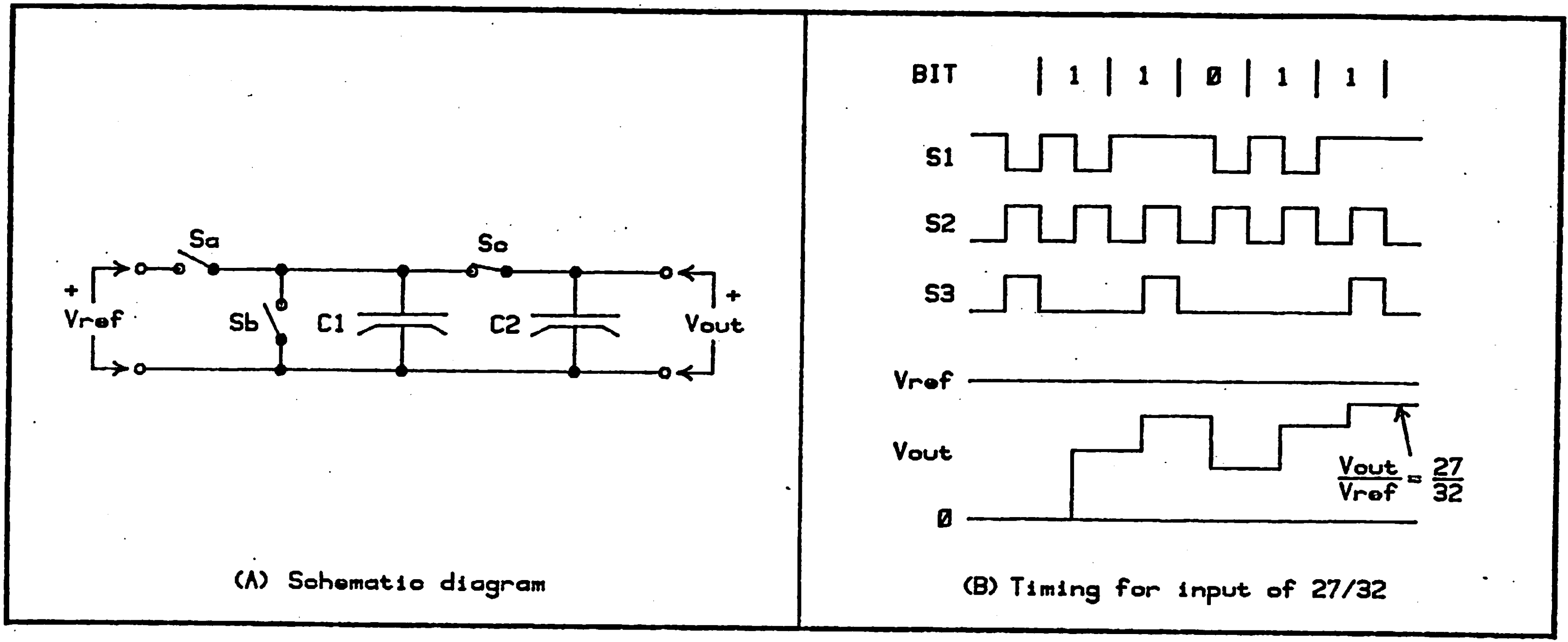
$$\sim \frac{C2}{C1*T} * \frac{1}{\frac{C2}{C1*T}+S}$$

(a) Schematic

(b) RC equivalent

(c) Timing

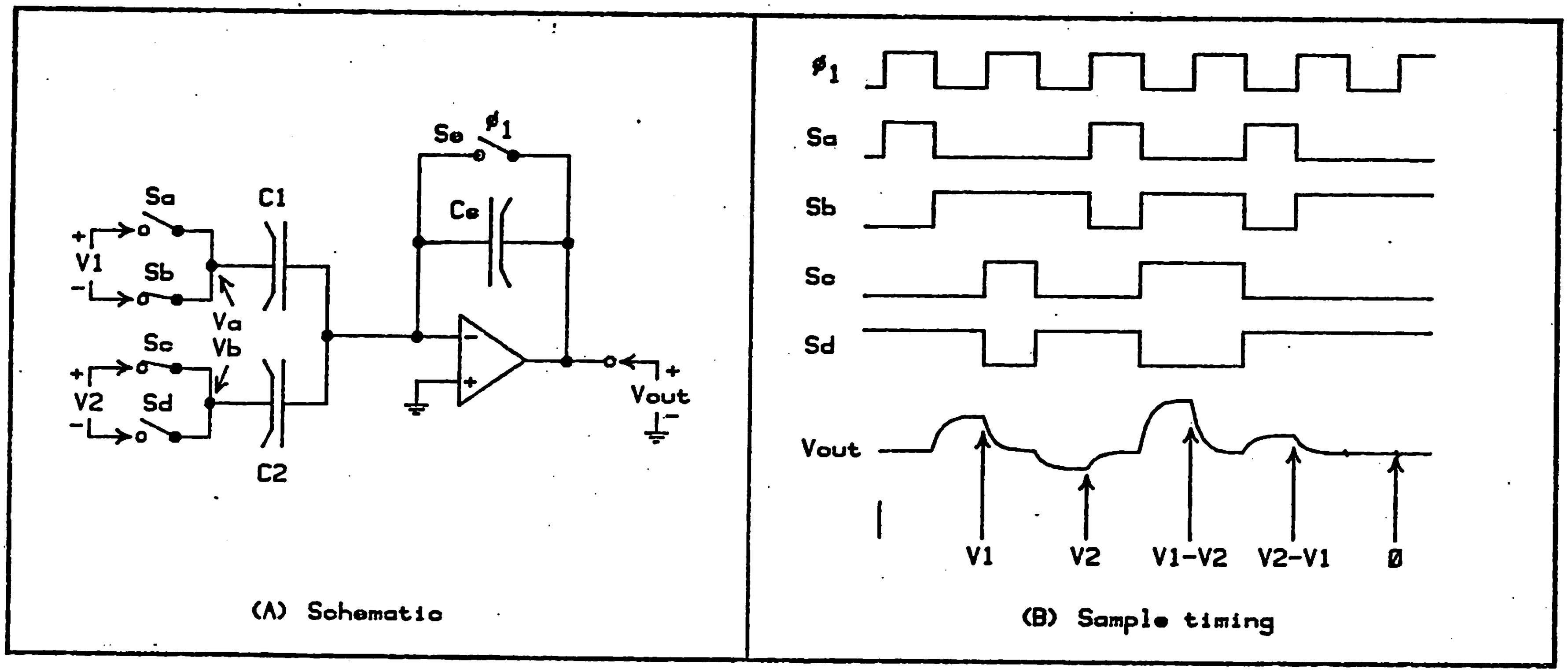
Fig. 1 -- Charge Circuit Filter



(A) Schematic diagram

(B) Timing for input of 27/32

Fig. 2 -- Charge Circuit D/A Converter



(A) Schematic

(B) Sample timing

Fig. 3 -- Clocked Analog Inverter

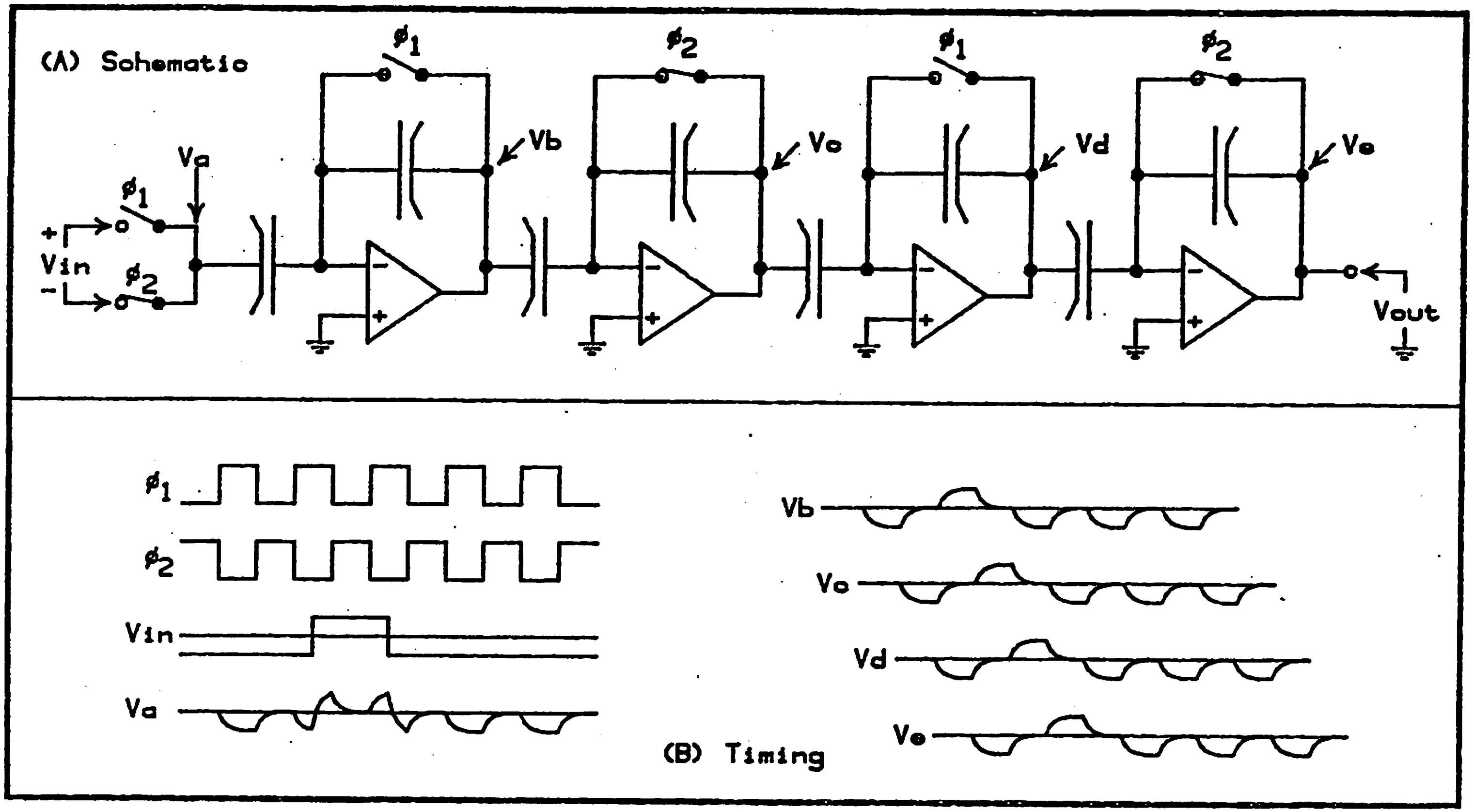


Fig 4 -- Analog Shift Register

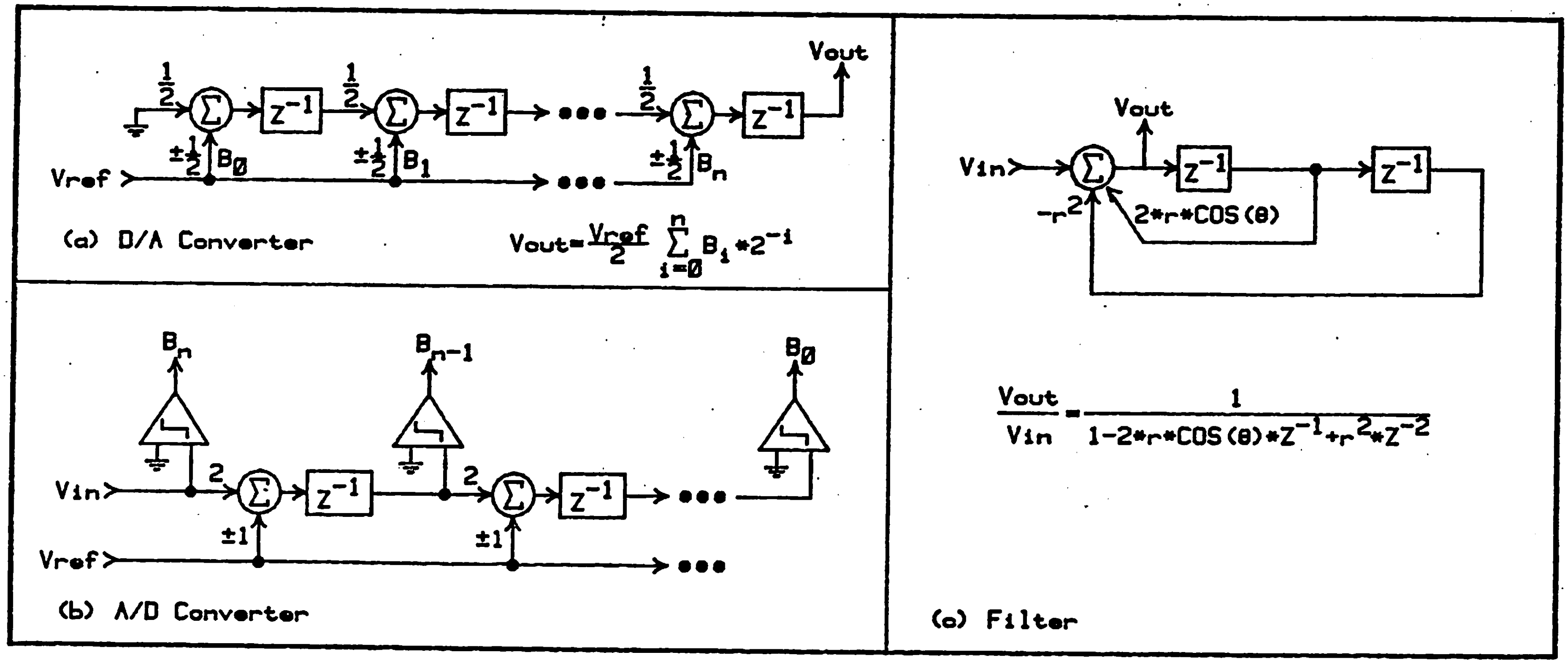


Fig 5 -- Signal Flow Diagrams

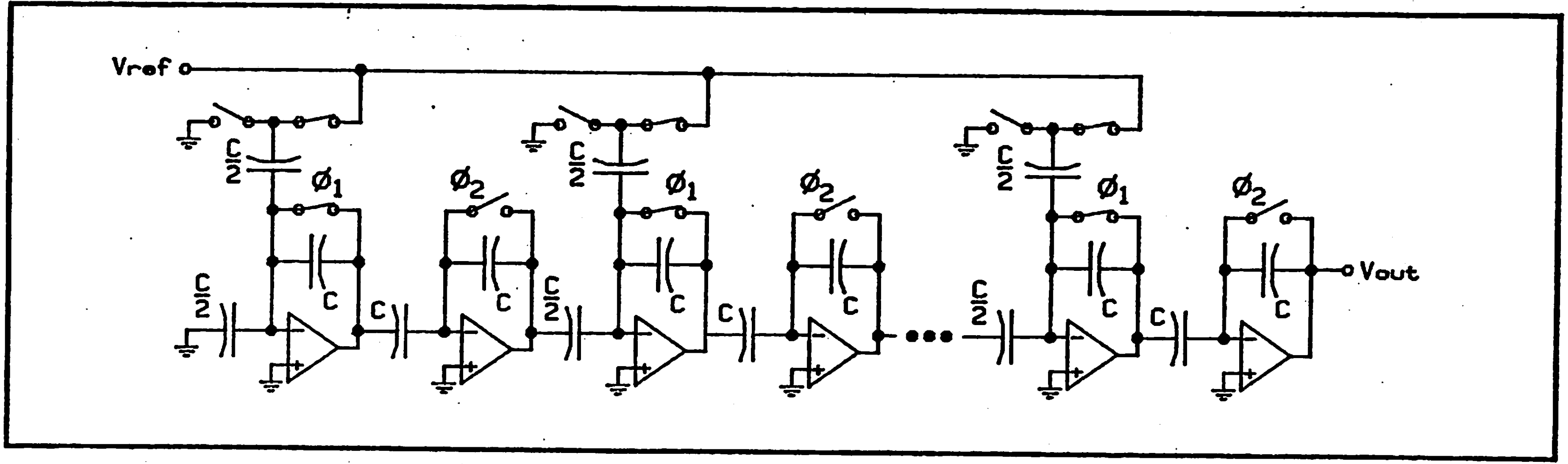


Fig 6 -- D/A Converter by Replication

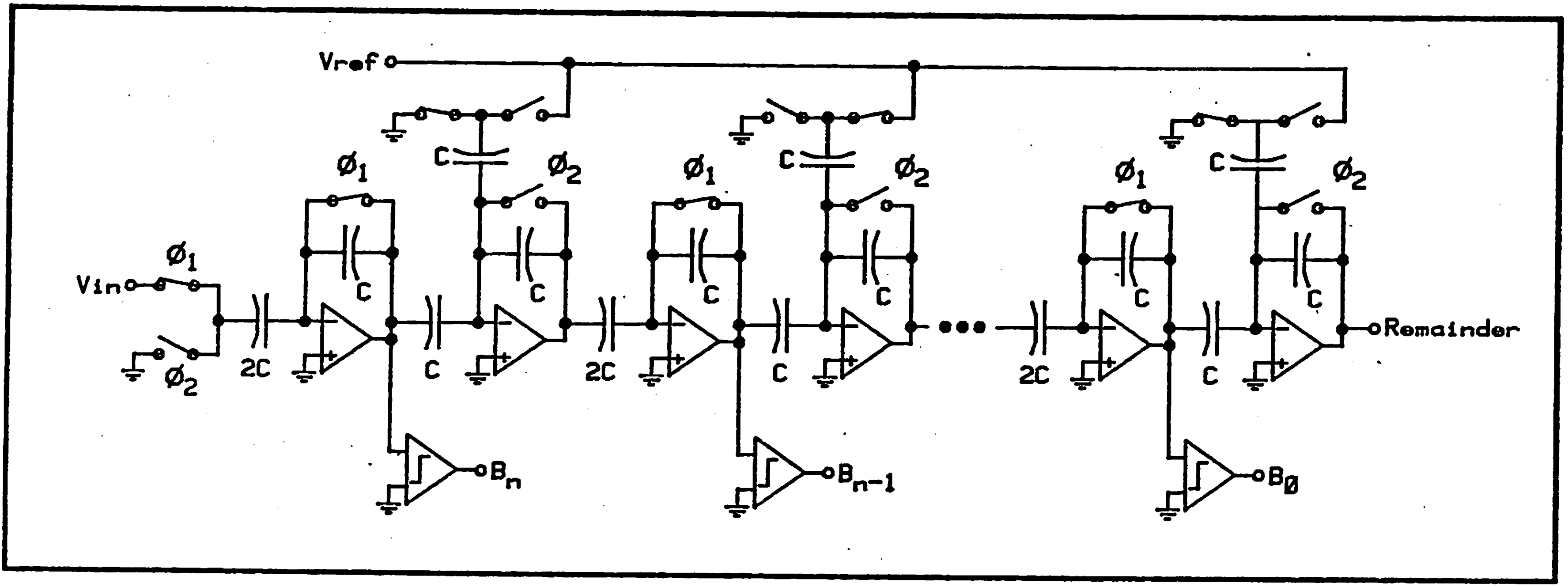


Fig 7 -- A/D Converter by Replication

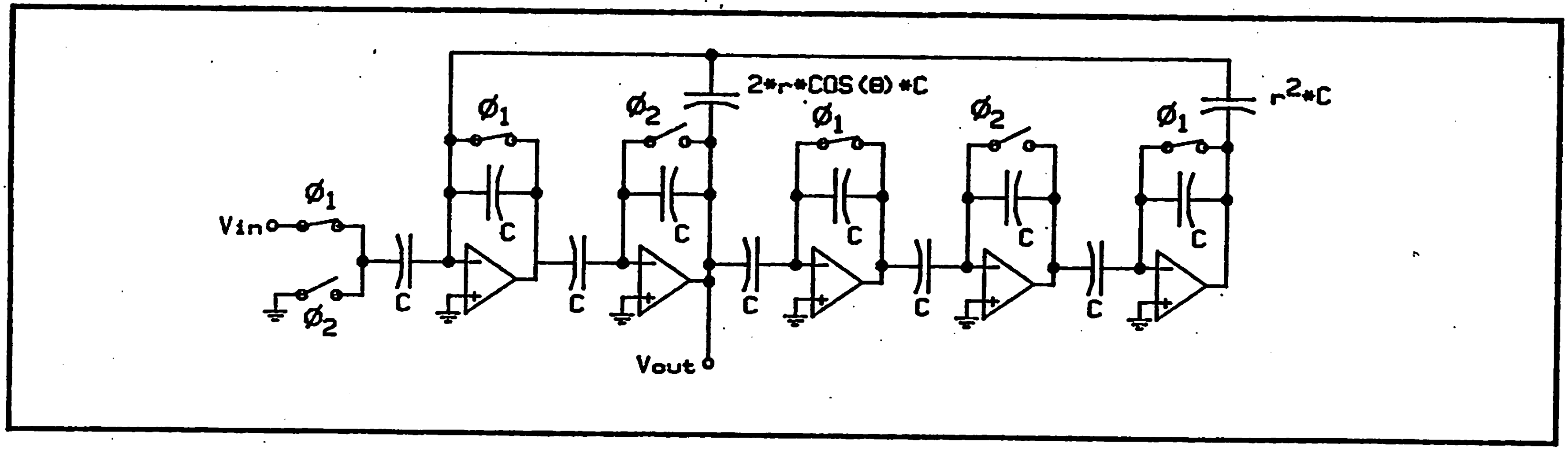
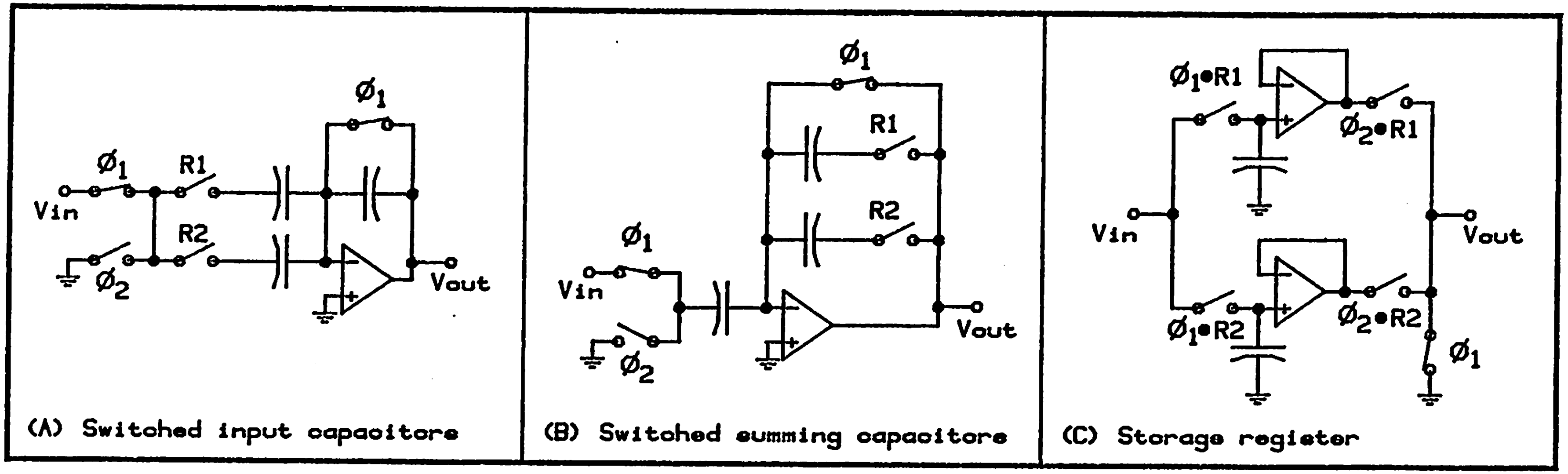


