The Design of High-Performance Analog Circuits on Digital CMOS Chips

ERIC A. VITTOZ, MEMBER, IEEE

Abstract—**Devices available in digital oriented CMOS processes are reviewed, with emphasis on the various modes of operation of a standard transistor and their respective merits, and on additional specifications required to apply devices in analog circuits. Some basic compatible analog circuit techniques and their related tradeoffs are then surveyed by means of typical examples. The noisy environment due to cohabitation on the chip with digital circuits is briefly evoked.**

I. Introduction

THE EVOLUTION of scaled-down digital processes will shift the boundary between digital and analog parts of systems [1]. However, analog circuits will remain irreplaceable components of systems-on-a-chip. Besides A/D conversion, they will always be needed to perform a variety of critical tasks required to interface digital with the external world, such as amplification, prefiltering, demodulation, signal conditioning for line transmission, for storage, and for display, generation of absolute values (voltages, currents, frequencies), and to implement compatible sensors on chip. In addition, analog will retain for a long while its advantage over digital when very high frequency or very low power is required.

Most of the limitations of analog circuits are due to the fact that they operate with electrical variables and not simply with numbers. Therefore, their accuracy is fundamentally limited by unavoidable mismatches between components, and their dynamic range is limited by noise, offset, and distortions.

For economical reasons, the analog part of a system-on-a-chip must be fully compatible with a process basically tailored for digital requirements, and this with a minimum number of additional specifications. Section II will review all active and passive devices available in digital CMOS processes, together with the additional specifications needed to use them for implementing analog circuits. Some basic analog circuit techniques will then be described in Section III by means of typical examples. Finally, the problems related to the noisy environment due to cohabitation on the chip with large digital circuits will be briefly evoked in Section IV.

Manuscript received November 2, 1984; revised December 24, 1984.
The author is with Centre Suisse d'Electronique et de Microtechnique, (CSEM, formerly CEH) Maladiere 71, 2000 Neuchâtel 7, Switzerland.

II. Devices Available for Analog Circuits

A. Transistors

A clear understanding of the various ways of biasing a normal MOS transistor, and of their respective merits, is a key factor in the design of optimum analog subcircuits. Fig. 1 illustrates the complete transfer characteristics $I_D(V_G)$ of a n-channel MOS transistor in saturation for various possible modes of operation. For the sake of symmetry, all potentials are defined with respect to that of the local substrate, in this case the p-well.

The general behavior of drain current in saturation I_D in the two basic modes of field effect operation can be described by two separate approximative models [2], [3] which sacrifice accuracy to clarity and simplicity:

Strong inversion ($I_D \gg \beta U_T^2$)

$$I_D = \frac{\beta}{2n}(V_G - V_{T0} - nV_S)^2, \quad \text{for } V_D > V_{D\,sat} = (V_G - V_{T0})/n. \quad (1)$$

Weak inversion ($I_D \ll \beta U_T^2$)

$$I_D = K\beta U_T^2 \exp((V_G - V_{T0} - nV_S)/nU_T), \quad \text{for } V_D > V_{D\,sat} = 3 \text{ to } 4U_T. \quad (2)$$

These models only include the three most important device parameters required for circuit design

$\beta = \mu C_{ox} W/L$ transfer parameter for strong inversion

V_{T0} gate threshold voltage for $V_S = 0$

n slope factor in weak inversion, which also describes approximately the effect of fixed charges in the channel in strong inversion. Its value depends slightly on V_S [3] and ranges usually from 1.3 to 2.

K is a factor somewhat larger than 1, which connects weak and strong inversion. Its exact value has no importance in circuit design, since transistors in weak inversion must be biased at fixed drain current I_D to avoid the very high sensitivity to $U_T = kT/q$ and V_{T0} for fixed gate voltage V_G.

In CMOS logic circuits, transistors usually operate with $V_S = 0$ as shown in heavy lines in Fig. 1. Their role is to

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-20, no. 3, pp. 657–665, June 1985.

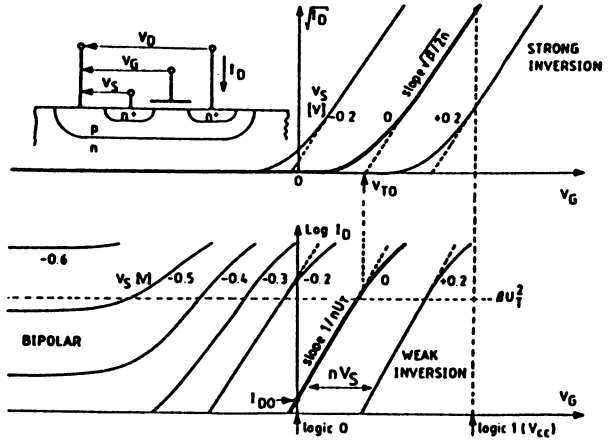


Fig. 1. General transfer characteristics $I_D(V_G)$ of a MOS transistor in saturation. Different modes of operation can be identified, namely strong inversion, weak inversion, and bipolar.

provide maximum drain current in the "on" state for $V_G = V_{cc}$, and minimum residual current I_{D0} in the "off" state for $V_G = 0$. The only requirements for digital circuits are thus a maximum possible value of transfer parameter β, and a value of threshold voltage V_{T0} as low as possible while ensuring acceptable value of residual current for $V_G = 0$

$$I_{D0} = K\beta U_T^2 \exp(-V_{T0}/nU_T). \tag{3}$$

This is only possible if slope factor n of weak inversion is not too large. These requirements are also favorable to analog circuits, since they allow a maximum value of transconductance g_m which can be easily derived from (1) and (2) as

$$g_m = ((2\beta I_D)/n)^{1/2} = 2I_D/(V_G - V_{T0} - nV_S) \quad \text{(strong inversion)}. \tag{4}$$

$$g_m = I_D/nU_T \quad \text{(weak inversion)}. \tag{5}$$

However, specifications on the maximum range of variation of β, V_{T0}, and n are usually necessary.

Transconductance g_m is proportional to drain current I_D in weak inversion, but only to the square root of I_D in strong inversion. If source voltage V_S is not zero, the gate voltage for constant drain current is shifted by nV_S for both modes of operation. Thus transconductance g_{ms} from source to drain is given by

$$g_{ms} = ng_m. \tag{6}$$

Fig. 1 also shows that when gate voltage V_G is sufficiently negative, it has not more effect on drain current, which means that gate transconductance g_m decreases to zero. However, I_D can still be controlled by negative values of source voltage V_S which corresponds to a forward-biased source junction. The device then operates as a lateral bipolar, with the flow of carriers pushed away from the surface by the negative gate potential [4]. The various flows of carriers in this mode of operation are shown in Fig. 2. Source, drain, and p-well have been renamed emitter E,

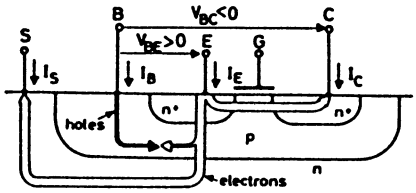


Fig. 2. Flows of carriers in bipolar operation.

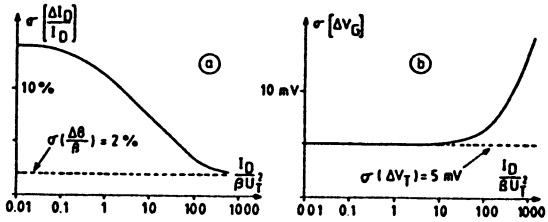


Fig. 3. Matching of a pair of MOS transistors as a function of drain current, for (a) same gate voltage and (b) same drain current. Uncorrelated components with mean standard deviations of 2 percent for $\Delta\beta/\beta$ and 5 mV for ΔV_T are assumed in this example.

collector C, and base B. Since a large fraction of emitter current I_E flows to substrate, the maximum alpha-gain of this lateral bipolar is only 0.2 to 0.6, depending on the process. However, owing to the low rate of recombination in the well, the β-gain can reach quite acceptable values ranging from 20 to 500. This high value of current gain is only obtained for the transistor implemented in the well, in this case a n-p-n. The n-p-n to substrate can be used without lateral collector, but only in common collector configurations.

For implementing analog circuits, it is necessary to specify the matching properties of similar adjacent transistors. Matching must be characterized by two independent statistical values: threshold mismatch ΔV_T, which may have in practice a mean standard deviation ranging from 1 to 20 mV, and $\Delta\beta/\beta$ mismatch which is usually in the range of 0.5–5 percent. Fig. 3 shows that when two transistors have the same gate voltage, as in a current mirror, the mismatch of their drain currents

$$\Delta I_D/I_D = \Delta\beta/\beta - (g_m/I_D)\Delta V_T \tag{7}$$

is maximum in weak inversion, for which g_m/I_D is maximum, and only comes down to $\Delta\beta/\beta$ when the transistors operate deeply in strong inversion. On the contrary, when they have the same drain current, as in a differential pair, the mismatch of their gate voltages

$$\Delta V_G = \Delta V_T - (I_D/g_m)\cdot\Delta\beta/\beta \tag{8}$$

is just ΔV_T in weak inversion, and increases in strong inversion where g_m/I_D is reduced.

Noise is a very important limitation of most analog circuits. As shown in Fig. 4, the noise of a transistor must also be characterized by at least two independent sources: White channel noise is independent of the process and corresponds to an equivalent input noise resistance R_N approximately equal to the inverse of transconductance g_m [5]. Gate interface $1/f$ noise dominates at low frequencies

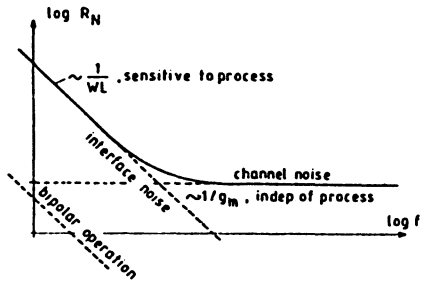


Fig. 4. Contributions to equivalent input noise resistance R_N of a MOS transistor.

TABLE I
SPECIFICATIONS ON TRANSISTORS FOR ANALOG AND DIGITAL APPLICATIONS.
(Analog Requires Specifications on Additional Parameters, and on Maximum and Minimum Values of Some Parameters.)

Parameter	Digital	Analog
β	min.	min.-max.
V_T	max.	min.-max.
n	max.	min.-max.
I_{D0}	max.	max.
β -bipolar	—	min.
mismatch	—	max.
1/f noise	—	(max.)
output resist.	—	min.

and is approximately independent of drain current. It is inversely proportional to gate area, and very sensitive to process quality. It should therefore be eliminated by circuit techniques such as chopping or autozeroing [6]–[8].

Both flicker noise and threshold mismatch are drastically reduced when the transistor operates in the lateral bipolar mode [4]. This is because the device is then shielded from all surface effects.

The respective qualitative specifications on transistors for digital and analog applications are summarized in Table I. An additional requirement for analog is a high value of output resistance which is approximately proportional to channel length. Designs should be made independent of the exact value of this parameter.

B. Passive Components

In digital CMOS circuits, passive components, namely capacitors and resistors, are only present as parasitics and should therefore by minimized. On the contrary, functional passive components of reasonable values and acceptable quality are required in most analog subcircuits.

Excellent precision capacitors can be implemented in a compatible way by using the silicon dioxide dielectric, provided both electrodes have a sufficiently low resistivity. Thin oxide gate capacitors are available in metal gate technologies, but they cannot be implemented in Si-gate processes without additional steps. For processes with a single polysilicon layer, the only reasonable choice is the capacitor between aluminum and polysilicon layers [9], which usually achieves rather low specific values. Many modern technologies provide two layers of polysilicon that can be used as electrodes for the capacitors [10].

Good resistors of less than 100 Ω/sq. can be obtained in the polysilicon layer. Higher values of few kiloohms per square are possible by using the well diffusion, but these resistors are slightly voltage dependent, and they are always associated with a large parasitic capacitance. Lightly doped polysilicon resistors such as those used to implement quasi-static RAM's [11] achieve very high values but they have a very poor accuracy.

Most of the modern design techniques for analog circuits are based on ratios of capacitances or resistances, and therefore only require specifications on matching and linearity of passive devices. If absolute values are needed as well, data on spread, temperature behavior, and aging must be available, and must be ensured by periodic statistical measurements.

No floating diode is usually available, except the base–emitter junction of the bipolar transistor to substrate. Some special micropower processes offer a lateral diode in the polysilicon layer [12].

III. BASIC ANALOG CIRCUIT TECHNIQUES

A. Optimum Matching

Most analog circuit techniques are based on the matching properties of similar components. For a given process, matching of critical devices may be improved by enforcing the set of rules that are summarized in Table II. These rules are not specific to CMOS and are applicable to all kinds of IC technologies. The relevancy and the quantitative importance of each of these rules depend on the particular process and on the particular device under consideration.

1) Devices to be matched should have the same structure. For instance, a junction capacitor cannot be matched with an oxide capacitor. This also means that the error due to parasitic junction capacitors cannot be compensated by adjusting the value of functional oxide capacitors.

2) They should have same temperature, which is no problem if power dissipated on chip is very low. Otherwise, devices to be matched should be located on the same isotherm, which can be obtained by a symmetrical implementation with respect to the dissipative devices.

3) They should have same shape and same size. For example, matched capacitors should have same aspect ratios, and matched transistors or resistors should have same width and same length, and not simply same aspect ratios.

4) Minimum distance between matched devices is necessary to take advantage of spatial correlation of fluctuating physical parameters.

5) Common-centroid geometries should be used to cancel constant gradients of parameters. Good practical examples

TABLE II
RULES FOR OPTIMUM MATCHING

1. Same structure
2. Same temperature
3. Same shape, same size
4. Minimum distance
5. Common-centroid geometries
6. Same orientation
7. Same surroundings
8. Non minimum size

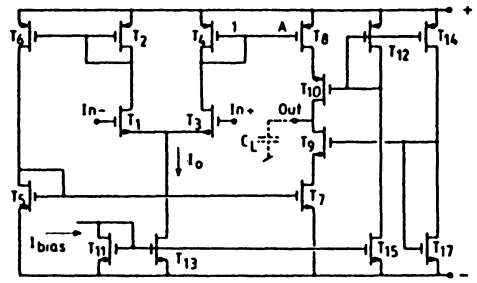


Fig. 5. Single-stage cascode operational transconductance amplifier (OTA) [13].

are the quad configuration used to implement a pair of transistors, and common-centroid sets of capacitors.

6) The same orientation on chip is necessary to eliminate dissymmetries due to unisotropic steps in the process, or to the unisotropy of the silicon substrate itself. In particular, the source to drain flows of current in matched transistors should be strictly parallel.

7) Devices to be matched should have the same surroundings in the layout. This to avoid for instance the end effect in a series of current sources implemented as a line of transistors, or the street effect in a matrix of capacitors.

8) Using nonminimum size is an obvious way of reducing the effect of edge fluctuations, and to improve spatial averaging of fluctuating parameters.

Matching can be extended to the realization of non-unity n/m ratios by separately grouping m and n matched devices. A slight alteration of one or many devices is necessary when intermediate ratios are required.

B. Amplifiers

The basic configurations and tradeoffs related to the realization of amplifiers can be discussed with the example of a single-stage cascode OTA represented in Fig. 5 [13].

Differential pair $T_1 - T_3$ converts the differential input voltage into a difference of currents which is integrated in load capacitance C_L. These transistors can have minimum channel length since they are loaded by the high input conductance of current mirrors. Remaining design parameters are then channel width W and value of tail current I_0. Optimization of this input pair therefore amounts to choosing the best possible point in the (W, I_0) plane, with respect to conflicting requirements. This plane is represented in Fig. 6, with the limit between weak and strong inversion which corresponds to a given value of W/I_0.

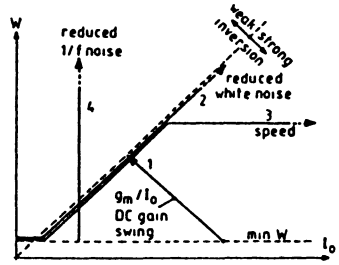


Fig. 6. Optimization of width W and tail current I_0 of input pair $T_1 - T_3$. Arrowed paths 1 to 4 indicate displacements in plane (W, I_0) for maximum improvement of various features of the amplifier.

Displacements in the plane for maximum improvement of various important features of the whole amplifier are represented in a qualitative way.

Transconductance for a given current, dc gain, and maximum possible swing are all improved by increasing W/I_0 to approach weak inversion, where they reach their maximum values (path 1). This also reduces input offset voltage.

The white noise spectral density is inversely proportional to transconductance g_m, which increases linearly with current I_0 up to the upper limit of weak inversion. To keep the advantages of weak inversion, a further increase of g_m requires a parallel increase of width W and current I_0 (path 2), which is only limited by size.

Speed (path 3) is proportional to g_m as long as parasitic capacitances of the transistors (proportional to W) are constant or negligible. A further increase in speed requires a progressive incursion into strong inversion, which results in progressive degradations of dc gain and of maximum possible swing. Speed in strong inversion only increases with the square root of current.

Low frequency $1/f$ noise is reduced by increasing channel width (path 4) and by choosing the better type of transistor, which is usually a p-channel.

If input current can be tolerated, very low $1/f$ noise and very high speed can be achieved by using transistors operated as lateral bipolars [4], [14].

The maximum differential current available from the pair is equal to tail current I_0, which puts a fundamental limit to slew rate. This problem can be circumvented by momentarily increasing current I_0 by a fixed amount each time an input step is anticipated, which yields a dynamic amplifier [15]. It can also be increased by an amount proportional to the difference of drain currents to realize an adaptive bias [16], [17], which provides operation in class AB.

Complementary pairs $T_1 - T_2$ and $T_3 - T_4$ can be viewed as common emitter amplifiers, each of which amplifies half of the total differential input voltage. Gain $g_{m1,3}/g_{m2,4}$ must be high enough, of the order of 3 to 10, to have noise and offset voltage limited to the only contribution of the differential input pair. This requires operation of transistors T_2 and T_4 deep enough into strong inversion. Too much gain reduces the stability phase margin.

As was illustrated by Fig. 3, the mismatch of current mirrors is maximum in weak inversion. It can be shown

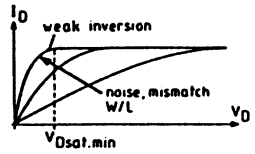


Fig. 7. Optimization of W/L of current mirrors.

that this is true as well for both white and $1/f$ noise [18]. However, according to relations (1) and (2), weak inversion provides the minimum possible value of drain saturation voltage of the order of 100 mV. Therefore, the optimization of W/L of current mirrors $T_{11}-T_{13}-T_{15}, T_2-T_6, T_4-T_8, T_5-T_7$ amounts to an acceptable compromise between small mismatch and low noise (V_G-V_{T0} large), and small saturation voltage (V_G-V_{T0} small) to permit large-signal swing (Fig. 7).

The overall transconductance of the amplifier can be multiplied by ratio A of mirror T_4-T_8, at the expense of a reduction in phase margin. Some of the mirrors can be avoided by using the folded cascode scheme [20].

Cascode transistors T_9 and T_{10} decrease the output conductance by a factor equal to g_{ms}/g_o. This is obtained without any noise penalty, and with only a very small reduction of phase margin. The resulting dc gain is thus higher than that of a two-stage noncascode amplifier which requires internal compensation. Gain may be further boosted by using double or triple cascode [19], until it becomes limited by the direct conductance to ground due to impact ionization in the drain depletion layers.

The reduction of maximum output swing due to the cascode transistors can be minimized by careful design of the bias circuitry $T_{12}-T_{14}-T_{15}-T_{17}$ [13], [20]. Drain voltages of transistors T_7 and T_8 can be made equal to their limit value V_{Dsat} for saturation, independently of bias current. Maximum output swing is then only reduced by $4V_{Dsat}$ with respect to total supply voltage, which only amounts to about 400 mV in weak inversion.

The circuit can be modified to provide differential output [20]. This doubles the maximum output swing, but requires a common mode feedback scheme.

All amplifiers based on a differential input pair suffer noise and speed penalties with respect to a simple CMOS inverter used as an amplifier with an adequate biasing scheme [21]. This kind of amplifier is furthermore free from any slew rate limitation. It represents a very attractive solution for very low power [7] or very high speed [22] applications, in spite of its poor intrinsic PSRR.

C. Switch and Sample-and-Hold

The realization of the analog switch, which is a very important component of CMOS analog circuits, is illustrated in Fig. 8. A n-channel transistor is switched on by connecting its gate to the positive power line V_B. However, its on-conductance g_n comes to zero if potential V_F at which the device floats is too high. The same is true if V_F is too low for on-conductance g_p of the p-channel transistor

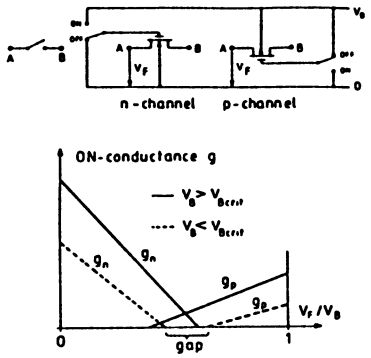


Fig. 8. Realization of an analog switch. On-conductances g_n and g_p of n-channel and p-channel transistors depend on floatation voltage V_F.

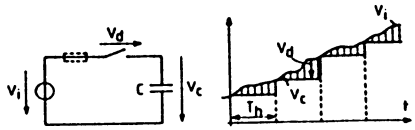


Fig. 9. Elementary sample-and-hold.

with gate connected to zero. If total supply voltage V_B is larger than a critical value V_{Bcrit}, the conductance of the switch can be ensured independently of V_F by connecting both types of transistors in parallel. Below this critical value, a gap of conduction appears at intermediate levels of V_F. Because of substrate effects, this critical supply voltage may widely exceed the sum $V_{T0p}+V_{T0n}$ of p and n thresholds for zero source voltages. For example, $V_{T0p}=0.6$ V and $V_{T0n}=0.7$ V may correspond to $V_{Bcrit}=2.3$ V [18], [23]. This very severe limitation to low-voltage operation of analog circuits may be circumvented by on-chip clock voltage multiplication [8], [24].

Leakage in the off state is due to residual channel current and to reverse currents of junctions. Care must be taken not to bootstrap the switch potential beyond that of the power supply lines, which would forward bias these junctions [25].

The combination of a switch and a capacitor provides a basic sample-and-hold shown in Fig. 9. Voltage V_C across capacitor C keeps a constant value equal to that of input voltage V_i at the last sampling instant. The value the of noise voltage at the sampling instant is also frozen in capacitor C; therefore, the total noise power is concentrated below the clock frequency. Voltage V_d across the switch is readjusted to zero each time the switch is closed, which corresponds to the transfer function for the fundamental signal (component of output signal V_d at the frequency of input signal V_i) shown in Fig. 10 [26]. At low frequency, this autozeroing by means of a sample-and-hold amounts to a differentiation, with a time constant equal to half the value of hold duration T_h. It may be used to cancel offset and to reduce low-frequency noise components generated in a circuit [6], [27].

Another source of sampling error is caused by the charge which is released from the channel into holding capacitor C when the transistor of the switch is blocked [28]. This problem has been analyzed in the general case shown in

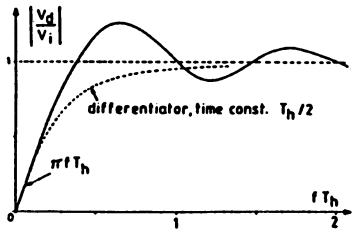


Fig. 10. Differentiating property of autozeroing obtained by sample-and-hold.

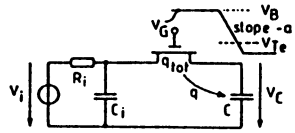


Fig. 11. Equivalent circuit of a practical sample-and-hold. Finite fall time of gate voltage V_G allows redistribution of charge through the transistor.

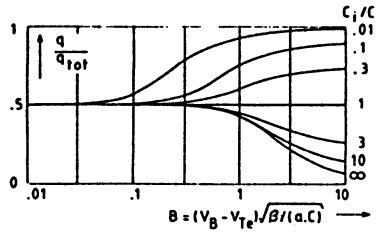


Fig. 12. Calculated fraction q of total charge q_{tot} left in holding capacitor C after switch-off [23].

Fig. 11 where the source of the signal is assumed to have an internal capacitance C_i [23], [26]. Finite fall time of V_G (slope $-a$) from initial value V_B to effective threshold voltage V_{Te} allows a redistribution through the transistor of total charge $q_{tot} = C_{gate}(V_B - V_{Te})$ between C and C_i. The result obtained by numerical integration of a normalized nonlinear equation describing this process, for time constant R_iC_i much larger than the switching time, is represented in Fig. 12. This figure shows the fraction q of total channel charge q_{tot} which goes into holding capacitor C for various values of ratio C_i/C. This fraction is a function of an intermediate parameter B which combines clock amplitude, clock slope, β of transistor, and value of holding capacitor C. These curves suggest various strategies for minimizing parasitic charge q.

A first possibility is to choose C_i very large and B much larger than 1. All charges released into C flow back into C_i during the decay of gate voltage, and q tends to zero (some easily calculable additional charge is due to the coupling through overlap capacitors after switching off). The drawback is the long period of time needed for switching off.

A second solution is to equilibrate the values of both capacitors [29]. By symmetry, half of the channel charge flows in each capacitor, and can be compensated by half-sized dummy switches that are switched on when the main switch is blocked [28].

The need for equal values of capacitors may be eliminated by choosing a value of B much smaller than 1 which also ensures equipartition of the total charge. Charge q is then compensated by a single dummy switch.

When complementary transistors are used to implement the switch, they partially compensate each other, although matching is very poor. The effect of charge injection can be drastically reduced by appropriate circuit techniques such as differential implementation [25], and active compensation by a low sensitivity auxiliary input [26], [30].

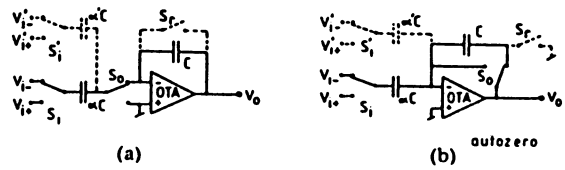


Fig. 13. Stray insensitive SC integrators.

D. Switched Capacitor Integrators

Switched capacitor integrators are the building blocks of all kinds of circuits, in particular SC filters. Two different implementations that are insensitive to parasitic capacitances to ground are shown in Fig. 13. Both provide a differential input and a time constant $1/\alpha f_c$, which only depends on clock frequency f_c and ratio of capacitors α. Version b includes autozeroing, which reduces low-frequency noise and compensates offset. It can therefore be realized with a nondifferential amplifier such as a CMOS inverter [7].

Output resetting can be obtained by additional switch S_r, and many differential signals can be separately weighted and summed by repeating the input circuitry, as shown in dotted lines. These integrators may be damped at will by connecting one of these additional inputs to output.

They can be transformed into amplifiers with controlled gain, either by resetting the output at every clock cycle, or by deleting integrating capacitor C in a damped configuration.

E. Comparators

Comparators must usually achieve a very low value of input offset voltage. An excellent solution is obtained by removing integrating capacitor C in the basic integrator of Fig. 13b [31]. Any difference $V_{i+} - V_{i-}$ will cause an output current to charge (or discharge) the parasitic output capacitance. The comparator thus behaves as an integrator of input error voltage, and sensitivity is proportional to the time alotted for comparison. It is ultimately limited by the finite dc gain of the amplifier. The speed-sensitivity ratio may be increased by achieving nth order integration along a cascade of n stages [26], as shown in Fig. 14.

The effects of charge injection and switching noise may be virtually cancelled by sequentially opening switches S_1 to S_n before toggling switch S_0: When S_1 is opened first, charge injection and sampled noise cause an error voltage across C_1. Since switch S_2 is still closed, a compensation voltage appears across C_2 after equilibration. The same is true when S_2 to S_{n-1} are then opened sequentially. The

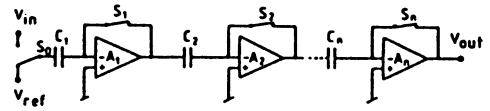


Fig. 14. Multistage comparator.

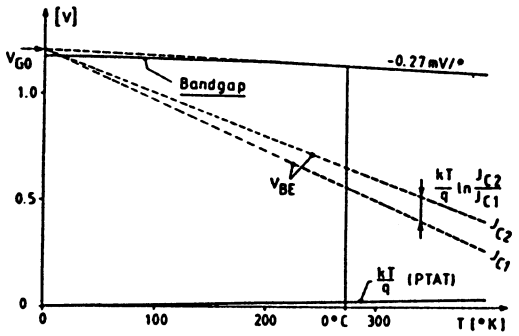


Fig. 15. "Built-in" voltages available in silicon.

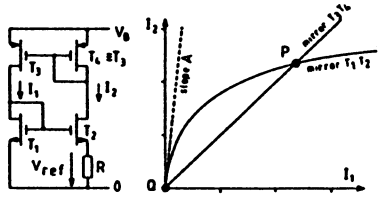


Fig. 16. Extraction of $U_T = kT/q$ with MOS transistors $T_1 - T_2$ operated in weak inversion.

only residual error at the input is thus that due to switch S_n divided by the gain of the $n-1$ first stages [32], [33]. Accuracy is further improved by using a fully differential implementation for which values of offset as low as 5 μV have been reported [34].

F. Voltage and Current References

The realization of absolute references must be based on intrinsic physical values, in order to reduce their sensitivity to process variations.

The various "built-in" voltages provided by silicon are represented in Fig. 15. They can be extracted by adequate circuits to implement voltage references.

Thermal voltage $U_T = kT/q$, proportional to absolute temperature (PTAT), can be extracted by two MOS transistors operated in weak inversion with different current densities, as shown in Fig. 16 [3], [35]. If $T_3 = T_4$, application of weak inversion model (2) to transistors T_1 and $T_2 = AT_1$ yields

$$V_{ref} = U_T \ln(A) \tag{9}$$

which in turn imposes current I_2 in the circuit.

Bandgap voltage V_{gap} decreases approximately linearly with temperature from extrapolated value V_{G0}, with a slight curvature. A possible technique for direct extraction of V_{gap} is shown in Fig. 17 [36]. Transistor T_1 is n-channel with a normal n^+-doped silicon gate. Transistor T_2 is also

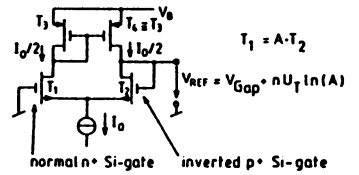


Fig. 17. Extraction of bandgap voltage V_{gap} by MOS transistors [36].

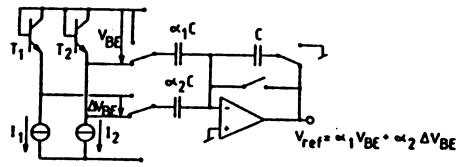


Fig. 18. Principle of SC-weighted bandgap reference.

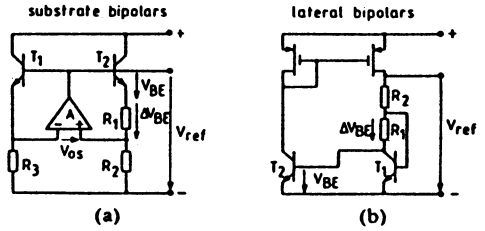


Fig. 19. Principle of R-weighted bandgap reference.

n-channel but with a p^+-doped gate, as is possible in some technologies. Their threshold voltages therefore differ by approximately V_{gap} which appears at the output of this simple amplifier connected in unity gain configuration. After compensation by a small PTAT voltage obtained by operating T_1 and T_2 in weak inversion at different current densities, a temperature coefficient lower than 30 ppm °C can be obtained. All references based on MOS operation have their accuracy degraded by large uncontrolled offset components due to surface effects.

Base–emitter voltage V_{BE} of bipolar transistors is not really a physical parameter, but it only depends very slightly on the process. The difference ΔV_{BE} of two bipolars operated at different current densities is strictly proportional to kT/q. After multiplication by an adequate factor, it can be added to V_{BE} to obtain a voltage of value V_{G0} independent of temperature. This principle of bandgap reference is well known in bipolar technology, and can be applied to the bipolars available in CMOS technology.

Weighting and summing V_{BE} and ΔV_{BE} can be achieved by the SC circuit shown in Fig. 18, which is derived from the integrator of Fig. 13(b) [23], [37], [38]. Transistors T_1 and T_2 are bipolars to substrate. Accuracy is mainly limited by charge injection from the feedback switch.

Another solution consists in using a resistive divider, as depicted in Fig. 19. Version *a* uses substrate bipolars and a CMOS amplifier [39]. The offset of this amplifier, which is multiplied by R_2/R_1 of the order of 10, causes independent errors of the value of V_{ref} and of its temperature coefficient. Adjustment at two different temperatures would thus be required to achieve an accuracy of a few millivolts. Version *b* uses compatible lateral bipolars and avoids any MOS amplifier [4]. The error due to the p-channel mirror is

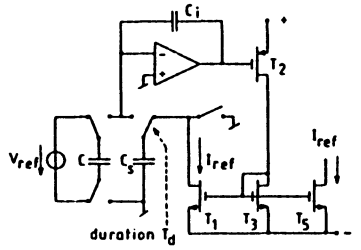


Fig. 20. SC voltage-to-current converter [40].

lowered by operating deeply in strong inversion. The offset of bipolars is small and only causes an error PTAT, which can be corrected at a single temperature.

Semiconductor physics do not provide any "built-in" current. Current references must thus be derived from voltage references by applying Ohm's law in voltage to current converters. Poor absolute precision and temperature coefficient of available resistors result in errors of many tens of percent. A good solution based on a SC scheme that takes advantage of the better accuracy of available capacitors is shown in Fig. 20 [40]. It is a closed loop system which forces equilibrium, for every clock cycle, between charge CV_{ref} poured into storage capacitor C_s and charge $I_{ref}T_d$ withdrawn from the same capacitor. Thus

$$I_{ref} = CV_{ref}/T_d. \tag{10}$$

Generation of an independent frequency on chip is not possible with a precision better than a few tens of percent. It is normally not required since an accurate clock frequency is usually provided by the system, from which synchronous signals of any frequency can be derived in a digital way. Totally asynchronous signals of accurate frequency can be produced by quasi-sinusoidal SC oscillators [41]–[43].

IV. Analog Circuits in a Digital Environment

If no precaution is taken, the dynamic range of analog circuits will be limited by the noise generated by digital circuits operating on the same chip. This problem is specially accute for sampled circuits which fold down high-frequency noise components by undersampling.

Coupling may occur through power lines, current in the common substrate, capacitive links, and possibly by minority carriers that are released when digital transistors are being blocked.

Various provisions can be suggested to improve the situation: Utilization of separate power lines, including pads, bondings, and possibly pins. Implementation of power supply filters or regulators. High value of PSRR, also at high frequencies for sampled circuits; care must be taken not to destroy the PSRR of amplifiers by the additional circuitry. Systematic implementation of fully differential structures. Avoidance of large current spikes in digital circuits, and of any digital transition during critical analog tasks. Provision of maximum distance on chip to digital lines, and of separate clean clocks for analog circuits. Critical nodes should be shielded from substrate and from digital lines by adequate layers, and analog circuits can be separated by special wells that collect parasitic minority carriers. Processes that provide an epitaxial layer on a highly doped substrate, to improve immunity to latch-up, allow to drain all parasitic currents to substrate.

V. Conclusion

Thanks to the versatility of the CMOS technology, all kinds of analog circuits can be combined on the same chip with digital circuits, without any process modification. However, additional parameters need to be specified and guaranteed, the number of which depends on the type of function and on design cleverness.

A standard MOS transistor can be operated in various modes which have their respective merits. Weak inversion provides maximum values of gain and signal swing, and minimum offset and noise voltages. Strong inversion is required to achieve high speed, and provides minimum relative values of offset and noise currents. The lateral bipolar mode exhibits excellent matching properties and very low $1/f$ noise, and can be used to implement a variety of schemes previously developed for normal bipolars transistors.

Mismatch of active and passive devices represents a major limitation to the accuracy of analog circuits. It can be minimized by respecting a set of basic rules. Designs must be based on sound concepts that take maximum advantage of all available components.

Single-stage cascoded OTA's should be preferred to multistage amplifiers. Their optimization amounts to choosing the best possible compromise with respect to various conflicting requirements. The excellent performance of the MOS transistor as a switch and the availability of high-quality capacitors are key elements in the implementation of a variety of analog functions. The elementary sample-and-hold that is made possible by the absence of any dc gate control current can be used to reduce low-frequency noise and to compensate offset. Its major limitation is due to charge injection from the switch, which can be evaluated and partially compensated by an adequate strategy.

Accurate absolute references are very critical circuits which must be based on intrinsic physical values to achieve low sensitivity to process. This is possible for voltage references, from which current references can be derived by applying Ohm's law.

Precautions must be taken to avoid degradation of analog performances by the noise generated by the digital part of the chip.

References

[1] R. W. Brodersen and P. R. Gray, "The role of analog circuits in future VLSI technologies," in *ESSCIRC '83 Dig. Tech. Papers*, Sept. 1983, pp. 105–110.

[2] J. D. Chatelain, "Dispositifs à semiconducteurs," *Traité d'Electricité EPFL*, vol. VII (Georgi, St-Saphorin, Switzerland), 1979.

[3] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 224–231, June 1977.
[4] E. Vittoz, "MOS transistors operated in the lateral bipolar mode and their applications in CMOS technology," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 273–279, June 83.
[5] J. Fellrath and E. Vittoz, "Small signal model of MOS transistors in weak inversion," in *Proc. Journées d'Electronique* 1977 (EPF-Lausanne), pp. 315–324.
[6] H. W. Klein and W. Engl, "Design techniques for low noise CMOS operational amplifiers," in *ESSCIRC '84 Dig. Tech. Papers* (Edinburgh), Sept. 1984, pp. 27–30.
[7] F. Krummenacher, "Micropower switched capacitor biquadratic cell," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 507–512, June 1982.
[8] M. Degrauwe and F. Salchli, "A multipurpose micropower SC-filter," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 343–348, June 1984.
[9] E. Vittoz and F. Krummenacher, "Micropower SC filters in Si-gate technology," in *Proc. ECCTD '80* (Warshaw), Sept. 1980, pp. 61–72.
[10] J. L. McCreary, "Matching properties, and voltage and temperature dependence of MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 608–616, Dec. 1981.
[11] Y. Uchida *et al.*, "A low power resistive load 64 kbit CMOS RAM," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 804–909, Oct. 1982.
[12] M. Dutoit and F. Sollberger, "Lateral polysilicon p-n diodes," *J. Electrochem. Soc.*, vol. 125, pp. 1648–1651, Oct. 1978.
[13] F. Krummenacher, "High voltage gain CMOS OTA for micropower SC filters," *Electron. Lett.*, vol. 17, pp. 160–162, 1981.
[14] S. Gustafsson *et al.*, "Low-noise operational amplifiers using bipolar input transistors in a standard metal gate CMOS process," *Electron. Lett.*, vol. 20, pp. 563–564, 1984.
[15] B. J. Hosticka, "Dynamic CMOS amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 887–894, Oct. 1980.
[16] M. G. Degrauwe *et al.*, "Adaptive biasing CMOS amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 522–528, June 1982.
[17] E. Seevinck and R. F. Wassenaar, "Universal adaptive biasing principle for micropower amplifiers," in *ESSCIRC '84 Dig. Tech. Papers* (Edinburgh), Sept. 1984, pp. 59–62.
[18] E. Vittoz, "Micropower techniques," *Advanced Summer Course on Design of MOS-VLSI Circuits for Telecommunications*, L'Aquila, Italy, June 1984, to be published by Prentice-Hall, 1985.
[19] P. W. Li *et al.*, "A ratio-independent algorithmic ADC technique," in *ISSCC Dig. Tech. Papers*, Feb. 1984, pp. 62–63.
[20] T. C. Choi *et al.*, "High-frequency CMOS switched-capacitor filters for communications application," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 652–664, Dec. 1983.
[21] F. Krummenacher, E. Vittoz, and M. Degrauwe, "Class *AB* CMOS amplifiers for micropower SC filters," *Electron. Lett.*, vol. 17, pp. 433–435, June 1981.
[22] S. Masuda *et al.*, "CMOS sampled differential push–pull cascode operational amplifier," in *Proc. ISCAS '84* (Montreal, Ont., Canada), May 1984, p. 1211.
[23] E. Vittoz, "Microwatt SC circuit design," "Summer course on SC circuits," KU-Leuven, Belgium, June 1981, republished in *Electrocomponent Sci. Technol.*, vol. 9, pp. 263–273, 1982.
[24] F. Krummenacher, H. Pinier, and A. Guillaume, "Higher sampling rates in SC circuits by on-chip clock-voltage multiplication," in *ESSCIRC '83 Dig. Tech. Papers* (Lausanne), pp. 123–126, Sept. 1983.
[25] D. J. Alstott, "A precision variable-supply CMOS comparator," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1080–1087, Dec. 1982.
[26] E. Vittoz, "Dynamic analog techniques," *Advanced Summer Course on Design of MOS-VLSI Circuits for Telecommunications*, L'Aquila, Italy, June 1984, to be published by Prentice-Hall, 1985.
[27] C. Enz, "Analysis of the low-frequency noise reduction by autozero technique," *Electron. Lett.*, vol. 20, pp. 959–960, Nov. 1984.
[28] E. Suarez *et al.*, "All-MOS charge redistribution analog-to digital conversion techniques," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 379–385, Dec. 1975.
[29] L. Bienstman and H. J. De Man, "An eight-channel 8-bit microprocessor compatible NMOS converter with programmable scaling," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1051–1059, Dec. 1980.
[30] M. G. Degrauwe, E. Vittoz, and I. Verbauwhede, "A micropower CMOS instrumentation amplifier," in *ESSCIRC '84 Dig. Tech. Papers* (Edinburgh), Sept. 1984, pp. 31–34.
[31] Y. S. Lee *et al.*, "A 1 mV MOS comparator," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 294–297, June 1978.
[32] R. Poujois *et al.*, "Low-level MOS transistor amplifier using storage techniques," in *ISSCC Dig. Tech. Papers*, 1973, pp. 152–153.
[33] A. R. Hamade, "A single chip all-MOS 8-bit ADC," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 785–791, Dec. 1978.
[34] R. Poujois and J. Borel, "A low drift fully integrated MOSFET operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 499–503, Aug. 1978.
[35] E. Vittoz and O. Neyroud, "A low-voltage CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 573–577, June 1979.
[36] H. Oguey and B. Gerber, "MOS voltage reference based on polysilicon gate work function difference," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 264–269, June 1980.
[37] O. Leuthold, "Integrierte Spannungsüberwachungsschaltung," presented at Meet. Swiss Chapter Solid-State Devices Circuits (Bern, Switzerland), Oct. 1981.
[38] B. S. Song and P. R. Gray, "A precision curvature-compensated CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 634–643, Dec. 1983.
[39] R. Ye and Y. Tsividis, "Bandgap voltage reference sources in CMOS technology," *Electron. Lett.*, vol. 18, pp. 24–25, 1982.
[40] H. W. Klein and W. L. Engl, "A voltage–current-converter based on a SC-controller," in *ESSCIRC '83 Dig. Tech. Papers*, (Lausanne, Switzerland), Sept. 1983, pp. 119–122.
[41] E. Vittoz, "Micropower SC oscillator," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 622–624, June 1979.
[42] B. J. Hosticka *et al.*, "Switched-capacitor FSK modulator and demodulator in CMOS technology," in *ESSCIRC '83 Dig. Tech. Papers*, pp. 231–216 (Lausanne, Switzerland), Sept. 1983.
[43] F. Krummenacher, "A high resolution capacitance-to-frequency converter," in *ESSCIRC '84 Dig. Tech. Papers* (Edinburgh), Sept. 1984, pp. 95–98.

MOS Operational Amplifier Design—A Tutorial Overview

PAUL R. GRAY, FELLOW, IEEE, AND ROBERT G. MEYER, FELLOW, IEEE

(Invited Paper)

Abstract—**This paper presents an overview of current design techniques for operational amplifiers implemented in CMOS and NMOS technology at a tutorial level. Primary emphasis is placed on CMOS amplifiers because of their more widespread use. Factors affecting voltage gain, input noise, offsets, common mode and power supply rejection, power dissipation, and transient response are considered for the traditional bipolar-derived two-stage architecture. Alternative circuit approaches for optimization of particular performance aspects are summarized, and examples are given.**

I. Introduction

THE rapid increase in chip complexity which has occurred over the past few years has created the need to implement complete analog-digital subsystems on the same integrated circuit using the same technology. For this reason, implementation of analog functions in MOS technology has become increasingly important, and great strides have been made in recent years in implementing functions such as high-speed DAC's, sampled data analog filters, voltage references, instrumentation amplifiers, and so forth in CMOS and NMOS technology [1]. These developments have been well documented in the literature. Another key technical development has been a maturing of the state of the art in the implementation of operational amplifiers (op amps) in MOS technology. These amplifiers are key elements of most analog subsystems, particularly in switched capacitor filters, and the performance of many systems is strongly influenced by op amp performance. Many of the developments in MOS operational amplifier design have not been as well documented in the literature, and the intent of this paper is to review the state of the art in this field. This paper is focused on the design of op amps for use within single-chip analog-digital LSI systems, and the particular problems of the design of stand-alone CMOS amplifiers are not addressed.

Manuscript received August 24, 1982; revised September 27, 1982. This work was supported by the Joint Services Electronics Program under Contract F49620-79-c-0178 and the National Science Foundation under Grant ENG79-07055.

The authors are with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720.

In Section II, the important performance requirements and objectives for operational amplifiers within a monolithic analog subsystem are summarized. In Section III, the performance of the basic two-stage CMOS operational amplifier architecture is summarized. In Section IV, alternative circuit approaches for the improvement of particular performance aspects are considered. In Section V, the particular problems associated with NMOS depletion load amplifier design are considered, and in Section VI, the design of output stages is considered. Finally, a summary and discussion of the design of amplifiers in scaled technologies are presented in Section VII.

II. Performance Objectives for MOS Operational Amplifiers

The performance objectives for operational amplifiers to be used within a monolithic analog subsystem are often quite different from those of traditional stand-alone bipolar amplifiers. Perhaps the most important difference is the fact that for many of the amplifiers in the system, the load which the output of the amplifier has to drive is well defined, and is often purely capacitive with values of a few picofarads. In contrast, stand-alone general-purpose amplifiers usually must be designed to achieve a certain level of performance independent of loading over capacitive loads up to several hundred picofarads and resistive loads down to 2 kΩ or less. Within a monolithic analog subsystem, only a few of the amplifiers must drive a signal off chip where the capacitive and resistive loads are significant and variable. In this paper, these amplifiers will be termed output buffers, and the amplifiers whose outputs do not go off chip will be termed internal amplifiers. The particular problems of the design of these output buffers are considered in Section VII.

A typical application of an internal operational amplifier, a switched capacitor integrator, is illustrated in Fig. 1. The basic function of the op amp is to produce an updated value of the output in response to a switching event at the input in which the sampling capacitor is charged from the source and discharged into the summing node. The output must assume the new updated value within the required accuracy, typically on the order of 0.1 percent, within one clock period, typically on the order of 1 μs for voiceband filters. Important performance

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-17, no. 6, pp. 969–982, Dec. 1982.

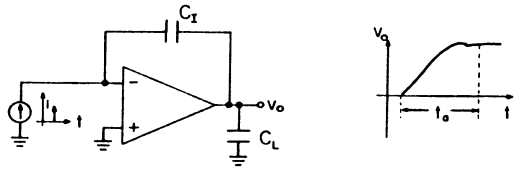


Fig. 1. Typical application of an internal MOS operational amplifier, a switched capacitor integrator.

TABLE I
TYPICAL PERFORMANCE, CONVENTIONAL TWO-STAGE CMOS INTERNAL OPERATIONAL AMPLIFIER (+/−5 V SUPPLY, 4 μm SI GATE CMOS)

dc gain (capacitive load only)	5000
Setting time, 1 V step, C_l = 5 pF	500 ns
Equiv. input noise, 1 kHz	100 nV/$\sqrt{\text{Hz}}$
PSRR, dc	90 dB
PSRR, 1 kHz	60 dB
PSRR, 50 kHz	40 dB
Supply capacitance	1 fF
Power dissipation	0.5 mW
Unity-gain frequency	4 MHz
Die area	75 mils^2
Systematic offset	0.1 mV
Random offset std. deviation	2 mV
CMRR	80 dB
CM range	within 1 V of supply

parameters are the power dissipation, maximum allowable capacitive load, open-loop voltage gain, output voltage swing, equivalent input flicker noise, equivalent input thermal noise, power supply rejection ratio, supply capacitance (to be defined later), and die area. In this particular application the input offset voltage, common-mode rejection ratio, and common-mode range are less important, but these parameters can be important in other applications. Because of the inherent capacitive sample/hold capability in MOS technology, dc offsets can often be eliminated at the subsystem level, making operational amplifier offsets less important. A typical set of values for the parameters given above for a conventional amplifier design in 4 μm CMOS technology are given in Table I. In the following section, the factors affecting the various performance parameters are evaluated for the most widely used amplifier architecture.

III. BASIC TWO-STAGE CMOS OPERATIONAL AMPLIFIER

Currently, the most widely used circuit approach for the implementation of MOS operational amplifiers is the two-stage configuration shown in Fig. 2(b). This configuration is also widely used in bipolar technology, and the bipolar counterpart is also illustrated in Fig. 2(a). The behavior of this circuit when implemented in bipolar technology has been reviewed in an overview article published earlier [2]. This circuit configuration provides good common mode range, output swing, voltage gain, and CMRR in a simple circuit that can be compensated with a single pole-splitting capacitor. While the implementation of this architecture in NMOS technology requires additional circuit elements because of the lack of a complementary device, many NMOS amplifiers commercially manufactured at the present time use a conceptually similar

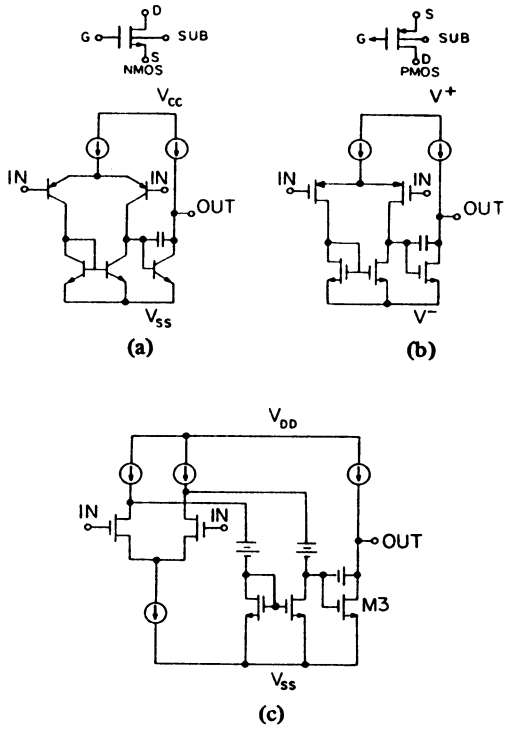


Fig. 2. Two-stage operational amplifier architecture. (a) Bipolar implementation. (b) CMOS implementation. (c) An example of an NMOS implementation with interstage coupling network.

configuration, as illustrated in Fig. 2(c) where a differential interstage level-shifting network composed of voltage and current sources has been inserted between the first and second stages so that both stages can utilize n-channel active devices and depletion mode devices as loads. The implementation of this circuit is discussed further in Section V.

In this section, we will analyze the various performance parameters of the CMOS implementation of this circuit, focusing particularly on the aspects which are different from the bipolar case.