Fig 8 -- Filter by Replication

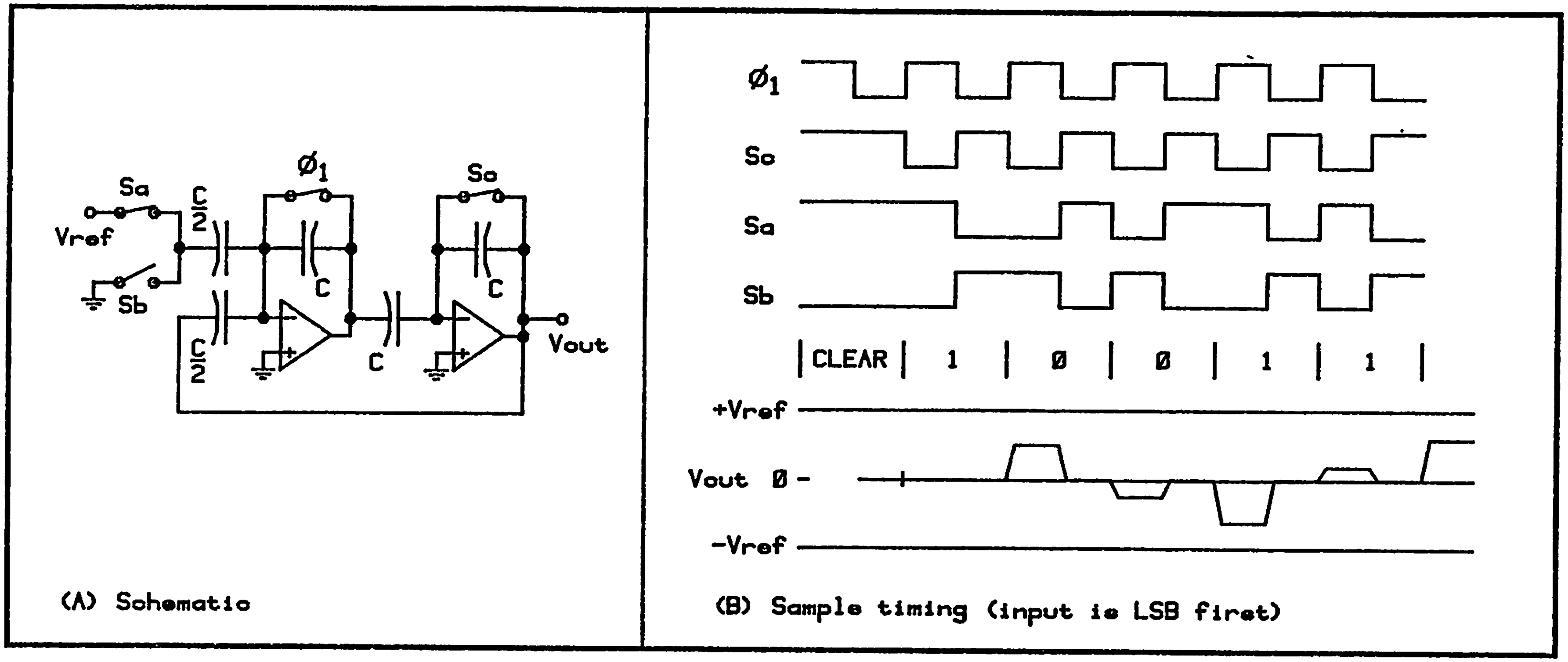


(A) Switched input capacitors

(B) Switched summing capacitors

(C) Storage register

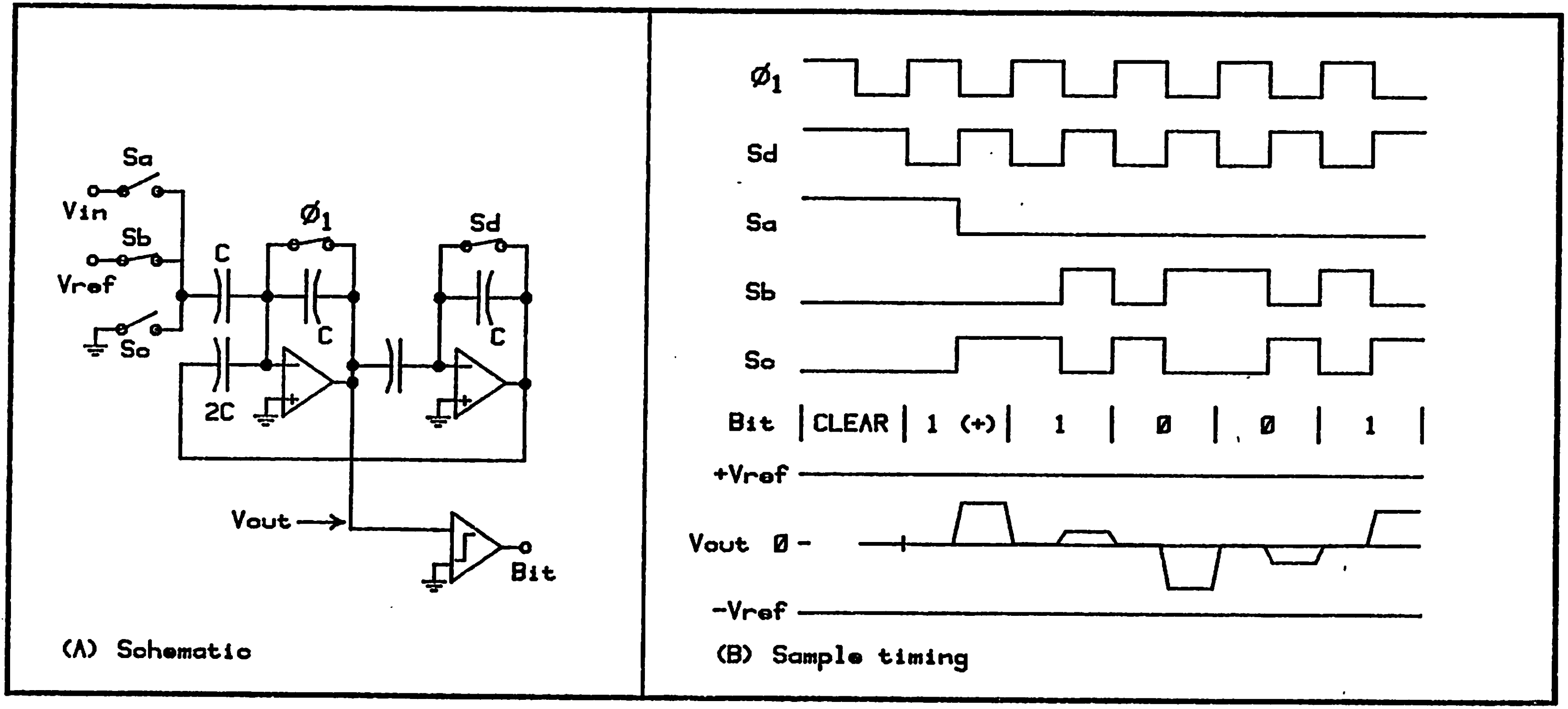
Fig 9 -- Analog Storage



(A) Schematic

(B) Sample timing (input is LSB first)

Fig 1Ø -- D/A Converter by Iteration



(A) Schematic

(B) Sample timing

Fig 11 -- A/D Converter by Iteration

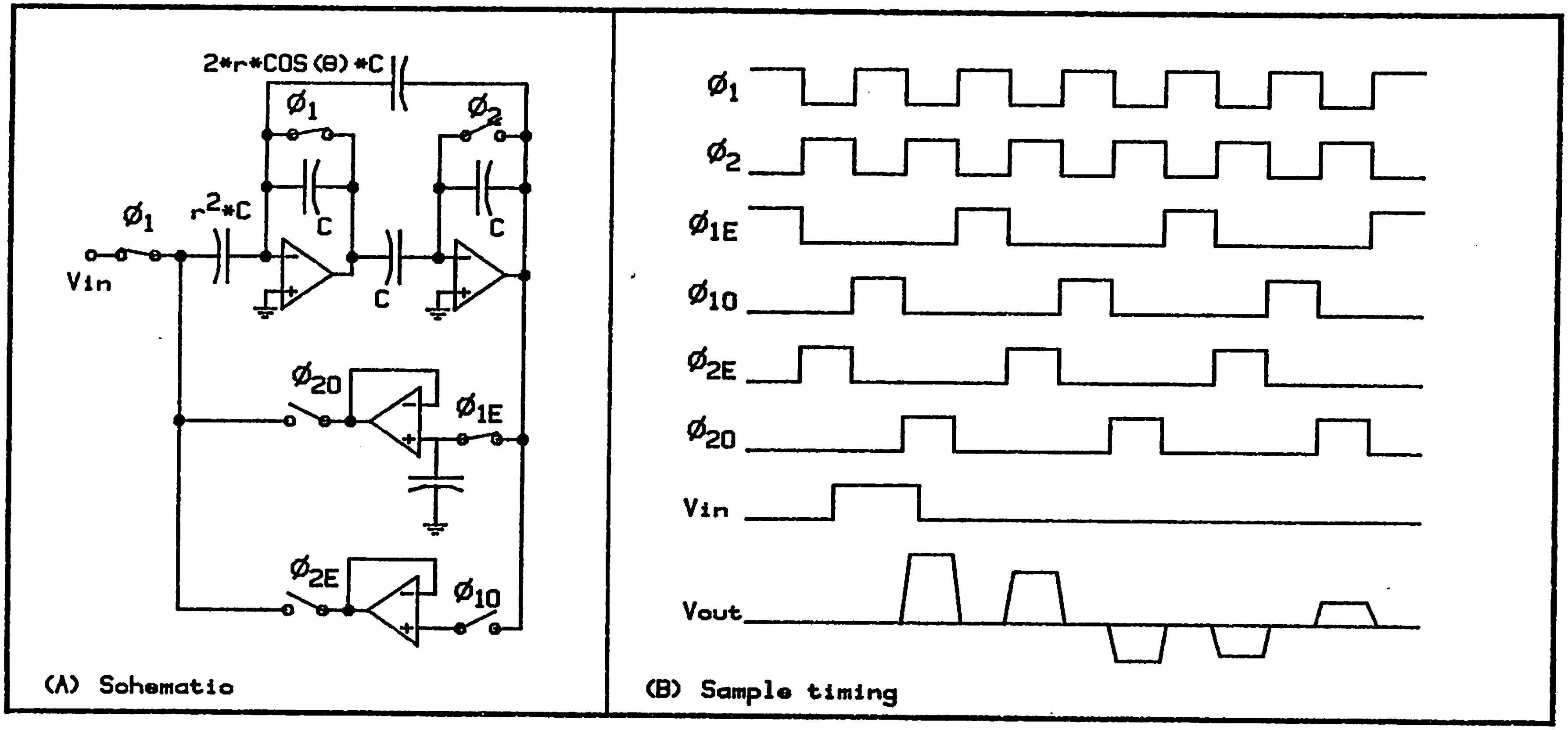
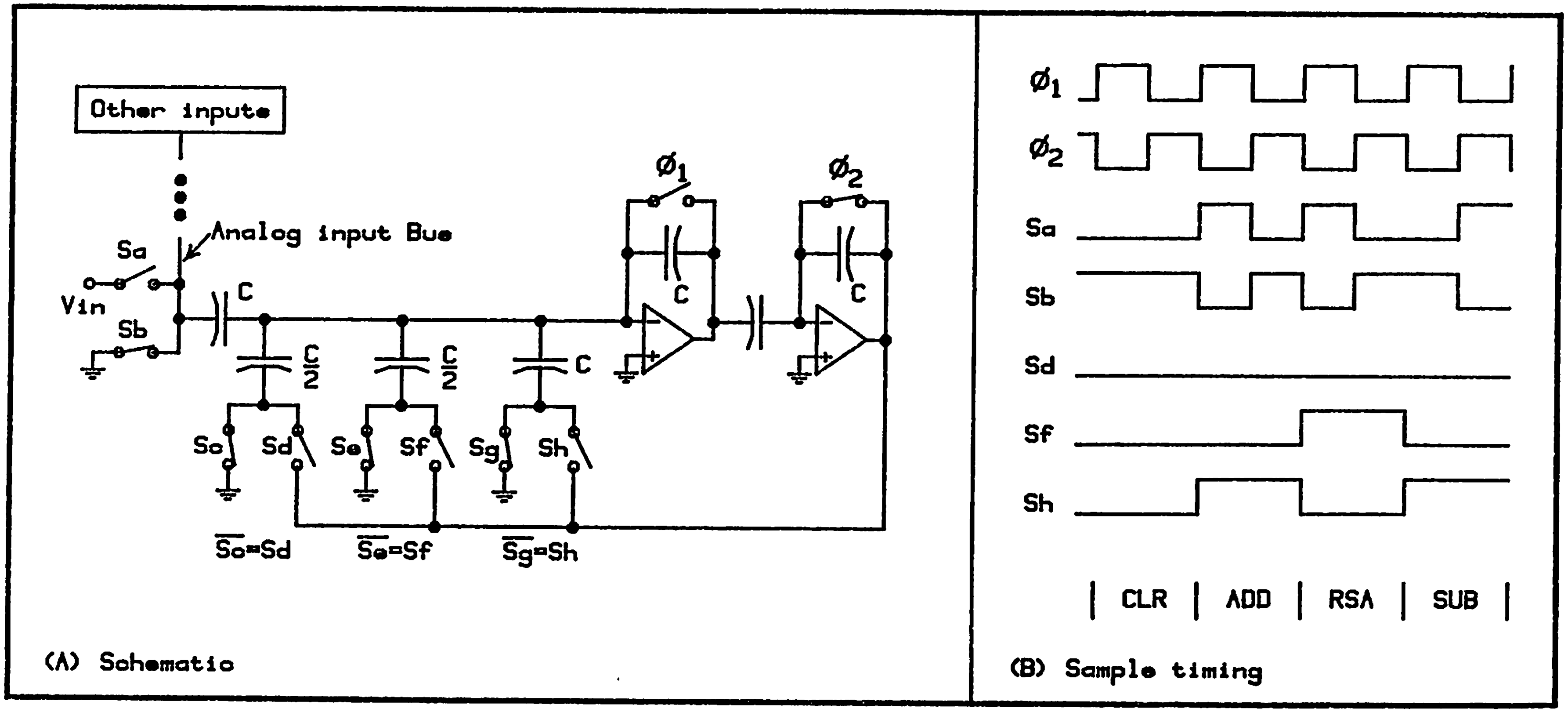


Fig 12 -- Filter by Iteration



(A) Schematic

(B) Sample timing

Fig 13 -- Analog Arithmetic Unit

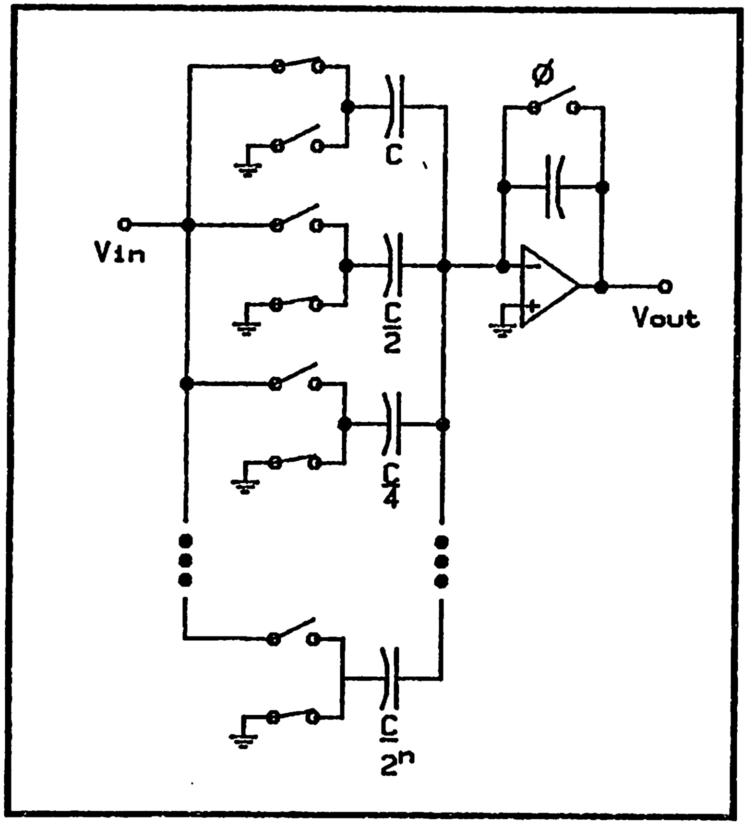
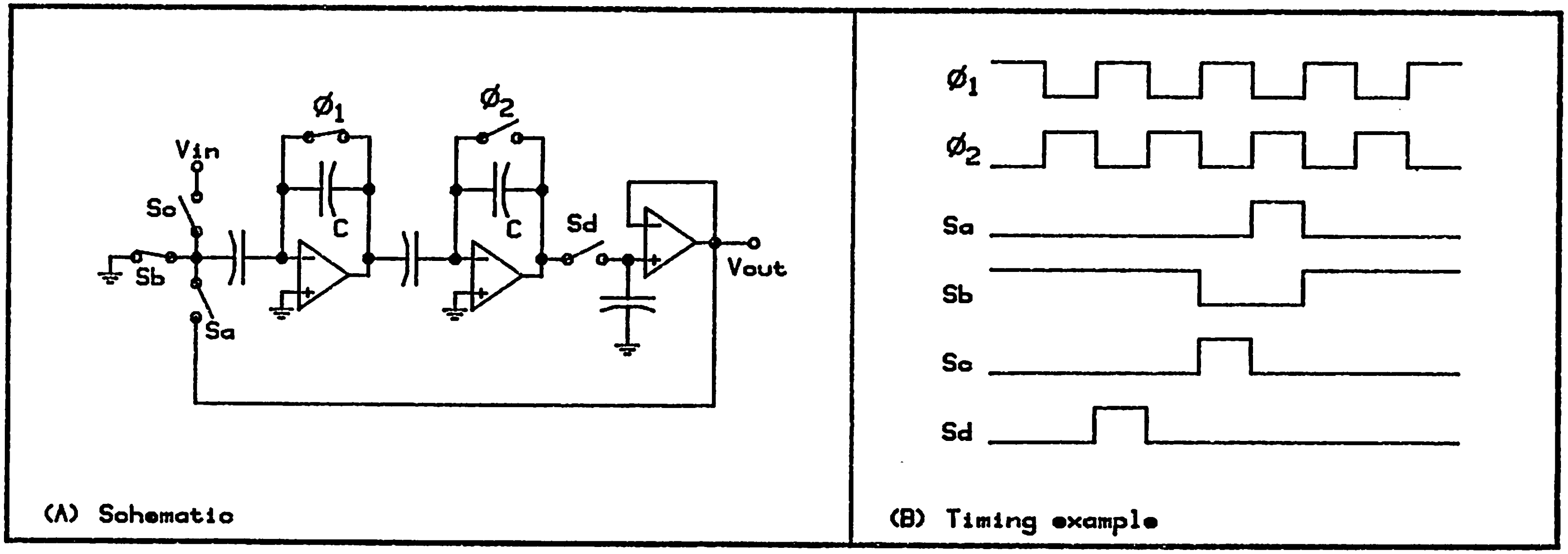


Fig 14 -- Variable Coefficients



(A) Schematic

(B) Timing example

Fig 15 -- Offset Cancellation

CHAPTER II

Understanding the MOS Transistor

Introduction

Charge circuits are composed of analog switches, capacitors and amplifiers. They can be implemented in bipolar, JFET or MOS technologies, but these technologies are not all equally suited for the task. The analog switches will require high-voltage logic to drive them. The thin oxide capacitance in an MOS process is carefully controlled and matched to device characteristics. The amplifiers must function with a minumum of input bias current. These requirements make MOS the technology of choice for implementing charge circuits.

The MOS circuit designer has more control over his device characterisitics than his bipolar counterpart. He can change channel length and substrate voltage as well as device area, operating current, voltage and temperature. To design analog circuits he must understand the MOS transistor.

To develop this understanding we will treat the MOS transistor as a circuit and make use of rules which are widely used in the modeling of parasitic effects. According to those rules, homogeneous regions are modeled as lumped capacitance or resistance while inhomogeneous regions are modeled as abrupt junctions. The resulting device picture is accurate over a wide range of currents and voltages. It provides simple and intuitive explanations for a variety of device effects including: subthreshold conduction, current saturation, nonlinear gate capacitance and charge pumping. Experimental agreement is obtained with a variety of laboratory and commercial devices.

MOS Transistor Model

We begin our discussion of the MOS transistor by observing the results from a two-dimensional simulation [1]. Plots of electron and potential distributions for a device in saturation are shown in fig 1. The device can be divided into three regions. Near the source the field is opposite to the direction of current flow. A central region contains a high carrier concentration near the silicon surface. This region is called the MOS channel. At the drain end of the channel there is a region where the field is high and the channel comes away from the surface. These three regions are intrinsic parts of the device and each of them plays an important role in determining its characteristics.

The device picture used here is sketched in fig 2. Linear capacitors C_1 and C_2 represent the gate oxide capacitance. The channel is modeled by a vanishingly thin layer of mobile charge carriers. This channel is isolated from the substrate by the depletion capacitance of junctions D_3 and D_4. Connections to the channel are made by gate-controlled diodes D_1 and D_2.

These diodes replace the strong inversion approximation in square-law models [2] or the channel boundary conditions of Pao-Sah [3] and El-Mansy [4]. They model the high-low junction which connects the channel to the source and drain diffusions [5]. Since the connection of any measuring instrument to the channel will create another junction, their characteristics cannot be measured directly. Nevertheless, they do exist.

The model presented here was developed independently, but it is similar to the Brews charge sheet model [6]. Compared to the Brews model, this model includes velocity saturation, channel-length modulation and some correction for vertical field mobility variations. We differ with Brews in neglecting diffusion current in the channel region. This simplifies the result, and it gives rise to a constant error of kT/q in the gate voltage and to other errors on the order of $(kT/q)^2$ in the applied potentials. With this in mind the devices-oriented reader may wish to refer to [6]. Additional background material may be found in [7,8].

In the following discussion we assume a surface-channel device. The concept can be extended to buried-channel devices, but the treatment is more complex. Potentials which can be measured will be denoted by a capital "V" and an appropriate subscript. Potentials which cannot be measured will be indicated by a "W" with a subscript. All potentials are referenced to the bulk silicon. For example, V_g denotes the applied gate-to-substrate potential, while W_g is the same potential after correction for work function differences and Q_{ss}. All signs will be taken as positive. This allows a uniform treatment for both P-channel and N-channel devices. The reader should have no difficulty in determining the correct signs. A glossary of the notation is shown in Table I.

We now consider separately each of the circuit elements used in this model.

GLOSSARY

C_d	Depletion Capacitance	F*Cm^{-2}V$^{-0.5}$
C_{ox}	Oxide Capacitance	F*Cm^{-2}
E	Electric Field (Parallel to current)	F*Cm^{-1}
E_d	E evaluated at pinch-off point	F*Cm^{-2}
I_{ds}	Drain-Source Current	A
I_{gcd}	Gate-Controlled Diode Current	A
k	Boltzmann's Constant	J
L	Channel Length	Cm
N_b	Bulk Impurity Concentration	Cm^{-3}
N_{eff}	Effective Impurity Concentration	Cm^{-3}
N_s	Surface Impurity Concentration	Cm^{-3}
Q_b	Depletion Charge Density (Under Channel)	Coul*Cm^{-2}
Q_d	Q_m Evaluated at Drain	Coul*Cm^{-2}
Q_m	Mobile Charge Density	Coul*Cm^{-2}
Q_s	Q_m Evaluated at Source	Coul*Cm^{-2}
Q_{sat}	Saturation Charge Density	Coul*Cm^{-2}
Q_{st}	Steady-State Charge Density	Coul*Cm^{-2}
R_d	Series Resistance at Drain	Ohm
R_s	Series Resistance at Source	Ohm
V_d	Applied Drain-to-Substrate Potential	V
V_s	Applied Source-to-Substrate Potential	V
V_{sat}	Carrier Saturation Velocity	Cm*Sec^{-1}
V_x	V_s or V_d	V
W_d	Internal Drain-to-Substrate Potential	V
W_g	V_g-V_{fb}	V
W_s	Internal Source-to-Substrate Potential	V
W_x	W_s or W_d	V
Z	Channel Width	Cm
Z_d	Channel Width at Drain	Cm
Z_s	Channel Width at Source	Cm
Δl	Channel Shortening from Drain Depletion Region	Cm
ϵ_{si}	Dielectric Constant of Silicon	F*Cm^{-1}
μ	Low-field (measured) mobility	Cm^{2}V^{-1}Sec^{-1}
ϕ_F	Built-in Potential of Bulk (kT/q LOG(N_b/N_i))	V

Channel Charge

If the suface potential is W_x, then the total charge induced in the semiconductor is:

$$Q_t=C_{ox}(W_g-W_x) \tag{1}$$

Some of this charge resides in a depletion region under the channel. If we assume a uniform impurity concentration in the substrate then this bulk charge is:

$$Q_b=C_d\sqrt{W_x} \tag{2}$$

Where C_d is given by:

$$C_d=\sqrt{2qN_b\epsilon_{si}} \tag{3}$$

The difference between the total charge and the bulk charge is the mobile charge which forms the channel:

$$Q_m=Q_t-Q_b \tag{4}$$

It is this mobile charge density which determines the conductance of the channel region.