Open Circuit Voltage Gain

An important difference between MOS and bipolar technology is the fact that the maximum transistor open circuit voltage gain g_m/g_o is much lower for MOS transistors than for bipolar transistors, typically by a factor on the order of 10–40 for typically used geometries and bias currents [3]. Under certain simplifying assumptions, voltage gain can be shown to be

$$g_m/g_o = \frac{2L}{V_{gs} - V_T} \left(\frac{dx_d}{dV_{ds}}\right)^{-1} \tag{1}$$

where x_d is the width of the depletion region between the end of the channel and the drain and L is the effective channel length. The expression illustrates several key aspects of MOS devices used as analog amplifiers. First, for constant drain current decreasing either the channel length or width results in a decrease in the gain, the latter because of the fact that V_{gs} increases. This fact, along with noise considerations, usually dictates the minimum size of the transistors that must be used in a given high-gain amplifier application. Usually, this is

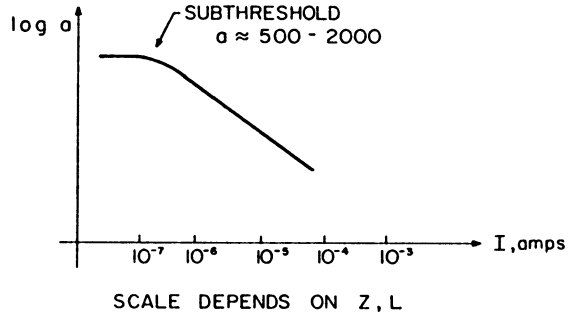


Fig. 3. Typical open circuit gain of an MOS transistor as a function of bias current.

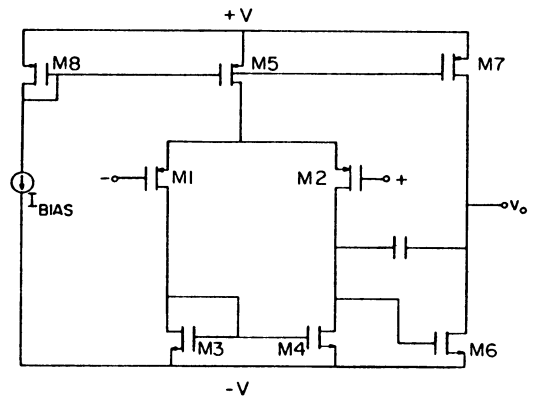


Fig. 4. Schematic of basic two-stage CMOS operational amplifier.

larger than the length and width used for digital circuits in the same technology.

Second, if the device geometry is kept constant, the voltage gain is inversely proportional to the square root of the drain current since $(V_{gs} - V_T)$ is proportional to the square root of the drain current. A typical variation of open circuit voltage gain as a function of drain current is shown in Fig. 3 [4]. The gain becomes constant at a value comparable to bipolar devices in the subthreshold range of current. This fact makes use of low current levels desirable, and at the same time complicates the design of high-speed amplifiers which must operate at high current.

Third, if device size and bias current are kept constant, the gain is an increasing function of substrate doping since dx_d/dV_{ds} decreases with increasing doping. Thus, devices which have received a channel implant to increase threshold voltage would display a higher open circuit gain than an unimplanted device whose channel doping was lower. Finally, the expression demonstrates that open circuit gain is not degraded by technology scaling in the constant field sense since all terms in the expression decrease in proportion. However, scaling in the quasi-constant field or constant voltage sense would result in a decrease in gain.

Turning to the operational amplifier, the voltage gain of the first stage of the circuit shown in Fig. 4 can be shown to be simply

$$A_{v1} = \frac{g_{m1}}{g_{o2} + g_{o4}} \tag{2}$$

where g_m is the device transconductance and g_o is the small signal output conductance, and assuming that $M1$ and $M2$ are identical and that $M3$ and $M4$ are identical. Similarly, the second stage voltage gain is

$$A_{v2} = \frac{g_{m6}}{g_{o6} + g_{o7}}. \tag{3}$$

For switched capacitor filter applications, the overall voltage gain required is on the order of several thousand [5], implying a gain in each stage on the order of 50. In order to achieve this level of gain per stage, transistor bias currents and channel lengths and widths are usually chosen such that the transistor $(V_{gs} - V_T)$ is several hundred millivolts, and the drain depletion region is on the order of one fifth or less of the effective channel length at the typical drain bias of several volts. Circuit approaches to achieving more voltage gain or, alternatively, achieving the same voltage gain with smaller devices, are discussed in Section IV.

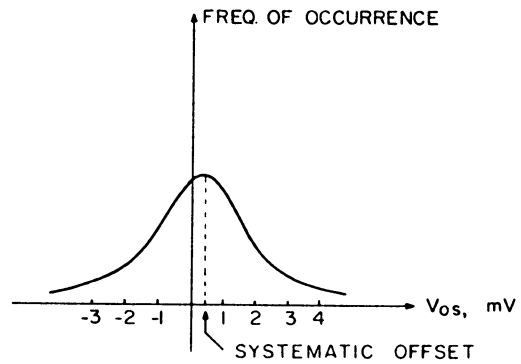


Fig. 5. Typical input offset distribution, MOS operational amplifier.

DC Offsets, DC Biasing, and DC Power Supply Rejection

The input offset voltage of an operational amplifier is composed of two components, the systematic offset and the random offset. The former results from the design of the circuit and is present even when all of the matched devices in the circuit are indeed identical. The latter results from mismatches in supposedly identical pairs of devices. A typical observed distribution of input offset voltages is shown in Fig. 5.

Systematic Offset Voltage

In bipolar technology, the comparatively high voltage gain per stage (on the order of 500) tends to result in a situation in which the input-referred dc offset voltage of an operational amplifier is primarily dependent on the design of the first stage. In MOS op amps, because of the relatively low gain per stage, the offset voltage of the differential to single-ended converter and second stage can play an important role. In Fig. 6, the operational amplifier of Fig. 4 has been split into two separate stages. Assuming perfectly matched devices, if the inputs of the first stage are grounded, then the quiescent output voltage at the drain of $M4$ is equal to the voltage at the drain of $M3$ ($M3$ and $M4$ have the same drain current and gate-source voltage, and hence must have the same drain-source voltage). However, the value of the gate voltage of $M6$ which is required to force the amplifier output voltage to zero may be different from the quiescent output voltage of the first stage. For a first stage gain of 50, for example, each 50 mV difference in these

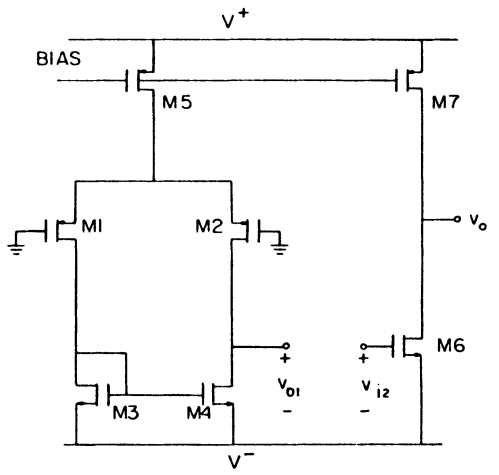


Fig. 6. Two-stage amplifier illustrating interstage coupling constraints.

voltages results in 1 mV of input-referred systematic offset. Thus, the W/L ratios of $M3$, $M4$, and $M6$ must be chosen so that the current density in these three devices are equal. For the simple circuit of Fig. 6, this constraint would take the form

$$\frac{(W/L)_3}{(W/L)_6} = \frac{(W/L)_4}{(W/L)_6} = \left(\frac{1}{2}\right)\frac{(W/L)_5}{(W/L)_7}. \tag{4}$$

In order that this ratio be maintained over process-induced variations in channel length, the channel lengths of $M3$, $M4$, and $M6$ usually must be chosen to be the same, and the ratios provided by properly choosing the channel widths. The use of identical channel lengths for the devices is at odds with the requirements (discussed later) that for low noise, $M3$ and $M4$ have low transconductance, and that for best frequency response under capacitive loading, $M6$ has high transconductance.

Systematic offset voltage is closely correlated with dc power supply rejection ratio. If a systematic offset exists, it is likely to display a dependence on power supply voltage, particularly if the bias reference source is such that the bias currents in the amplifier are not supply independent.

Random Input Offset Voltage

Source-coupled pairs of MOS transistors inherently display somewhat higher input offset voltage than bipolar pairs for the same level of geometric mismatch or process gradient. The reason for this is perhaps best understood intuitively by means of the conceptual circuit shown in Fig. 7. Here, a differential amplifier is made up of an identical pair of unilateral active devices biased at a current I and displaying a transconductance g_m. If the load elements, in this case assumed to be resistors, are assumed to mismatch by a percentage Δ, then in order for the output voltage of the differential amplifier to be zero, the absolute difference in the currents in the two devices must be equal to ΔI. This in turn requires that the dc input difference voltage applied to bring about this difference be

$$V_{gs} = \frac{I}{g_m}\Delta. \tag{5}$$

Thus, the input offset in this case depends on the I/g_m ratio of the active devices and the fractional mismatch in the

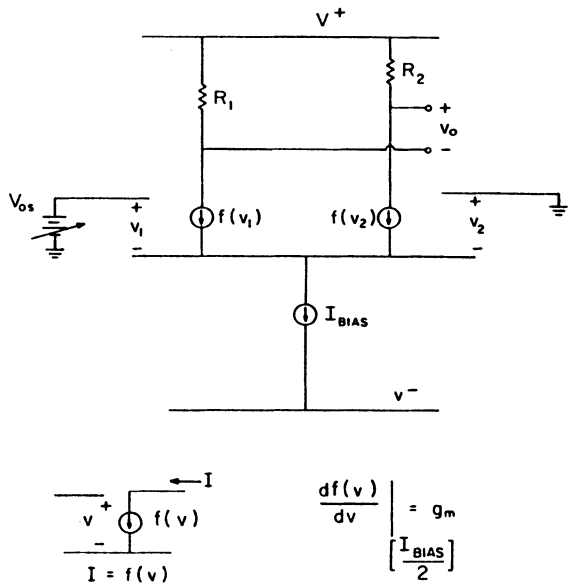


Fig. 7. Conceptual circuit for calculation of random offset voltage.

matched elements. A similar dependence is found for mismatches in many of the parameters of the active devices themselves, such as area mismatches in bipolar transistors and channel length and width mismatches in MOS transistors.

For bipolar devices, the I/g_m ratio is equal to kT/q or 0.026 V at room temperature. For MOS transistors, the ratio is $(V_{gs} - V_T)/2$, a bias-dependent quantity which is normally in the 100-500 mV range. While the offset voltage can be substantially improved by operating at low values of V_{gs}, the result is typically a somewhat larger offset voltage than in the bipolar case [2]. As discussed in a later section, the I/g_m ratio also directly effects the slew rate for class A input stages, so that often transient performance requirements place a lower limit on the allowable value of this parameter.

One mismatch component present in MOS devices which is not present in bipolar transistors is the mismatch in the threshold voltage itself. This component does not obey the above relationship, and results in a constant offset component which is bias current independent. Threshold mismatch is a strong function of process cleanliness and uniformity, and can be substantially improved by the use of common centroid geometries. Published data indicate that large-geometry common-centroid structures are capable of achieving threshold match distributions with standard deviations on the order of 2 mV in a silicon gate MOS process of current vintage [6].

Frequency Response, Compensation, Slew Rate, and Power Dissipation

The compensation of the two-stage CMOS amplifier can be carried out much as in the case of its bipolar equivalent using a pole-splitting capacitor C_c as shown in Fig. 2. However, important differences arise because of the much lower transconductance of the MOS transistor relative to bipolar devices [7]. The circuit can be approximately represented by the small-signal equivalent circuit of Fig. 8(a) if the nondominant poles due to the capacitances at the source of $M1$-2, the capacitance at the gate of $M3$, and any other nondominant poles which may exist on the circuit are neglected. This circuit has been analyzed by many authors [2], [8] because it occurs so fre-

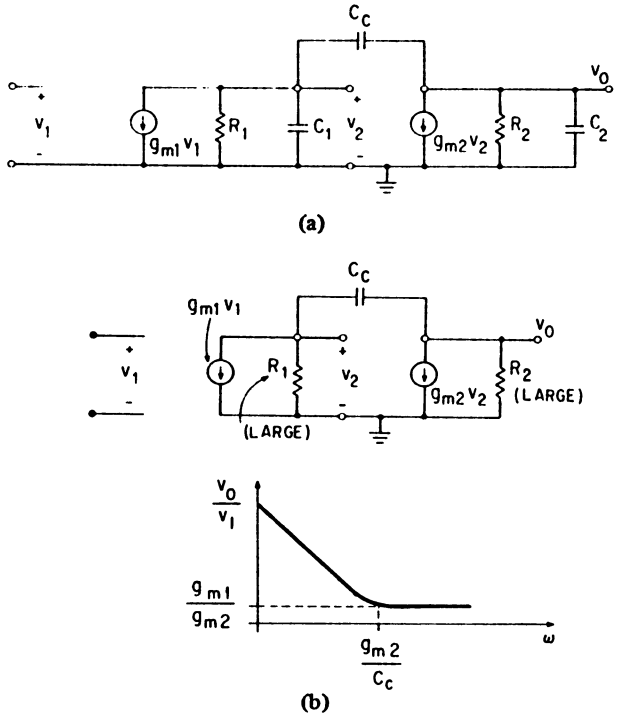


Fig. 8. (a) Small-signal equivalent circuit for two-stage amplifier. (b) Small-signal equivalent circuit with C_1 and C_2 set to zero, and gain of the circuit versus frequency.

quently in bipolar amplifiers. The circuit displays two poles and a right half-plane zero, which under the assumption that the poles are widely spaced, can be shown to be approximately located at

$$p_1 = \frac{-1}{(1+g_{m2}R_2)C_cR_1} \tag{6}$$

$$p_2 = \frac{-g_{m2}C_c}{C_2C_1 + C_2C_c + C_cC_1} \tag{7}$$

$$z = +\frac{g_{m2}}{C_c}. \tag{8}$$

Note that the pole due to the capacitive loading of the first stage by the second, p_1, has been pushed down to a very low frequency by the Miller effect in the second stage, while the pole due to the capacitance at the output node of the second stage, p_2, has been pushed to a very high frequency due to the shunt feedback. For this reason, the compensation technique is called pole splitting.

A unique problem arises when attempting to use pole splitting in MOS amplifiers. Analytically, the problem is illustrated by considering the location of the second pole p_2 and the right half-plane zero z relative to the unity-gain frequency g_{m1}/C_c. Here we make the simplifying assumption that the internal parasitic C_1 is much smaller than either the compensation capacitor C_c or the load capacitance C_2. This gives

$$\left|\frac{p_2}{\omega_1}\right| = \frac{g_{m2}C_c}{g_{m1}C_2} \tag{9}$$

$$\left|\frac{z}{\omega_1}\right| = \frac{g_{m2}}{g_{m1}}. \tag{10}$$

Note that the location of the right half-plane zero relative to the unity-gain frequency is dependent on the ratio of the transconductances of the two stages.

Physically, the zero arises because the compensation capacitor provides a path for the signal to propagate directly through the circuit to the output at high frequencies. Since there is no inversion in that signal path as there is in the inverting path dominant at low frequencies, stability is degraded. The location of the zero can best be conceptually understood by considering a case in which C_1 and C_2 are zero as illustrated in Fig. 8(b). For low frequencies, this circuit behaves like an integrator, but at high frequencies, the compensation capacitor behaves like a short circuit. When this occurs, the second stage behaves like a diode-connected transistor, presenting a load to the first stage equal to $1/g_{m2}$. Thus, the circuit displays a gain at high frequencies which is simply g_{m1}/g_{m2}, as illustrated in Fig. 8(b). The polarity of this gain is opposite to that at low frequencies, turning any negative feedback that might be present around the amplifier into positive feedback.

In bipolar technology, the transconductance of the second stage is normally much higher than the first because it is operated at relatively high current and the transconductance of the bipolar device is proportional to current level. In MOS amplifiers, the transconductances of the two stages tend to be similar, in part because the transconductance varies only as the square root of the drain current. Also, the transconductance of the first stage must be kept reasonably high for thermal noise reasons.

Fortunately, two effective means have evolved for eliminating the effect of the right half-plane zero. One approach has been to insert a source follower in the path from the output back through the compensation capacitor to prevent the propagation of signals forward through the capacitor [7]. This works well, although it requires more devices and dc bias current. An even simpler approach is to insert a nulling resistor in series with the compensation capacitor as shown in Fig. 9 [9]. In this circuit, note that at high frequencies, the output current from the first stage must flow principally as drain current in the second stage transistor. This, in turn, gives rise to voltage variation at the gate of the second stage which is proportional to the small-signal current from the first stage and inversely proportional to the transconductance of the second stage. In the circuit of Fig. 8, this voltage appears directly at the output. However, if a resistor of value equal to $1/g_{m2}$ is inserted in series with the compensation capacitor, the voltage across this resistor will cancel the small-signal voltage appearing on the left side of the compensation capacitor, resulting in the cancellation of the feedthrough effect.

Using an analysis similar to that performed for the circuit of Fig. 8, one obtains pole locations which are close to those for the original circuit, and a zero location of

$$z = \frac{1}{C_c\left(\frac{1}{g_{m2}} - R_z\right)}. \tag{11}$$

As expected, the zero vanishes when R_z is made equal to $1/g_{m2}$. In fact, the resistor can be further increased to move the zero into the left half-plane to improve the amplifier phase margin [10]. The movement of the zero for increasing values of R_z is illustrated in Fig. 10.

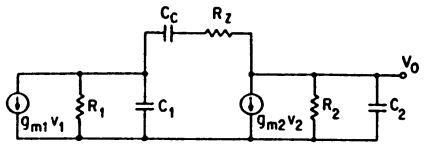


Fig. 9. Small-signal equivalent circuit of the basic amplifier with nulling resistor added in series with the compensation capacitor.

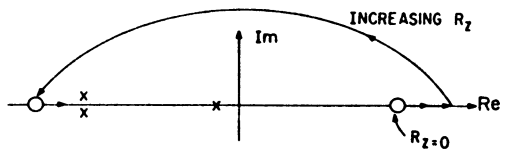


Fig. 10. Pole-zero diagram showing movement of the transmission zero for various values of R_Z.

A second problem in compensation involves the effects of capacitive loading. From (9), the location of the nondominant pole due to capacitive loading on the output node relative to the unity-gain frequency is determined by the ratio of the second-stage transconductance to that of the first and the ratio of the load capacitance to the compensation capacitance. Since the stage transconductances tend to be similar, this implies that the use of load capacitances of the same order as the compensation capacitance will tend to degrade the unity-gain phase margin because of the encroachment of this nondominant pole. This is of considerable practical significance in switched capacitor filters where large capacitive loads must be driven, and the use of an output stage is undesirable for power dissipation and noise reasons.

Slew Rate

As in its bipolar counterpart, the CMOS op amp of Fig. 4 displays a relationship among slew rate, bandwidth, input stage bias current, and input device transconductance of

$$SR = \frac{I_{D1}}{g_{m1}} \omega_1 \tag{12}$$

where g_{m1} is the input transistor transconductance, I_{D1} is the bias current of the input devices, and ω_1 is the unity-gain frequency of the amplifier. For the MOS case, this gives

$$SR = \frac{(V_{gs} - V_T)_1}{2} \omega_1. \tag{13}$$

In effect, the $(V_{gs} - V_T)$ of the input stage is the range of differential input voltage for which the input stage stays in the active region. If the bandwidth is kept constant and this range is increased, the slew rate improves. Because this range is usually substantially higher in MOS amplifiers than in bipolar amplifiers, MOS amplifiers usually display relatively good slew rate. In micropower or precision applications where the input transistors are operated at very low $(V_{gs} - V_T)$, this may not be the case, however.

Power Dissipation

Even for the simple circuit of Fig. 4, the minimum achievable power dissipation is a complex function of the technology used and the particular requirements of the application. In sampled data systems such as switched capacitor filters, the requirement is that the amplifier be able to settle in a certain time to a certain accuracy with a capacitive load of several picofarads. In this application, the factors determining the minimum power dissipation tend to be the fact that there must be enough standing current in the amplifier class *A* second stage such that the capacitance can be charged in the allowed time, and the fact that the amplifier must have sufficient phase margin to avoid degradation of the settling time due to ringing and overshoot. The latter requirement dictates a certain minimum g_m in transistor *M6* for a given bandwidth and load capacitor. This, in turn, usually dictates a certain minimum bias current in *M6* for a reasonable device size. If a class *A* source follower output stage is added, then the same comment would apply to its bias current since its g_m, together with the load capacitance, contribute a nondominant pole.

The preceding discussion is predicted on the use of class *A* circuitry (i.e., circuits whose available output current is not greater than the quiescent bias current). Quiescent power dissipation can be greatly reduced through the use of dynamic circuits and class *B* circuits, as discussed later.

Noise Performance

Because of the fact that MOS devices display relatively high $1/f$ noise, the noise performance is an important design consideration in MOS amplifiers. All four transistors in the input stage contribute to the equivalent input noise, as illustrated in Fig. 11. By simply calculating the output noise for each circuit and equating them (11),

$$V_{eqTOT}^2 = V_{eq1}^2 + V_{eq2}^2 + \left(\frac{g_{m3}}{g_{m1}}\right)(V_{eq3}^2 + V_{eq4}^2) \tag{14}$$

where it has been assumed that $g_{m1} = g_{m2}$ and that $g_{m3} = g_{m4}$. Thus, the input transistors contribute to the input noise directly, while the contribution of the loads is reduced by the square of the ratio of their transconductance to that of the input transistors. The significance of this in the design can be further appreciated by considering the input-referred $1/f$ noise and the input-referred thermal noise separately.

Input-Referred 1/f Noise

The equivalent input noise spectrum of a typical MOS transistor is shown in Fig. 12. The dependence of the $1/f$ portion of the spectrum on device geometry and bias conditions has been studied by many authors [12]-[14]. Considerable discrepancy exists in the published data on $1/f$ noise, indicating that it arises from a mechanism that is strongly affected by details of device fabrication. Perhaps the most widely accepted model for $1/f$ noise is that for a given device, the gate-referred equivalent mean-squared voltage noise is approximately independent of bias conditions in saturation, and is inversely proportional to the gate capacitance of the device. The following analytical results are based on this model, but it should be emphasized that the actual dependence must be verified for each process technology and device type [12], [15]. Thus,

$$V_{1/f}^2 = \frac{K}{C_{ox}WL}\left(\frac{\delta f}{f}\right). \tag{15}$$

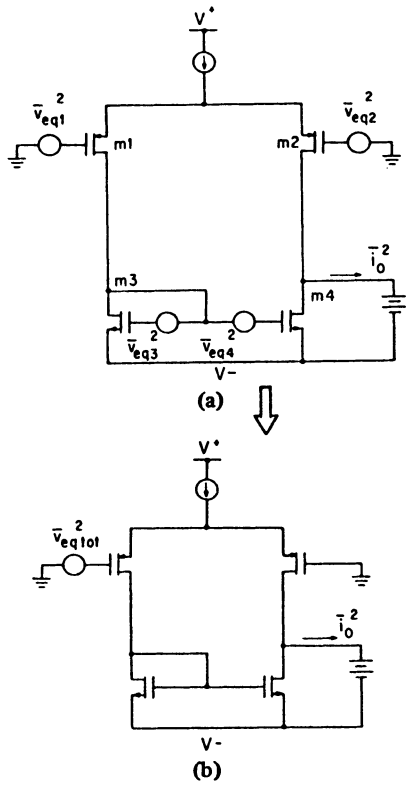


Fig. 11. CMOS input stage. (a) Device noise contributions. (b) Equivalent input noise.

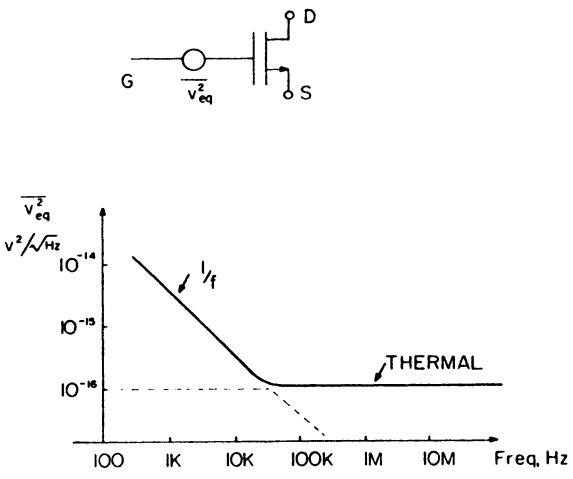


Fig. 12. Typical equivalent input noise, MOS transistor.

Utilizing this assumption, we obtain for the equivalent input noise

$$V_{1/f}^2 = \frac{2K_p}{W_1 L_1 C_{ox}} \left(1 + \frac{K_n \mu_n L_1^2}{K_p \mu_p L_3^2}\right)\left(\frac{\delta f}{f}\right) \tag{16}$$

where K_n and K_p are the flicker noise coefficients for the n-channel and p-channel devices, respectively. Depending on processing details, these may be comparable or different by a factor of two or more. Note that the multiplying term in front is the input noise of the input transistors, and the second term is the increase in noise due to the loads. It is clear from this second term that the load contribution can be made small by simply making the channel lengths of the loads longer than that of the input transistors by a factor of on the order of two or more. The input transistors can then be made wide enough to achieve the desired performance. It is interesting to note that increasing the width of the channel in the loads does not improve the $1/f$ noise performance.

Thermal Noise Performance

The input-referred thermal noise of an MOS transistor is given by [8]

$$V_{eq}^2 = 4kT\left(\frac{2}{3g_m}\right)\delta f. \tag{17a}$$

Utilizing the same approach as for the flicker noise, this gives

$$V_{eq}^2 = 4kT\,\frac{4}{3\sqrt{2\mu_p C_{ox}(W/L)_1 I_D}}\left(1 + \sqrt{\frac{\mu_n (W/L)_3}{\mu_p (W/L)_1}}\right). \tag{17b}$$

Again, the first term represents the thermal noise from the input transistors, and the term in parentheses represents the fractional increase in noise due to the loads. The term in parentheses will be small if the W/L's are chosen so that the transconductance of the input devices is much larger than that of the loads. If this condition is satisfied, then the input noise is simply determined by the transconductance of the input transistors.

Power Supply Rejection and Supply Capacitance

Power supply rejection ratio (PSRR) is a parameter of considerable importance in MOS amplifier design. One reason for this is that in complex analog-digital systems, the analog circuitry must coexist on the same chip with large amounts of digital circuitry. Even though separate analog and digital supply buses are often run on chip, it is hard to avoid some coupling of digital noise into the analog supplies. A second reason is that in many systems, switching regulators are used which introduce power supply noise into the supply voltage lines. If these high-frequency signals couple into the signal path in a sampled data system such as a switched capacitor filter or high-speed A/D converter, they can be aliased down into the frequency band where the signal resides and degrade the overall system signal-to-noise ratio. The parameters reflecting susceptibility to this phenomenon in the operational amplifier are the high-frequency PSRR and the supply capacitance.

The PSRR of an operational amplifier is simply the ratio of the voltage gain from the input to the output (open loop) to that from the supply to the output. It can be easily demonstrated that for frequencies less than the unity-gain frequency, if the operational amplifier is connected in a follower configuration and an ac signal is superimposed on one of the power supplies, the signal appearing at the output is equal to the applied signal divided by the PSRR for that supply. The basic circuit of Fig. 4 is particularly poor in terms of its high-frequency rejection from the negative power supply, as illustrated in Fig. 13. The primary reason is that as the applied frequency increases, the impedence of the compensation capacitor decreases, effectively shorting the drain of $M6$ to its gate for ac signals. Thus, the gain from the negative supply to the output

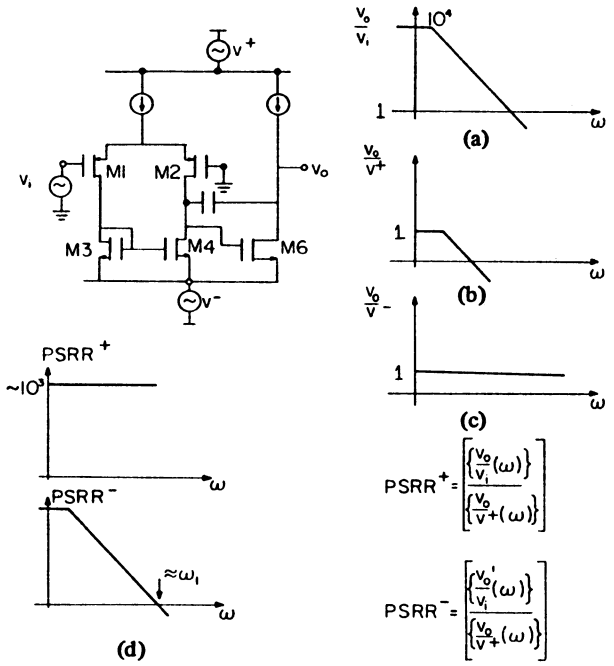


Fig. 13. High-frequency PSRR of bipolar-derived op amp. (a) Gain from input to output. (b) Gain from positive supply to output. (c) Gain from negative supply to output. (d) Positive and negative PSRR.

approaches unity and stays there out to very high frequencies. The same phenomenon causes the gain from the positive supply to fall with frequency as the open-loop gain does, so that the positive PSRR remains relatively flat with increasing frequency. The negative supply PSRR falls to approximately unity at the unity-gain frequency of the operational amplifier. Several alternative amplifier architectures have evolved which alleviate this problem, and they are discussed in a later section.

A second important contribution to coupling between the power supply and the signal path at high frequency is termed supply capacitance [10], [17]. This phenomenon manifests itself as a capacitive coupling between one or both of the power supplies and the operational amplifier input leads. The effect of this capacitance is illustrated in Fig. 14 for a switched capacitor integrator. Since the op amp input is connected to the summing node, then the power supply variations will appear at the integrator output attenuated by the ratio of the supply capacitance to the integrator capacitance. The result can be quite poor power supply rejection in switched capacitor filters and other sampled data analog circuits.

Supply capacitance effects can occur in several ways, but four important ones are given below.

1) Variation in drain voltage on $M1$, $M2$ with negative supply voltage. If the op amp inputs are grounded and the negative supply voltage changes, then a displacement current flows into the summing node because of the resulting change in voltage across the drain-gate capacitance of the input transistors. This is usually eliminated by the use of cascode transistors in series with the drains of the input transistors.

2) Variation of drain current in $M1$, $M2$ with supply voltage. Use of a bias reference which results in bias current variations with supply voltage will, in turn, cause the $V_{gs} - V_T$ of the input devices to change with supply voltage. This will cause a

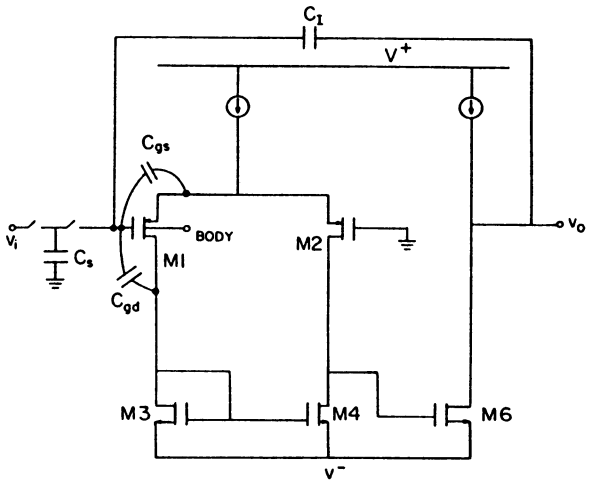


Fig. 14. Supply capacitance in a CMOS amplifier.

displacement current to flow through the gate-source capacitance of $M1$, $M2$ onto the summing node. The usual solution to this problem is the use of a supply-independent bias reference.

3) Variation of body bias on $M1$, $M2$ with supply voltage variations. If the substrate terminal of the input transistors is tied to a supply or supply-related voltage, then as the supply voltage changes, the substrate bias changes. This, in turn, changes the threshold, which changes V_{gs}. The resulting displacement current in C_{gs} flows into the summing node. In CMOS operational amplifiers, the usual solution to this problem is to put the input transistors in a well and tie the well to the sources of the input transistors. This dictates, for example, that in a p-well process, the input devices be n-channel devices, and vice versa. In NMOS, the use of fully differential circuitry is probably the only way to fully alleviate the problem since the substrate must be tied to a supply. A second alternative is capacitive decoupling of the substrate so that it does not follow high-frequency supply variations [17].

4) Interconnect crossovers in the amplifier layout and in the system layout can produce undesired supply capacitance. This can usually be overcome with careful layout.

IV. Alternative Architectures for Improved Performance

The bipolar-derived amplifier discussed above is widely used at the present time, although with many variations, in a variety of applications. However, many alternative circuit approaches have been investigated and, in many cases, utilized in commercial products in order to achieve performance which is superior to that available from the basic circuit in some respect. In this section, we first consider variations on the basic circuit, and then alternative architectures.

Variations on the Basic Two-Stage Amplifier

Use of Cascodes for Improved Voltage Gain: In precision applications involving large values of closed-loop gain, the voltage gain available from the basic circuit shown in Fig. 4 may be inadequate. One approach to improving the voltage gain without adding an additional common-source stage with its

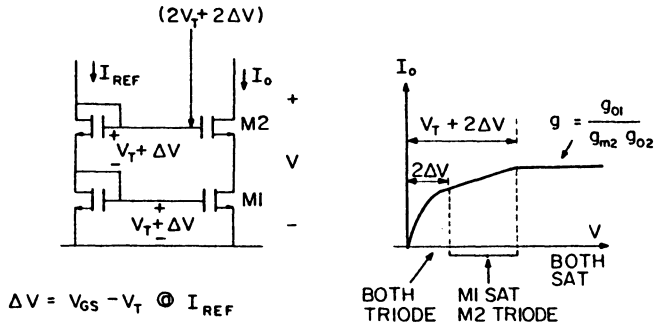


Fig. 15. Cascode current source.

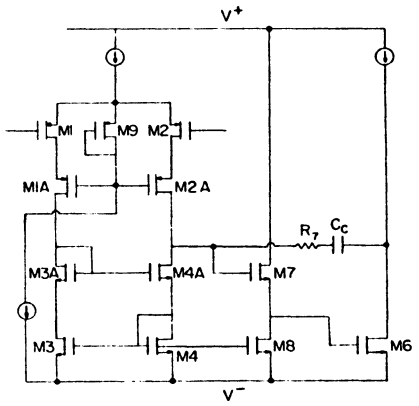


Fig. 16. Two-stage amplifier with cascoded first stage.

associated high impedence node and pole is to add a common-gate, or cascode, transistor to increase the output resistance of the common-source transistors in the basic amplifier. The basic cascode circuit is shown in Fig. 15. It is easily demonstrated that the incremental output resistance of this current source is equal to

$$r_o = r_{o2}[1 + g_{m2}r_{o1}] + r_{o1}. \tag{18}$$

The output resistance is increased by an amount equal to the open circuit gain of the cascode transistor. This circuit may be directly applied to the basic two-stage amplifier in either the first stage, second stage, or in both stages. The circuit of Fig. 16 illustrates the use of cascodes in the first stage. One disadvantage of this circuit is a substantial reduction in input stage common mode range, but this can be alleviated by optimizing the biasing of the cascodes, to be discussed in Section V.

Improved PSRR Grounded-Gate Cascode Compensation: Read and Wieser [18] have recently described a technique for improving the negative supply PSRR of the circuit of Fig. 4. Conceptually, if the left end of the capacitor could be connected to a virtual ground, then the capacitor voltage would not have to change whenever the negative supply voltage changed in order to have the output remain constant. This was accomplished by inserting a cascode device in this loop with the gate connected to ground, as shown in Fig. 17. The displacement current from the capacitor flows into the source of this transistor and out the drain into the compensation point. An additional current source and current sink of equal values must be added to bias the common-gate device in the active region and so as not to contribute any systematic offset.

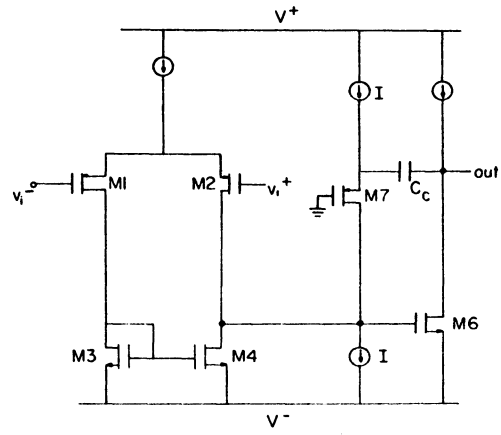


Fig. 17. Schematic of basic amplifier with cascode feedback compensation.

The resulting negative PSRR at high frequencies is greatly improved at the cost of a slight increase in complexity, random offset, and noise [16].

Common-Source Common-Gate Amplifiers

The basic amplifier considered thus far is actually a cascade of two common-source stages. An alternative approach is to use a cascade of a common-source stage and a common-gate stage, often called a cascode amplifier. An example of an amplifier utilizing this architecture is shown in Fig. 18. The voltage gain of this circuit at dc is approximately the same as that of the basic two-stage circuit. The small-signal impedance at the output node is increased by $g_m r_o$ relative to the output node of the two-stage circuit, and the voltage gain is simply the product of the transconductance of the input transistors and the impedance at the output mode:

$$A_v = \frac{g_{m1}}{\dfrac{g_{o2}+g_{o9}}{g_{m4}r_{o4}} + \dfrac{g_{o7}}{g_{m5}r_{o5}}}. \tag{19a}$$

The principal reasons for considering this architecture are twofold. First, the compensation capacitor and load capacitor are the same element in this circuit. The first nondominant pole comes from the g_m/C_{gs} time constant of the n-channel cascode devices, and gives a pole frequency approximately at the f_t of these devices. A second nondominant pole results from the differential to single-ended converter. However, the nondominant pole due to the load capacitance present in the two-stage circuit, is not present in this circuit. Thus, this circuit is capable of achieving higher stable closed-loop bandwidth with large capacitive load. The principal application of this architecture to date has been in high-frequency switched capacitor filters [20], [21].

An important advantage of this circuit is that it does not suffer from the degradation of the high-frequency power supply rejection problem inherent in the pole-split compensated two-stage architecture, assuming that the load capacitance or part of it is not tied to a power supply.

Because of the fact that cascode transistors are used at the output, the output swing of this circuit is lower than the common-source common-source amplifier. This problem can

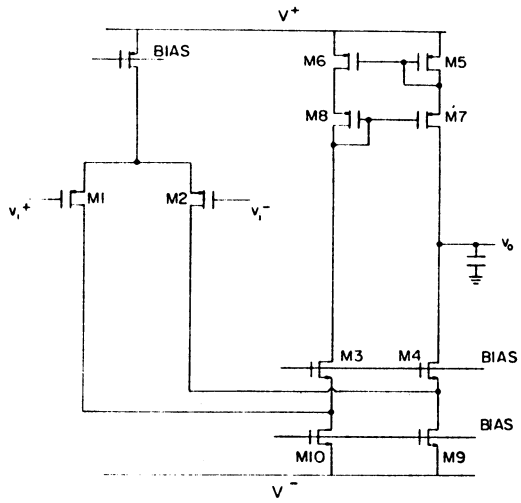


Fig. 18. One-stage amplifier schematic.

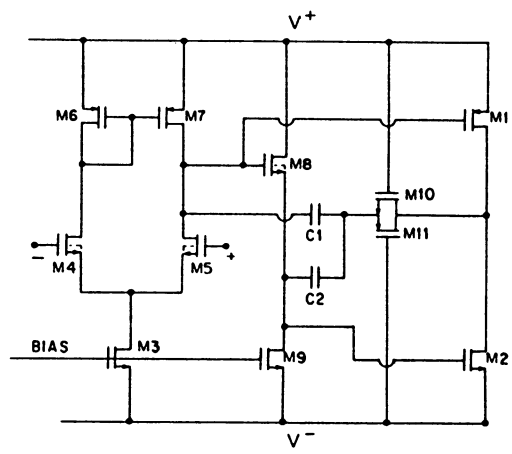


Fig. 19. Example circuit illustrating class *AB* second stage.

be minimized by modifying the bias generator such that the lower transistors in the cascode current source are biased on the edge of saturation (i.e., $V_{gd} = V_T$). This results in an available output voltage swing within $2(V_{gs} - V_T)$ of each supply, or perhaps 0.4 to 0.8 V in voiceband filters. MOS transistors actually display a rather indistinct transition from triode to saturation as the drain depletion region forms, and as a result, the bias point must actually be chosen so as to bias the lower transistor a few hundred millivolts into saturation if the predicted value of incremental output resistance is to be obtained.

A second disadvantage of this circuit is that more devices contribute to the input-referred voltage and input offset voltage. Assuming that transistors *M5*-*M8* are biased at the same current as the input devices, the input-referred flicker noise can be shown to be

$$v_{eq}^2 = \frac{2K_p}{W_1L_1}\left[1 + \frac{2K_n\mu_n}{K_p\mu_p}\left(\frac{L_1}{L_9}\right)^2 + \left(\frac{L_1}{L_5}\right)^2\right]\frac{\delta f}{f}. \qquad (19b)$$

In this case, the current sources *M9* and *M10* contribute an additional term not present in (16). However, as in the case of the common-source common-source amplifier, the equivalent input noise can be made almost equivalent to the noise of the input transistors alone by choosing the channel lengths of the input transistors to be short compared to those of *M5*, *M6*, *M9*, and *M10*. The same considerations apply for the thermal noise.

Class AB Amplifiers, Dynamic Amplifiers, and Dynamic Biasing

Many, if not all, MOS analog circuits commercially produced utilize class *AB* circuitry in some form. Here the term class *AB* is taken to mean a circuit which can source and sink current from a load which is larger than the dc quiescent current flowing in the circuit. The most widespread application is in output buffers, but if an important objective is the minimization of chip power, then the philosophy of using class *AB* operation can be extended to the internal amplifiers. The motivation for doing so is that one of the factors that dictates the value of the quiescent current with an MOS amplifier is the value of current required to charge the load and/or the compensation capacitance in the required time. However, it is relatively rare that the operational amplifier outputs actually have to change the maximum amount in one clock cycle. Large power savings can be effected if only that current is drawn which is required to charge the capacitance on that particular cycle. An example of an amplifier utilizing a class *A* first stage and a class *AB* second stage is shown in Fig. 19 [22]. In a conventional circuit, the gate of *M2* would be connected to a level-shifted version of the stage input voltage. Thus, when the first stage output swings positive, reducing the current in *M1*, the current in *M2* increases above its quiescent dc value. An example of a single-stage amplifier that operates on this principle is shown in Fig. 20 [23]. This particular circuit can be used in the inverting mode only. With the input grounded, the quiescent current in the input transistors is determined by the bias voltages shown. Upon the application of a voltage to the input, the current in one side of the input stage increases monotonically with the applied voltage until the power supply is reached, while the other side of the input stage turns off. The amount of current available at the output is much larger than the quiescent current, and the circuit, as a result, does not follow the relationship of (12). In fact, the circuit does not display slew rate limiting in the usual sense. Another aspect of this circuit is the fact that the small-signal voltage gain in the quiescent mode can be quite high because of the low current level, and the fact that the voltage gain falls off during transients because of the high current levels is of little consequence. Similar circuits have been used extensively in bipolar technology [24].

Degrauwe *et al.* [25] have recently described a novel approach to the same objective. A conventional class *A* amplifier configuration is used, but an auxiliary circuit is used to detect the presence of large differential signals at the input. The bias current in the class *A* circuitry is then increased when such signals are present. Experimental versions of such amplifiers have yielded quiescent power dissipation of less than 10 μW.

A second class of amplifiers has been explored by several authors, beginning with Copeland [26], in which the quiescent current in the absence of signals is allowed to decay to zero. Such amplifiers are fully dynamic in the sense that no dc paths

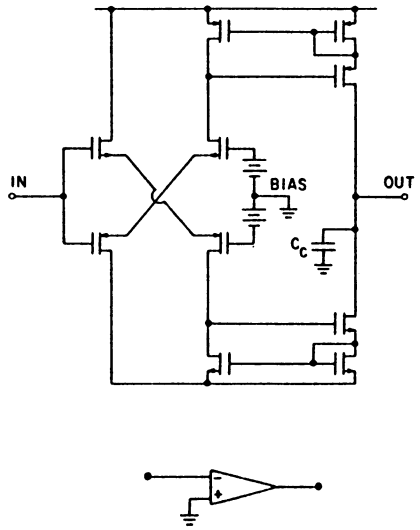


Fig. 20. Examples of a single-stage class *AB* op amp.

exist for current to flow from the supply. While very low values of power dissipation can be obtained, difficult problems of settling time and power supply rejection remain to be solved with these amplifiers.

Hostica [27] has described a third approach to micropower amplifier design for switched capacitor applications which utilizes a time-varying periodic bias current which is synchronous with the master clock in the filter. In contrast to the approaches described above, the power supply current is independent of signal amplitude, and is made large during the early part of the clock period for fast slewing and small during the later portion for high gain and power savings. The bias current waveform is generated by discharging a capacitor into the input of a current mirror. While this technique, in principle, dissipates more power than the other approaches under low signal conditions, it can be implemented with relatively simple circuitry, and it has demonstrated good experimental results for both one-stage and two-stage amplifiers [28].

V. Differential Output Amplifiers

As has been mentioned, power supply rejection is an important performance parameter for amplifiers to be used in complex analog/digital systems. In addition, one inevitable result of technology scaling is a reduction in power supply voltage with an accompanying reduction in internal signal swings and dynamic range. These two considerations make use of fully differential signal paths throughout the analog portions of the system attractive for some systems [29], [30]. The inherently differential nature of the circuit tends to give very high PSRR since the supply variations appear as a common mode signal. Also, the effective output swing is doubled, while the magnitude of the input-referred operational amplifier noise remains the same, giving a 6 dB improvement in operational amplifier noise-limited dynamic range.

A typical implementation of a differential switched capacitor integrator is shown in Fig. 21. The operational amplifier is required to produce two analog outputs which are symmetric about ground, in contrast to the single-ended case where only one is produced. An equivalent circuit for a differential op amp is shown in Fig. 22. An example of a CMOS differential output operational amplifier is shown in Fig. 23.

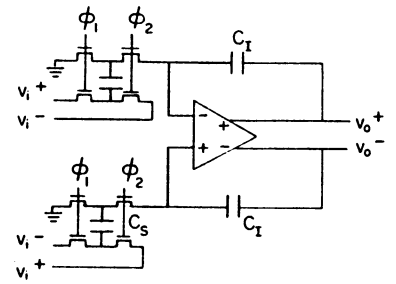


Fig. 21. Differential switched capacitor integrator.

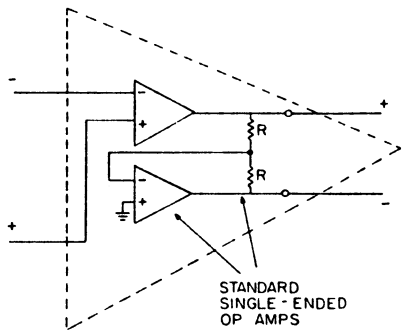


Fig. 22. Equivalent circuit for a differential output operational amplifier.

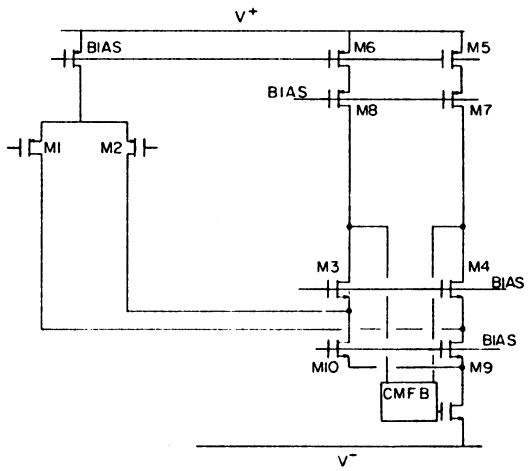


Fig. 23. Example of a differential output amplifier. The block labeled *CMFB* serves to keep the common-mode output voltage near ground.

An important problem in such amplifiers is the design of a feedback loop to force the common mode output voltage to be ground or some other internal reference potential. This feedback path can be implemented with transistors in a continuous-time circuit or can be implemented with switched capacitor circuitry. The continuous approach is potentially simpler, but presents a difficult design problem in making the common-mode output voltage independent of the differential mode signal voltage [21], [29]. Switched capacitor circuitry can make use of the linearity of MOS capacitors to achieve this goal [30]. The choice between the two techniques depends on the sensitivity of the particular application to variations in common-mode voltage.

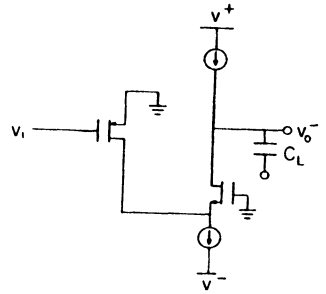


Fig. 24. Small-signal differential half circuit for the amplifier in Fig. 23.

Another important advantage of differential output amplifiers is that the differential single-ended converter with its associated nondominant poles is eliminated. The small-signal equivalent circuit for the circuit in Fig. 22, for example, is a simple common-source common-gate cascade, as shown in Fig. 24. This circuit has only one nondominant pole, at the f_t of the common-gate device. Thus the configuration is particularly well suited to the implementation of high-frequency switched capacitor filters. A configuration of this type has been used in recently reported work yielding high-Q switched capacitor filters clocked at 4 MHz with center frequencies of 250 kHz in a 4 μm silicon gate CMOS technology [21].

VI. NMOS Operational Amplifiers

The design of an operational amplifier of a given performance level in NMOS depletion load technology is a much more difficult task than in CMOS technology. The absence of a complementary device makes the implementation of level shifters which track supply voltage variations much more complex. The level of body effect found in most depletion load devices makes the realization of large gains per stage difficult. Assuming that the basic architecture is similar to that illustrated in Fig. 2(c), the small-signal properties, voltage gain, transient response and slew rate, input noise, and power supply rejection considerations are basically similar to the two-stage CMOS amplifier. The key additional considerations are the shunting effects of the incremental output conductance of the depletion load current sources and the impedance and power supply variation of the floating level-shifting voltage sources and the resulting degradation of power supply rejection ratio. Nonetheless, creative circuit design has resulted in NMOS amplifiers which nearly match CMOS amplifiers in most performance aspects, albeit at the cost of somewhat more complexity, die area, and power dissipation. Circuit techniques used to achieve this include replica biasing for tracking level shifters [9], [17], positive feedback for high voltage gain [31], [30], differential configurations for power supply rejection [29], [30], and others.

While there will no doubt always be a need for NMOS amplifiers for some applications, the emergence of CMOS as a key VLSI digital technology has resulted in the widespread adoption of CMOS for new mixed analog-digital designs.

VII. Output Buffers

In amplifier applications involving either a large capacitive or resistive load, an output stage must be added to the basic amplifier to prevent the load from degrading the voltage gain or closed-loop stability. This situation most often arises when signals must be supplied off the chip to an external environment. The key requirements on such stages is that they be sufficiently broad band with heavy capacitive loading such that they do not degrade the loop stability of the operational amplifier, and such that the output is able to supply a large enough voltage swing to the load with the maximum load conductance. While class A source follower or emitter follower circuits can be used in some applications, quiescent power dissipation considerations usually dictate a class AB implementation of the circuit. This discussion is limited to class AB output buffers.

In bipolar operational amplifier design, the complementary emitter follower class AB configuration is used in the vast majority of cases. In contrast, class AB CMOS output stage implementations tend to vary widely, depending on the specific devices available in the particular technology used. The CMOS complementary source follower class AB output buffer stage shown in Fig. 25 is a direct analog of its bipolar counterpart. The primary drawback of this circuit is that the output voltage swing is limited by the gate-source voltage of the output transistors. This occurs because the transistors used for logic functions on the chip have thresholds in the 0.5-1 V range, so that the amount of swing lost due to threshold voltage plus the $(V_{gs} - V_T)$ drop is too large for many applications. However, many technologies have an extra device type with very low threshold voltage, and in this case, this low threshold device can be used for one of the two output transistors. It is rare that both p-channel and n-channel low threshold devices are available in the same technology.

In many CMOS technologies, a bipolar transistor follower is available and can be used in place of one of the output followers. This provides very low output resistance and good output swing. In processes with light substrate doping, potential latchup problems can make the use of such devices in off-chip driver stages impractical because of the fact that the collector current of the transistor flows in the substrate and can cause voltage drops which cause a junction to be forward biased. An example of the use of a bipolar device in an MOS output stage together with a low threshold device is illustrated in Fig. 26.