The MOS Channel

The MOS channel consists of a thin layer of charge carriers connecting the source and drain junctions. Its conductance is proportional to the total mobile charge in the channel. To first order the channel current is given by:

Ids=[Sheet conductance]*[# squares]*[Applied voltage]

$$I_{ds}=\mu\frac{Q_s+Q_d}{2}*\frac{Z}{L}*(W_d-W_s) \qquad (5)$$

Where Q_s and Q_d are the mobile surface charge densities given by:

$$Q_s=Q_t(W_s, W_g)-Q_b(W_s) \qquad (6)$$

$$Q_d=Q_t(W_d, W_g)-Q_b(W_d) \qquad (7)$$

These equations for charge density are valid only for positive results. If the result is negative, the channel does not exist, and the mobile surface charge density is zero.

The above equation for current is a lumped approximation to the standard result. It gives good results for devices with 1000 angstrom oxides and impurity concentrations up to 10^{16}. For heavier substrate concentrations or thicker oxides the standard result is obtained by integration:

$$I_{ds}=\mu\frac{Z}{L}*P(W_d, W_s, W_g) \qquad (8)$$

Where the function $P(W_d, W_s, W_g)$. is given by:

$$P(W_d, W_s, W_g)=C_{ox}\left[W_g-\frac{W_d+W_s}{2}\right](W_d-W_s) \qquad (9)$$
$$-\frac{2}{3}C_d\left[W_d^{3/2}-W_s^{3/2}\right]$$

Given the potentials W_s and W_d, we can calculate the channel current. If we set $W_s=V_s+2*\phi_F$ and $W_d=V_d+2*\phi_F$ then these models correspond to the strong-inversion approximation in the linear region. An example of these two equations is plotted in fig 3. The numbers used in this example are typical for an NMOS device in a CMOS process. The difference between the lumped and integrated characteristic curves is less than five percent. For most devices the difference between the lumped and integrated channel equations will be insignificant.

Another case of interest is the annular device. For this device the channel equation can be shown to be:

$$I_{ds}=\mu\frac{A}{LOG(Z_d/Z_s)}P(W_d, W_s, W_g) \qquad (10)$$

Where A is the angular width of the device in radians and Z_s and Z_d are the source and drain widths.

Velocity Saturation

It is physically impossible for the velocity of the charge carriers to increase without limit as the field (applied W_d-W_s) increases. It is observed, for example, that the maximum velocity for electrons in silicon is approximately 10^7 cm/sec. A commonly used approximation for carrier velocity as a function of applied field is [9]:

$$Velocity=\frac{\mu E}{1+\mu E/V_{sat}} \qquad (11)$$

A solution which includes velocity saturation can be derived from any low-field model which fits the following form:

$$I_{ds}=\mu*G(L)*F(W_d, W_s, W_g) \qquad (12)$$

F(.,.,.) and G(.) are arbitrary functions. The complete solution is given by:

$$I_{ds}=\mu*G(L+\frac{\mu(Wd-Ws)}{Vsat})*F(W_d, W_s, W_g) \qquad (13)$$

The proof of this solution is obtained by substitution, and it applies to all of our channel equations. As the carrier velocity increases, the apparent channel length increases. This reduces the channel current.

Some channel characteristics which include velocity saturation are shown in fig 4. Annular devices are affected less than rectangular ones if the source is interior. This is due to their logarithmic dependance on channel length.

Field-Dependant Mobility

It has long been remarked that mobility varies with the vertical field across the channel [10,11,12,13,14,15,16]. But measurements of this mobility variation are subject to differing interpretations [17,18]. To make an accurate measurement of the actual mobility in an MOS transistor it is first necessary to measure the applied potential and mobile charge density in the channel. This has not been possible.

This variation in mobility is electrically similar to series resistance [19,20,21,22,23]. We model the field dependance of the mobility as resistors at both ends of the device. This allows the mobility variation to be added to the source and drain contact resistance. The equation used for determining this resistance is:

$$R_s = R_d = \frac{L}{2Z\mu Q_{sat}} \qquad (14)$$

This is an empirical equation. Q_{sat} was 0.65 microcoulombs per square centimeter for all devices tested.

The mobility used in our model is a measured process constant. Its value ranges from 50 to 75 percent of the bulk mobility [24] for an equivalent substrate impurity concentration.

The Gate-Controlled Diode

As the potential at the drain end of the channel is increased, the corresponding mobile charge density decreases to zero. From a physical standpoint this is impossible. Because of velocity saturation, the mobile charge at the drain end of the channel must be greater than the channel current divided by the carrier saturation velocity. This constraint prevents us from obtaining any solution for channel current in the saturation region. Another problem with the channel model is that it does not give the observed exponential characteristics at low V_g (subthreshold region). These problems are removed by including a model for the current-voltage characteristics of the channel-to-diffusion gate-controlled diodes.

To model these gate-controlled diodes we want an equation with some special characteristics. When no current is flowing between the channel and the diffusion we must satisfy the boundary conditions used by Brews [25], El Mansy [26] and Sze [27]. This gives rise to the exponential current behavior in the subthreshold region. To be physically reasonable current must always be less than the saturation velocity times the mobile charge density. The simplest possible form for this equation would be:

$$I_{gcd}=ZV_{sat}(Q_{st}-Q_t) \tag{15}$$

Q_{st} is the surface charge conventionally associated with the gate-controlled diode. Q_t is the actual induced surface charge as in equation (1). The steady-state charge density is given by the boundary conditions referenced above:

$$Q_{st}=\sqrt{Q_b^2+C_d^2\frac{kT}{q}\mathrm{EXP}\left(\frac{W_x-V_x-2\phi_b}{\frac{kT}{q}}\right)} \tag{16}$$

Combining this expression with the previous one gives:

$$I_{gcd}=ZV_{sat}\left[\sqrt{Q_b^2+C_d^2\frac{kT}{q}EXP\left(\frac{W_x-V_x-2\phi_b}{\frac{kT}{q}}\right)}-C_{ox}(W_g-W_x)\right] \qquad (17)$$

A plot of some I-V characteristics computed from this equation with gate voltage as a parameter is shown in fig 5.

This model gives the correct surface potential under forward bias (according to Boltzmann statistics) and the correct current under reverse bias. It is qualitatively similar to the actual device. These features are sufficient for our purpose.

The equation used for the gate-controlled diode in our model is a simplified version of the previous one:

$$I_{gcd}=ZV_{sat}\left[\frac{C_d}{2W_x}\frac{kT}{q}EXP(W_x-V_x)-Q_m\right] \qquad (18)$$

Fig 6 compares I-V characteristics for this equation with the the previous equation.

Channel Length Modulation

Under reverse bias the gate-controlled diode will have a depletion region extending into the channel. This channel shortening is the principal cause of output conductance in the saturation region. An expression for this channel-length modulation can be derived from electrostatics:

$$\Delta l=\frac{\varepsilon_{si}}{qN_s}\left[\sqrt{E_d^2+\frac{2qN_s}{\varepsilon_{si}}(V_d-W_d)}-E_d\right] \tag{19}$$

The principle fault of this equation is that it is not a circuit equation. This equation is often oversimplified as follows:

$$\Delta l=\sqrt{\frac{2\varepsilon_{si}}{qN_s}(V_d-W_d)} \tag{20}$$

This expression simply neglects the field in the drain depletion region. It tends to overestimate the extent of channel shortening for large gate voltages.

To calculate the channel shortening caused by the drain depletion region we use the simple expression (20) above and modify it by increasing the effective substrate impurity concentration in the following way:

$$N_{eff}=N_s\frac{Q_d+\sqrt{2kT\varepsilon_{si}N_s}}{\sqrt{2kT\varepsilon_{si}N_s}} \tag{21}$$

Substituting (21) for $N_s$in equation 20 gives better results in many cases. In other cases it tends to underestimate the degree of channel shortening for large gate voltages.

Neglecting the field in the drain region will often give reasonable answers, but it may also cause serious problems [28]. A special-purpose program was written which solves the modeling equations using equation (19) instead of (20) and (21). This program solves for current and then plots the potential in the channel from the source to the drain. The boundary between the channel and drain-depletion regions was determined by the requirement that the lateral field be smooth and continuous. A comparison of results for the simplified equation and the more complete one is shown in fig 7. Both the saturation voltage and the amount of channel shortening are reduced by including the field. Neglecting the field in the drain region is a serious limitation in calculating the output conductance of the MOS transistor.

The MOS Transistor

A DC model for the MOS transistor is shown in table I together with a table for the modeling equations. To obtain a solution for device current we must satisfy Kirchoff's laws, the current equations for the channel, gate-controlled diodes and series resistance; and we must iterate to find the correct value for channel-length modulation. We will now illustrate this model by discussing some features of the MOS transistor.

Saturation Region

In this model the saturation region results from the characteristics of the reverse biased gate-controlled diode, which will not conduct unless there is mobile charge in the channel. This can be visualized with the help of the graphical solutions shown in fig 8. In this sketch the channel potential is swept, and the drain diode forms a nonlinear load line over the channel characteristics. Curve "a" shows the device far into the linear region. In this region, the load line changes drain potential, and the resulting solution is also a strong funtion of drain potential. Curve "b" shows the device on the edge of saturation. Curve "c" shows the device far into saturation. In this region the load line no longer moves as V_d is increased. Current is nearly constant at a value of $Z*Q_d*V_{sat}$. What we have shown is that the saturation region does not arise from a defect in the channel equations, but rather from the characteristics of the connection to the drain.

Subthreshold Region

The subthreshold region [29,30,31,32] can also be explained by the use of graphical methods. In fig 9 we have assumed that the drain voltage is high; the transistor is in saturation. The dashed curves are the channel characteristics and the solid lines are the characteristics of the source diode as the potential at the source end of the channel is varied. The gate-controlled diode is now the driver, and the channel now forms the load line. Corresponding curves intersect at solutions for device current. At V_g=5 Volts the transistor is far into the square law region. The characteristic curves for the diode change little with increasing gate voltage. The slope of the diode curves is much greater than the slope of corresponding channel curves, and the channel potential is almost constant at approximately 2.5 ϕ_F. As the gate voltage is decreased, the source potential begins to change; between V_g=5 and V_g=2 volts the diode potential changes by 60 millivolts. Below V_g=2 Volt, the solution is dominated by the diode characteristics, and the current becomes exponential with gate voltage.

In this subthreshold region:

$$\Delta W_s = \Delta W_g \frac{C_{ox}}{C_d/(2W_s)+C_{ox}} \tag{22}$$

and the logarithmic slope becomes:

$$\frac{d}{dV_g} \mathrm{LOG}_{10}(I_{ds}) = \frac{C_{ox}}{2.3\frac{kT}{q}(C_{ox}+C_d/(2W_s))} \tag{23}$$

If we neglect surface states, this agrees with theoretical and experimental work referenced above.

Nonlinear Gate Capacitance

This model also provides a qualitative explanation for the nonlinear gate capacitance. A typical measured C-V curve for an MOS transistor under bias is shown in fig 10. An equivalent circuit for our model is also shown. In regions "a" and "b" of the C-V curve, the gate-controlled diodes are reverse-biased, and the equivalent resistance of the diodes is high. The resulting capacitance is that of the oxide capacitance in series with the depletion capacitance. In region "a" the depletion depth is zero (surface in accumulation), and so the depletion capacitance is infinite. In region "b" the total capacitance decreases as the depletion depth increases. In region "c" the source diode becomes forward biased. Capacitors C_1 and C_2 are now connected to the source. The gate capacitance per unit area near the source is equal to the oxide capacitance. The capacitance at the drain is still decreasing due to the widening of the depletion region at the drain end of the transistor. As the gate voltage is increased further, the drain diode becomes forward biased, and the transistor moves into the linear region. In this region, the channel potential is clamped by the diodes thus shielding the depletion region from further increases in gate potential. The gate capacitance is now equal to C_{ox} and is connected at both ends of the channel to the diffusions.

Charge Pumping

We can also provide a simple explanation for charge pumping [33,34].

Fig 11 shows an MOS transistor connected to a pulse generator and a capacitor. The parasitic junctions in the device form a voltage clamp circuit. The channel potential V_{ch} will never rise more than 2.5 ϕ_F above the source and drain potential. When the clock rises this forces charge to flow from the diffusions into the channel. If the clock fall time is fast compared to the time it takes charge to flow out of the transistor, the source and drain diodes will turn off before the channel can discharge to the source and drain junctions. This will cause the junction between the channel and the substrate to forward bias when the clock goes low, and charge will flow from the channel to the substrate.

Substrate Nonuniformity

We have so far assumed a uniform substrate impurity concentration. Nonuniform impurity concentrations are common in MOS transistors, and they can have substantial effect on the device characteristics. Some nonuniformity exists in every device due to impurity redistribution. These nonuniformities tend to be small and concentrated near the surface. If the depth of the disturbance in impurity concentration is small compared to the depletion width under the channel, then the effect of the nonuniformity will be to give an apparent shift in the flatband voltage of Q_{nu}/C_{ox}. Q_{nu} is the total amount of charge contained in the impurity disturbance.

Nonuniform impurity concentrations are often produced deliberately by implanting dopant species directly in silicon. If the implants are shallow and not deeply diffused then we can approximate their effect in the same as above.

If the depth of the depletion region is large in comparison with the depth of the implant, the flat-band shift will be less. The width of the depletion region under the channel will also be in error. This will cause an error in the calculation of mobile charge density. This error is particularly significant in the saturation region where mobile charge density is small in comparison to the depletion charge density.

Nonuniform impurity concentrations will also affect the vertical field at the surface. This field determines the low (lateral) field mobility used in the model. For this reason implanted devices such as depletion loads will not be accurately modeled.

Experimental Results

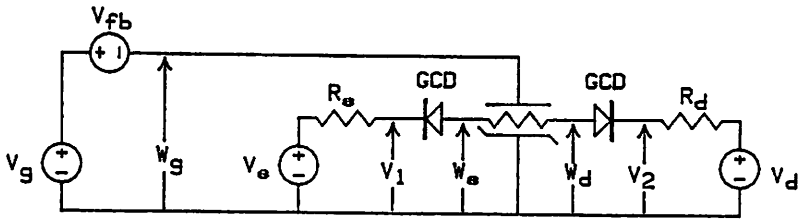
Drain characteristics for several different MOS transistors were obtained. Points from these characteristics were transferred by hand to corresponding theoretical device plots made on a computer. The effective flatband voltage was adjusted for each device. The surface mobility factor was fitted to each process. Other parameters were taken from process data adjusted within the limits of experimental error. The parameters used for each device are shown in table II.

Fig 12 shows results for a long-channel NMOS device which was fabricated in the Berkeley IC facility. Theoretical and experimental points match extremely well at both high and low currents. Fig 12 also shows results from a similar long channel PMOS device. These results are also quite good.

This two figure also shows results for shorter channel devices fabricated on the same wafers as the previous two devices. In these plots we see some errors in the saturation region. Because the field in the drain region is much higher for these devices we cannot use the same simple expression for channel-length modulation as for the longer devices. Despite these problems agreement for device current is within 10 per cent.

TABLE I - DC Modeling Equations

Mobile charge: $W_g = V_g - V_{fb}$ $Q_s = C_{ox}(W_g - W_s) - \sqrt{2q\epsilon_{si}N_bW_s}$ $Q_d = C_{ox}(W_g - W_d) - \sqrt{2q\epsilon_{si}N_bW_d}$	**Channel-length modulation:** $L_{eff} = L - \Delta_l + \mu(W_d - W_s)/V_{sat}$ $\Delta_l = \sqrt{\frac{2\epsilon_{si}}{qN_{eff}}(V_2 - W_d)}$ $N_{eff} = N_s\frac{Q_d + \sqrt{2kT\epsilon_{si}N_s}}{\sqrt{2kT\epsilon_{si}N_s}}$
Device current: $I_{ds} = \frac{\mu Z}{L_{eff}}\frac{Q_s + Q_d}{2}(W_d - W_s)$ $I_{ds} = ZV_{sat}\left[\sqrt{\frac{\epsilon_{si}qN_s}{2W_s}}\frac{kT}{q}\text{EXP}\left[\frac{W_s - V_1 - 2\phi_b}{kT/q}\right] - Q_s\right]$ $-I_{ds} = ZV_{sat}\left[\sqrt{\frac{\epsilon_{si}qN_s}{2W_d}}\frac{kT}{q}\text{EXP}\left[\frac{W_d - V_2 - 2\phi_b}{kT/q}\right] - Q_d\right]$	**Series resistance:** $R_s = R_d = \frac{L}{2\mu Q_{sat}Z}$



When the substrate impurity concentration is increased the field-dependant term in the channel-length modulation again becomes less significant. Fig 13 shows results from a commercial NMOS device which is fabricated in a heavily doped p-type well. Despite the shorter channel length excellent agreement is again obtained. Above 10 volts of drain potential we see low-level avalanche currents. These currents are caused by impact ionization in the drain-to-channel junction [35,36,37,38,39] and they flow from the drain to the substrate rather than the source. These substrate currents are important in the design of high-gain amplifiers and in floating-substrate processes such as SOS, where they can cause the substrate to become forward biased.

The complementary PMOS device is also shown in fig 13. At low currents the experimental and theoretical points match closely. The upward curvature in the drain characteristics appears to be punch-through, but it is accurately modeled in this case by channel length modulation. At higher currents the agreement is not good due to the lightly doped substrate and the relatively short channel length.

The next two devices show how the substrate impurity profile can affect the device I-V characteristics. These devices were fabricated using an n-well CMOS process [40] An unusually deep double-dose boron implant was performed which gave a step in the implanted boron concentration at a depth of approximately 0.6 microns. We expect our model to work well as long as the depletion region boundary stays in an area of uniform impurity concentration.

Fig 14 shows results for an NMOS device. Below about 10 microamperes the transconductance is lower than predicted here. Between 10 microamperes and 1 milliampere of drain current the depletion region extends into the step in impurity concentration. In this region agreement is excellent. Above 1 milliampere the depletion region near the drain extends beyond the step, so that the depletion region charge is higher than we would expect. This causes the model to overestimate the saturation current.

Because of the large increase in impurity concentration near the surface, we expect prediction for the PMOS device (fig 15) to be too high at low gate voltages. At device currents ranging from 10 microamperes to 500 microamperes the model gives excellent predictions. Above this current level, the drain depletion region extends beyond the step. In this case the effective substrate concentration decreases causing the model to underestimate the saturation current.

Our last two sets of data are taken from published results for some other models. These models use the empirical mobility variations discussed earlier. Fig 16 shows a comparison of our model with the data of Merckel, et al [41]. Again we have excellent results for the longer device and good results for the shorter device. This figure also compares our model with the data of Sibbert [42]. These results are also good. Replacing mobility variations with series resistance works well for normal operating conditions.

Limitations

Any model is an idealized picture. Many limitations of this model have been pointed out above. For the sake of clarity we summarize these limitations.

The influence of the source and drain depletion regions on channel charge has been omitted. This "short channel effect" can be included as in [43,44,45].

The lumped resistor model for mobility variation with (vertical) field is accurate only for constant V_s. For substrate impurity concentrations greater than $10^{16}/cm^3$ this variation will be five per-cent or less. For lightly-doped substrates the low (lateral) field mobility needs to be varied with (vertical) field as in [46,47].

Substrate current has been neglected. The most significant component of substrate current is avalanche current generated in the drain depletion region. This current is proportional to both drain current and to an exponential of the peak field. Its most serious effect for analog circuits is to degrade the drain conductance of the MOS transistor at drain voltages well below breakdown. This effect is seen on N-channel devices fabricated on heavily-doped substrates. In a typical commercial N-channel CMOS device this effect will dominate the drain conductance in saturation above 8 Volts V_d . Such a device is shown at the top of fig 13.

We have assumed that the MOS transistor is a circuit. This implies that the current is controlled only by the potential and not by the field. In the drain depletion region this assumption breaks down. However, the conventional expression for channel-length modulation gives the same result. The empirical correction used in this model does not always work. This is also true of the empirical corrections used in [23]. Explicit inclusion of the field in a one-dimensional model can give good results, but it is computationaly cumbersome, and extensive comparison with experimental data has not been done. At the present time accurate determination of the output conductance requires a two-dimensional model.

Thermal effects can be important. The power dissipated in the MOS device will cause a local rise in temperature which causes a decrease in mobility. This can cause the output conductance to become negative [48].

We have assumed that the channel is a thin sheet of charge at the surface. Because we have treated the drain depletion region separately from the channel this assumption will be valid for many devices. This assumption is not valid for depletion devices. For these devices the details of the impurity profile become important.

Our equations are not valid when the gate potential is near the flat-band voltage or when the surface is accumulated; MOS devices are not normally operated in this region.

We have ignored the fast surface state charge. This will affect the subthreshold conduction but it is not significant for present-day oxide thicknesses and processing techniques.

We have shown that the combined effect of the gate-controlled diode and channel-length modulation gives results similar to punch-through, but our model does not account for subsurface punch-through currents.

Conclusion

We have presented circuit model which accurately predicts the I-V characteristics of MOS transistors. By adapting the conventional square-law model to include substrate-to-channel and diffusion-to-channel junctions, simple explanations can be found for otherwise obscure device properties. It is hoped this model will be useful for hand analysis and for computer simulation.

Acknowlegement

I am grateful to those who shared their candid criticisms and offered suggestions. I am especially grateful to Richard Coen and Charlie Sodini for sharing their experimental results; and to Jim Eaton and Robert Wu of HP Labs for providing the two-dimensional MOS simulation results.

References for Chapter II

[1] Wu, R. and Eaton, J.

2-D MOS Simulation Stanford/Hewlett-Packard Research Review: Two-Dimensional Model. September 19,1980

[2] Grove, A. S.,

Physics and Technology of Semiconductor Devices. John Wiley & Sons, 1967. Chapter 11, pp. 317 ff.

[3] Pao, H. C. and Sah, C. T.

Effects of Diffusion Current on Characteristics of Metal-Oxide Semiconductor Transistors Solid-State Electronics. Great Britain: Pergamon Press, Vol 9, pp. 927-937.

[4] El-Mansy, Y. A. and Boothroyd, A. R.

A New Approach to the Theory and Modeling of Insulated-Gate Field-Effect Transistors IEEE Transactions on Electron Devices, Vol ED-24, no. 3, March 1977, pp. 241 ff.

[5] Johnson, E. O.,

The Insulated-Gate Field-Effect Transistor -- A Bipolar Transistor in Disguise. RCA Review, Vol 34, March 1973, pp. 80-93.

[6] Brews, J. R.

A Charge-Sheet Model of the MOSFET Solid-State Electronics, Vol 21, 1978, pp. 345-355.

[7] Loeb, H. W., Andrew, R., and Love, W.

Application of 2-Dimensional Solutions of the Shockley-Poisson Equation to Inversion-Layer M.O.S.T. Devices Electronics Letters, Vol 4 No 17, 1968, pp. 352-355.

[8] Armstrong, G. A., Magowan, J. A., and Ryan, W. D.

Two-Dimensional Solution of the D.C. Characteristics for the M.O.S.T. Electronics Letters, Vol 5 No 17, 1968, pp. 406-409.

[9] Murphy, B. T.

Unified Field-Effect Transistor Theory Including Velocity Saturation. IEEE Journal of Solid-State Circuits, Vol SC-15, no. 3, June 1980.

[10] Chen, J. T. C. and Muller, R. S.

Carrier Mobilities at Weakly Inverted Silicon Surfaces. Journal of Applied Physics, Vol 45 no. 2, February 1974, pp 828 ff.

[11] Sabnis, A. G. and Clemens, J. T. *Characterization of the Electron Mobility in the Inverted <100> Si Surface*. 1979 IEDM, pp. 18-21.

[12] Rutledge, J. L. and Armstrong, W. E. *Effective Surface Mobility Theory* Solid-State Electronics, Great Britain: 1972, Vol 15, pp. 215-219.

[13] Martinot, H. et al. *Mobility Parameters and Metal-Oxide-Semiconductors-Transistor Properties* Electronics Letters, October 24, 1972.

[14] Frohman-Bentchkowsky, D. *On the Effect of Mobility Variation on MOS Device Characteristics*. Proceedings of the IEEE, February 1968, pp. 217-218.

[15] Berger, J. and Lisiak, K. *Electron Mobility in Silicon Surface Inversion Layers*. 1977 Semiconductor Interface Specialist Conference Record.

[16] Sodini, C. *Mobility Measurements* Hewlett-Packard Co. internal memo dated April 3, 1980.

[17] Brews, J. R.
Comments on "A New Approach to the Theory and Modeling of IGFET's" IEEE Transactions on Electron Devices, Vol. ED-24 no. 12, December 1977 pp. 1369,1370

[18] El-Mansy, Y. A. and Boothroyd, A. R.
Authors' Reply to "Comments on 'A New Approach to the Theory and Modeling of IGFET's'" IEEE Transactions on Electron Devices, Vol ED-25 no. 3, March 1978.

[19] Mellor, J. T.
Drain Series Resistance in MOS Transistors Proceedings of the IEE, Vol 118 no. 10, October 1971, pp. 1393 ff.

[20] Johnson and Harwick (Editors)
Field-Effect Transistors: Physics, Technology and Applications New Jersey: Prentice-Hall, Chapter 5, pp. 148-155.

[21] Richman, P.
MOS Field-Effect Transistors and Integrated Circuits. John Wiley & Sons, 1973, pp. 158-165.

[22] Lee, T. F.
Effect of External Series Resistance on MOS Carrier Surface Mobility Measurement Hewlett-Packard Co.

internal memorandum dated July 2, 1979.

[23] Suciu, P. and Johnston, R.
Experimental Derivation of the Source and Drain Resistance of MOS Transistors IEEE Transactions on Electron Devices, Vol. ED-27, no. 9; September 1980, pp. 1846 ff.

[24] Jacoboni, C. et al.
A Review of Some Charge Transport Properties of Silicon. Solid-State Electronics, Great Britain: Vol 20, 1977, pp. 77-89.

[25] Reference [6], equation (12).
The second term inside the square root is not significant in weak or strong inversion.

[26] Reference [5], equation (8).
This equation can be rearanged to give the same result.

[27] Sze S.M.,
The Physics and Technology of Semiconductor Devices. New York: Wiley Interscience, 1969. pg. 511 equation (5). The first and third terms inside the square root are not significant in weak or strong inversion.

[28] Frohman-Bentchkowsky, D. and Grove, A. S. Conductance of MOS Transistors in Saturation IEEE Transactions on Electron Devices, Vol. ED-16 no. 1, January 1969, pp. 108-113.

[29] Rideout, V. L. Device Design Considerations for Ion Implanted n-Channel MOSFETS. IBM Journal of Research and Development. January 1975, pp. 50-59.

[30] Gosney, W. M. Subthreshold Drain Leakage Currents in MOS Field-Effect Transistors. IEEE Transactions on Electron Devices, Vol. ED-19 no. 2, February 1972. pp. 213-219.

[31] Troutman, R. R. and Chakravarti, S. N. IEEE Transactions on Circuit Theory, Vol. CT-20 no. 6, November 1973, pp. 659-665.

[32] Swanson, R. M. and Meindl, J. D. Ion-Implanted Complementary MOS Transistors in Low-Voltage Circuits. IEEE Journal of Solid-State Circuits, Vol. SC-7 no. 2, April 1972. pp. 146-153.

[33] Brugler, J. S.
Charge Pumping in MOS Devices. IEEE Transactions on Electron Devices, Vol. ED-16 no. 3, March 1969. pp. 297-302.

[34] Backensto, W. V. and Viswanathan, C. R.
The Utilization of Charge Pumping Techniques to Evaluate the Energy and Spacial Distribution of Interface States of an MOS Transistor.

[35] Troutman, R. R.
Low-Level Avalanche Multiplication in IGFET's IEEE Transactions on Electron devices, Vol ED-23 no 4, April 1976, pp. 419-425.

[36] Lattin, W. W. and Rutledge, J. L.
Impact Ionization Current in MOS Devices. Solid-State Electronics. Great Britain: Pergamon Press, Vol 16, pp. 1043-1046.

[37] Tihanyi, J. and Schlotterer, H.
Properties of ESFI MOS Transistors Due to the Floating Substrate and the Finite Volume. IEEE Transactions on Electron Devices, Vol. ED-22 no. 11, November 1975, pp. 1017-1023.

[38] Abbas, S. A.

Substrate Current -- A Device and Process Monitor.

[39] El-Mansy, Y. A. and Caughey, D. M.

Modelling weak Avalanche Multiplication Currents.

[40] Black, W. C. et al.

CMOS Process for High-Performance Analog LSI. IEDM Technical Digest, 1976, Washington, D. C. pp. 331-334.

[41] Merckel, et al.

An Accurate Large-Signal MOS Transistor Model for Use in Computer-Aided Design. IEEE Transactions on Electron Devices, Vol. ED-19, no. 5, May 1972, pp. 681-690

[42] Sibbert, H.

Modellierung ung Netzwerkanalyseprogramm fur MOS-Schaltungen mit hoher Leistungfahigkeit. PhD. dissertation, West Germany: University of Dortmund, 1977.

[43] Yau, L. D.

A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFET's Solid-State Electronics, Vol. 17, pp 1059-1063, 1974. Pergamon Press, Great

Britain.

[44] Sun, E.
A Short Channel MOS Model Hewlett-Packard internal memorandum dated April 4,1978.

[45] Ratnakumar, K. N. and Meindl, J. D.
Performance Limits of E/D NMOS VLSI 1980 IEEE ISSCC Digest of Technical Papers. pp 72-73.

[46] Sun, S.C. and Plummer, J. D.
Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces IEEE Transactions on Electron Devices, Vol ED-27 No 8, August 1980, pp. 1497-1508.

[47] Sabnis, A.G. and Clemens, J. T.
Characterization of the Electron Mobility in the Inverted <100> Si Surface 1979 IEDM Digest of Technical Papers, pp. 18-21.

[48] Sharma, D., Gautier, J. and Merckel, G.
Negative Dynamic Resistance in MOS Devices. IEEE Journal of Solid-State Circuits, Vol. SC-13 no. 3, June 1978, pp. 378-380.

Table II

Device Parameters

Process	L (μM)	Z (μM)	μ/μ_b	C_{ox} (nF/cm^2)	N_s*10^{14} cm^{-3}	N_b*10^{14} cm^{-3}	V_{fb} (V)
Coen's NMOS	19.9 (a) 7.2	256 (a) 129	0.67 (c)	39 (a)	2.3 (d)	7.05 (a)	-0.97 (e) -1.02
Coen's PMOS	21.9 (a) 9.2	256 (a) 129	0.52 (c)	41 (a)	9.6 (d)	7.7 (a)	0.32 (e) 0.27
CD4007 NMOS CD4007 PMOS	5.0 (b) 5.0	180 (b) 400	0.50 (c) 0.75 (c)	29 (b)	200. (d) 11.	340. (b) 4.	-2.51 (e) 0.97
McC's NMOS	17.0 (b)	80 (b)	0.52 (c)	55 (b)	40. (d)	60 (b)	-1.24 (e)
McC's PMOS	16.0 (b)	80 (b)	0.72 (c)	55 (b)	100. (d)	30 (b)	-0.42 (e)
Merc's PMOS	20.9 (a) 10.8	100 (a) 100	0.68 (c)	27.7 (a)	40. (d)	30 (a)	-0.40 (e) -0.40
Sibb's NMOS	19.5 (a) 7.0	10 (a) 10	0.66 (c)	42 (a)	20. (d)	20 (a)	-0.9 (e) -1.00

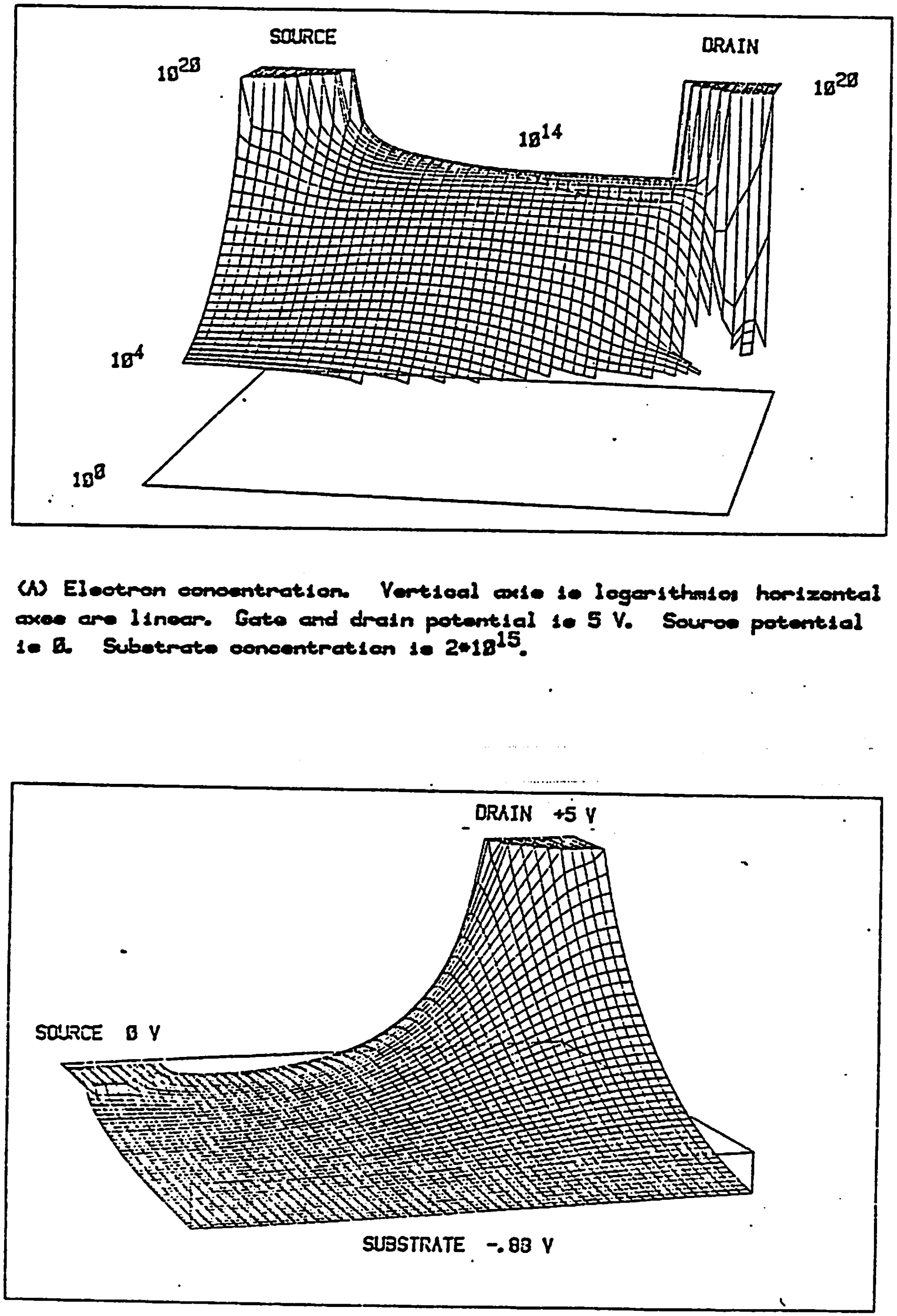
(a) Published elsewhere or independantly measured.

(b) Based on physical dimensions and process data

(c) Fitted to process. Match Ids.

(d) Fitted to process. Match Gds.

(e) Fitted to each device.



(A) Electron concentration. Vertical axis is logarithmic; horizontal axes are linear. Gate and drain potential is 5 V. Source potential is 0. Substrate concentration is $2*10^{15}$.

HPL/IRL

J. Eaton 7/80

(B) Potential. Same conditions as above.

Fig 1 -- 2-D MOS Simulation Results

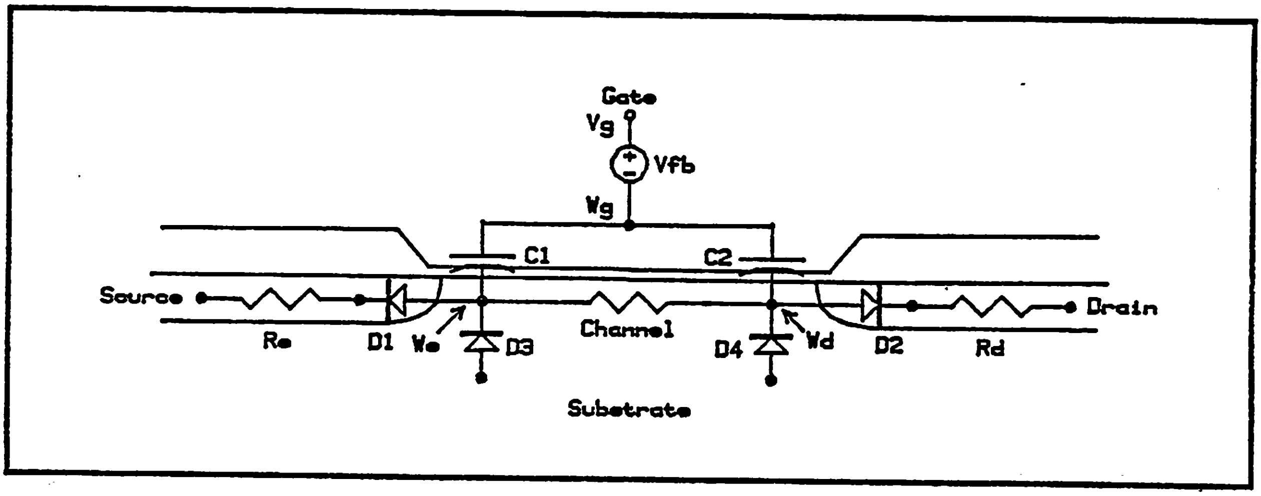


Fig 2 -- MOS Transistor Model

C_1 and C_2 represent gate oxide capacitance. D_1 and D_2 are gate-controlled diodes which connect the channel to the source and drain diffusions. D_3 and D_4 are parasitic substrate diodes. These parasitic diodes are normally represented only by a depletion capacitance. In weak and strong inversion they isolate the channel from the substrate.

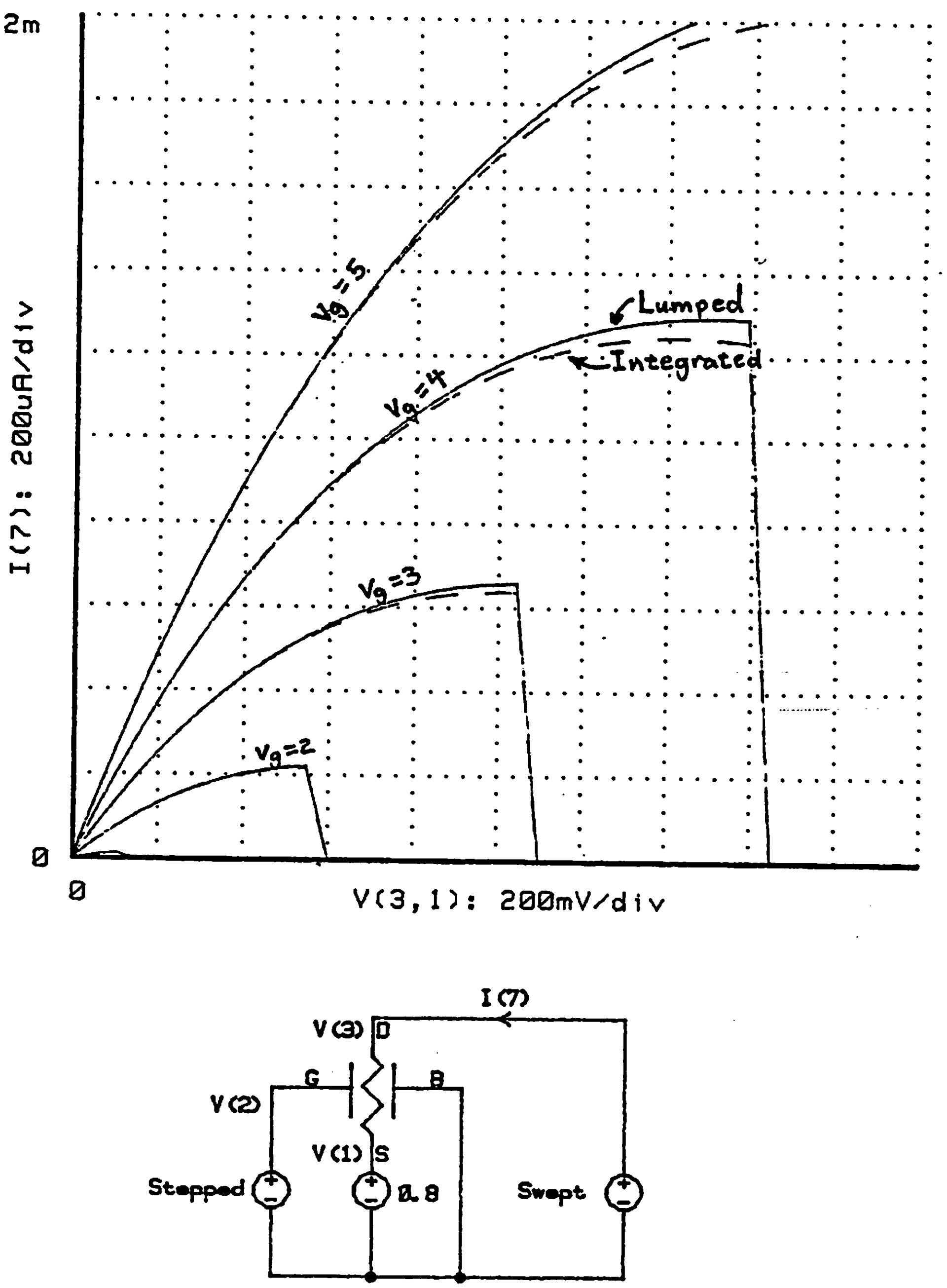


Fig 3 -- Channel Characteristics

A lumped resistor model gives almost the same results as the standard integrated model. In this and following examples $N_s = N_b = 2*10^{16}$ cm^{-3}, $C_{ox} = 3.45*10^{-8}$ F/cm^2, Z=100 microns, L=4 microns and mobility is 700 cm^2/V-sec.

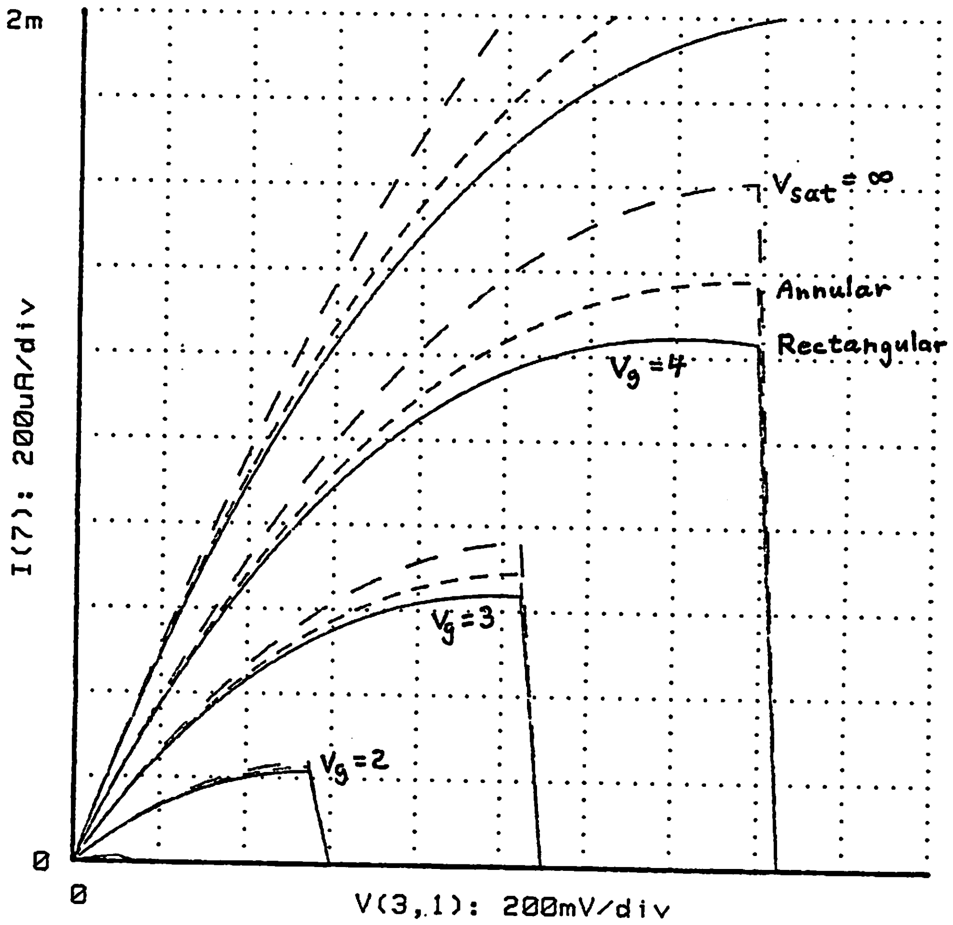


Fig 4 -- Effect of Velocity Saturation

Velocity saturation lowers the saturation current in MOS devices. Annular devices are affected less than rectangular ones if the source is interior. Channel characteristics which include velocity saturation have a region of negative resistance. This region is removed by the connections to the channel. This negative resistance region is the region where the mobile charge density at the drain is less than channel current divided by saturation velocity; it is not physically possible for this to occur.