A third alternative is the use of quasi-complementary configurations in which a common-source transistor together with an error amplifier is used in place of one or both of the follower devices. This circuit is shown conceptually in Fig. 27. The combination of the error amplifier and the common-source device mimics the behavior of a follower with high dc transconductance. Such quasi-complementary circuits provide excellent dc performance with voltage swings approaching the supply rails, but since the amplifier must be broad band to prevent crossover distortion problems, they present difficult problems in compensation of the local feedback loop in the presence of large capacitive loads. Proper control of the quiescent current is also a key design constraint.

Low threshold devices, bipolar devices, and quasi-complementary devices can be used in any combination, depending on what devices are available in the particular technology

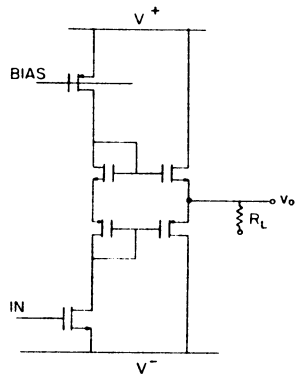


Fig. 25. Complementary source follower CMOS output stage based on the traditional bipolar implementation.

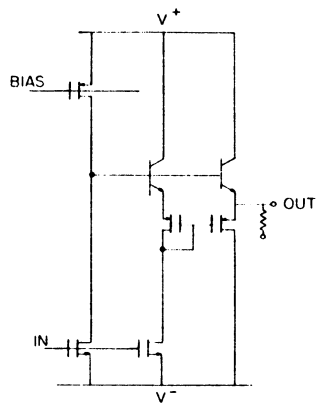


Fig. 26. Example of a CMOS output stage using a bipolar emitter follower and a low-threshold p-channel source follower.

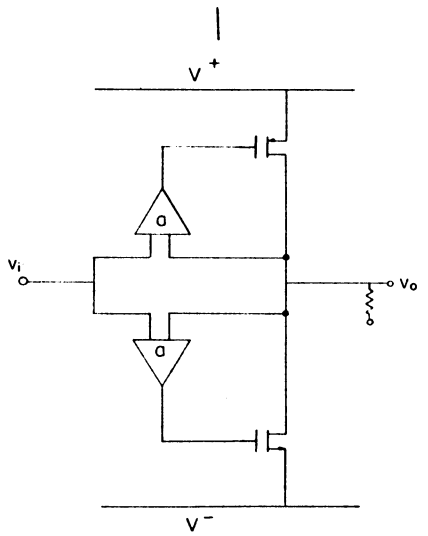


Fig. 27. Example of a complementary class *B* output stage using compound devices with imbedded common-source output transistors.

being used. Whereas in the bipolar case the vast majority of output stage applications can be satisfied using the traditional complementary class *B* emitter follower stage, no single circuit approach has yet emerged as the standard for CMOS output stages.

VIII. Summary and Conclusions

In this paper, we have attempted to summarize the various techniques and architectures which have been applied in the design of MOS operational amplifiers in the past several years. An important question is the extent to which these amplifier designs can be scaled as minimum feature sizes continue to decrease. As pointed out in a recent study [32], dc parameters such as voltage gain are generally unaffected for constant-field scaling, although they are degraded for quasi-constant voltage or constant voltage scaling. Perhaps the most difficult problem results from the fact that the effective dynamic range of the amplifier falls in scaled technologies. This occurs fundamentally because of the fact that analog signal swings fall with reductions in power supply voltage. Input-referred thermal noise remains constant because of the fact that the device transconductance remains constant under constant-field scaling. The input-referred $1/f$ noise increases, but this does not appear to be a fundamental limitation on system dynamic range because the signal can always be translated to a higher portion of the spectrum using techniques like chopper stabilization [29]. Also, newer technologies have demonstrated continuing reductions in $1/f$ noise as a result of better process control.

In sampled data analog amplifiers, filters and data converters, the primary limitation on dynamic range, assuming that $1/f$ noise has been removed, is the kT/C noise contributed by the analog switches making up the filter. The kT/C limited dynamic range also falls as the technology is scaled, and since for practical clock rates and capacitor sizes this noise source is dominant over op amp thermal noise, there appears to be no barrier to constant-field scaling of operational amplifiers for this application, assuming that $1/f$ noise is removed by circuit or technological means. Thus, the adaptation of the circuit approaches described in this paper to lower supply voltages and scaled devices, and the removal of $1/f$ noise from the signal path in such circuits, are important objectives in future work.

References

[1] D. A. Hodges, P. R. Gray, and R. W. Broderson, "Potential of MOS technologies for analog integrated circuits," *IEEE J. Solid-State Circuits*, pp. 285-293, June 1978.

[2] J. E. Solomon, "The monolithic op amp, A tutorial study," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 314-332, Dec. 1974.

[3] Y. P. Tsividis, "Design considerations in single-channel MOS analog circuits—A tutorial," *IEEE J. Solid-State Circuits*, pp. 383-391, June 1978.

[4] P. R. Gray, "Basic MOS operational amplifier design—An overview," in *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980, pp. 28-49.

[5] R. W. Broderson, P. R. Gray, and D. A. Hodges, "MOS switched capacitor filters," *Proc. IEEE*, pp. 61-75, Jan. 1979.

[6] O. H. Shade, Jr., "BiMOS micropower integrated circuits, *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 791-798, Dec. 1978. See also O. H. Schade, Jr. and E. J. Kramer, "A low-voltage BiMOS op amp," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 661-668, Dec. 1981.

[7] Y. P. Tsividis and P. R. Gray, "An integrated NMOS operational amplifier with internal compensation," *IEEE J. Solid-State Circuits*, vol. SC-11, pp. 748-753, Dec. 1976.

[8] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*. New York: Wiley, 1977.

[9] D. Senderowicz, D. A. Hodges, and P. R. Gray, "A high-performance NMOS operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 760-768, Dec. 1978.

[10] W. C. Black and D. J. Allstott, "Low power CMOS channel filter," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 929-938, Dec. 1980.

[11] J. C. Bertails, "Low frequency noise considerations for MOS am-

plifier design," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 773-776, Aug. 1979.
[12] M. B. Das and J. M. Moore, "Measurements and interpretation of low-frequency noise in FET's," *IEEE Trans. Electron Devices*, vol. ED-21, Apr. 1974.
[13] N. R. Mantena and R. C. Lucas, "Experimental study of flicker noise in MIS transistors," *Electron. Lett.*, vol. 5, pp. 697-603, 1969.
[14] H. Mikoshiba, "1/*f* noise in n-channel silicon gate MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-29, June 1962.
[15] E. Vittoz and J. Fellrath, "CMOS integrated circuits based on weak inversion operation," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 214-231, June 1977.
[16] B. Ahuja, Intel Corporation, private communication.
[17] H. Ohara, W. M. Baxter, C. F. Rahim, and J. L. McCreary, "A precision low power PCM channel filter with on-chip power supply regulation," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1005-1013, Dec. 1980.
[18] R. Read, private communication.
[19] P. R. Gray and R. G. Meyer, "Recent advances in monolithic operational amplifier design," *IEEE Trans. Circuits Syst.*, pp. 317-327, May 1974.
[20] P. R. Gray, R. W. Broderson, D. A. Hodges, T. C. Choi, R. Kaneshiro, and K. C. Hsieh, "Some practical aspects of switched capacitor filter design," in *Dig. Tech. Papers, 1981 Int. Symp. Circuits Syst.*
[21] T. Choi, R. Kaneshiro, R. W. Broderson, and P. R. Gray, "High frequency CMOS switched capacitor filters for communications applications," in *Dig. Tech. Papers, 1983 Int. Solid-State Circuits Conf.*
[22] Y. A. Haque, R. Gregortan, D. Blasco, R. Mao, and W. Nicholson, "A two-chip PCM codec with filters," *IEEE J. Solid-State Circuits*, vol. SC-24, pp. 961-969, Dec. 1979.
[23] W. Black, personal communication.
[24] P. C. Davis and V. Saari, "A high slew rate monolithic op amp using compatible complementary PNPs," in *Dig. Tech. Papers, IEEE Int. Solid-State Circuits Conf.*, Philadelphia, PA, Feb. 1974.
[25] M. G. Degrauwe, J. Rijmenants, E. A. Vittoz, and H. J. De Man, "Adaptive biasing CMOS amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 522-528, June 1982.
[26] M. A. Copeland and J. M. Rabaey, "Dynamic u amplifiers for MOS technology," *Electron. Lett.*, vol. 15, pp. 301-302, May 1979.
[27] B. J. Hosticka, "Dynamic CMOS amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 887-894, Oct. 1980.
[28] B. J. Hosticka, D. Herbst, B. Hoefflinger, U. Kleine, J. Pandel, and R. Schweer, "Real-time programmable low-power SC bandpass filter," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 499-506, June 1982.
[29] K. C. Hsieh, P. R. Gray, D. Senderowicz, and D. Messerschmitt, "A low-noise differential chopper stabilized switched capacitor filtering technique," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 708-715, Dec. 1981.
[30] D. Senderowicz, S. F. Dreyer, J. M. Huggins, C. F. Rahim, and C. A. Laber, "Differential NMOS analog building blocks for PCM telephony," in *Dig. Tech. Papers, 1982 Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 1982. Also appears in full length form in this issue, pp. 1014-1023.
[31] J. Guinea and D. Senderowicz, "High frequency NMOS switched capacitor filters using positive feedback techniques," this issue, pp. 1029-1038.
[32] S. Wong and C. A. T. Salama, "Scaling of MOS analog circuits for VLSI applications," in *Dig. Tech. Papers, 1982 Symp. VLSI Technology*, Tokyo, Japan, Sept. 1982.

Additional References on MOS Operational Amplifiers

[33] F. H. Musa and R. C. Huntington, "A CMOS monolithic 3½ digit A/D converter," in *Dig. Tech. Papers, 1976 Int. Solid-State Circuits Conf.*, Philadelphia, PA, Feb. 1976, pp. 144-145.
[34] A. G. F. Dingwall and B. D. Rosenthall, "Low-power monolithic COS/MOS dual-slope 11-bit A/D converters," in *Dig. Tech. Papers, 1976 Int. Solid-State Circuits Conf.*, Philadelphia, PA, Feb. 1976, pp. 146-147.
[35] Y. P. Tsividis and D. Fraser, "A process insensitive NMOS operational amplifier," in *Dig. Tech. Papers, 1979 Int. Solid-State Circuits Conf.*, Philadelphia, PA, Feb. 1979, pp. 188-189.
[36] S. Kelley and D. Ulmer, "A single-chip PCM codec," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 54-58, Feb. 1979.
[37] B. J. White, G. M. Jacobs, and G. Landsburg, "Monolithic dual tone multifrequency receiver," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 991-997, Dec. 1979.
[38] I. A. Young, "A high performance all-enhancement NMOS operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 1070-1076, Dec. 1979.

A Programmable CMOS Dual Channel Interface Processor for Telecommunications Applications

BHUPENDRA K. AHUJA, MEMBER, IEEE, PAUL R. GRAY, FELLOW, IEEE, WAYNE M. BAXTER, AND GREGORY T. UEHARA, MEMBER, IEEE

Abstract—**A CMOS analog VLSI chip for telecommunication applications has been designed with many desirable line card features, which are programmable through a unique digital interface from the central switching office. The paper emphasizes the circuit innovations of some key analog functions realized on the chip, specifically, the operational amplifier family, the precision bandgap reference circuit, and the line balancing function. The die size of the analog VLSI is approximately 50000 mils^2, and the active power dissipation is 80 mW with a 1 mW standby mode.**

Manuscript received June 15, 1984; revised August 3, 1984.

B. K. Ahuja, W. M. Baxter, and G. T. Uehara are with the Intel Corporation, Chandler, AZ 85224.

P. R. Gray is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

I. INTRODUCTION

THE advent of single-chip codec/filter integrated circuits [1], [2] has greatly reduced the cost of per-line electronics in digital switching and transmission systems. However, one drawback of such chips has been the necessity of adding external components to the line card to perform functions such as line balancing. A second drawback has been the difficulty of adapting the line card to the specific requirements of different system applications. This paper describes a CMOS VLSI interface circuit incorporating all of the low-voltage functions for a subscriber line

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, pp. 892-899, Dec. 1984.

card on the same monolithic chip. The circuit provides programming of all the features through a unique digital interface from the central switching office.

This paper does not attempt to describe each functional block on the chip, but gives an overview of the system architecture followed by details on the design of three analog blocks which incorporate some novel ideas, namely, the operational amplifier family, the precision bandgap reference circuit, and the programmable balance networks. In conclusion, some system attributes and device performance are also provided.

II. System Architecture

The most important objective in defining the architecture of this analog VLSI was to make all its features programmable through software control from the central switching office. Fig. 1 shows the internal organization of the chip. Along with the basic function of the codec/filter for the voice channel, the chip also provides gain control for transmit and receive directions, balance networks and 2- to 4-wire conversion, three-party conferencing, and a secondary analog channel which may be used for performing loop tests, loop monitor and control, or any other low-bandwidth application. For the transmit voice path, the input signal (VFX) first goes through an antialiasing filter (AAF), followed by a low-noise amplifier (LNA), which can be used to set the transmit gain with external resistors at TG_1 and TG_2. Another programmable gain stage (XPG) is provided in the voice signal path to allow gain control of $+6--6$ dB in 0.5 dB steps under software control. The signal is then band limited to 3.2 kHz in accordance with CCITT requirements by an eighth-order switched-capacitor transmit filter and an auto zero network (XF/AZ) prior to conversion into A or u law PCM words by an all capacitive charge redistribution A/D converter. This encoder is also used to perform another 8 bit PCM conversion for the secondary analog inputs SAI_1 and SAI_2, which can be encoded at an 8 kHz rate in single-ended or differential modes.

On the receive side, the D/A converter performs decoding of three A or u law PCM words every 125 μs. The primary and third-party voice PCM bytes are summed and filtered by a fifth-order switched-capacitor filter (RLPF). A receive programmable gain (RPG) circuit provides gain control of 0 to -12 dB in 0.5 dB steps under software control. Two on-chip power amplifiers (PA's) can drive up to 300 Ω transformer loads. Information decoded from the receive data byte appears at the secondary analog output (SAO), which is capable of driving up to a 10 kΩ load. No on-chip filtering is included for the secondary channels, and thus either of these should be externally filtered, or their applications should be limited to low bandwidths only.

The 2- to 4-wire conversion necessary for the subscriber interface is implemented on-chip. The option of using either internal or external balance networks (BNW's) provides flexibility for any application.

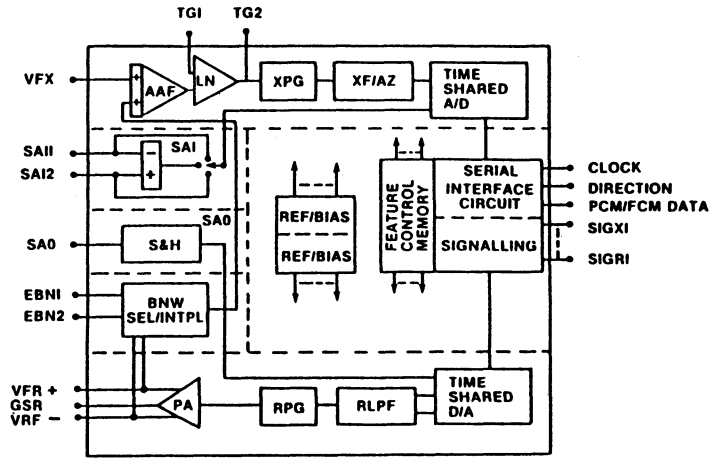


Fig. 1. Block diagram of the chip.

TABLE I
DESIGN OBJECTIVES OF THE OPERATIONAL AMPLIFIER FAMILY

Type	AVMIN	RLMIN	CLMAX	SWING	PDMAX	PSRR @ 100KHZ
Internal	5000	100K	10PF	±4V	1mW	60dB
10k Buffer	5000	10K	10PF	±4V	2mW	50dB
1k Buffer	5000	1K	100PF	±4V	3mW	50dB
300Ω Buffer	5000	300	100PF	±4V	6mW	50dB

Self and system test capabilities for diagnostic purposes have been integrated. A unique bidirectional serial interface to a line card controller chip provides further capabilities to the system designer by allowing features like time slot assignment through a microprocessor or an HDLC port and software control of the analog VLSI processor. This paper will describe the analog VLSI processor only.

III. Operational Amplifier Family

The family of operational amplifiers used in the circuit is summarized in Table I. For internal applications involving purely capacitive loads, a conventional single-ended output two-stage amplifier is used, incorporating the cascode compensation scheme for improved high-frequency supply rejection from the negative supply. This amplifier has been discussed elsewhere [3] and will not be discussed further here.

For applications on the chip requiring the amplifier to drive finite values of load resistance such as resistive programmed internal gain stages and off-chip loads, a composite operational amplifier is used which consists of a common core preamplifier together with one of a family of three class AB output stages of different drive capability. The design of the output buffers was particularly challenging because of the requirements for low quiescent power dissipation together with an output voltage swing to within 1 V of each power supply while driving resistive loads.

Some key requirements of the core preamplifier are that it realize good power supply rejection at high frequencies, that its nondominant poles be at sufficiently high frequencies so that the excess phase budget can be used entirely by the output stage with its capacitive loading, and that its

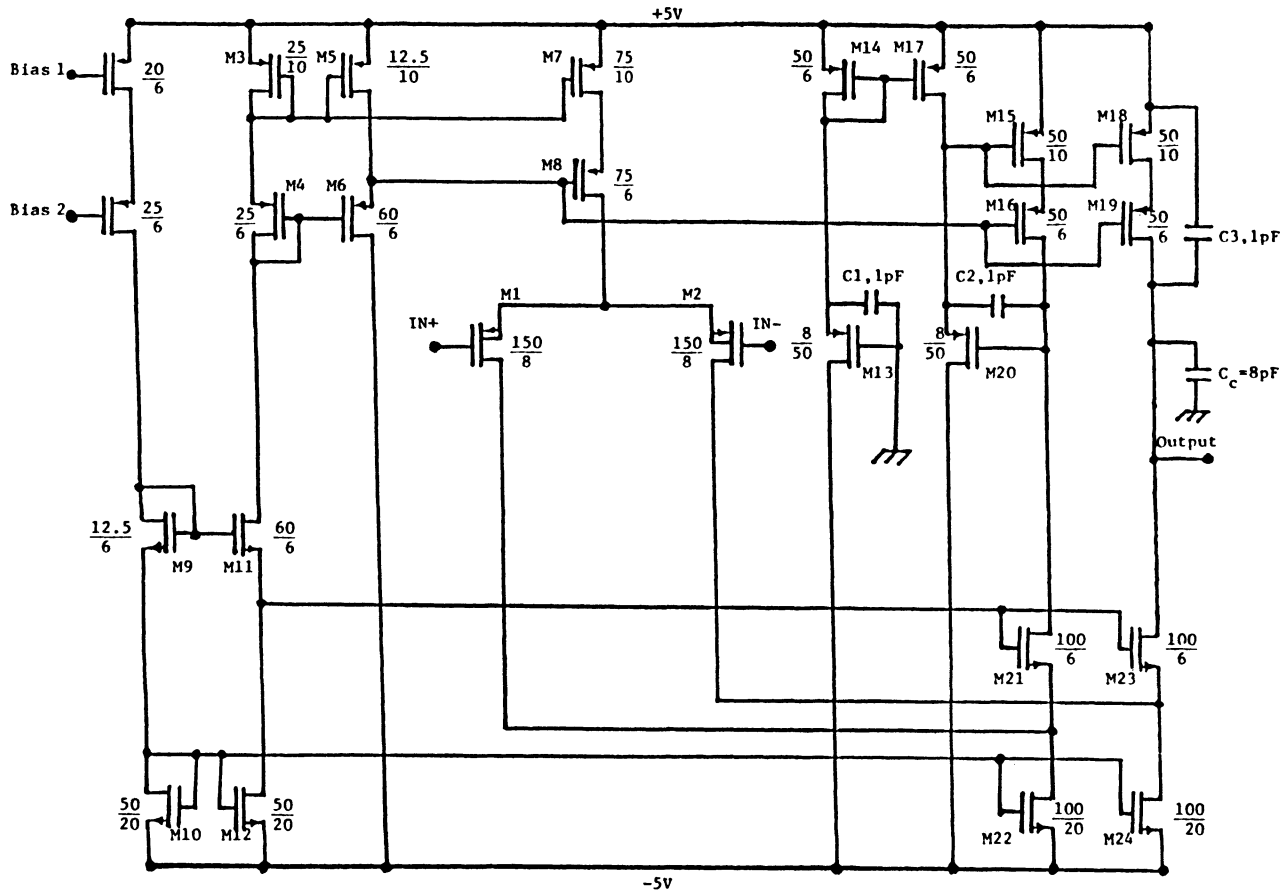


Fig. 2. Single-stage common-source–common-gate amplifier schematic.

power dissipation be kept to a minimum. In order to achieve this, the common-source–common-gate configuration, as shown in Fig. 2, was chosen. In this circuit, transistors M_9, M_{10}, M_{11}, and M_{12} are used to develop bias voltages for M_{21}, M_{23}, M_{22}, and M_{24} such that the latter two are biased on the edge of the triode region. The same function is provided by $M_3 - M_6$. This allows the output to swing to within 2 V_{dsat} of the positive and negative supplies while maintaining a high incremental voltage gain. At the bias currents used in this circuit, the value of V_{dsat} is approximately 0.2 V. The differential-to-single-ended conversion is performed by the current source $M_{13} - M_{14} - M_{17}$ in conjunction with the follower M_{20}. Since the gate of M_{13} is returned to ground, the resultant V_{gs} of M_{20} is the correct value needed to give a nominal zero output voltage and hence a negligible input-referred systematic offset. However, the low transconductance of M_{20} in combination with the gate capacitances of M_{15} and M_{18} result in poor frequency response in the differential-to-single-ended converter. To alleviate this problem, the feedforward capacitor C_2 is included to bypass this signal path at high frequencies. Because high-frequency power supply rejection is a critical requirement, capacitors C_3 and C_1 are also included to balance the displacement current flowing through C_2 when signals are present on the positive power supply.

Compensation of the circuit is achieved with a single capacitor from the output node to ground. The actual value of this capacitance depends on the bandwidth desired in the particular application.

One of the most challenging problems in achieving the performance objectives of the overall chip was the realization of the output stage so as to drive low impedance resistive and capacitive loads to voltages near the power supply while achieving good linearity, good supply rejection, low noise, and low quiescent power dissipation.

The approach taken in the class AB buffer is illustrated in Fig. 3(a). Here, the output swing is limited by the on-resistance of the output transistors which enter the triode region at the extremes of the output swing. Bipolar transistors cannot be used reliably because of the occurence of load faults such as short circuits and the possibility of chip latchup due to the substrate currents. In this configuration, the error amplifiers must meet a number of constraints. First, since the amplifiers will have input-referred offset voltages on the order of several millivolts, relatively modest values of open loop voltage gain on the order of 10 must be used. This ensures that the input offset voltages, when referred t o the output, will not cause gross variations in the quiescent current in the amplifier. This problem can also be attacked by including a crossover circuit [4]. A second constraint is that the amplifiers have an output swing near the supplies so as to provide good drive capability for the output transistor gate and an input common-mode range that extends over the same range of voltages as the output swing of the buffer. Finally, when

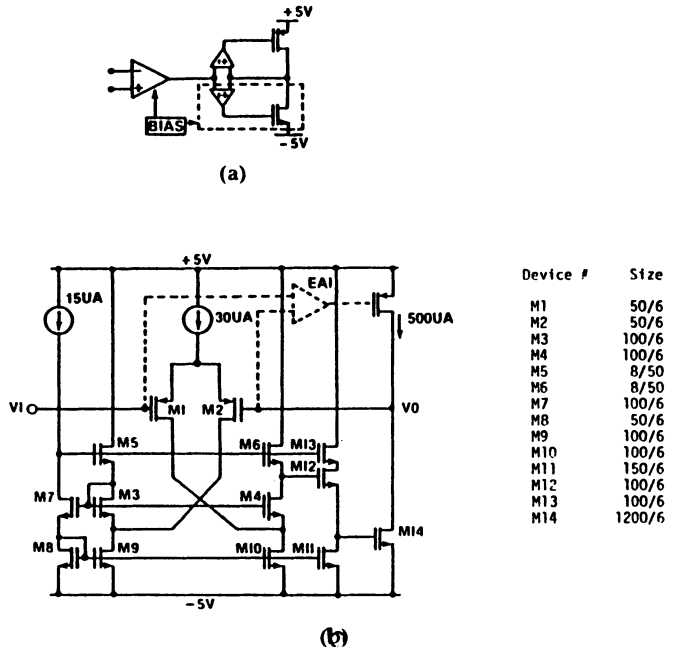


Fig. 3. (a) Complimentary class AB output buffer. (b) Circuit schematic of the class AB output buffer.

the differential dc input voltage is zero, the error amplifiers must provide a quiescent dc output voltage which is the correct value to give the desired quiescent bias current in the output transistors.

The implementation used to satisfy these requirements is illustrated in Fig. 3(b). The device sizes shown are those used for the 300 Ω transformer driver output buffer. Only the error amplifier driving the n-channel output transistor is shown; a dual of this circuit is used to drive the p-channel output transistor. Transistors M_1-M_2 form a source-coupled pair driving enhancement load transistors M_5-M_6 through common-gate transistors M_3-M_4. This combination produces a voltage gain in the error amplifier of about 8. Source follower M_{12} drives the output transistor gate. Transistors M_7, M_8, M_9, M_{10}, and M_{11} provide differential-to-single-ended conversion and produce the desired quiescent voltage at the gate of the output transistor.