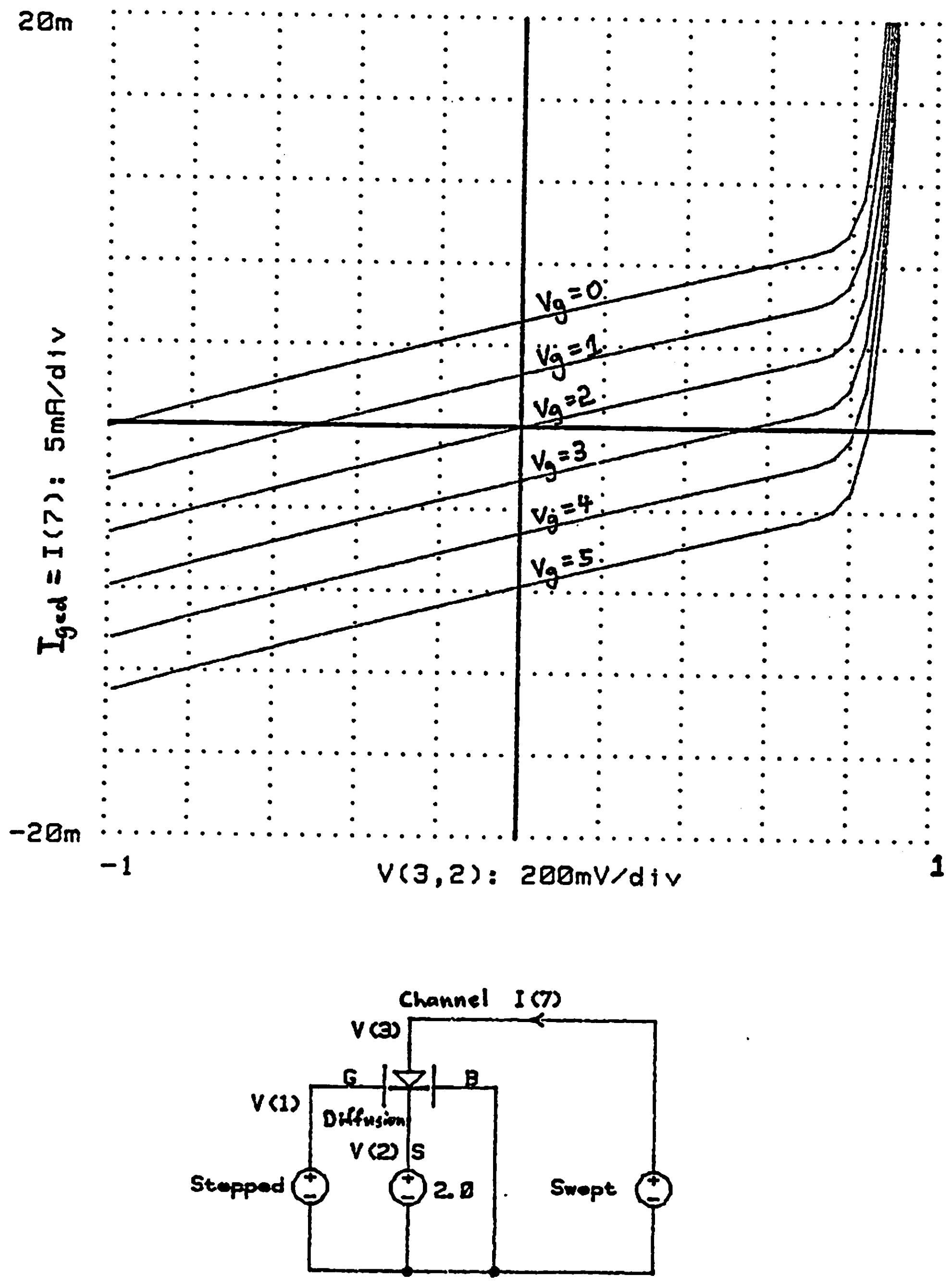


Fig 5 -- Gate-Controlled Diode Characteristics

I-V characteristics of a gate-controlled diode are similar to a conventional junction. Reverse current is controlled by the gate voltage.

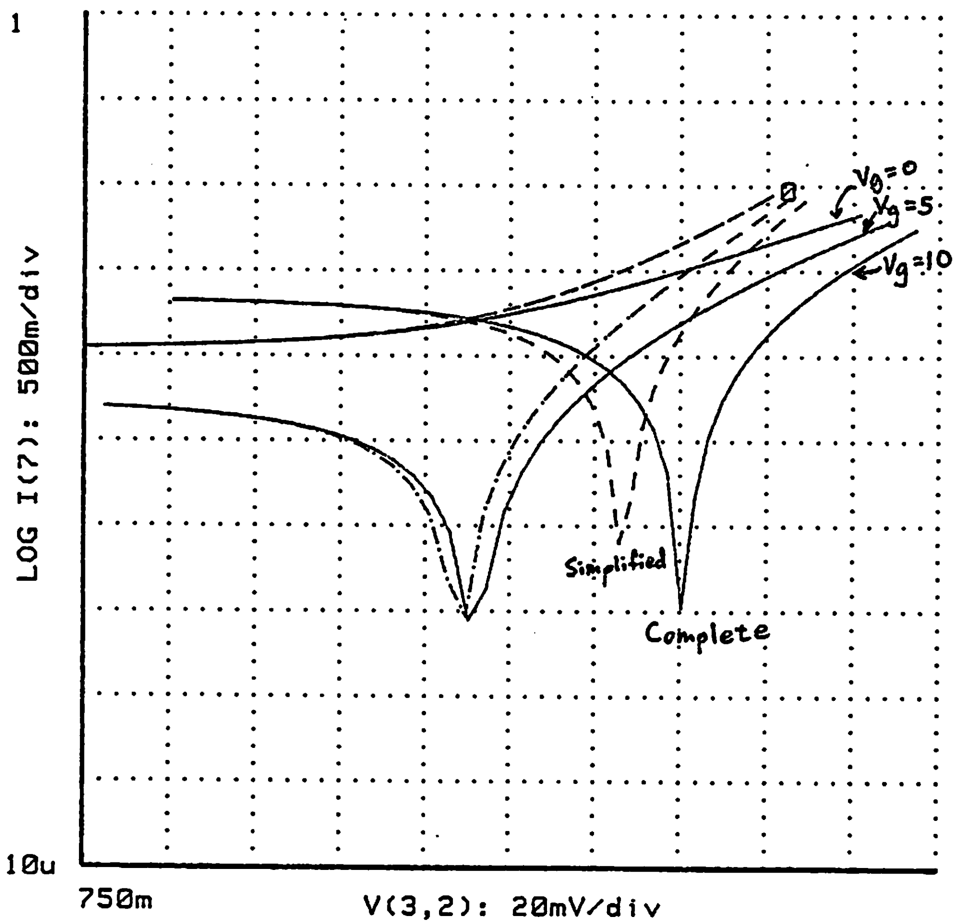


Fig 6 -- Simplified and Complete GCD Models

Simplified (18) and complete (17) gate-controlled diode models give nearly identical results. Worst case difference is approximately 20 mV. The worst case error occurs for high gate voltage, and for this reason it is not significant. Note that except for the log (base 10) scale the conditions are the same as for the previous figure.

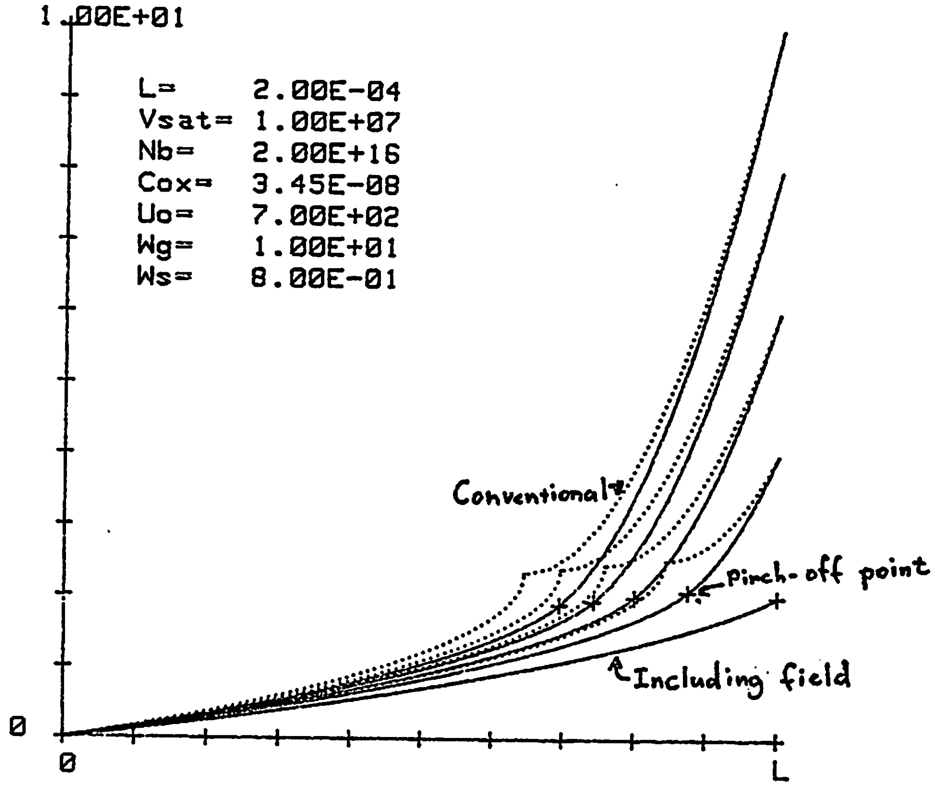


Fig 7 -- Potential Distribution Implied by Channel-Length Modulation

Potential distribution inside the (surface) channel. Source is at the left and drain is on the right. Drain voltage is stepped from 2 to 10 volts in one volt steps. Conventional channel-length modulation equation implies a discontinuity in the field at the pinch-off point. Removing this discontinuity reduces the amount of channel-length modulation and the pinch-off voltage. This will lower the saturation current and drain conductance for high gate voltages.

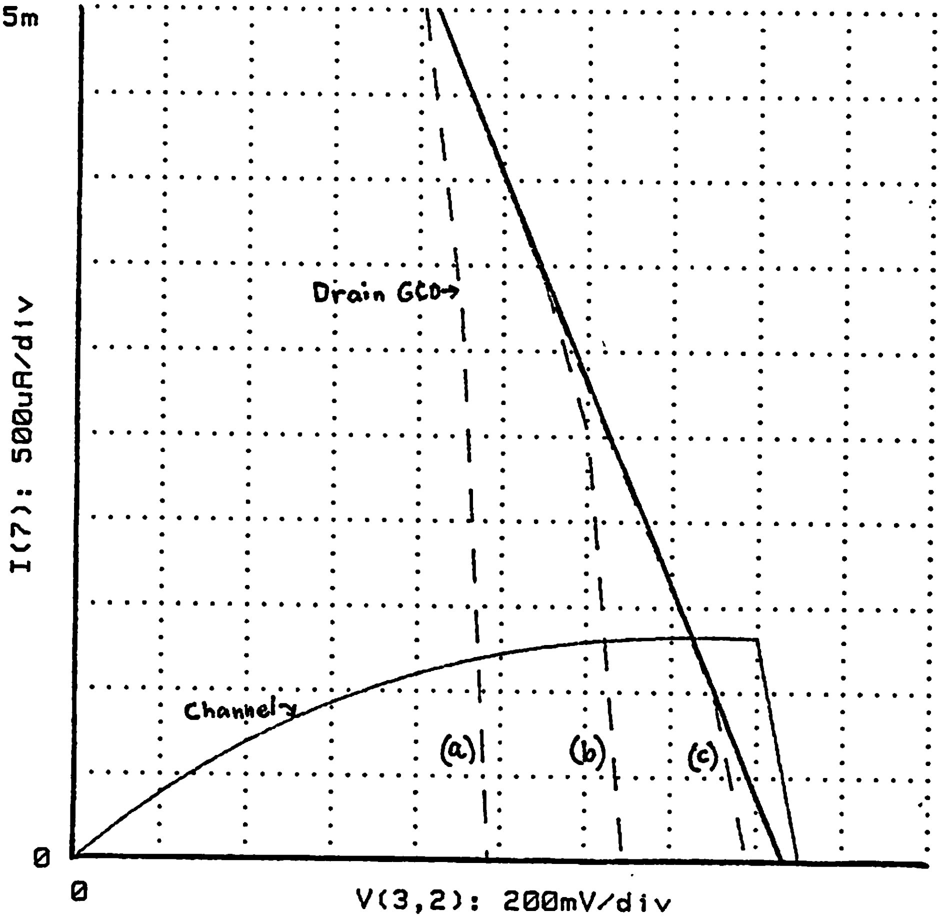
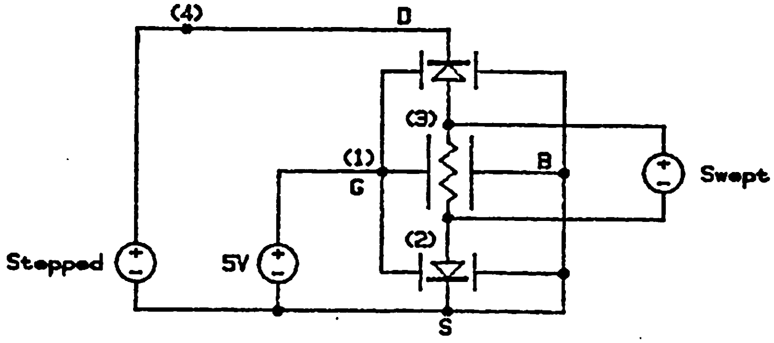


Fig 8 -- Saturation Region Solutions



Graphical solution for drain current. Dashed curves a, b and c represent different drain voltages. As the drain voltage is incresed the potential at the drain end of the channel becomes independant of drain voltage.

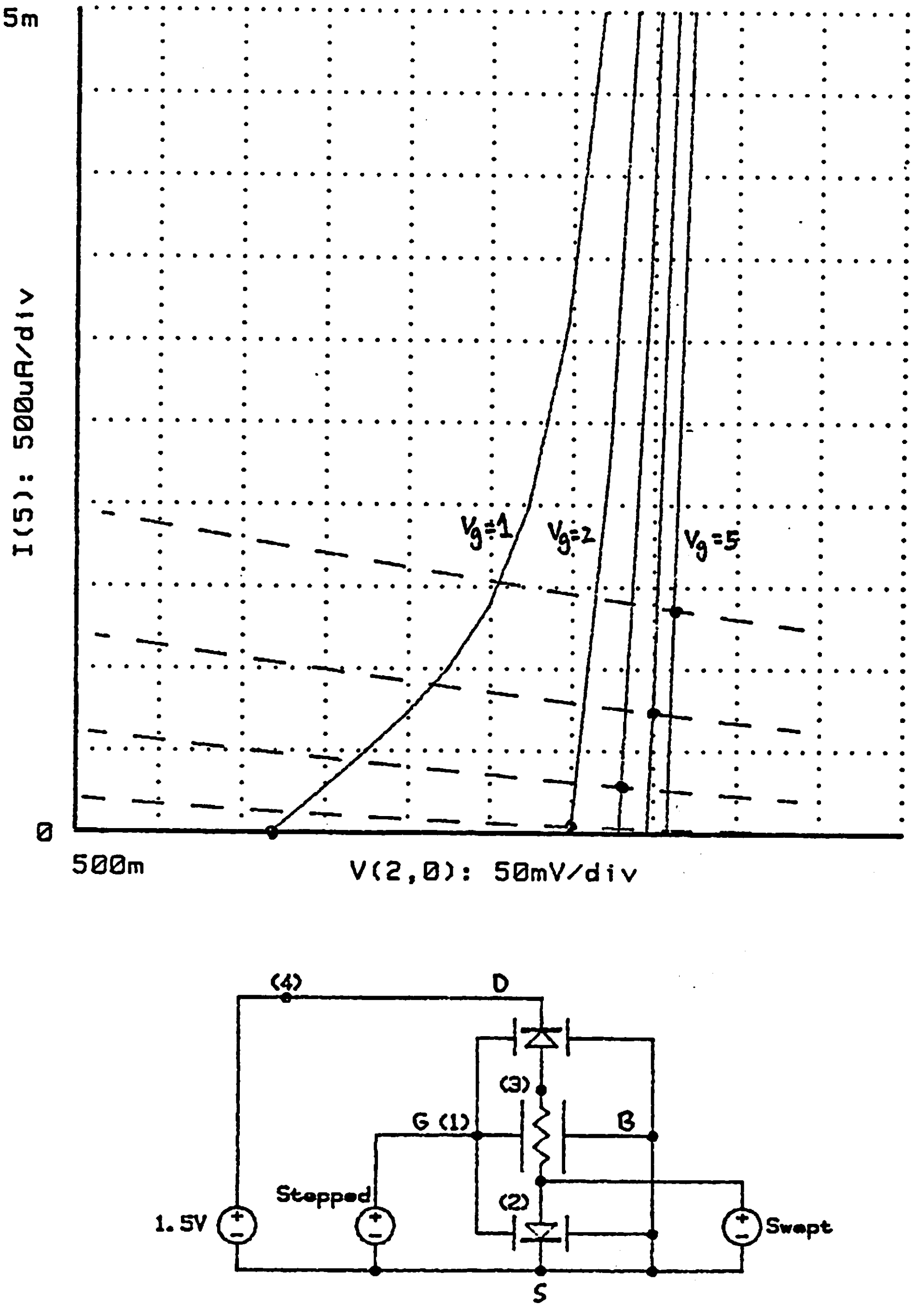


Fig 9 -- Subthreshold Solutions

Graphical solutions for drain current. Vertical scale is drain current; Horizontal scale is potential at source end of channel. Dashed curves are channel characteristics; solid curves are gate-controlled diode characteristics. Dots indicate intersection of corresponding curves.

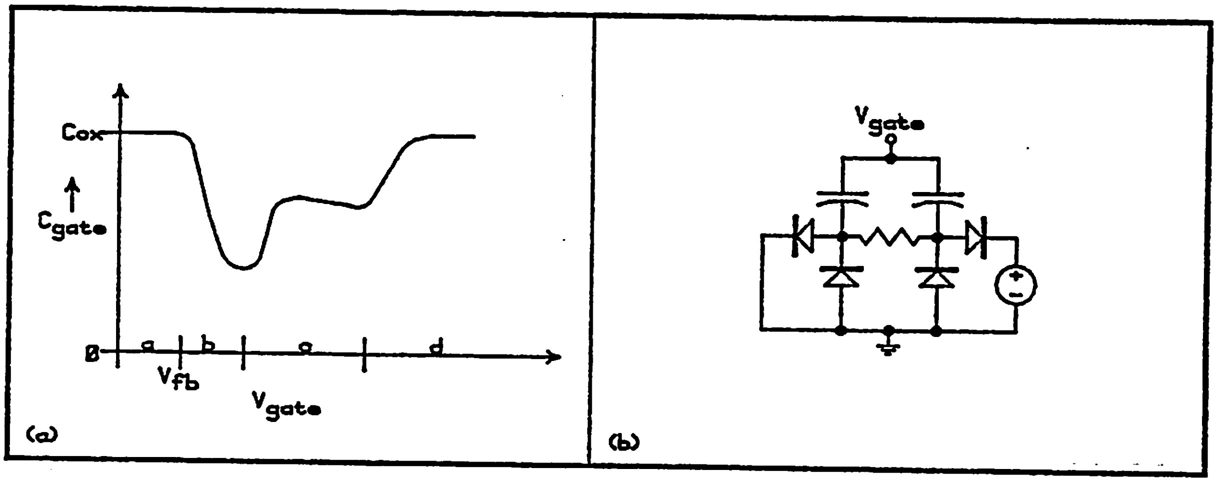


Fig 10 -- C-V Characteristics

C-V characteristics of the MOS transistor come from the gate-controlled and parasitic substrate diodes. In region "a" the parasitic substrate diodes are forward biased, and the capacitance is equal to the oxide capacitance. In region "b" the substrate depletion capacitance is decreasing causing the apparent gate capacitance to decrease. In region "c" the source gate-controlled diode is forward biased; in region "d" both the source and drain gate-controlled diodes are forward biased.

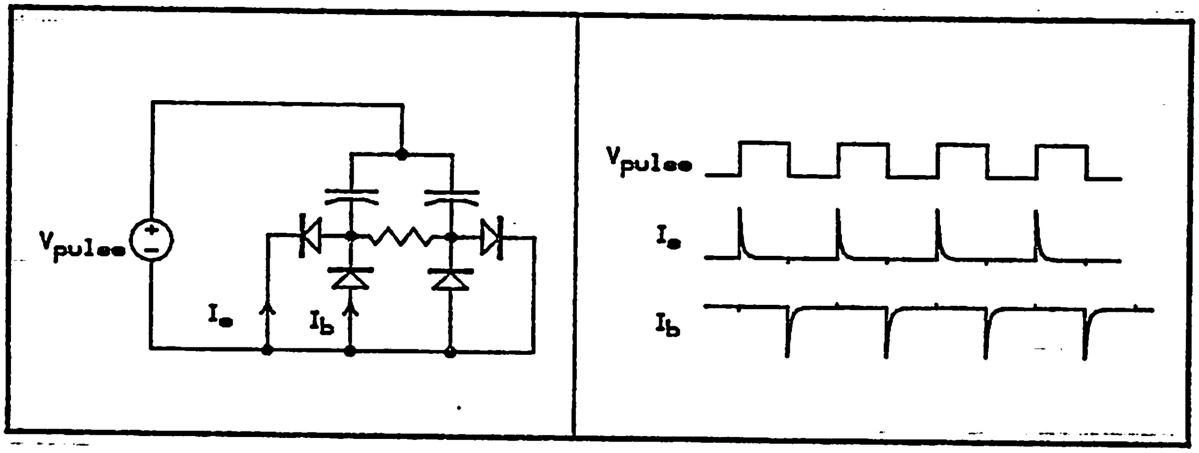
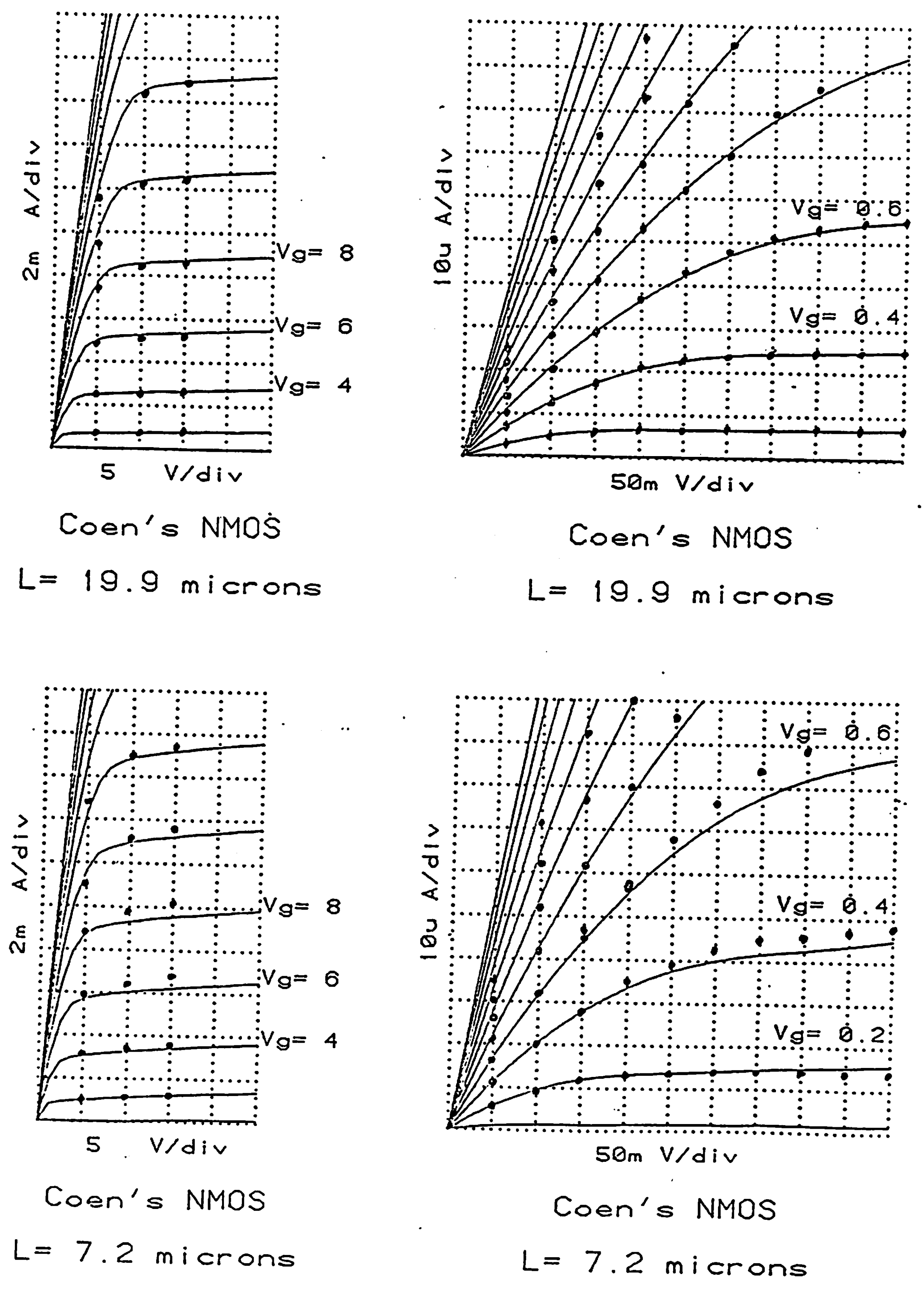


Fig 11 -- Charge Pumping

In "charge pumping" charge flows through the source and drain gate-controlled diodes when the gate voltage rises and through the parasitic sübstrate diodes when the gate voltage falls. The gate-controlled and parasitic substrate diodes form a "voltage clamp" circuit.



Coen's NMOS

L= 19.9 microns

Coen's NMOS

L= 19.9 microns

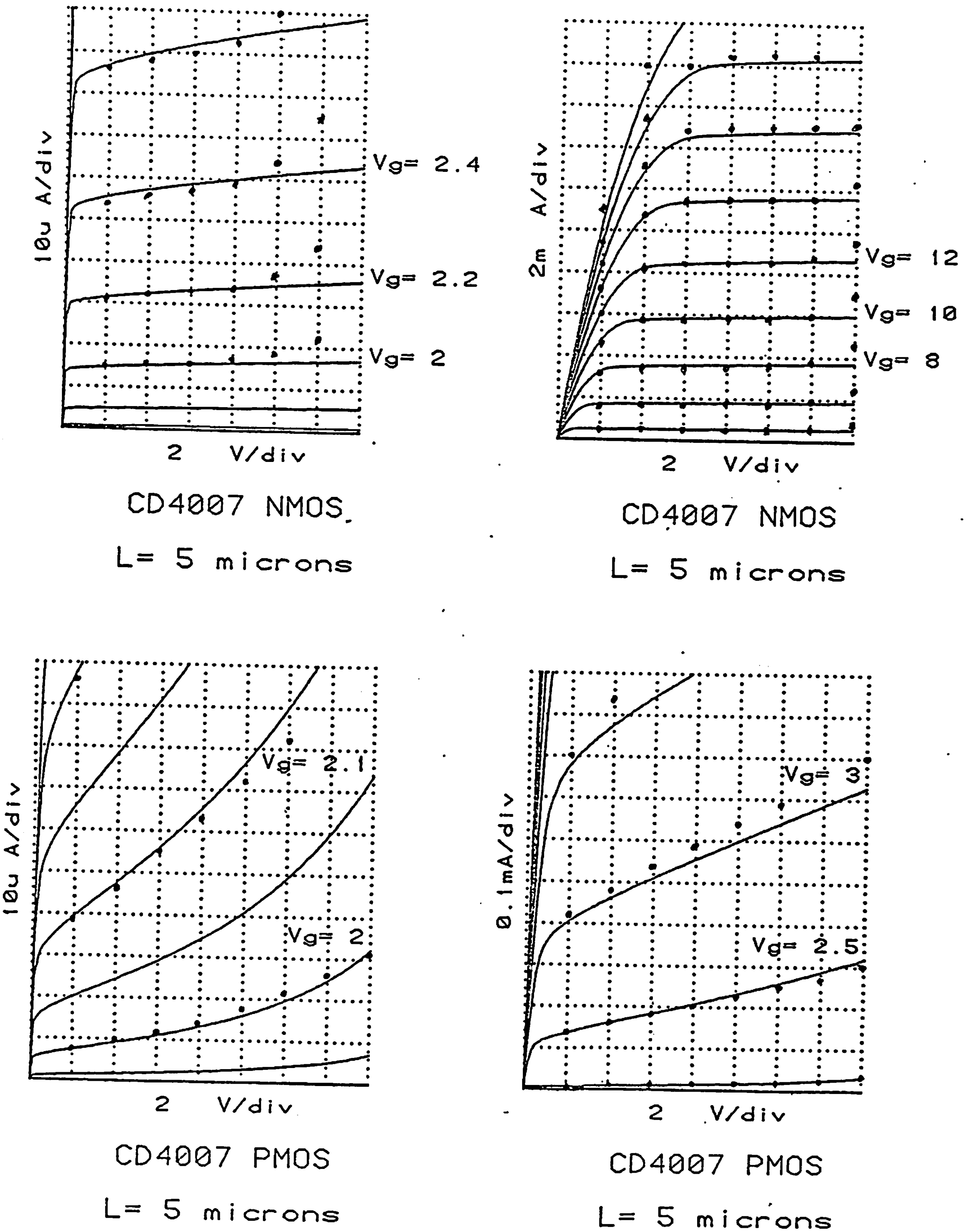
Coen's NMOS

L= 7.2 microns

Coen's NMOS

L= 7.2 microns

Fig 12



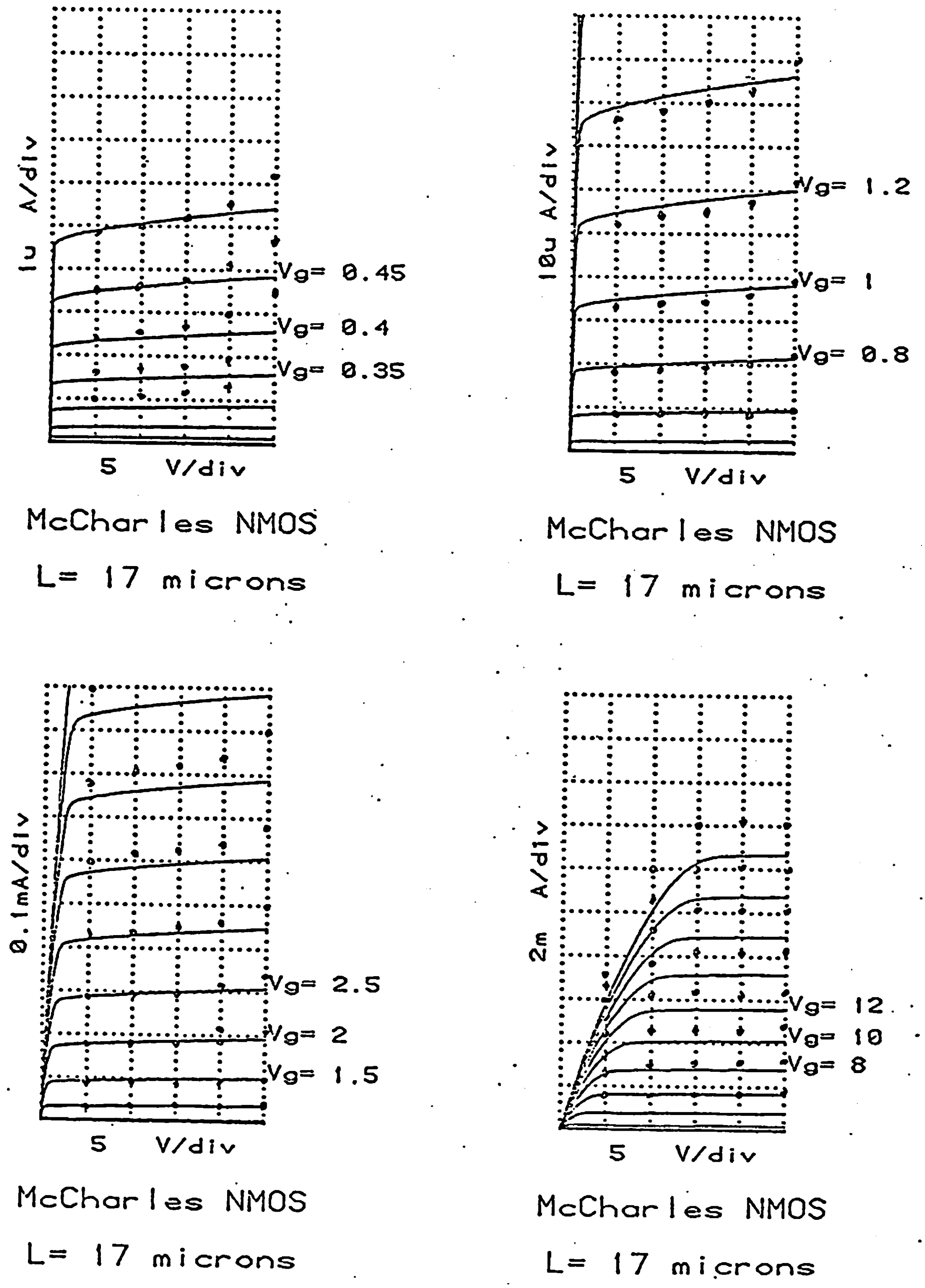
CD4007 NMOS.
L= 5 microns

CD4007 NMOS
L= 5 microns

CD4007 PMOS
L= 5 microns

CD4007 PMOS
L= 5 microns

Fig 13



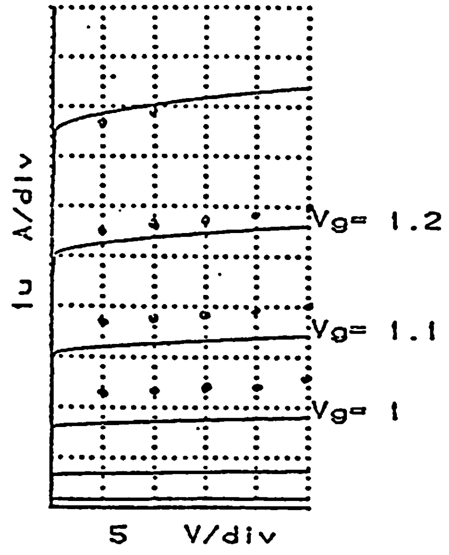
McCharles NMOS
L= 17 microns

McCharles NMOS
L= 17 microns

McCharles NMOS
L= 17 microns

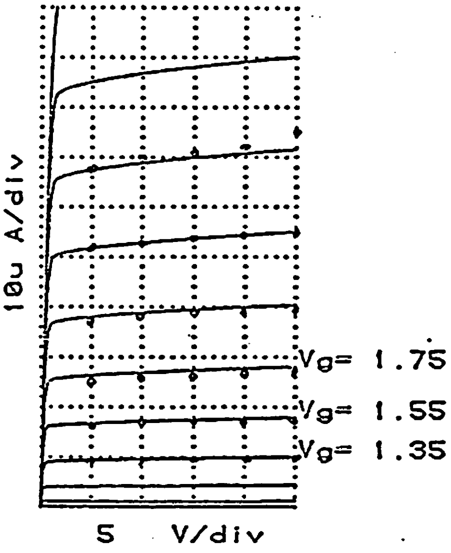
McCharles NMOS
L= 17 microns

Fig 14



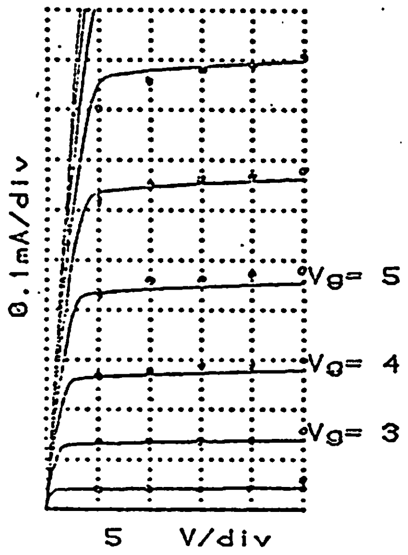
McCharles PMOS

L= 16 microns



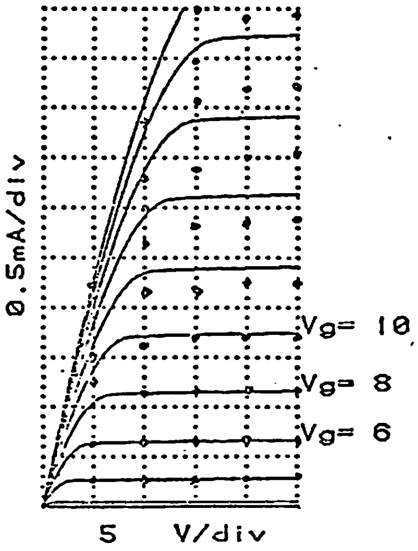
McCharles PMOS

L= 16 microns



McCharles PMOS

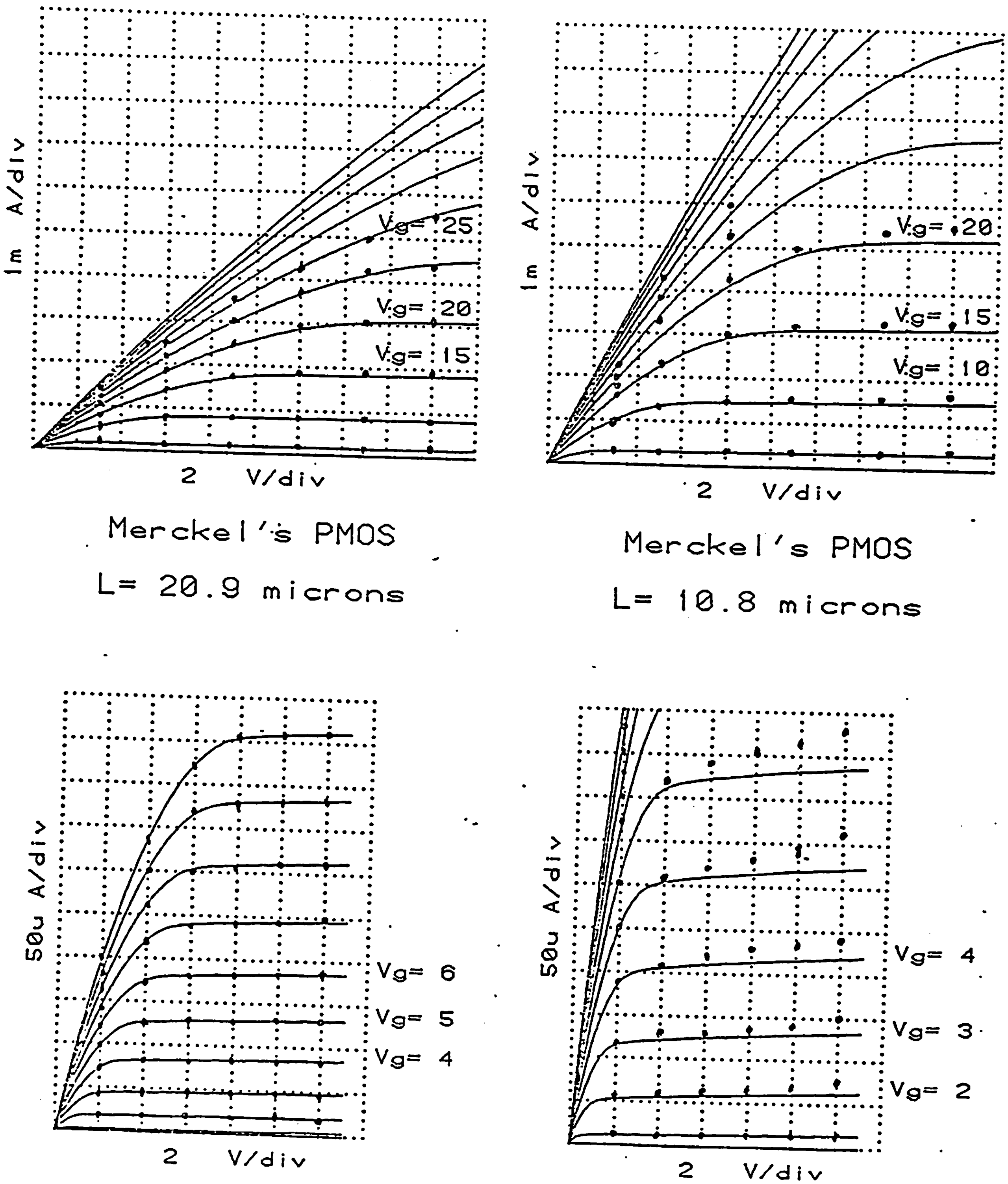
L= 16 microns



McCharles PMOS

L= 16 microns

Fig 15



Merckel's PMOS

L= 20.9 microns

Merckel's PMOS

L= 10.8 microns

Sibbert's NMOS

L= 19.5 microns

Sibbert's NMOS

L= 7 microns

Fig 16

Chapter III

Cyclic A/D Converters and Error Correction

This chapter deals with cyclic A/D converters constructed with clocked analog inverters. We will use the non-restoring divide algorithm discussed in chapter 1 to study the effect of analog circuit errors on converter accuracy. We will introduce and prove the infinite resolution theorem , which shows when these errors are correctable. Two error correction algorithms will be introduced together with experimental results.

Introduction

In chapter 1 we showed how digital algorithms can be converted to analog algorithms using clocked analog arithmetic. A natural use of this technique is to perform A/D conversion. An A/D converter can be constructed using the clocked analog inverters shown in fig 1. All analog parts for this converter can be constructed in a 2 millimeter square of die area. This converter is potentially capable of unlimited resolution, but its accuracy is limited by analog circuit errors.

Analog circuit errors can be reduced by careful attention to design and layout, but if the highest accuracy is required fabrication may prove difficult. Analog circuit errors can also be reduced by adding trim tabs or externally adjustable parts. This increases testing cost, and it will not correct for temperature changes or for aging. Another way of reducing analog circuit errors is by using analog control loops [1].

This chapter discusses a way of reducing analog circuit errors using digital arithmetic. This method has inherent advantages compared to analog techniques. Because critical accuracy requirements are removed from the analog parts, design, layout and fabrication are less critical. Digital calibration data may be stored indefinitely, or it may be updated as often as desired. This allows us to correct for temperature variations or for aging. We will call our approach digital error correction.

The Analog Loop

We will now introduce a notation for quantifying the analog circuit errors. This will allow us to calculate error bounds and prove a crucial theorem.

The model used for the cyclic A/D converter is shown in fig 2. $\{A_i\}$ is the sequence of analog variables presented to the comparator. To simplify the algebra we will assume a one-volt analog reference. For the same reason we will use an unconventional binary representation for the digital output:

$$D=\sum_{i=\emptyset}^{N} B_i 2^{-i-1} \qquad (1)$$

D is the numerical value of the digital output, and $\{B_i\}$ is the sequence of comparator outputs. The individual B_i each have a value of +1 or -1. The value of D always lies in the interval (-1,1).

T(.) denotes the loop transfer function. The ideal loop transfer function is T(x)=2x. The actual loop transfer function can be represented in the following way:

$$T(x)=2x+\sum_{i=\emptyset}^{N} E_i x^i+N_x \qquad (2)$$

Where E is the i^{th} order error in the loop transfer function, and N_x is a stochastic variable with peak magnitude N_p which accounts for noise. The non-restoring divide algorithm can be stated recursively as follows:

$$A\emptyset=V_{in} \qquad (3)$$

$$A_i=T(A_{i-1})-B_{i-1} \tag{4}$$

We will call the noiseless loop transfer function $T_x(.)$ as:

$$T_x(x)=2X+\sum_{i=\emptyset}^{\infty}E_i x^i \tag{5}$$

Several properties of this function will be needed. These properties are physically obvious. $T_x(.)$ is strictly increasing and continuous. The derivative of $T_x(.)$ exists and will lie in the interval $[2-D_{max},2+D_{max}]$, where D_{max} is much less than one.

In the following treatment we will neglect the error introduced by the initial sample-and-hold operation.

Uncorrected Errors

Let us denote the maximum error which can be made in any clock cycle by E_{max}. E_{max} is defined by:

$$E_{max} = \sum_{i=\emptyset}^{\infty}|E_i|+N_P \tag{6}$$

Now suppose we know the value of A_i. Then we can calculate limits on the previous value as follows:

$$\left| A_i - (2A_{i-1} - B_{i-1}) \right| \leqslant E_{max} \tag{7}$$

$$\left| A_{i-1} - \frac{A_i + B_{i-1}}{2} \right| \leqslant \frac{E_{max}}{2} \tag{8}$$

The equivalent reflected error in A_0 after N cycles is then bounded by:

$$\varepsilon \leqslant E_{max} \sum_{i=1}^{N} 2^{-i} \tag{9}$$

As the number of comparisons goes to infinity this becomes

$$\varepsilon \leqslant E_{max} \tag{10}$$

The worst case error in the A/D conversion is approximately equal to the worst case error in the loop transfer function. Figs 3 ,4 ,5 and 6 show simulated error characteristics for A/D conversion error for several different E_i. Each type of error has its own distinctive error plot.

Error Correction

In this section we present an intuitive description of digital error correction. The mathematical description follows.

Under certain circumstances the sequence of comparator results may not contain enough information to allow correction to take place. If two different analog inputs can result in the same sequence of comparator results then no subsequent processing can distinguish between them. This means that the mapping from the analog domain into the sequence of comparator results must be one-to-one. If there is any error in the A/D conversion process then the amount of information in the sequence of comparator output must be greater than the converter resolution. In cyclic A/D converters only N binary results are available in an N-bit result. These A/D converters must be modified to allow redundancy in the output before digital error correction can take place.

One way of introducing redundancy into a digital word is to change the numeric representation from base 2 to some smaller base, such as base 1.9. This increases the number of bits required to represent an input to a given resolution by the ratio of the log of the bases. For example, eight per-cent more bits are required to represent a number in base 1.9 compared to base 2.

The Infinite Resolution Theorem , which is proved in the next section, gives a quantitative statement of how much redundancy is required to tolerate a given amount of analog innacuracy.. In order to perform digital error correction, the A/D converter must first satisfy the infinite-resolution criterion. Then the digital logic or processor must convert from the A/D converter representation (for example, base 1.9) into the desired base 2 representation.

Many sources of analog error cause an uncertainty in the base of numeric representation. These analog errors include capacitor mismatch, finite amplifier gain and slow-settling transients. If the actual base can be determined by digitizing a known input (for example, the reference), then these sources of error can be removed directly using a base conversion algorithm. Most of the other sources of error give rise to an offset error. If the infinite-resolution criterion is satisfied, then these sources can be summarized by an equivalent offset reflected at the input. Similarly, noise in the A/D conversion process can be summarized by a reflected noise at the input. Other sources of error, such as third order distortion (E_3), give rise to errors which can be corrected only in principle.

The Infinite Resolution Theorem

Let the number of comparisons go to infinity. Under what circumstances and within what limits can we reconstruct the analog signal?

If A_0 is the analog input, and D is the digital output, and the following infinite resolution criterion is satisfied:

$$N_P + \sum_{i=0}^{\infty} \left[|E_{2i}| + E_{2i+1} \right] \leqslant 0 \qquad (11)$$

Then the infinite-resolution theorem says there exists a function G(.) such that:

$$|G(D) - A_0| \leqslant \frac{N_P}{1 - E_{max}} \qquad (12)$$

This is a precise statement of the infinite-resolution theorem. To prove it we will need the following lemma:

Lemma

If the infinite-resolution criterion is satisfied then:

$$|A_i| \leqslant 1 \Rightarrow |A_{i+1}| \leqslant 1 \qquad (13)$$

<u>Proof of lemma</u>:

T_x is continuous and strictly increasing. Therefore we only need to show that $T(-1) \geqslant -1$ and $T(1) \leqslant 1$.

$$T(1) = 1 + \sum_{i=\varnothing}^{\infty} E_i + N \qquad (14)$$

$$\leqslant 1 + \sum_{i=\varnothing}^{\infty} E_i + N_P \qquad (15)$$

From the infinite-resolution criterion:

$$\varnothing \geqslant \sum_{i=\varnothing}^{\infty} \left[|E_{2i}| + E_{2i+1} \right] + N_P \qquad (16)$$

$$\geqslant \sum_{i=\varnothing}^{\infty} E_i + N_P \qquad (17)$$

Substituting (17) into (15) gives:

$$T(1) \leqslant 1 \qquad (18)$$

Now for T(-1):

$$T(-1) = -1 + \sum_{i=\varnothing}^{\infty} E_i (-1)^i + N_i \qquad (19)$$

Therefore:

$$T(-1) \geqslant -1 + \sum_{i=0}^{\infty} (E_{2i} - E_{2i+1}) - N_P \tag{20}$$

From the infinite-resolution criterion:

$$0 \leqslant \sum_{i=0}^{\infty} \left[-|E_{2i}| - E_{2i+1} \right] - N_P \tag{21}$$

$$\leqslant \sum_{i=0}^{\infty} (E_{2i} - E_{2i+1}) - N_P \tag{22}$$

Substituting (22) into (20) gives:

$$T(-1) \geqslant -1 \tag{23}$$

And the lemma is proved.