The quiescent current in the output transistor can be calculated using the equivalent circuit shown in Fig. 4. Here, it has been assumed that the differential input voltage to the M_1-M_2 pair is zero, so that the drain currents of M_1 and M_2 are each equal to $I_o/2$. From the symmetry of the circuit it is clear that voltages V_1 and V_2 at the drains of M_3 and M_4 are equal. Thus, the voltage V_2 is given by

$$V_1 = V_2 = 2V_T + \sqrt{\frac{2I_o}{\mu C_o\left(\frac{W}{L}\right)_8}} + \sqrt{\frac{2I_o}{\mu C_o\left(\frac{W}{L}\right)_7}} \quad (1)$$

where μ = carrier mobility; C_o = gate oxide capacitance density; V_T = threshold voltage; and $(W/L)_n$ = aspect ratio of device n.

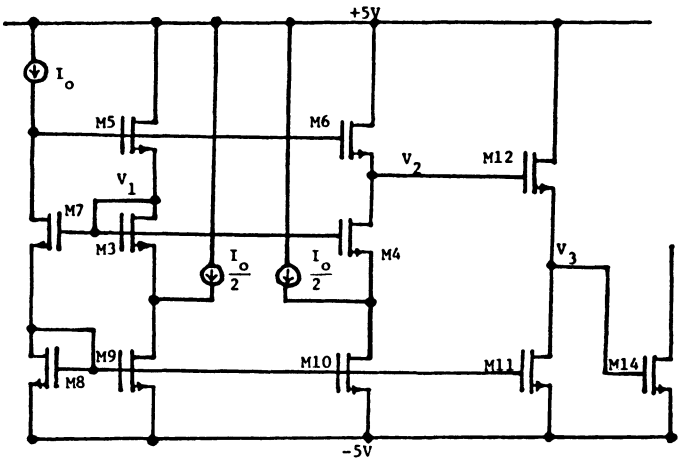


Fig. 4. Equivalent circuit of negative error amplifier for balanced quiescent condition.

The voltage applied to the gate of the output transistor is this voltage minus the gate-source drop of transistor M_{12}. The drain current of M_{12} is equal to the drain current of M_{11}, which is given by

$$I_{D11} = I_{D12} = I_o\frac{(W/L)_{11}}{(W/L)_8}. \quad (2)$$

Thus, the gate-source voltage of the output transistor is given by

$$V_{GS14} = V_T + \sqrt{\frac{2I_o}{\mu C_o}}\left(\frac{1}{\sqrt{(W/L)_8}} + \frac{1}{\sqrt{(W/L)_7}} - \sqrt{\frac{(W/L)_{11}}{(W/L)_{12}(W/L)_8}}\right). \quad (3)$$

Here it has been assumed that body effect in the various transistors can be neglected, and the threshold voltages are equal. Because the sources of M_7, M_4, and M_{12} are at almost the same potential, this assumption produces little error. Finally, using the preceding two equations, the quiescent current in the output transistor is given by

$$I_{\text{out}} = I_o\left(\frac{W}{L}\right)_{14}\left[\frac{1}{\sqrt{(W/L)_8}} + \frac{1}{\sqrt{(W/L)_7}} - \sqrt{\frac{(W/L)_{11}}{(W/L)_{12}(W/L)_8}}\right]. \quad (4)$$

The circuit produces a quiescent current which is directly proportional to the bias current. Furthermore, it is possible to adjust the W/L ratio of M_{11} and M_{12} so as to reduce the quiescent current in the output transistor to any desired value. In the particular case of the circuit described here, the bias current in the output transistor is set at about $10I_o$. This value of approximately 300 μA is a compromise between power dissipation and bandwidth. This bias point corresponds to a typical V_{dsat} of 150 mV in the output transistor. Thus, with a gain of 8 in the error amplifiers, an

TABLE II
300 Ω BUFFER AMPLIFIER EXPERIMENTAL RESULTS @ +5 V, 40°C

Open Loop Gain	5500
Voltage Swing	±4.1V @ 300Ω Load
Quiscent Power	6mW
VCC PSRR @ 1 kHz	−74dB
@ 100 kHz	−44dB
VBB PSRR @ 1 kHz	−62dB
@ 100 kHz	−55dB
Input Noise Density	
@ 1kHz	95 nV/√Hz
Active Area For	
300Ω Buffer	628 Sq. Mil.
1 k Buffer	563 Sq. Mil.
10 k Buffer	372 Sq. Mil.

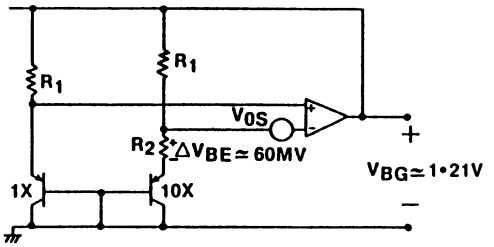


Fig. 5. Basic bandgap reference circuit implementation.

input-referred offset voltage of 3 mV would result in an output-referred offset of 24 mV, which would give shifts in the quiescent current on the order of 30 percent.

The maximum gate drive to the negative output transistor results when enough differential drive is applied to the error amplifier to turn off transistor M_2. The combination of the bias current I_o and the W/L of M_5 are chosen such that when this occurs the gate of M_5 assumes a potential near the positive supply. In this mode, the gate voltage of M_{14} is approximately equal to the supply voltage minus two threshold voltages. Alternatively, by choosing a smaller value of quiescent V_{gs} for M_5, the circuit can be operated in a current-limiting mode in which the maximum gate drive on the output transistor is programmed by the value of I_o in conjunction with the W/L of M_5.

Table II summarizes the measured performance of the 300 Ω power amplifier.

IV. Precision Bandgap Reference Circuit

A basic CMOS bandgap reference circuit is shown in Fig. 5 [5], [6]. Here the output voltage is

$$V_{BG} = V_{BE} + (\Delta V_{BE} + V_{OS}) \cdot \left(1 + \frac{R_1}{R_2}\right) \qquad (5)$$

where V_{OS} is the amplifier offset voltage. However, this approach has two basic disadvantages. First, the amplifier offset voltage adds directly to the difference in base–emitter voltages ΔV_{BE} of the bipolar transistors. Offset voltages of typical CMOS operational amplifiers range between ±15 mV and, when amplified by the resistor ratio gain factor, lead to a large variation in the reference voltage. This increases the reference voltage trimming requirements for a desired precision. Second, the offset voltage of a CMOS op amp drifts with time and has a temperature coefficient of

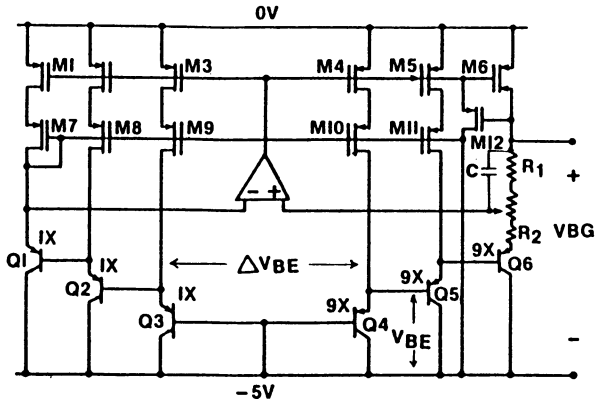


Fig. 6. Improved CMOS bandgap reference voltage circuit schematic.

around 20 μV/°C. These variations are amplified and degrade the reference stability and performance.

One possible solution is to use chopper stabilization to null out the op amp offset voltage [7]. This is effective for a reference voltage valid only during a portion of the clock period. However, the analog signal paths of this VLSI processor require a continuous and stable reference to generate various bias voltages rendering this solution unsuitable.

Another approach is to reduce the relative effect of the offset voltage by increasing the contribution of the bipolar transistor base–emitter voltages. By using an area-ratioed stack of three closely matched bipolar transistors, the circuit, shown in Fig. 6, produces a basic reference voltage which is three times the silicon bandgap voltage. This reduces the effect of the offset by a factor of 3. The bandgap voltage is given by

$$V_{BG} = 3V_{BE} + (3\Delta V_{BE} + V_{OS}) \cdot \left(1 + \frac{R_1}{R_2}\right). \qquad (6)$$

Transistors $M_1 - M_6$ are matched current sources, each of which forces a current equal to $3\Delta V_{BE}/R_2$ into each bipolar transistor. The transistors $M_7 - M_{11}$ drop the necessary voltage required to match the currents in $M_1 - M_6$ to within 0.5 percent. The resulting output voltage V_{BG} is 3.8 V.

In this approach, there are both positive and negative feedback paths around the amplifier. The negative feedback path must remain dominant at all frequencies to avoid instability. To ensure this, a capacitor C shunts the resistor R_1, increasing negative feedback at high frequency with feedforward action. The resistors R_1 and R_2 are made of p diffusion and $Q_1 - Q_6$ are vertical p-n-p transistors with a common collector.

Besides the normal desired operating state, this circuit also has a degenerate state where all the devices are off. Transistors M_{12} is used to start up the feedback loop and turns off in normal operation.

The bandgap voltage is referenced to the negative supply. The primary reason for this is to avoid collector-base reverse bias modulation of the emitter currents, which would degrade the reference power supply rejection. How-

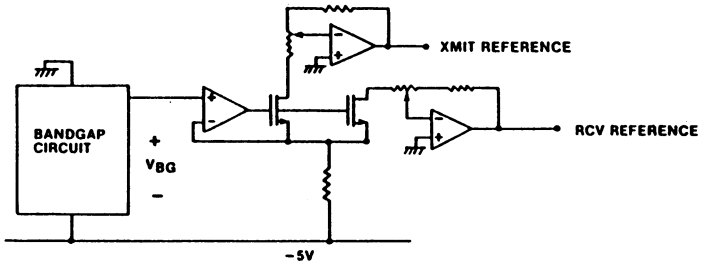


Fig. 7. Reference level shifting and gain adjust circuitry.

TABLE III
PERFORMANCE OF THE REFERENCE VOLTAGE CIRCUIT

Parameter Measured Data	
Temp. Coeff. (0–70°C)	< 100 ppm/°C
Power Supply Rejection	
+5 V @ 1 kHz	−75 dB
+5 V @ 100 kHz	−36 dB
−5 V @ 1 kHz	−70 dB
−5 V @ 100 kHz	−37 dB
Power Dissipation	8 mW
Active Area	3000 mils^2

ever, the A/D and D/A converters need their reference voltages with respect to ground. Fig. 7 shows the circuits which provide the level shifting and gain adjust.

The reference voltage must be trimmed to have a zero temperature coefficient at room temperature. Because the relative offset voltage has been reduced by a factor of 3, only 4 bits are required to trim the bandgap voltage under all process variations. Also, to account for manufacturing variations, 3 bits in the level shifter and gain adjust circuit trim the transmit and receive gains to within ±1 percent accuracy. The complete trim procedure involves: 1) measurement of the amplifier offset voltage; 2) measurement and adjustment of the bandgap voltage V_{BG} and trimming it to three times the silicon bandgap voltage plus the amplifier offset voltage with the 4 bits of trim provided; and 3) adjustment of the transmit and receive reference voltages to nominally 3.2 V with the additional 3 bits of trim.

Table III presents the measured performance of the bandgap reference circuit.

V. Line Balancing Function

In a typical telecommunications line card design, the subscriber interface uses 2-wire transmission while the switching and transmission equipment employs 4-wire. The circuit which performs the 2- to 4-wire conversion requires a balance network to improve the return loss of the receive channel into the transmit channel. This line card function usually is implemented with discrete components and only recently has it been integrated into low-voltage MOS technology [8], [9]. This analog VLSI processor provides the system designer with two different options for implementing the 2- to 4-wire conversion and line balancing on the same chip. The first option provides three of the most commonly used balance networks, which

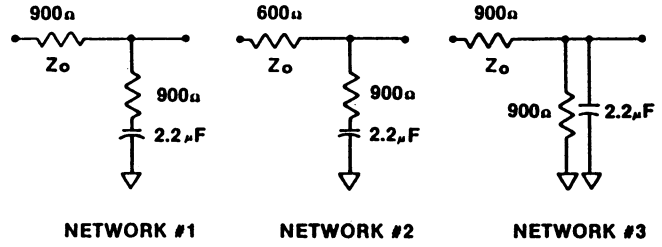


Fig. 8. Equivalent circuits of the three internal balance networks.

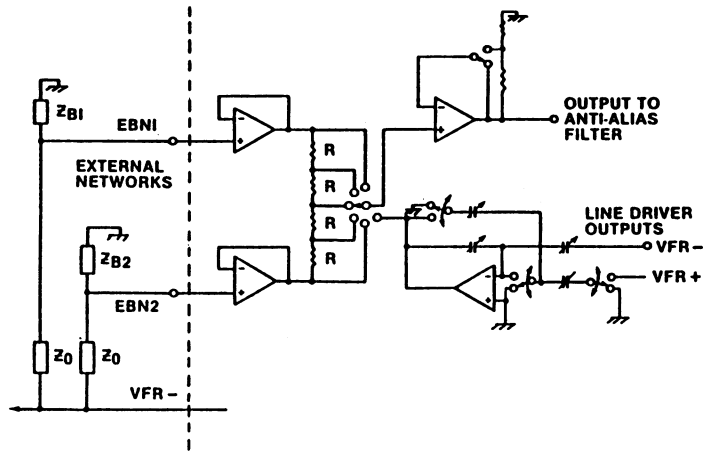


Fig. 9. Programmable balance network.

are shown in Fig. 8. These three networks use bilinear z-transform techniques and are implemented with one amplifier and a programmable switched-capacitor array clocked at 32 kHz. The user can select any of these three networks under software control.

However, there may be different system requirements calling for a special balance network. The second option allows the user to connect two different off-chip balance networks and select either of these or provide an interpolation between these two frequency characteristics under software control [10]. By connecting these networks to the analog VLSI, the user can design a compromise balance network transfer characteristic for different loop lengths by programming the interpolation coefficient a to be 0, 0.25, 0.5, 0.75, or 1.0. The resulting frequency characteristic will be

$$H(f) = aH_1(f) + (1-a)H_2(f) \tag{7}$$

where

H_1 = the transfer function of external Network 1

H_2 = the transfer function of external Network 2.

This will cover many possible loop lengths, improving and easing the line balancing function in the line card. Also, because of different interface requirements of either transformer based hybrids or the new generation of integrated SLIC's, a programmable gain of 0 or 6 dB is provided in the path of the return signal.

Fig. 9 shows the programmable internal balance networks and how the two external balance networks are

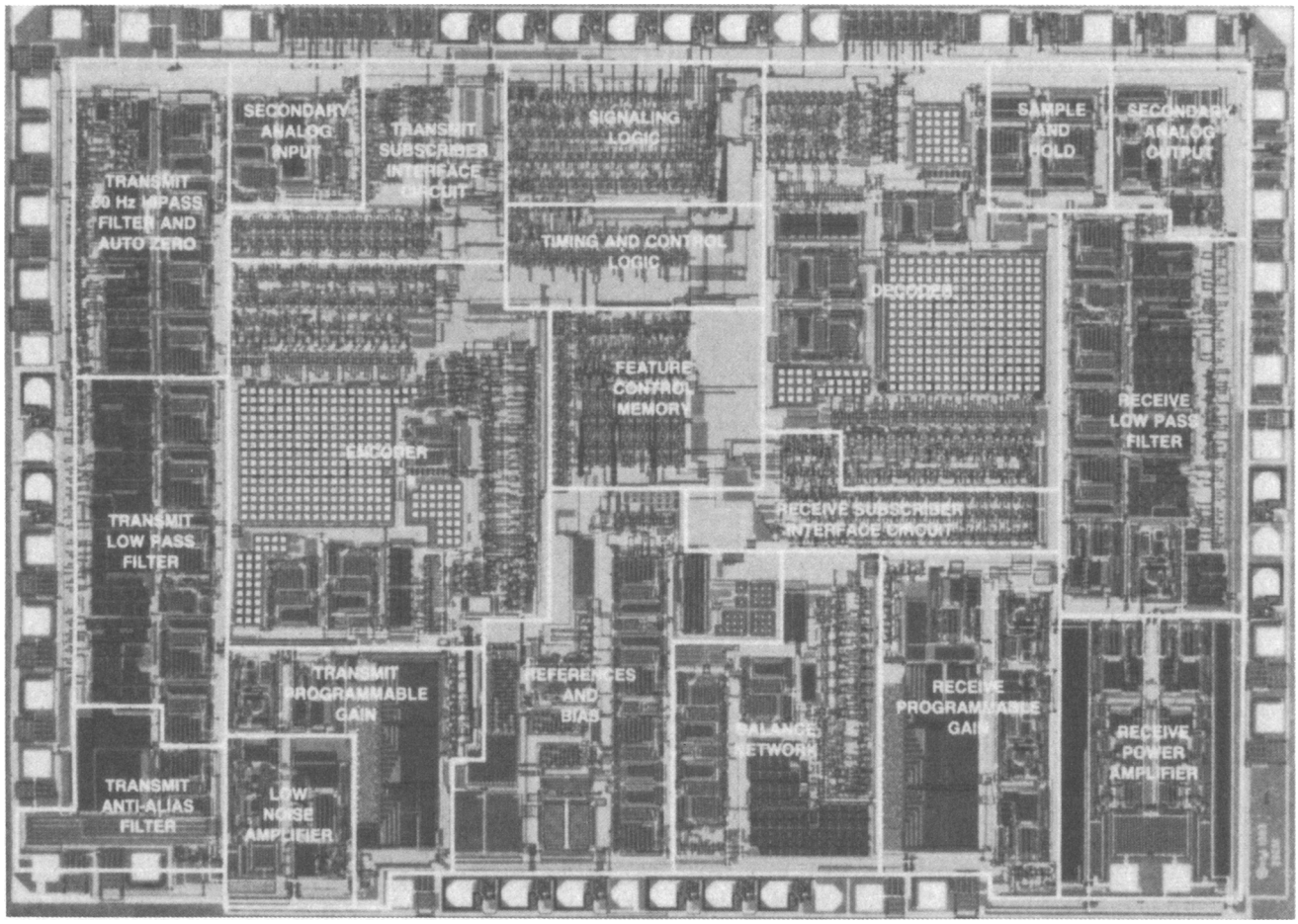


Fig. 10. Photomicrograph of the chip.

TABLE IV
SUMMARY OF THE CHIP PERFORMANCE

Parameter	Transmit Channel	Receive Channel
Gain Tracking @ 1020 Hz		
+3- -40 dBmO	> -0.12, < +0.05 dB	< +0.10 dB
-40- -55 dBmO	> -0.30, < +0.50 dB	< +0.20 dB
Sig./Dist. (C-msg) @ 1020 Hz		
+3- -30 dBmO	> 37 dB	> 37 dB
-30- -40 dBmO	> 31 dB	> 33 dB
-45 dBmO	28 dB	28 dB
Idle Channel Noise	15 dBrnCo	4 dBrnCo
Filter Frequency Response		
300-3000 Hz	> -0.01, < +0.08 dB	< +0.05 dB
Above 4600 Hz	< -35 dB	< -48 dB
Power Supply Rejection		
+5 V @ 1 kHz	-69 dB	-50 dB
+5 V @ 50 kHz	-60 dB	-28 dB
-5 V @ 1 kHz	-45 dB	-51 dB
-5 V @ 50 kHz	-40 dB	-28 dB

buffered by op amps. A resistive network and switches provide the selection and interpolation features.

VI. System Attributes/Performance

The photomicrograph of the chip is shown in Fig. 10. The die size is approximately 50000 mils^2, and it dissipates 80 mW active power with a 1 mW standby mode. A power trim circuit on the chip can adjust the power dissipation to within 10 percent of these typical numbers. The programmable features are controlled by 30 bits. The chip is fabricated in a 4 μm n-well CHMOS process and is packaged in either a 28 or 22 pin DIP. Table IV summarizes the overall chip performance.

VII. Conclusions

A CMOS VLSI analog/digital interface circuit has been described. Its unique system architecture provides much higher level integration of the line card functions on a single chip than previously reported. The analog VLSI employs several novel circuit design techniques to achieve superior performance.

References

[1] D. Senderowicz, S. F. Dreyer, J. M. Huggins, C. F. Rahim, and C. A. Laber, "A family of differential NMOS analog circuits for a PCM CODEC," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1014-1023, Dec. 1982.

[2] B. K. Ahuja, M. R. Dwarkanath, T. E. Seidel, D. G. Marsh, "A single chip CMOS CODEC with filters," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 1980, pp. 242-243.

[3] B. K. Ahuja, "An improved frequency compensation technique for CMOS operational amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 629-633, Dec. 1983.

[4] K. E. Brehmer and J. B. Weiser, "Large swing CMOS power amplifier," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 624-629, Dec. 1983.

[5] K. E. Kujik, "A precision reference voltage source," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 222–226, June 1973.
[6] R. J. Widlar, "New developments in IC voltage regulators," *IEEE J. Solid-State Circuits*, vol. SC-6, pp. 2–7, Feb. 1971.
[7] B. S. Song and P. R. Gray, "A precision curvature compensated CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 634–643, Dec. 1983.
[8] M. Foster, H. El Sissi, V. Korsky, K. Silmens, R. Wallace, and W. Sin, "A monolithic NMOS filter and line balancing chip," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 1980, pp. 182–183.
[9] R. Apfel, H. Ibrahim, and R. Ruebush, "Signal-processing chips enrich telephone line-card architecture," *Electronics*, pp. 113–118, May 5, 1982.
[10] A. De la Plaza, "Hybrid with automatic selection of balance networks," in *Proc. Int. Symp. Circuits Syst.*, Apr. 1981, pp. 725–728.

A Precision CMOS Bandgap Reference

JOHN MICHEJDA AND SUK K. KIM

Abstract — This paper describes the design of a precision on-chip bandgap voltage reference for applications with CMOS analog circuits. The circuit uses naturally occurring vertical n-p-n bipolar transistors as reference diodes. P-tub diffusions are used as temperature-dependent resistors to provide current bias, and an op-amp is used for voltage gain. The circuit is simple. Only two reference diodes, three p-tub resistors, and one op-amp are necessary to produce a reference with fixed voltage of −1.3 V. An additional op-amp with two p-tub resistors will adjust the output to any desired value.

The criteria for temperature compensation are presented and show that the properly compensated circuit can *in principle* produce thermal drift which is less than 10 ppm/°C. Process sensitivity analysis shows that in practical applications it is possible to control the output to better than 2 percent, while keeping thermal drift below 40 ppm/°C. Test circuits have been designed and fabricated. The output voltage produced was −1.30± 0.025 V with thermal drift less than 7 mV from 0°C to 125°C. Significant improvements in performance, at modest cost in circuit complexity, can be achieved if the op-amp offset contribution to the output voltage is reduced or eliminated.

I. Introduction

In a large and complex LSI-CMOS analog circuit, the voltage reference is often a potentially most troublesome component since it must produce a temperature stable, process invariant, and precisely controlled output. In the past, most of efforts in voltage reference design have emphasized temperature compensation at the expense of output precision. The commonly used references based on the difference between gate/source voltages of enhancement and depletion mode MOS transistors realize low thermal drift; however, the absolute magnitude of output is poorly controlled because it depends on the accuracy of depletion and enhancement implants [1]. In the bandgap references, where the output is derived from the voltage difference of two diodes forward biased by ratioed currents, both the thermal drift and the absolute value of the output can be controlled with precision [2]–[4].

A CMOS bandgap voltage reference which uses bipolar-like source-to-drain transfer characteristics of MOS transistor in weak inversion was reported [5], [6]. The output voltage exhibited relatively low thermal drift and tight voltage spread from sample to sample. Another approach [7] used precision curvature-compensated switched capacitor CMOS bandgap reference. It required trimming and used a complex circuitry for generation of bias currents, thus consuming a large area.

Manuscript received August 12, 1983; revised July 13, 1984.

J. Michejda is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

S. K. Kim is with Solid State Electronics Division, Honeywell, MN 55441.

This paper describes the design of another simple bandgap circuit that can be conveniently implemented in CMOS technology. The output of this circuit is both temperature stable and precise. The circuit configuration which follows that given by Kuijk [8] uses temperature dependent p-tub resistors to provide bias currents to the reference diodes, which are the emitter–base junctions of the bipolar transistors formed by the n^+ diffusion inside the p-tub. First, the basic circuit and the criteria for temperature compensation are presented, followed by discussion of the characteristics of the reference diodes and biasing resistors. The sensitivity of the output voltage to the most common process variation, and the power supply fluctuations will then be discussed in detail. Finally, the experimental results to verify circuit performance will be presented to illustrate the precision and the stability of the circuit.

II. The Bandgap Circuit

In the bandgap circuit the output voltage is derived from the voltage difference across two identical diodes forward-biased by two unequal, precisely ratioed currents. The positive temperature coefficient of this difference is then cancelled by the negative temperature coefficient of voltage across one of the diodes. If the voltage across diode 1 is $V_1(T)$ and across diode 2 is $V_2(T)$, then the output can be expressed as

$$\begin{aligned} V_{\text{out}} &= A(V_1(T)-V_2(T))+B(V_2(T)) \\ &= aV_1(T)-bV_2(T) \end{aligned} \tag{1}$$

where constants a and b are chosen to obtain a voltage V_{out} that has a minimum variation over the temperature range of interest.

The principles of bandgap references and the criteria for derivation of constants a and b are given in detail in [4]. These are derived for devices biased by either temperature independent constant current I, or currents which vary with temperature as T^{α}, where α is a constant. Neither one of these temperature variations of currents can be easily implemented in a CMOS analog circuit.

The bandgap circuit configuration used to derive the function in (1) is illustrated in Fig. 1. The first op-amp with transistors $Q1$ and $Q2$, and resistors $R1$, $R2$, and $R3$ is the bandgap circuit which produces a fixed voltage $V_{\text{out}} = -1.3$ V. The second op-amp with resistors $R4$ and $R5$ is a gain stage to adjust the output V_{ref} to a desired value. The discrete version of this bandgap circuit was first

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, pp. 1014–1021, Dec. 1984.

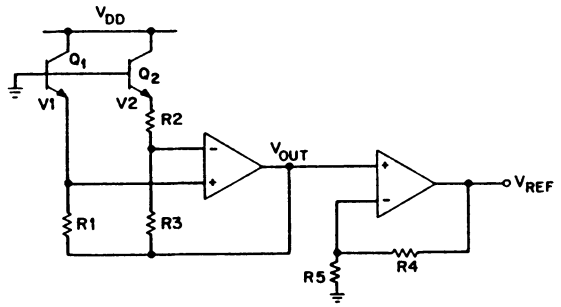


Fig. 1. Schematic of the CMOS bandgap circuit to produce negative output voltage. A bandgap source developing positive output with respect to ground is illustrated in [9].

proposed by Kuijk [8], who used integrated diode pairs and thin film resistors. More recently, Ye and Tsividis [9] have demonstrated this configuration, and a configuration producing a positive output voltage, using vertical n-p-n bipolar transistors and discrete external resistors. They also suggested using diffusion or polysilicon resistors in a fully integrated version of this circuit.