Now for the main theorem. $T_x(.)$ is continuous and strictly increasing. Therefore $T_x(.)$ is invertible. Let its inverse be $S(.)$. Consider the following sequence of functions:

$$G_0(A_0, D) = A_0 \tag{24}$$

$$G_1(A_1, D) = S(A_1 + B_0)$$

$$\vdots$$

$$G_i(A_i, D) = S(A_i + B_{i-1})$$

Each G_i has a derivative with respect to A_i, and this derivative will be in the range:

$$(2+D_{max})^{-i} \leqslant \frac{\partial G_i}{\partial A_i} \leqslant (2+D_{max})^{-i} \tag{25}$$

By induction from the lemma:

$$|A_\varnothing| \leqslant 1 \quad \Rightarrow \quad |A_i| \leqslant 1 \tag{26}$$

Therefore as i aproaches infinity, $G_i(A_i,D)$ converges to the function G(D) which is independant of the analog variable A_i. Because both the noise and the derivative are bounded we know that:

$$|G_i\,(A_i,D)-A_\varnothing| \leqslant \frac{N_p}{(2-D_{max})^i} + \frac{N_p}{(2-D_{max})^{i-1}} + \bullet\bullet\bullet + \frac{N_p}{(2-D_{max})^1} \tag{27}$$

So by using the formula for geometric series we get:

$$|G_i\,(A_i,D)-A_\varnothing| \leqslant \frac{N_p\left[1-\dfrac{1}{(2-D_{max})^i}\right]}{\left[1-\dfrac{1}{2-D_{max}}\right]\left[2-D_{max}\right]} \tag{28}$$

As i aproaches infinity:

$$|G(D)-A\emptyset| \leqslant \frac{N_p}{1-D_{max}} \tag{29}$$

And the theorem is proved.

What we have shown is that if the loop transfer function $T_x(.)$ is known, and the infinite-resolution criterion is satisfied then we can recover the initial analog signal A_0 within a limit set by the noise term by looking only at the sequence of comparator outputs.

First Order Error Correction

Having answered the question of when we can apply error correction, we now turn our attention to how error correction may be applied. The infinite-resolution criterion can be satisfied by deliberately introducing some negative first order error to ensure that the inequality in (9) is satisfied. This means that in each clock cycle we will multiply by $2+E_1$ instead of 2. In effect, we will be digitizing A_0 in base $2+E_1$ instead of base 2. The correct numerical evaluation of the digital word D then becomes:

$$D=\sum_{i=\emptyset}^{N} B_i\,(2+E_1)^{-i-1} \tag{30}$$

If the actual value of E_1 is known, equation (30) can be used to correct for it.

The above algorithm requires one multiply per bit. These multiplies can be eliminated by taking advantage of the fact that E_1 is small. The error between a base 2 interpretation of D and the correct base $2+E_1$ interpretation is:

$$\epsilon_1 = \sum_{i=0}^{N} B_i \left[(2+E_1)^{-i-1} - 2^{-i-1} \right] \tag{31}$$

rewriting this:

$$\epsilon_1 = \sum_{i=0}^{N} B_i 2^{-i-1} \left[(1+E_1/2)^{-i} - 1 \right] \tag{32}$$

and performing a binomial expansion:

$$\epsilon_1 = \sum_{i=0}^{N} B_i 2^{-i-1} (1 - iE_1/2 + OE_1^2 - 1) \tag{33}$$

$$\simeq - \sum_{i=0}^{N} iE_1 B_i 2^{-i-2} \tag{34}$$

This expression for the error can be evaluated with only two adds per bit. An algorithm for performing A/D conversion while correcting for E_1 is shown in fig 7.

Even Order Cancellation

The above techniqes correct for errors in the multiply operation. Such errors could be caused by capacitor mismatch, finite amplifier gain or slow settling transients. It also eliminates the need for a high resolution comparator. However, the first order error correction algorithm does not correct for other sources of error.

Clock feedthrough and substrate leakage give rise to zero order errors which are not improved. Amplifier nonlinearity gives rise to second and higher order errors which are also not improved.

A different technique can be used to remove zero and second order errors. The error in the digital result introduced by even order E is an even function of Vin. By performing a subtract instead of an add to perform the initial sample-and-hold operation, an A/D conversion can also be performed on -Vin. The even-order errors can be removed by performing A/D conversion twice and taking the mean of the results for -Vin and Vin. This algorithm for performing even order cancellation is shown in fig 8.

Determining E_1

To perform first order error correction we must know the value of E_1. The correct value for E_1 can be determined by digitizing V_{ref} and adjusting E_1 to give a corrected result of 1. The algorithm of fig 9 was used in an experimental evaluation of E_1.

Experimental Circuit

An experimental circuit was fabricated by the author in the Berkeley IC facility [1]. This circuit consisted of four ten-picofarad capacitors, a CMOS transconductance amplifier and an analog switch. A schematic of the circuit is shown in fig 10, and a die photo is shown in fig 11.

To form a clocked analog inverter the top plate of the capacitors was internally bonded to the -In connection of the amplifier. The bias input of the amplifier was connected through a resistor to V_{ss}. Bottom plate connects C0, C1 and C2 could be connected to amplifer output, or they could be used as clocked analog inverter inputs.

A schematic of the CMOS Amplifier is shown in fig 12. This circuit is similar to the bipolar RCA CA3080 transconductance amplifier. The input differential pair is formed with P-channel devices P1 and P2, and and they are biased by the current source P7. Devices N2 and N3 reduce the differential pair output currents in N4 and N5. This allows the input stage to run at a higher current level to reduce broadband noise.

Devices N4 and N6 form a current mirror which reflects one of the differential pair outputs into P3. Devices N5, N7 and N8 form an MOS version of the Wilson current source which reflects the other differential pair output to the amplifier output. A similar function is performed by P3, P4 and P5.

Some of the circuits were bonded without the connection the the top plate of the capacitor. This permitted the minus input of the amplifier to be bonded so that the amplifier could be tested seperately. The measured performance characteristics of the amplifier are shown in table I. With an 820 k-ohm bias resistor total supply current was 268 microamperes with a +/- 15 Volt supply. Input offset voltage was 120 mV. No special precautions (such as common-centroid layout) were taken to minimize the offset.

The small-signal differential-mode gain was approximately 80,000, and the unity-gain bandwidth was 800kHz with a 110 pF load. Slew rate was 2 Volts per microsecond. The output voltage range was from -14 to +12

Volts; but gain deteriorated rapidly as the ouput voltage rose above ground. Experimental investigation led to the conclusion that this was caused by substrate avalanche currents in the n-channel device N8. Substrate currents may also have contributed to the high offset voltage. In the results reported below the supply voltage was reduced to +/-9 Volts. The bias resistors were reduced to 72 K-ohm to minimize broadband noise. This also decreased the open-loop gain to approximately 20000.

Experimental Results

The clocked analog inverters described above were used to construct the basic non-restoring divide A/D converter. This converter was interfaced to an S-100 microprocessor system. An interface for a sixteen-bit resolution D/A converter was also constructed. The A/D conversion was performed on the output of the D/A converter. Using a Z-80 microprocessor, the difference between the DAC input and the A/D converter output was taken. This difference was then multiplied by one hundred, and sent back to the D/A converter. The analog result was sampled and viewed on an oscilloscope or plotter. The experimental apparatus is block-diagramed in fig 13.

The uncorrected results from the basic converter are shown in fig 14. The offset from clock feedthrough was 25 mV. The combined analog errors give a worst case error of 25 mV in the loop transfer function. The reference voltage was 5 V. Capacitor mismatch was insignificant.

The dominant sources of error are clearly E_0 and E_1. The worst case error occurs when the input is equal to V_{ref}, and is slightly less than 50 mV. The worst case nonlinearity occurs at -1/2 V_{ref}, and it is approximately 25 mV. These corresond to 1/2 lsb accuracy of 8-bits and a linearity of 9-bits.

Next, the infinite-resolution criterion was satified by placing 220 k-ohm resistors at the outputs of the clocked analog inverters. This introduced a gain error of -2%, making the numeric base 1.96. The resulting uncorrected error plot is shown in fig 15. Notice that while the accuracy is reduced, the spikes in the previous plot have been removed. These spikes are the result of offset when the infinite-resolution criterion is not satisfied.

The results after first-order error correction are shown in fig 16. The worst-case error occurs for $-V_{ref}$ and is about 15 mV, while the worst-case nonlinearity has been reduced to 12 mV. Thes provided 9-bit accuracy and 10-bit linearity. The same results are shown in expanded scale in fig 17. Two plots are superimposed here to show a low-frequency drift which was observed. This time-varying offset had a frequency of a fraction of a hertz and a magnitude of about 4 mV. This offset may be due to surface conduction or 1/f noise in the analog switch. The remaining errors are clearly dominated by offset and quadratic distortion.

The results for first-order correction combined with even-order cancellation are shown in fig 18. The worst-case linearity and accuracy are each 2.5 mV, corresponding to 12-bit overall accuracy. The linearity and monotonicity of the D/A converter used in testing was 13-bits, so a significant fraction of the remaining error may be attributed to it. Other perceptable sources of error include second-order effects of E_1 and third-order harmonic distortion.

Conclusion

We have derived some error properties of cyclic A/D converters. We have shown simple digital algorithms which can be used to improve the accuracy of this type of converter. An experimental A/D converter was constructed using CMOS clocked analog inverters. This converter gave a basic accurcy of 8-bits. The accuracy of this converter was improved to 12-bits using these error correcting techniques. Based on these results, it is possible to construct a 12-bit accurate successive-approximation A/D converter using conventional CMOS fabrication. This converter would have a conversion time of 200 microseconds or less, and would have an active die area of 3.0 mm square or less using 10-micron layout rules.

Acknowlegment

I would like to remember here two people who not only made this work possible, but made my stay in Berkeley tolerable. One of those people is Stew Taylor, whose desire to learn about MOS miraculously coincided with my desire to learn about operational amplifiers. But the person to whom I owe the most is Bill Black. His desire to develop a CMOS process miraculously coincided with my need to fabricate a CMOS circuit.

References for Chapter III

[1] Hornak, T. and Corcoran, J.

A High Precision Component-Tolerant A/D Converter IEEE Journal of Solid-State Circuits, Vol. SC-10, No. 6, December 1975.

[2] McCharles, R., Saletore, V., Black, W. C., and Hodges, D. A.

An Algorithmic A/D Converter 1977 IEEE ISSCC Digest of Technical Papers, pp. 96,97.

Table I

CMOS Amplifier Measurements

(Vdd=+15; Vss=-15; R_{bias}=820KΩ; C_{load}=110pF)

Supply Current		270	μA
Transconductance		2	m℧
Output Impedance		40	MΩ
Output Range	(Plus)	+12	V
	(Minus)	-14	V
Input Offset Voltage		120	mV
Slew Rate		2	V/μsec

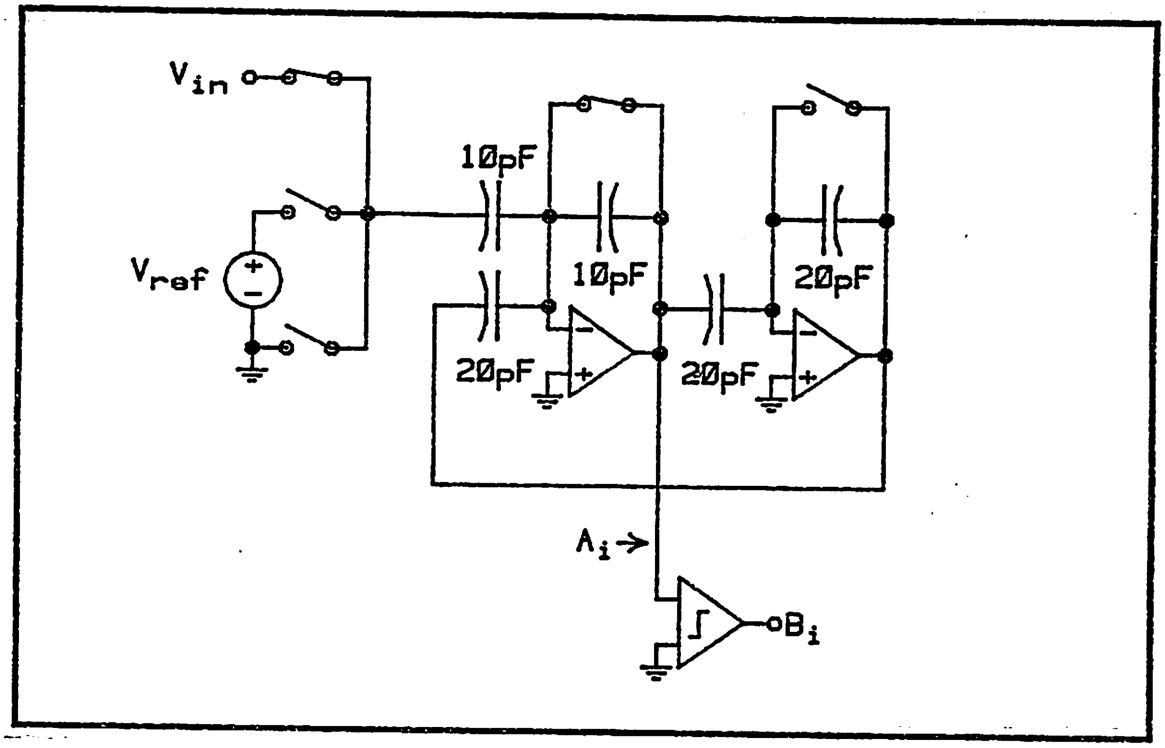


Fig 1 -- Cyclic A/D Converter

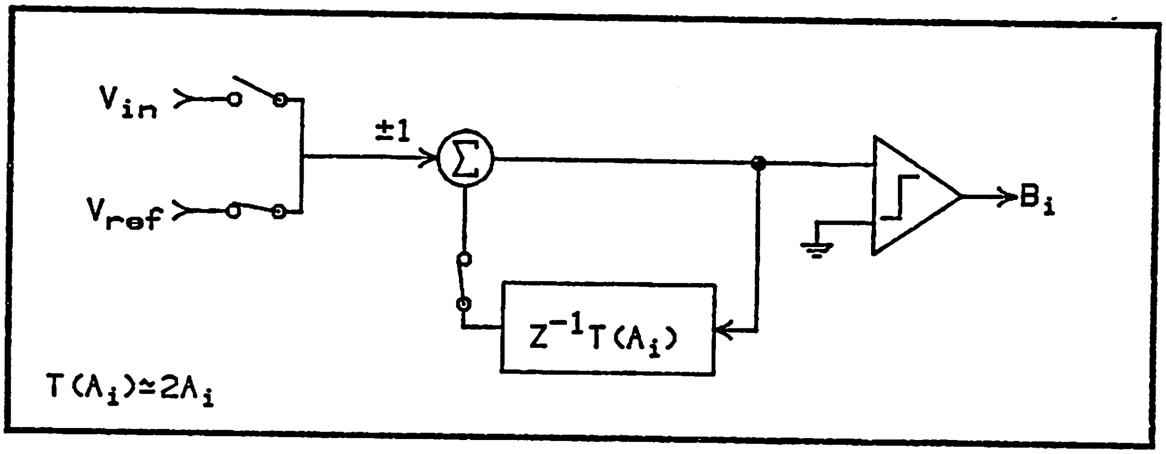


Fig 2 -- A/D Block Diagram

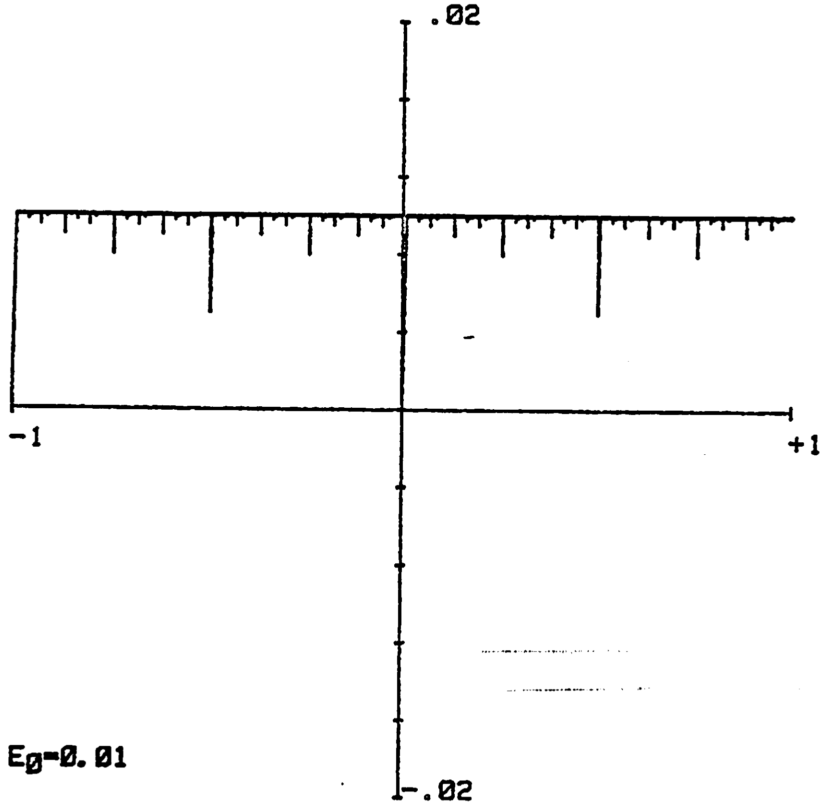


Fig 3 -- Error Characteristics of EØ

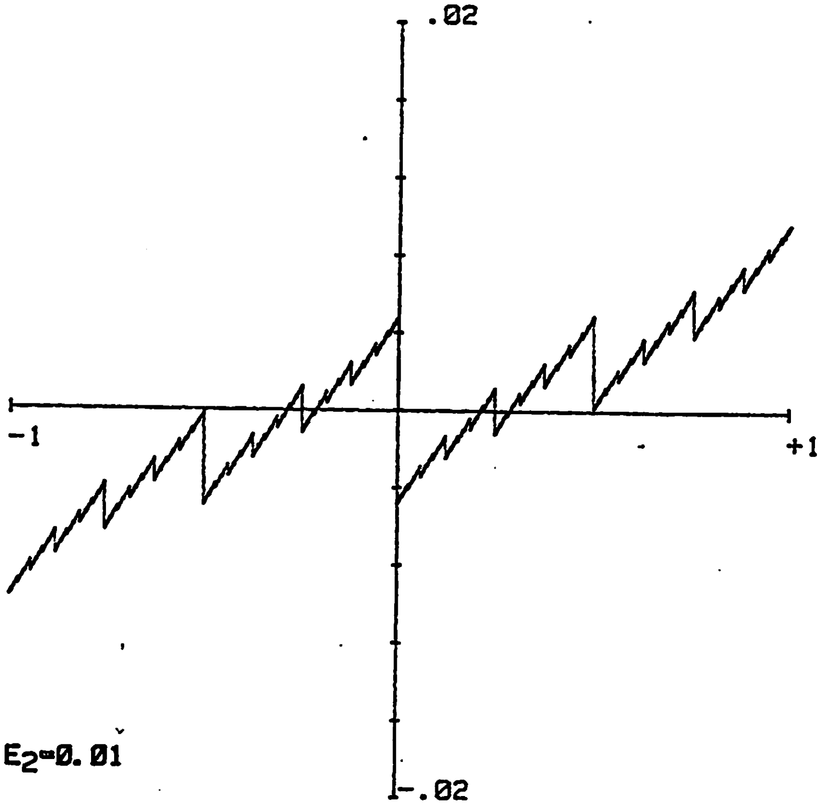


Fig 4 -- Error Characteristics of E_1

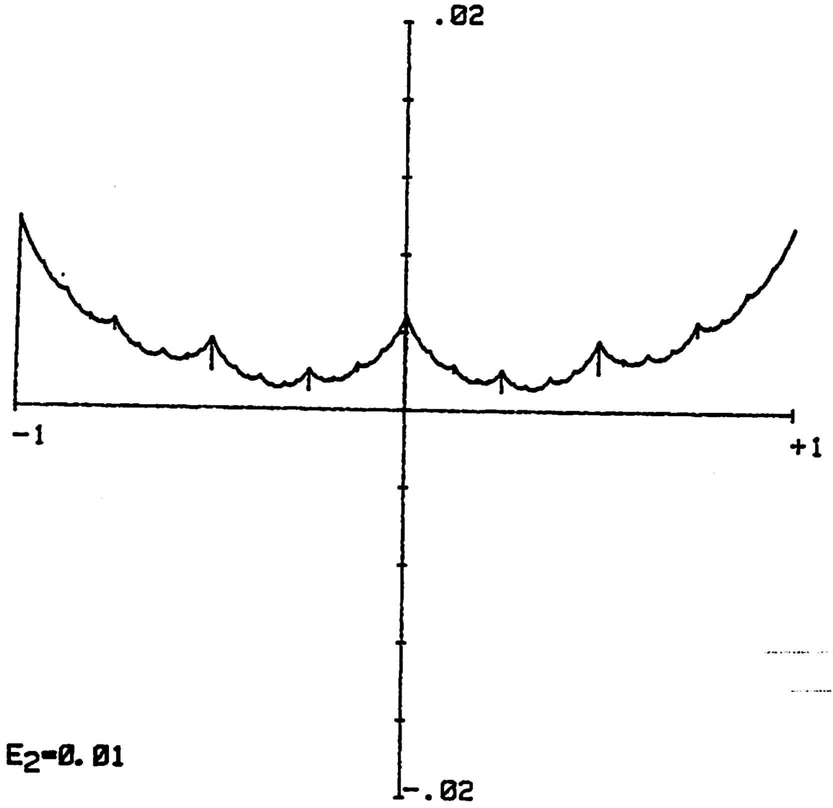


Fig 5 -- Error Characteristics of E_2

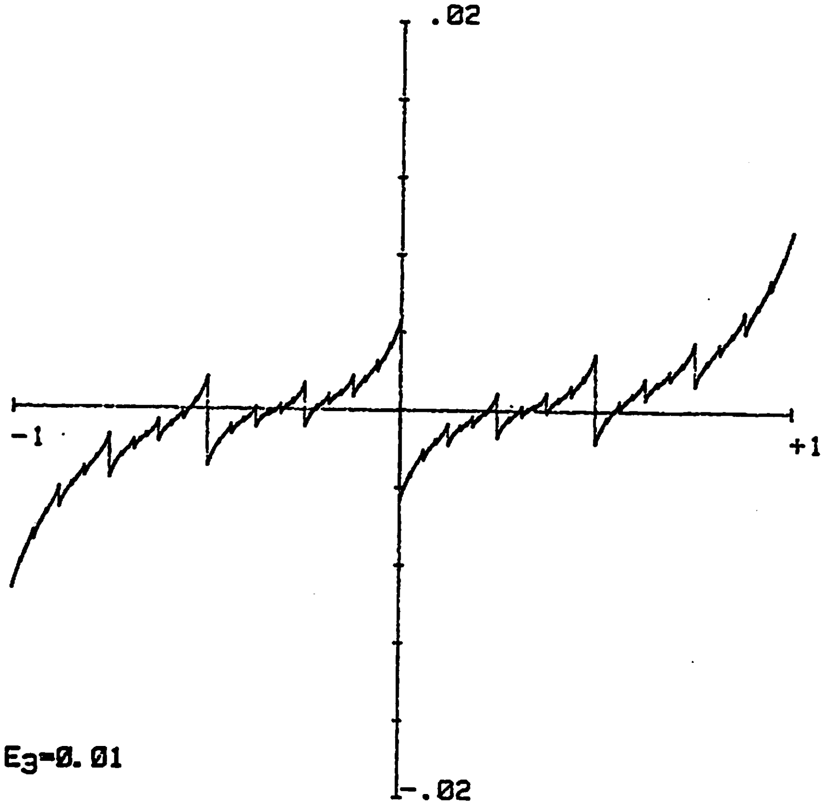


Fig 6 -- Error Characteristics of E_3

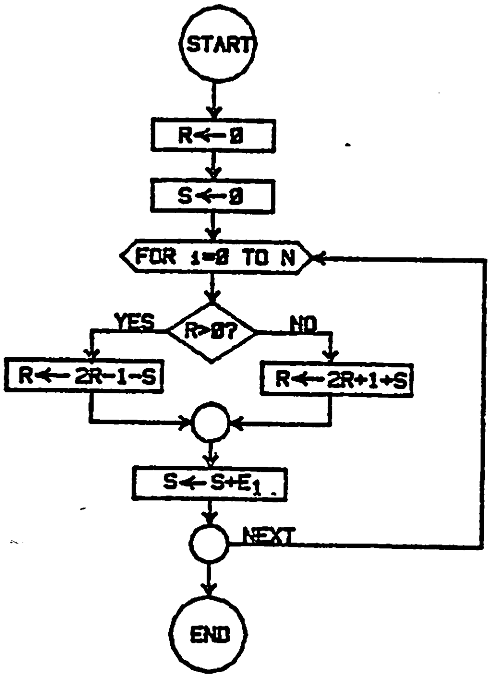


Fig 7 -- Arithmetic Operations Performed in First Order Error Correction

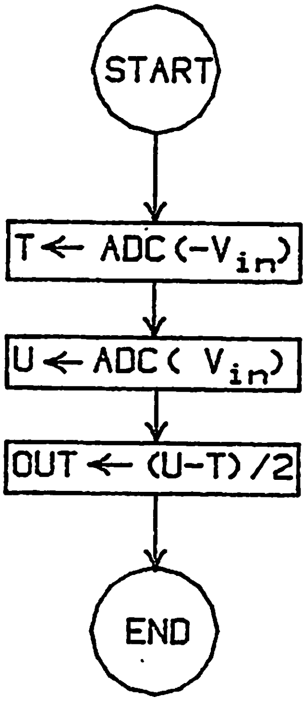


Fig 8 - Algorithm for Even Order Cancellation

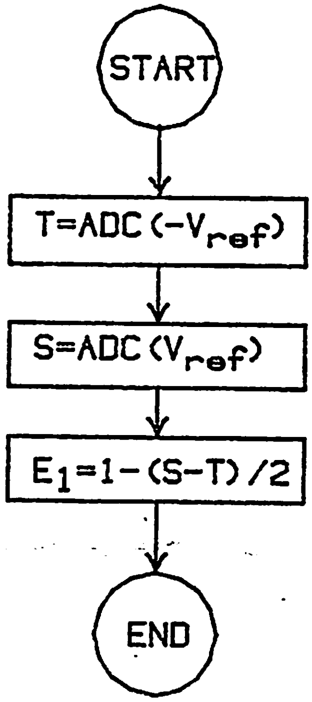


Fig 9 -- Algorithm for Determining E_1

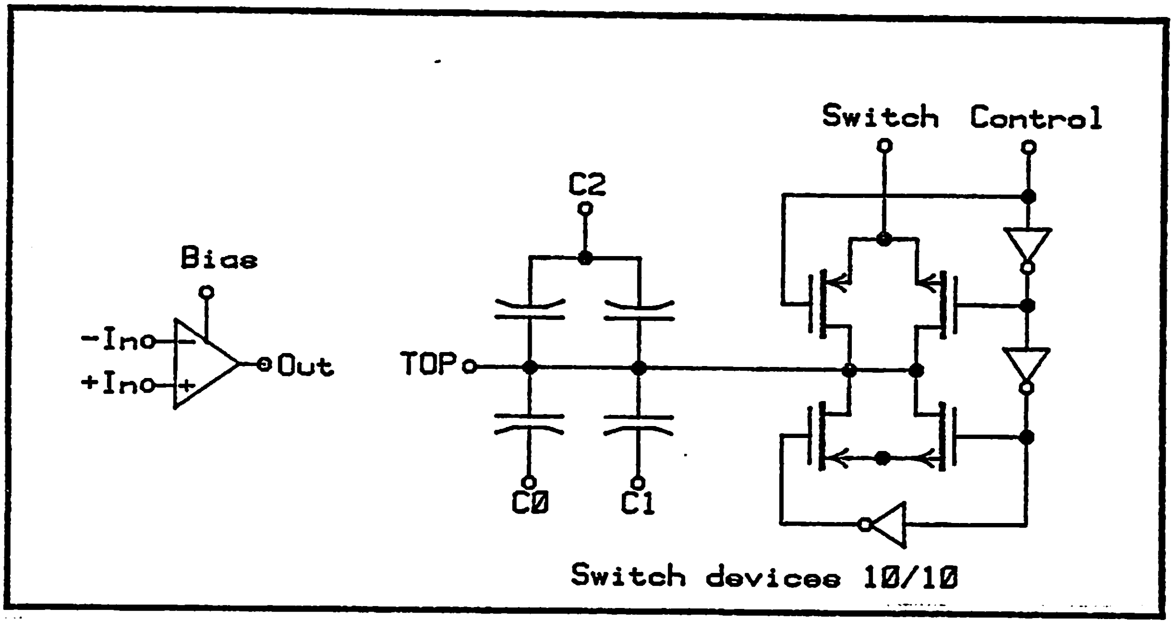
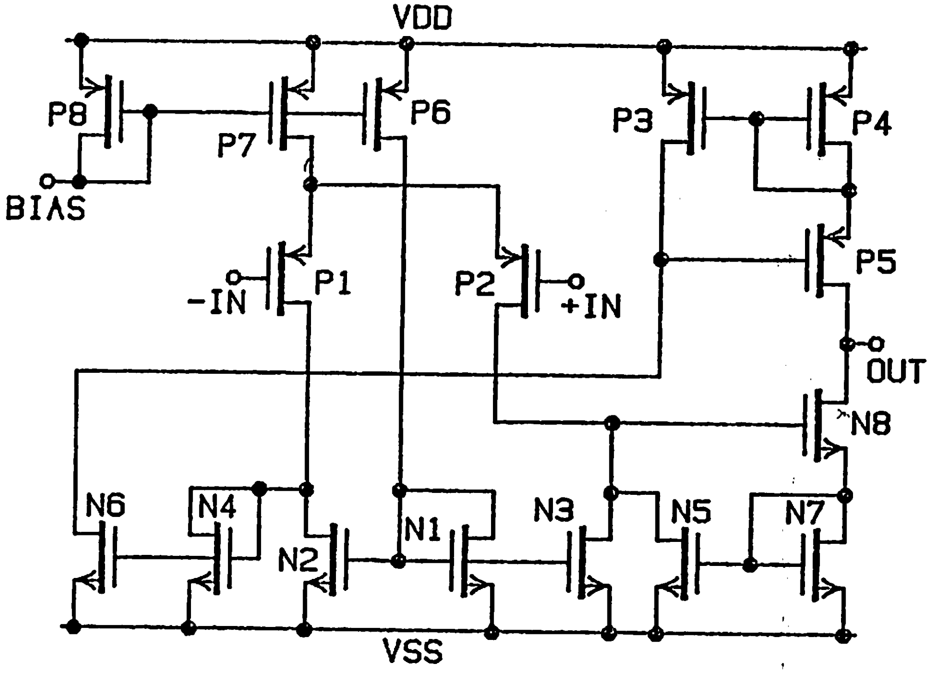


Fig 10 - Experimental IC



Device	Z/L	Device	Z/L
N1	40/20	P1	500/20
N2	80/20	P2	500/20
N3	80/20	P3	100/20
N4	80/20	P4	500/20
N5	80/20	P5	500/20
N6	40/20	P6	500/20
N7	200/20	P7	100/20
N8	200/20	P8	250/20

Fig 11 - CMOS Amplifier

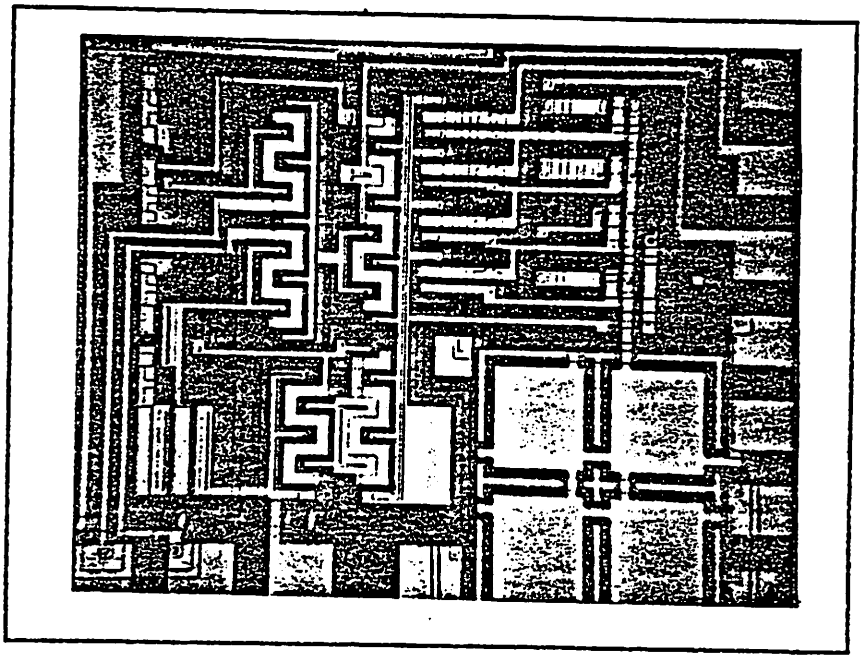


Fig 12 -- Die Photo of Experimental IC.

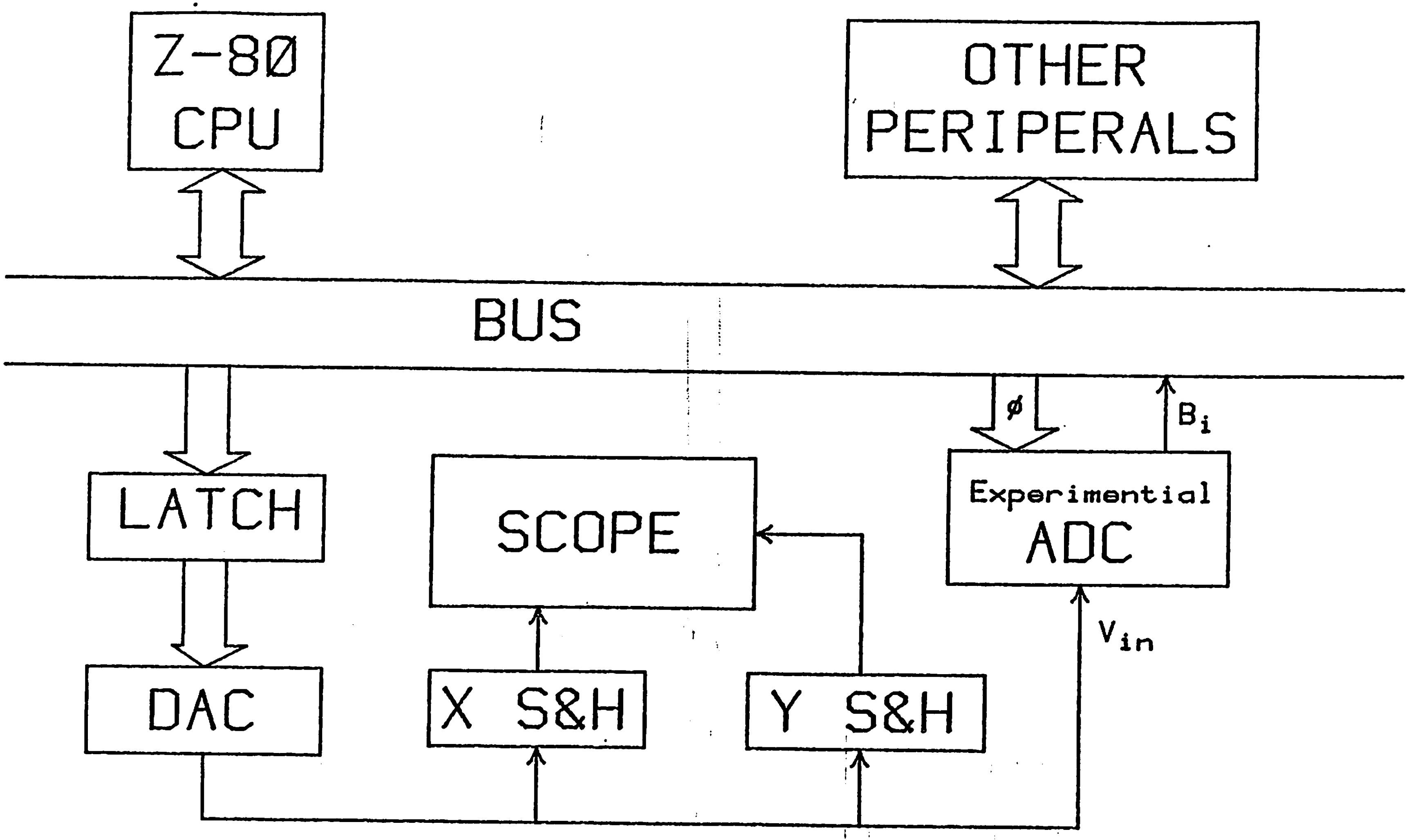


Fig 13 -- Experimental Apparatus

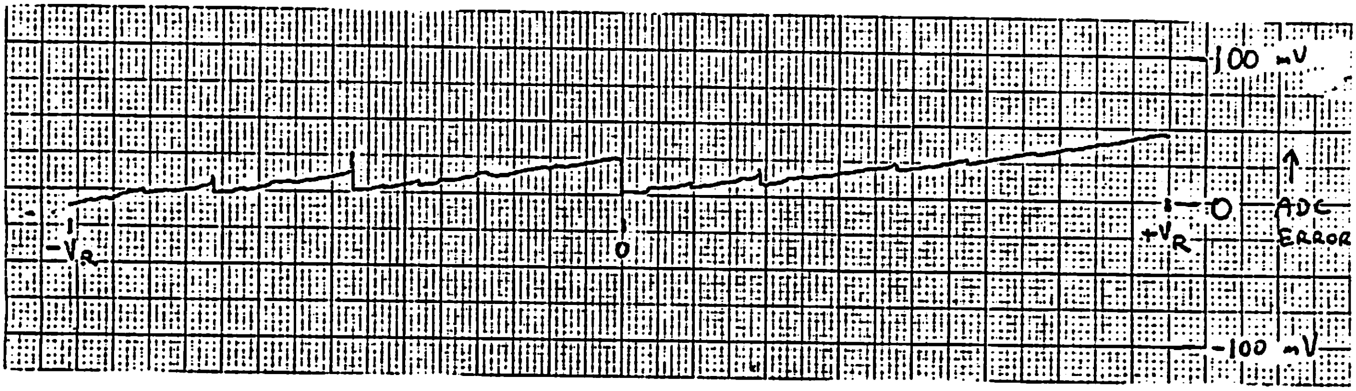


Fig 14 - Uncorrected cyclic A/D conversion error. Reference is 5 Volts.

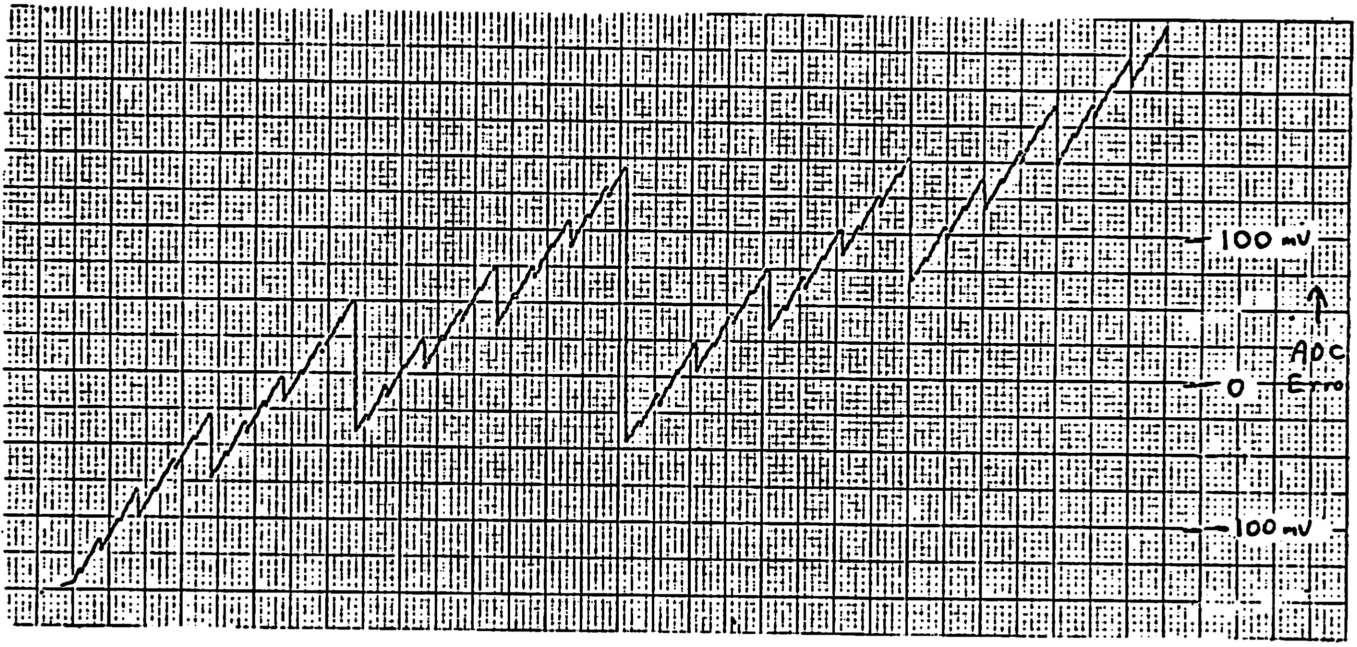


Fig 15 - Uncorrected A/D conversion error when infinite-resolution criterion is satisfied. Note the abscence of spikes in the curve compared to fig 14.

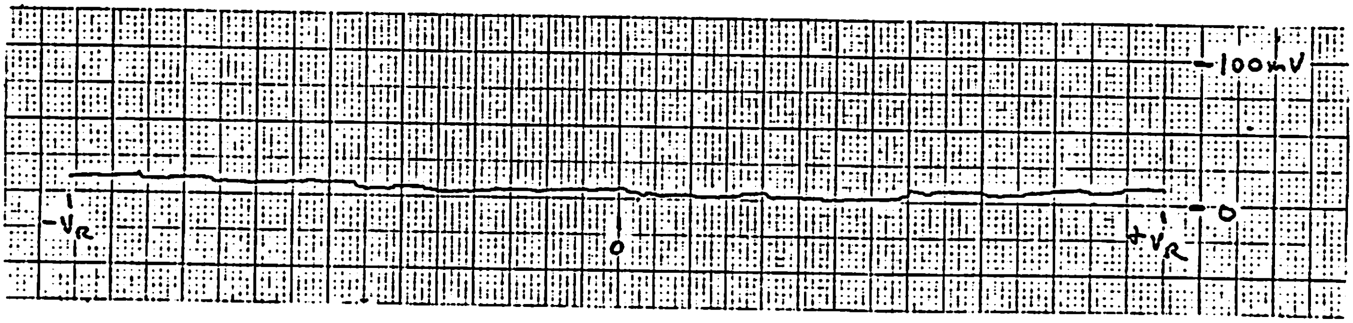


Fig 16 - A/D conversion error after applying first-order error correction.

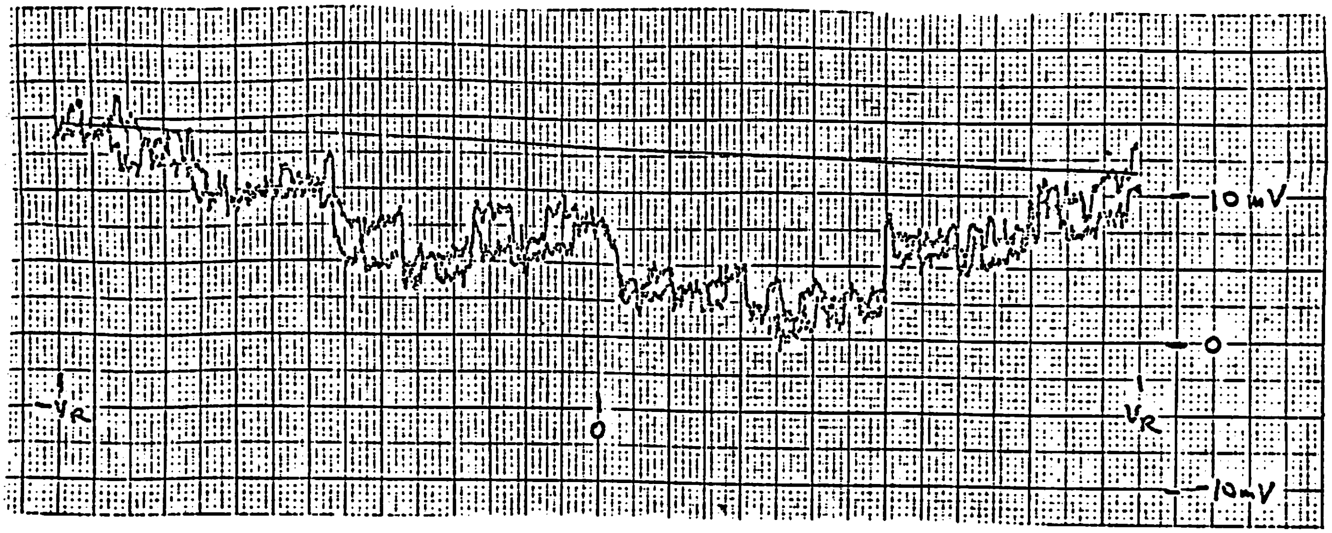


Fig 17 - Expanded scale of fig 16. Two error plots are superimposed to show effect of low-frequecy drift.

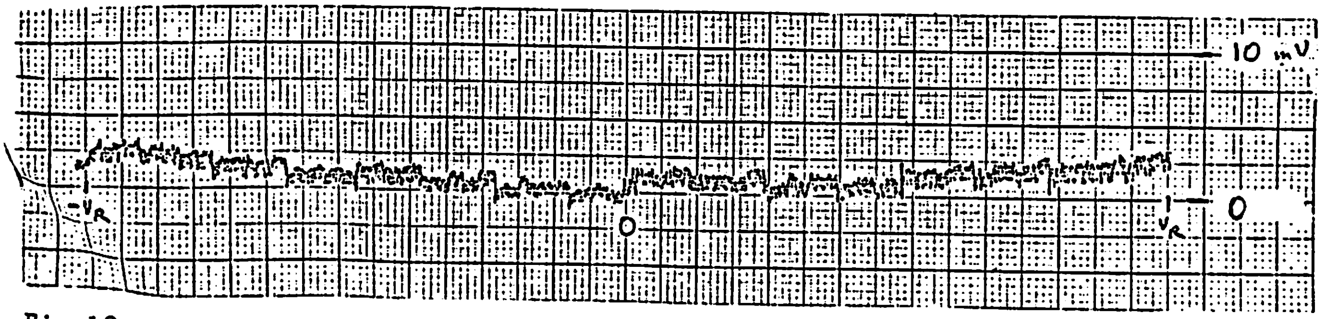


Fig 18 - Expanded scale of A/D conversion error after first-order correction and even-order cancellation.