The gain constants a and b of circuit in Fig. 1 are given by

$$a = R_3/R_2 + 1 \tag{2}$$

$$b = R_3/R_2. \tag{3}$$

In addition to the gain of the circuit, which is determined by the ratio of resistors R_3 and R_2, the magnitude and the ratio of the bias currents is determined by the ratio of resistors R_1 and R_3:

$$I_1(T) = \frac{V_{out} - V_1(T)}{R_1} \tag{4}$$

$$I_2(T) = \frac{V_{out} - V_1(T)}{R_3} = \frac{V_{out} - V_2(T)}{R_3 + R_2}. \tag{5}$$

In this implementation, however, the biasing currents $I_1(T)$ and $I_2(T)$ are temperature dependent because of the variation of $V_1(T)$ and $V_2(T)$, and to a lesser extent because of the variation of V_{out} with respect to temperature. The conditions for the temperature compensation of this circuit with temperature independent R_1, R_2, and R_3 are given in [8].

In CMOS technology the situation is even further complicated because the only on-chip conductors that have large enough resistance values for proper biasing of the reference diodes are also temperature-dependent. The p-tub resistors approximately double their resistance for a temperature increase from 0 to 100°C.

The detailed derivation of the temperature compensation for the circuit illustrated in Fig. 1 is algebraically tedius and is briefly summarized in this section.

For a reference diode whose current I is given by Shockley's equation, for $qV \gg nkT$,

$$I = I_0 e^{qV/nkT}. \tag{6}$$

The voltage drop at temperature T is given by

$$V(T) = n\left[V_G(T) + \left(\frac{kT}{q}\right)\ln\left(\frac{I}{AT^{\beta}}\right)\right] \tag{7}$$

where n is the nonideality factor, $V_G(T)$ is the bandgap voltage at temperature T, k is the Boltzmann constant, q is the electron charge, A is a normalizing constant related to the geometry of the device, and β is a constant related to the fabrication process.

$V_G(T)$ is itself a function of temperature. Reference [10] gives the empirical expression for the bandgap value extrapolated from physical measurement over temperature range from 300 to 400 K.

$$V_G(T) = V_{G0} + \frac{dV_G}{dT}T \tag{8}$$

where $V_{G0} = 1.20595$ V, and $dV_G/dT = -2.7325 \times 10^{-4}$ V/K.

For a circuit whose function is given by (1), subject to the condition that the temperature coefficient at temperature $T = T_0$ is zero,

$$\left.\frac{dV_{out}}{dT}\right|_{T=T_0} = 0. \tag{9}$$

The value of the output voltage $V_{out}(T_0)$ is unique and given by

$$V_{out}(T_0) = n\left[V_{G0} + \left(\frac{kT_0}{q}\right) \cdot \left(\beta - 1 + T_0\left(\frac{1}{R(T_0)}\right)\left(\left.\frac{dR}{dT}\right|_{T=T_0}\right)\right)\right] \tag{10}$$

where $dR/dT|_{T=T_0}$ is the derivative of resistance of biasing resistors with respect to temperature at T_0.

The temperature response to the bandgap equation can be described by the following differential equation:

$$T\frac{dV_{out}}{dT} - V_{out} + \frac{nkT^2}{q}\left(\frac{1}{R}\right)\left(\frac{dR}{dT}\right) + \frac{nkT}{q}(\beta - 1) + nV_{g0} = 0. \tag{11}$$

It is important to notice that the value of $V_{out}(T_0)$ as given in (10), and the bandgap temperature response as given in (11) depends only on physical diode parameters n and β, and resistor temperature coefficient $(1/R)(dR/dT)$.

III. The Reference Diode in a CMOS Bandgap Circuit

It is well known that nearly ideal diode characteristics can be obtained from the base–emitter voltage V_{be} of a bipolar transistor. In the twin tub CMOS technology the vertical n-p-n bipolar devices are readily available with n^- substrate collector, p-tub base, and n^+ emitter.

In the layout of the bipolar devices five unit transistors are connected in parallel to make one reference diode. In this approach, similar to the one given in [6], the reference device can operate at larger biasing current, and in addition better matching of references can be achieved.

These devices, each unit transistor with 20 μm$\times$20 μm emitter, manufactured in the 3.5 μm linear twin tub CMOS process, were characterized to obtain the value of parameters n and β necessary to predict the voltage drop across the reference device as given in (7).

The value of n was determined by measuring $I-V$ characteristics of the reference devices at room temperature between currents of 0.5 and 500 μA, and then fitting the measured voltage using (7), with n being the adjustable parameter. The voltage drop of the diode in the fit was normalized to the value at the lowest current. The best fit was obtained for $n=1.01$ over a range of currents from 0.5 to 25 μA. At currents above 25 μA, the differences between measured voltage and the fit were greater than 0.5 mV.

To measure the β parameter directly, the precise $I-V$ characteristics of reference devices over a wide range of temperatures are needed. Such a measurement is tedious and difficult to do precisely, because a minor inaccuracy (as little as 0.5°C) in temperature measurement of the references, can lead to large errors in estimate of β. When these measurements are made on devices placed on a wafer prober, an uncertainty in temperature between the thermocouple in the wafer chuck, and the wafer itself can also cause significant errors in the value of β.

A new indirect method of measurement was used in determining the value β. A precision op-amp with low input offset and high open-loop gain, and a set of precision discrete resistors whose values were individually measured, were externally connected to the bipolar devices on the wafer. The connections were identical to those of the bandgap circuit, and the values of discrete resistors were selected to minimize the output variations with temperature. The voltage produced by the circuit was measured as a function of temperature of the reference diodes. A fit to the data as a function of β was made using the computer simulation of the bandgap circuit. Fig. 2 illustrates the measured and predicted responses of the circuit. The best overall agreement is for $\beta=1.775$.

The simulator used to obtain the best fit was written in Fortran. The program contains appropriate diode models given in (7), and (8), to generate the $I-V-T$ characteristics of the reference diodes. Given diode characteristics, and the resistance values R_1, R_2, and R_3, the program solves iteratively for bias points V_1 and V_2, and output voltage V_{out} until,

$$\frac{V_{out}-V_1(R_1(T),T)}{R_1(T)}=\frac{V_{out}-V_2(R_3(T),T)}{R_2(T)+R_3(T)}\cdot\frac{R_3(T)}{R_1(T)}. \tag{12}$$

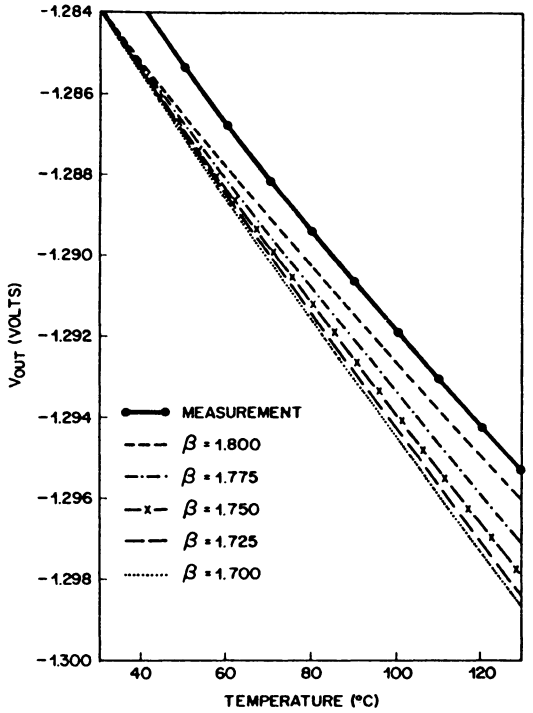


Fig. 2. Comparison of temperature response between simulated and measured fixed resistor bandgap circuit to determine the value of β.

This method of measurement is much less sensitive to vagaries in temperature measurement since the discrete resistors are chosen to minimize temperature dependence of the output voltage. This decreased output voltage sensitivity to temperature enables easier, and more precise determination of β.

IV. Biasing Resistors in the CMOS Bandgap Circuit

The results of the diode characterizations described in the previous section illustrate that the proper operation of reference devices requires small biasing currents in the microampere range. To provide these small currents, resistance values of the order of 10^5 Ω or higher are needed. The only on-chip conductor available in CMOS technology that has sufficiently high sheet resistance to render these resistors practical is the p-tub diffusion. The sheet resistance of the p-tub diffusion in the 3.5 μm twin tub CMOS technology is ~ 3 k$\Omega/\square$. Thus, resistor values up to 0.5 MΩ can be readily realized.

P-tub resistors exhibit temperature dependent behavior due to mobility changes over the temperature range of interest. The measurements of temperature effect on mobility variation of p-type silicon samples [11] yielded mobility dependence of $T^{-2.2}$. The measurements of temperature dependence of p-tub resistance, illustrated in Fig. 3, yielded essentially the same result.

Therefore, for a p-tub resistor,

$$R(T)=R_0T^{2.2} \tag{13}$$

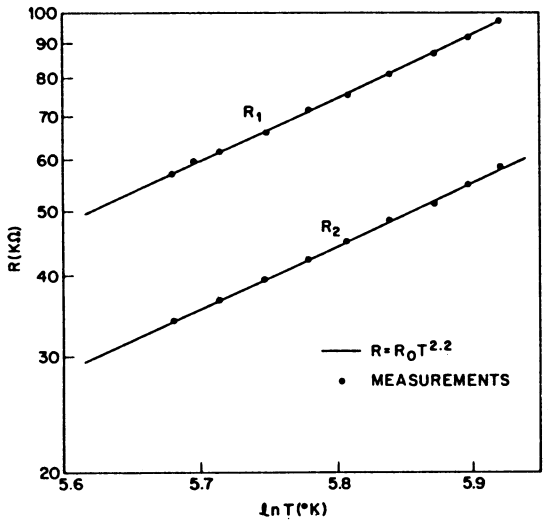


Fig. 3. Temperature characteristic of the p-tub resistors.

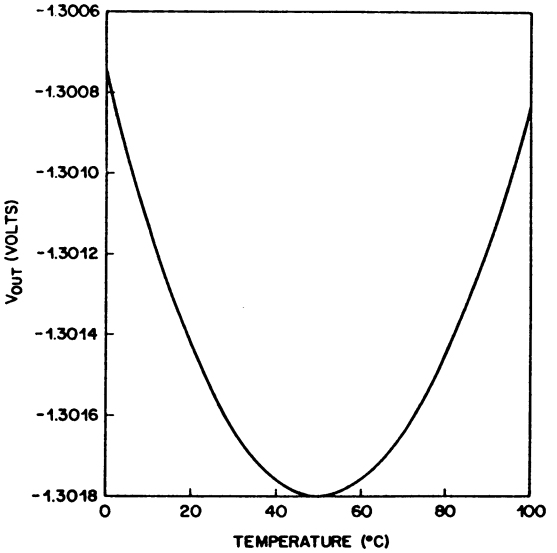


Fig. 4. Predicted temperature response of the bandgap circuit.

where R_0 is the normalizing constant. Therefore,

$$\left(\frac{1}{R}\right)\left(\frac{dR}{dT}\right)=\frac{2.2}{T} \tag{14}$$

and the proper compensation, according to formula (10), occurs when V_{out} at T_0 is

$$V_{out}(T_0)=n\left[V_{G0}+\left(\frac{kT_0}{q}\right)(\beta+1.2)\right]. \tag{15}$$

Fig. 4 illustrates the predicted temperature response of the bandgap circuit from 0 to 100°C, where the resistance of biasing resistors varies as $T^{2.2}$, the value of $\beta=1.775$, and the value of T_0 is 50°C. The value of $V_{out}(T_0)$ is -1.3018 V, and the temperature variation of V_{out} over 100°C is approximately 1 mV. The temperature coefficient that can be obtained using this approach is less than 8 ppm/°C.

V. Process Sensitivity of the Output Voltage

One of the advantages of the bandgap references over the threshold differencing scheme is that both magnitude and temperature compensation of the output voltage are relatively tolerant of processing variations. In this section an approximate analysis of the sensitivity of the output voltage to the processing variants will be presented. The current density ratio between reference diodes is 25:1 yielding voltage difference $(V_1-V_2)\sim 80$ mV at 25°C. The value of constants a and b in (1) is 10 and 9, respectively.

The main processing parameters that affect the output of the bandgap circuit are: p-tub doping, resistor mismatch, reference diode mismatch, and the threshold mismatch of the op-amp input devices. The threshold mismatch, which results in the op-amp offset error, mainly affects the magnitude of the output voltage if the offset itself is not a function of temperature. All other processing variations affect magnitude, as well as the temperature compensation of the output.

The doping of the p-tub affects the resistance of the biasing resistors and the V_{be} drop across the reference diodes. Both the value of the biasing resistors, and the V_{be} drop across the reference devices affect the voltages V_1 and V_2 of the circuit.

The resistance of the p-tub resistor uniformly doped by ion implantation is inversely proportional to the implant dose of boron N_s.

$$R\sim\frac{1}{N_s}. \tag{16}$$

Therefore, for a fractional error in implant dose dN_s/N_s, the fractional error on the p-tub resistance dR/R is;

$$\frac{dR}{R}\sim\frac{-dN_s}{N_s}. \tag{17}$$

For the n-p-n bipolar device biased with fixed V_{be}, the collector current is proportional to the number of impurities/unit area in the base [12] (also known as Gummel number). For the device with ion implanted p-tub base, this number is equal to the ion implant dose N_s. Therefore,

$$I\sim\frac{1}{N_s}e^{qV_{be}/kT} \tag{18}$$

and the V_{be} drop across the reference device is

$$V_{be}\sim\frac{kT}{q}\ln(IN_s). \tag{19}$$

The change in V_{be} of the reference device due to the p-tub ion implant error is

$$dV_{be}\approx\frac{kT}{q}\frac{dN_s}{N_s}. \tag{20}$$

The doping of the p-tub in the twin tub CMOS process can be controlled to ± 10 percent. This variation in the ion

implant dose should result in a $\pm 0.1\ kT$, or $\pm$ 2.5 mV error in V_{be} at room temperature.

From (19) the variation of the voltage across the reference device biased by the p-tub resistor resulting from the changes in V_{be} and the bias current is

$$dV_1 \approx dV_2 \approx \frac{kT}{q}\left[\frac{dN_s}{N_s} + \frac{dI}{I}\right]. \tag{21}$$

For bias current I in the circuit,

$$I = \frac{V_{\text{out}} - V_{be}}{R}, \tag{22}$$

the change in current dI is

$$dI = \frac{dV_{\text{out}} - dV_{be}}{R} - \frac{(V_{\text{out}} - V_{be})\,dR}{R^2} \tag{23}$$

for 10% resistance variation, the first term in (23) is small. Therefore,

$$\frac{dI}{I} \approx -\frac{dR}{R} \tag{24}$$

and using (17), (21), and (24)

$$dV_1 \approx dV_2 \approx \frac{2kT}{q}\left(\frac{dN_s}{N_s}\right). \tag{25}$$

The bandgap output voltage change is then

$$dV_{\text{out1}} = adV_1 - bdV_2 \approx (a-b)\,dV_1 \approx dV_1. \tag{26}$$

The total variation of the output voltage due to variation in the p-tub doping is

$$dV_{\text{out1}} \approx \frac{2kT}{q}\left(\frac{dN_s}{N_s}\right). \tag{27}$$

The ± 10 percent error in the p-tub implant will result in ± 5 mV error in the output voltage at room temperature. This is significantly better than the accuracy of the output voltage produced by the threshold differencing circuit due to the ion implant variation.

The V_{be} voltage produced by the reference devices is very uniform, and the mismatch of the voltage across different devices of the same size on the same chip is small. The measurements of the reference devices biased with constant current on the different chip sites of the same wafer yielded the maximum mismatch of less than 0.2 mV. Such a mismatch would result in a output voltage change of

$$dV_{\text{out2}} \approx adV_1 \approx bdV_2 \approx 2\text{ mV} \tag{28}$$

where the value of $a = 10$ is used.

The mismatch of the resistor ratio R_3 and R_2 affects the gain of the circuit. The mismatch of the ratio R_1 and R_3 influences the current ratio supplied to the references, and thus the difference between V_1 and V_2.

The resistances R_1 and R_3 can be matched accurately because they can be ratioed by an exact integer factor, and also because R_1 and R_3 have an identical voltage across them thus eliminating problems due to any nonlinearities of p-tub resistance. The mismatch of R_2 and R_3 is more likely to occur since these resistors are ratioed by a noninteger number, and because R_2 is biased at a different potential from the substrate than R_1 and R_2. Therefore, only R_3/R_2 mismatch will be considered here.

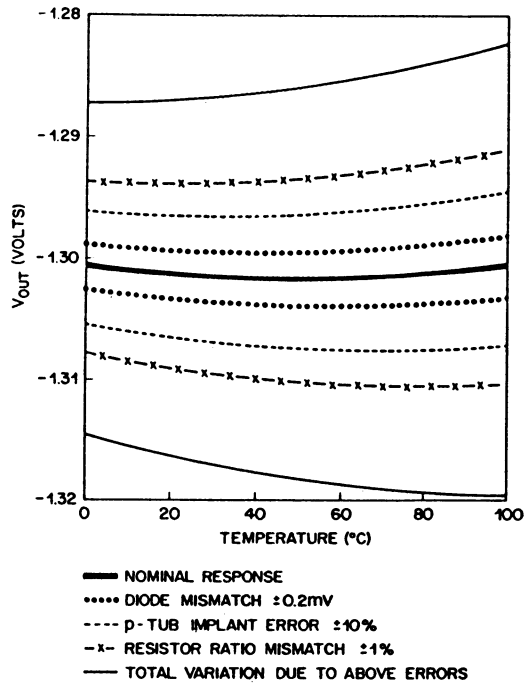


Fig. 5. Sensitivity of output voltage to processing variations.

The variation of the output voltage of the bandgap circuit due to mismatch of R_3, and R_2 is

$$dV_{\text{out3}} = d\left(\frac{R_3}{R_2}+1\right)V_1 - d\left(\frac{R_3}{R_2}\right)V_2 = d\left(\frac{R_3}{R_2}\right)(V_1 - V_2). \tag{29}$$

In a careful layout the resistance values of the p-tub resistors can be matched better than 1 percent. Therefore, for 1 percent resistance mismatching and 80 mV difference between V_1 and V_2, with $R_3/R_2 = 10$ the output voltage error is

$$dV_{\text{out3}} = 0.01\left(\frac{R_3}{R_2}\right)(V_1 - V_2) \approx 8\text{ mV}. \tag{30}$$

The above discussion illustrated that excluding the offset error of the op-amp the various processing variations can affect the output voltage of the bandgap circuit by about ± 15 mV at room temperature in the worst-case analysis. This is only about 1.2 percent of the total voltage produced by the bandgap circuit. The estimate of the effect of these parameters on temperature compensation is considerably more difficult and was done using the numerical bandgap simulator discussed in Section III. Fig. 5 illustrates the predicted worst-case behavior of the bandgap circuit and how each processing variation contributes to errors in the output voltage temperature compensation.

In this simulation the output of each reference diode voltage was varied by ±2.5 mV, reference device mismatch was varied by ±0.2 mV, the sheet resistivity of biasing resistors was varied by ±10 percent and the ratio of resistors was mismatched by ±1 percent. The worst-case analysis yields the temperature compensation of the output voltage of about 5 mV over the temperature range from 0 to 100°C.

The input offset error of the summing op-amp can have significant and detrimental effect on the control of the magnitude of the output voltage. An analysis of the bandgap circuit shown in Fig. 1, which includes the offset error of the op-amp, gives the following relation for V_{out}.

$$V_{out} = aV_1 - bV_2 + cV_{os} \tag{31}$$

where a and b are given in (2) and (3), V_{os} is the offset, and

$$c = -\left(1 + \frac{R_3}{R_2}\right) = -a. \tag{32}$$

The offset of the op-amp is therefore multiplied by the factor c. In a typical bandgap circuit, $c \approx 10$; thus, small offset value of the op-amp can contribute a large error to the output voltage.

VI. Sensitivity of Output Voltage to Power Supply Variations

The primary effect of the power supply variation on the output voltage of the bandgap circuit comes from change of the reverse bias on the biasing p-tub resistors R_1, R_2, and R_3. This, in turn changes their resistance values resulting in modified gain factors and bias currents to reference diodes.

In the circuit, resistors R_1 and R_2 are identically biased with respect to the substrate, although they operate at different current densities. R_2 is biased approximately 80 mV more positive than the other two. Therefore, neglecting the influence of bias current, the voltage coefficients of R_1 and R_3 should be identical

$$\frac{1}{R_1}\frac{dR_1}{dV_{sup}} = \frac{1}{R_3}\frac{dR_3}{dV_{sup}}. \tag{33}$$

After tedious algebra, using (1), (2), (3), (4), (5), (7), and (33) one can compute the total variation of the output voltage to be

$$\frac{dV_{out}}{dV_{sup}} = \frac{R_3}{R_2}\left[\frac{1}{R_3}\frac{dR_3}{dV_{sup}} - \frac{1}{R_2}\frac{dR_2}{dV_{sup}}\right](V_1 - V_2)$$
$$\frac{-nkT}{q}\left[\frac{1}{1 + \frac{nkT}{q}\left(\frac{1}{V_{out} - V_1}\right)}\right]\left(\frac{1}{R_1}\right)\left(\frac{dR_1}{dV_{sup}}\right). \tag{34}$$

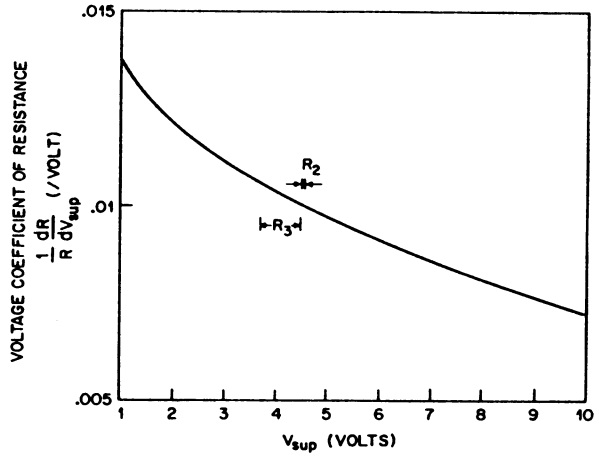


Fig. 6. Voltage coefficient of p-tub resistance as a function of substrate bias (V_{sup}).

The first term in brackets depends on differences between voltage coefficients of resistance between R_1 (or R_3) and R_2 due to different reverse biasing conditions. Fig. 6 illustrates the normalized voltage coefficient of the p-tub resistor as a function of reverse bias. (This curve is shown for an arbitrary current approximating operating point of R_3 resistor. The shape of the curve varies slightly at different currents.) On this curve the bias voltages are marked for resistors R_3 and R_2. Since the difference between the two coefficients of those points is small ($< 10^{-3}$), the first term in (34) is less than 0.8 mV/V.

The second term in (34) results from modified current through reference diodes. Again, a quick computation shows this value to be ~0.2 mV/V. The total expected variation of the output is therefore expected to be about 0.6 mV/V.

VII. Preliminary Results

Fig. 7 shows the photomicrograph of the circuit designed to test the performance of the resistor bandgap circuit. The operational amplifiers used in the circuit are described in [13]. The tester uses a second op-amp with p-tub resistors R_4 and R_5 to adjust the output voltage to a desired value. Because the gain of this stage is determined only by resistance ratio and not by resistance values themselves, it is reasonably precise and independent of temperature. The total circuit size, including the second op-amp and resistors R_4 and R_5, is 0.4 mm^2. The power consumption is 2 mW. In this paper only the results relating to the output of the first op-amp will be discussed. The data are based on measurements obtained from three device lots fabricated in the 3.5 μm linear twin tub CMOS process.

Fig. 8 shows the measured temperature response of one sample circuit along with the predicted response obtained from computer simulation. The output voltage shown has been compensated for the input offset error of the op-amp by measuring the offset contribution and subtracting it from the measured output of the circuit. The amplified

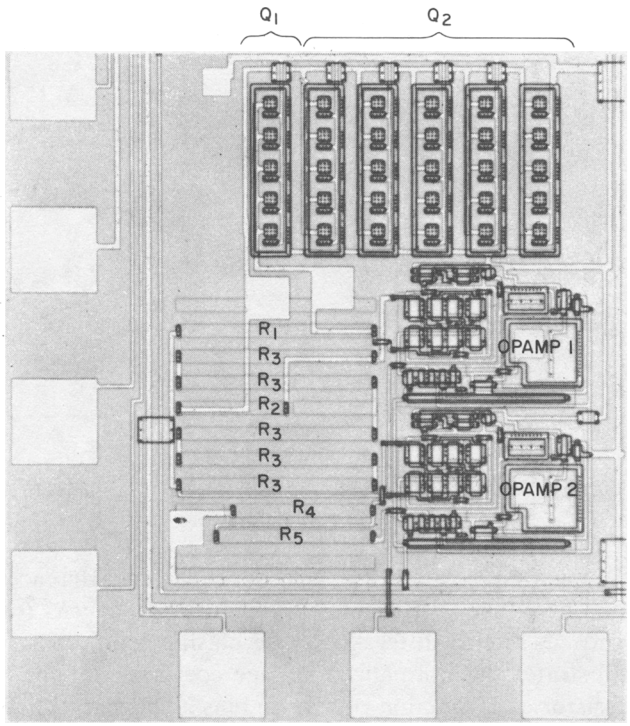


Fig. 7. Photomicrograph of the bandgap test circuit.

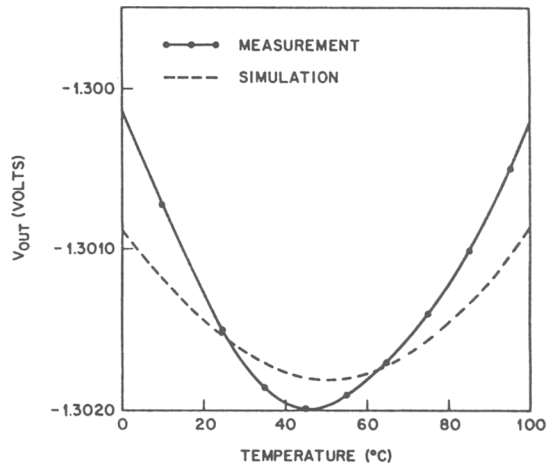


Fig. 8. Measured and predicted temperature responses of the bandgap circuit after compensation for the op-amp offset.

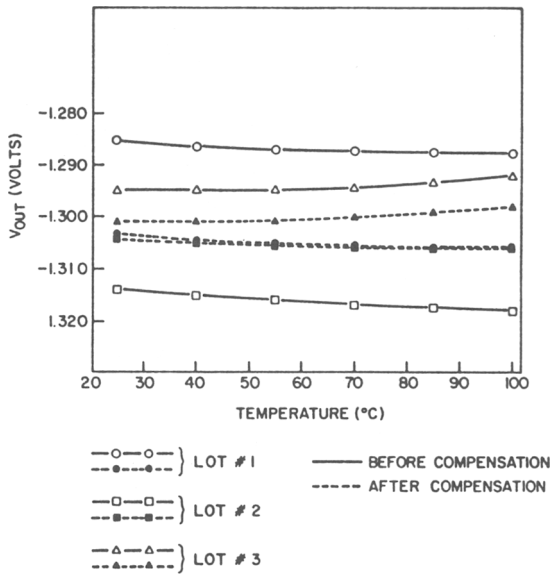


Fig. 9. Measured temperature response of three randomly selected bandgap circuits from three separate wafer lots.

offset contribution was measured by grounding emitters of Q_1 and Q_2 through internal pads and measuring the output. The predicted temperature response and the value of the output voltage agree well with the compensated measured values. The temperature stability of the output is better than 2 mV from 0 to 125°C.

More complete measurements of circuits on different chip sites of a single wafer show that the offset compensated output voltage variation is 3 mV at room temperature. The wafer to wafer variation of the offset compensated output is 7 mV while the temperature stability of the output of most the circuits is better than 5 mV from 25 to 125°C.

As expected, the largest contribution to the output error comes from the input offset of the CMOS summing op-amp. Fig. 9 shows the typical temperature responses of three randomly selected circuits from three separate wafer slices from three wafer lots. For comparison, the output of these circuits with op-amp offset subtracted are also shown. The bulk of the output voltage variation in these samples is due to the offset of the op-amp. More extensive measurements indicate that the output voltage of individual circuits may vary by ± 15 mV due to the offset error alone. In the extreme cases of large offset errors, the offset itself may be temperature-dependent, and may add 2 mV to the temperature instability of the bandgap circuit.

The sensitivity of the output to power supply variations was measured at room temperature with power supply voltage varied from 4.75 to 5.25 V. The output of most of the circuits changed by less than 0.3 mV, the mean was 0.4 mV. The maximum variation, observed on small percentage of samples, was 1.1 mV. The output variation is larger than predicted in Section VI. The differences are likely due to omission in analysis of the influence of bias current on the voltage coefficient of resistance of R_1 and R_3.

The preliminary data suggest that without any offset cancelling technique, the output voltage of the bandgap circuit is $1.30 \pm .025$ V, while the worst-case temperature drift is 7 mV from 0 to 125°C. Dramatic improvement in performance at modest cost in circuit complexity can be achieved if input error contribution is reduced either by cascading reference devices, or by offset cancelling techniques [14].

VIII. Conclusions

The design of a simple and practical precision CMOS bandgap reference circuit which uses p-tub temperature dependent resistors and naturally occurring n-p-n bi-polar transistors is described. The criteria for the proper temperature compensation of the output voltage are derived and

are shown to be independent of the design parameters such as current values and their ratios, resistor values, or diode bias points. The diodes manufactured in the 3.5 μm twin tub linear CMOS process are shown to be acceptable for the references in the bandgap circuit. The magnitude and temperature stability of the output voltage is shown to be tolerant to the most common variations of the CMOS process. The performance of the test circuits matches well the predictions of the bandgap response made by the bandgap computer simulations. The output voltage of the circuit is $-1.30 \pm .025$ V with temperature stability better than 7 mV from 0 to 125°C. A version of this circuit, which produces positive output is shown in [9]. A dramatic improvement in the performance can be achieved if op-amp offset error contribution is reduced by using offset correcting techniques.

ACKNOWLEDGMENT

The authors wish to acknowledge the support offered by H. J. Boll and J. G. Ruch, and to Y. P. Tsividis for many fruitful discussions. They are thankful to P. B. Smalley for performing the testing.

REFERENCES

[1] R. A. Blauschild, P. A. Tucci, R. S. Muller, and R. G. Meyer, "A new NMOS temperature-stable voltage reference," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 767–774, Dec. 1978.

[2] R. J. Widlar, "New developments in IC voltage regulators," *IEEE J. Solid-State Circuits*, vol. SC-6, pp. 2–7, Feb. 1971.

[3] A. P. Brokaw, "A simple three-terminal IC bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 388–393, Dec. 1974.

[4] P. R. Gray and R. G. Meyer, Analysis and Design of Analog Integrated Circuits. New York: Wiley, 1977.

[5] E. Vittoz and O. Neyrund, "A low voltage CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-14, June 1979.

[6] E. Vittoz, "MOS transistors operated in the lateral bipolar mode and their application in CMOS technology," *IEEE J. Solid-State Circuits*, vol. SC-18, June 1983.

[7] B. S. Song and P. R. Gray, "A precision curvature-compensated CMOS bandgap reference," in *ISSCC Dig. Tech. Papers*, vol. 26, Feb. 1983, pp. 240–241.

[8] K. Kuijk, "A precision reference voltage source," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 222–226, June 1973.

[9] R. Ye and Y. Tsividis, "Bandgap voltage reference sources in CMOS technology," *Electron. Lett.*, vol. 18, no. 1, pp. 24–25, Jan. 1982.

[10] Y. P. Tsividis, "Accurate analysis of temperature effects in $I_c - V_{be}$ characteristics with applications to bandgap reference devices," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1076–1084, Dec. 1980.

[11] C. Jacobini *et al.*, "A review of some charge transport properties of silicon," *Solid-State Electron.*, vol. 20, p. 77, 1977.

[12] S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. New York: Wiley, 1981.

[13] V. R. Saari, "Low-power high-drive CMOS operational amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 121–127, Feb. 1983.

[14] K. C. Hsiech and P. R. Gray, "A low-noise chopper-stabilized differential switched-capacitor filtering technique," in *ISSCC Dig. Tech. Papers*, vol. 24, Feb. 1981, pp. 128–129.

A Precision Curvature-Compensated CMOS Bandgap Reference

BANG-SUP SONG, STUDENT MEMBER, IEEE, AND PAUL R. GRAY, FELLOW, IEEE

Abstract — **A precision curvature-compensated switched-capacitor bandgap reference is described which employs a standard digital CMOS process and achieves temperature stability significantly lower than has previously been reported for CMOS circuits. The theoretically achievable temperature coefficient approaches 10 ppm/°C over the commercial temperature range utilizing a straightforward room temperature trim procedure. Experimental data from monolithic prototype samples are presented which are consistent with theoretical predictions. The experimental prototype circuit occupies 3500 mils^2 and dissipates 12 mW with ±5 V power supplies. The proposed reference is believed to be suited for use in monolithic data acquisition systems with resolutions of 10 to 12 bits.**

I. Introduction

An essential element of the analog and digital interface function is a voltage reference to control the scale factor of conversion. The temperature stability of a reference source is a key factor in the accuracy of the overall data acquisition function. Therefore, the ability to integrate an entire data acquisition system within a single CMOS VLSI chip is contingent upon the ability to realize a CMOS compatible voltage reference with a very low temperature drift. Since its introduction by Widlar [1], the bandgap referencing (BGR) technique has been widely employed for implementing a voltage reference source in bipolar integrated circuits. The temperature stability of the bandgap reference has been continuously improved via new circuit and technology innovations such as curvature compensation and laser trim [2]–[5]. In CMOS technology, the BGR technique has been directly applied [6]–[8]. However, the development of a high-performance CMOS bandgap reference has been hindered by several limiting factors attributable to the peculiarities of the bipolar devices available in a standard CMOS process, the high offset and drift of CMOS op amps that make up the circuit and the inherent curvature problem in the bandgap reference.

This paper will describe one circuit implementation of a precision CMOS bandgap reference which overcomes some of the drawbacks of a standard CMOS process, and embodies curvature compensation and differential offset cancellation to achieve experimental typical temperature drifts of 13.1 and 25.6 ppm/°C over the commercial and military temperature ranges, respectively. In the proposed reference, a temperature-stable voltage is developed by adding linear and quadratic temperature correction voltages to the forward-biased diode voltage which is obtained from the substrate p-n-p transistor available in CMOS processes. The linear temperature correction voltage is proportional to the absolute temperature (commonly called PTAT) while the quadratic temperature correction voltage is proportional to the absolute temperature squared (PTAT^2). They are independently adjustable to set the reference output voltage for a minimum temperature drift.

The offset voltage of the CMOS op amp is eliminated using the correlated-double sampling (CDS) technique [9]. The base current and base spreading resistance of the native substrate p-n-p transistor are cancelled to the first order and the amplification ratio is set by a capacitor ratio rather than a resistor ratio. Due to the cyclic behavior of the offset cancellation in this technique, the output reference voltage is not available at all times. However, the reference can be operated synchronously with other elements of the systems. Since the periodic offset sample and subtraction cycle effectively removes the low-frequency $1/f$ noise component of a CMOS op amp along with its offset, the dominant noise source is the thermal noise of a CMOS op amp which is designed to be on the order of 100 μV (rms) at the output in 500 kHz bandwidth.

In Sections II and III, the primary limitations in a conventional CMOS BGR implementation and the temperature curvature in bandgap references are discussed. In Section IV, a curvature-compensated switched-capacitor CMOS bandgap reference is introduced. In Section V, experimental results measured from monolithic prototype samples are presented and the problems related to the design of p-n-p transistors are discussed. The theoretical analysis of BGR temperature compensation techniques is included in the Appendix.

II. Conventional CMOS BGR Implementation

One example of a conventional CMOS BGR implementation in an n$^-$-well CMOS process is shown in Fig. 1. Transistors Q_1 and Q_2 are substrate p-n-p transistors whose collectors are always tied to the most negative power

Manuscript received April 6, 1983; revised August 3, 1983. This work was supported by the National Science Foundation under Grants ECS-8023872 and ECS-8120012, IBM Corporation, and the MICRO Project.

B.-S. Song is with Bell Laboratories, Murray Hill, NJ 07974.

P. R. Gray is with the Department of Electrical Engineering and Computer Science, Electronics Research Laboratory, University of California, Berkeley, CA 94720.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-18, no. 6, pp. 634–643, Dec. 1983.

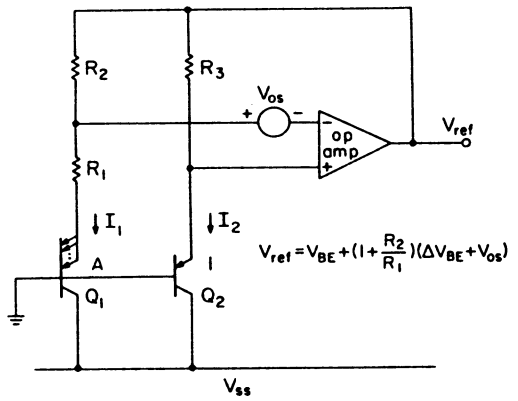


Fig. 1. Example of a conventional CMOS bandgap reference.

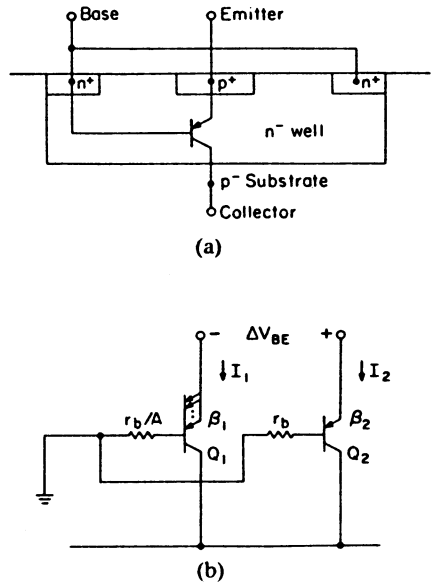


Fig. 2. (a) Substrate p-n-p transistor profile. (b) Nonideal parameters in the PTAT correction voltage generation circuit.

supply because, in an n^--well CMOS process, the p^+ diffusion in the n^--well, the n^--well itself and the p^- substrate form a vertical p-n-p structure as shown in Fig. 2(a). Therefore, it is not possible to sense the collector current directly as in the bipolar bandgap reference so as to reduce the error due to the finite current gain [3], [4]. In a p^--well process, a dual circuit incorporating n-p-n transistors would be used. While many other circuit implementations are possible, this circuit appears to be as good as any. Therefore, individual error sources will be described one by one for this circuit in the rest of this section. All resistors are the p^+-diffusion resistor in the n^--well and the CMOS op amp is assumed to have an infinite gain with the offset voltage of V_{os}. This assumption is justified because CMOS op amps usually have enough gains such that the error due to finite-gain effects is negligible for this application.

Assuming that transistor Q_1 in Fig. 1 has an area that is larger by a factor A than transistor Q_2, and both are in the forward active region, the output voltage of the reference is given by

$$V_{\text{ref}} = V_{BE} + \left(1 + \frac{R_2}{R_1}\right)(\Delta V_{BE} + V_{os}) \tag{1}$$

where V_{BE} is the emitter–base voltage of transistor Q_1, ΔV_{BE} is the difference between the emitter–base voltages of transistors Q_1 and Q_2, and V_{os} is the input offset voltage of the operational amplifier. The value of this expression is influenced by the nonidealities of the bipolar transistors as illustrated in Fig. 2(b). If these are taken into account, the transistor emitter–base voltage is given by

$$V_{BE} = V_T \ln \frac{I_1}{I_{s1}} + V_T \ln \frac{1}{1 + \frac{1}{\beta_1}} + \frac{r_b I_1}{A\beta_1} \tag{2}$$

where V_T is the thermal voltage kT/q, I_1 is the emitter current of transistor Q_1, I_{s1} is the saturation current of transistor Q_1, β_1 is the current gain of transistor Q_1 and r_b is the effective series base resistance of Q_2. The second term in this expression results from the fact that while the collector current is a well-defined function of the emitter–base voltage, the current sensed and controlled by this circuit is the emitter current. The third term results from the voltage drop in the finite series base resistance. The difference between the two emitter–base voltages is given by

$$\Delta V_{BE} = V_T \ln A + V_T \ln \frac{I_2}{I_1} + V_T \ln \frac{1 + \frac{1}{\beta_1}}{1 + \frac{1}{\beta_2}} + r_b \left(\frac{I_2}{\beta_2} - \frac{I_1}{A\beta_1}\right) \tag{3}$$

where I_2 is the emitter current of transistor Q_2 and β_2 is the current gain of transistor Q_2. If the bipolar transistors used to implement the reference are ideal in the sense that they have infinite current gain and zero base resistance, and if the emitter currents of the transistors are in fact equal, then only the first terms in (2) and (3) are nonzero. However, because of the relatively poor performance of CMOS-compatible devices, these terms can strongly influence the performance of the reference. The presence of the operational amplifier offset voltage in the output, multiplied by the gain factor $(1 + R_2/R_1)$, which is typically on the order of 10, is also an important degradation. Finally, the variation of the bias currents I_1 and I_2 with temperature must be carefully considered. In the following subsections, the effects of these nonidealities are examined in more detail.

A. Operational Amplifier Offset

The operational amplifier offset is the biggest error source that causes the nonreproducibility in the output voltage temperature coefficient. Normally, a bandgap reference is trimmed to an output voltage which is predetermined to give a near-zero temperature coefficient of the output. Large, non-PTAT components in the output due to

the op amp offset cause the trimming operation to give an erroneous result. If we assume the offset voltage V_{os} is independent of temperature, the resulting temperature coefficient error due to a 5 mV V_{os}, for example, is approximately

$$\text{TC error} = \frac{\left(1+\frac{R_2}{R_1}\right)V_{os}}{V_{\text{ref}}T_o} \approx \frac{10\times 5\text{ mV}}{1.26\text{ V}\times 300\text{ K}}$$
$$\approx 132\text{ ppm/°C}. \quad (4)$$

That is, a temperature coefficient on the order of 132 ppm/°C will result in the reference output temperature coefficient from a 5 mV temperature-independent offset voltage in the operational amplifier if the reference is trimmed assuming the offset is zero. This offset error contribution can be reduced by making ΔV_{BE} bigger, in effect decreasing the gain factor $(1 + R_2/R_1)$ as implied by (1). One way to achieve this is to obtain ΔV_{BE} by taking the difference of two cascaded transistor strings discussed later in Section IV. In this work, however, offset cancellation technique is employed to further reduce this error while the cascading scheme is used for the PTAT current generation.

B. Bias Current Variation

If the resistors R_1, R_2, and R_3 have a zero temperature coefficient, then the bias currents in transistors Q_1 and Q_2 must be PTAT since the voltage across R_1 is PTAT. The finite temperature coefficient of actual resistors formed from the source-drain diffusion or from polysilicon layers results in non-PTAT variation of the bias current. This in turn causes an additional component in the temperature variation of the V_{BE} term in the output. If only the first two terms of (2) and (3) are taken, the V_{BE} is given by

$$V_{BE} = V_T \ln \frac{I_1}{I_{s1}} = V_T \ln \frac{V_T \ln A}{R_1 I_{s1}}$$
$$= V_T \ln \frac{V_T \ln A}{R_1(T_o) I_{s1}} + V_T \ln \frac{R_1(T_o)}{R_1(T)} \quad (5)$$

where T_o is the reference temperature, usually room temperature. Note that the first term is the V_{BE} variation that results when there is no resistor temperature coefficient, and the second is that which results when a temperature coefficient is present. Further insight can be obtained by expanding this second term as a Taylor series in temperature about T_o, and neglecting higher order terms:

$$V_{BE} = V_{BE}|_{\text{ideal}} - V_T \cdot \frac{1}{R}\frac{dR}{dT}\bigg|_{T_o}(T-T_o) - V_T \frac{1}{2R}\frac{d^2R}{dT^2}\bigg|_{T_o}(T-T_o)^2 - \cdots. \quad (6)$$

From this relation it can be seen that even a purely linear variation in resistor value with temperature results in an output temperature variation with both PTAT and PTAT^2 temperature variation components. Assuming the resistor temperature behavior is known and reproducible, the PTAT portion can be compensated by simply changing the target trim value of the output voltage. For example, (6) can be used to show that a 1000 ppm/°C resistor TC would result in a −21 ppm/°C reference output TC, which could be removed by simply raising the output voltage trim target by approximately 8 mV. However, cancellation of PTAT^2 term precisely requires curvature compensation discussed later.

C. Other Effects

The effects of various nonidealities, including base resistance, β mismatch, β variations with temperature, and β variations with collect current can be evaluated with the use of (1), (2), and (3). Which of these effects is the most important in a given circuit application is strongly dependent on the nature of the bipolar transistors in the particular technology used. If the well (base) doping is particularly light, as is often the case, then the intrinsic base resistance effect, represented by the last term in (3), may well be the most important. The temperature coefficient in the output due to this term is given by

$$\text{TC error} = \left(1+\frac{R_2}{R_1}\right)\frac{r_b I_2}{V_{\text{ref}}\beta_2}\left(\frac{1}{r_b}\frac{dr_b}{dT} + \frac{1}{I_2}\frac{dI_2}{dT} - \frac{1}{\beta_2}\frac{d\beta_2}{dT}\right). \quad (7)$$

For example, assuming a 2 kΩ base resistance with 1000 ppm/°C TC, a 30 μA PTAT bias current level, a β of 150, and a β TC of 7000 ppm/°C, an output TC of −8.6 ppm/°C results. As in the case of the bias current variation, this can be partly compensated by modifying the trim target voltage if these parameters are reproducible. The other errors mentioned above are negligible if transistor performance is reasonably good. Another error source results from the temperature coefficient of the ratio of the diffused resistors R_1 and R_2. Data presented later show that for resistors used in this experimental work a differential ratio TC of 1 to 2 percent is achieved. Fortunately, this error is negligible compared to those already discussed.

III. Curvature in Bandgap References

For a bandgap reference which is ideal in the sense that the operational amplifier is ideal, the bipolar transistors have infinite current gain, zero intrinsic base resistance, and have perfect exponential junction relationships, and the bandgap of silicon varies linearly with temperature, the output voltage is typically given by [10]

$$V_{\text{ref}} = V_{go} + V_T(4-n-\alpha)\left(1+\ln\frac{T}{T_o}\right) \quad (8)$$

where V_{go} is the extrapolated silicon bandgap at 0 K, n is the exponent of the mobility variation in the base of the bipolar transistor (typically about 0.8), α is the exponent of the temperature variation of the bias current (1 for PTAT bias current, for example), and T_o is the temperature at which the reference output temperature coefficient is zero,

usually chosen to be near room temperature. Equation (21) derived in the Appendix is the more general form of (8). This relation illustrates the well-known fact that even for an ideal bandgap with an optimally chosen T_o, the output voltage as a function of temperature displays a curvature which causes it to decrease both for temperatures higher or lower than T_o. Usually, the practical performance aspect of interest is the maximum total variation of the output voltage over the range of temperatures. A further complication is the fact that the bandgap of silicon in fact does vary with T^2, as well as linearly with temperature [11], [12]. This higher order temperature variation adds a curvature term, increasing the effective TC even for optimally adjusted references. These effects together combine to give a best achievable TC of about 25 ppm/°C for a temperature range of −55°C to 125°C.

Several approaches have been suggested for curvature correction. If the quantity $(4-n-\alpha)$ in (8) could be made zero, then there would be no curvature other than the curvature of the silicon bandgap. This could be achieved for example by using a very strongly temperature-dependent bias current or by linearizing V_{BE} directly [5]. Several authors have proposed simply adding in higher order temperature-dependent terms in the output to cancel the PTAT^2 term of the output voltage variation [4]. This is the approach used in this paper to be described in the next section.

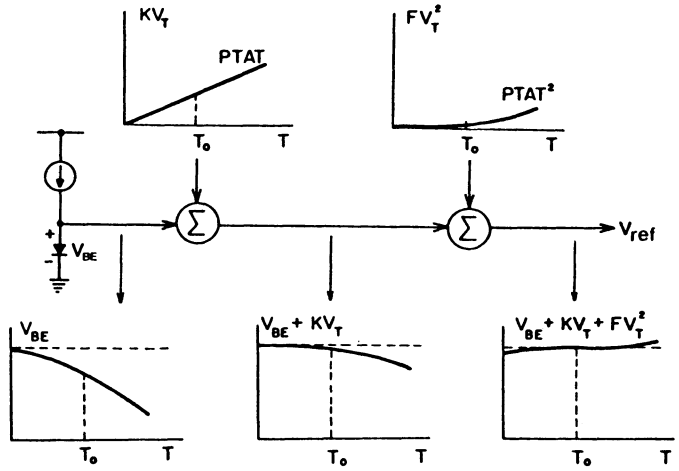


Fig. 3. Curvature-compensation concept (not scaled).

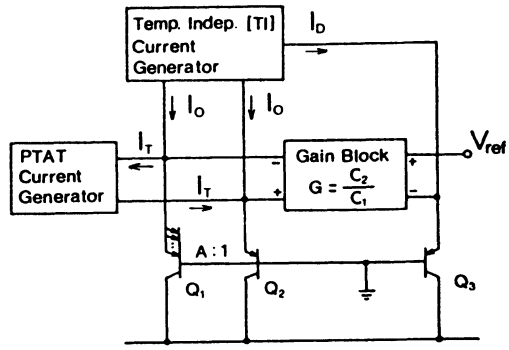


Fig. 4. Overall schematic of the curvature-compensated switched-capacitor CMOS bandgap reference.

IV. Precision Switched-Capacitor CMOS BGR Technique

In order to implement voltage references in CMOS technology which have performance approaching that achievable in bipolar technology, special steps must be taken to counteract the relatively poor performance of CMOS op amps and CMOS compatible bipolar transistors. In addition, curvature compensation, already widely applied in bipolar references, must be incorporated. In this section, the implementation of the techniques in CMOS technology is described.

A. Curvature Compensation

The overall concept of BGR temperature compensation is illustrated in Fig. 3 and general BGR temperature compensation techniques are described in the Appendix. The first step is to add a PTAT correction voltage KV_T to V_{BE} to cancel out the linear temperature variation of V_{BE}. After the PTAT correction voltage is added, the reference output V_{ref} will exhibit mostly the quadratic temperature variation as shown in Fig. 3. If a PTAT^2 correction voltage FV_T^2 is added to that to cancel out the quadratic temperature variation of V_{BE}, the final reference output V_{ref} should drift only due to higher order temperature variations and a zero temperature coefficient is achieved at T_o. One implementation of a switched-capacitor bandgap reference which embodies curvature compensation as well as offset-cancelled amplification is illustrated in Fig. 4. The gain G of the gain block is determined by the capacitor ratio C_2/C_1. The current I_o is temperature-independent (TI) while I_T is PTAT. The bias current I_D is also TI.

If the effects of the base current and the base spreading resistance are neglected, the reference output V_{ref} is given by

$$V_{\text{ref}} = V_{BE} + \frac{C_2}{C_1}\Delta V_{BE} \tag{9}$$

where

$$\Delta V_{BE} = V_T \ln A \frac{I_o + I_T}{I_o - I_T}$$

$$= V_T \ln A + 2V_T\left(\frac{I_T}{I_o}\right) + \frac{2}{3}V_T\left(\frac{I_T}{I_o}\right)^3 + \cdots. \tag{10}$$

The inclusion of the PTAT^2 voltage means that the trim procedure for the reference consists of two steps, one to give the correct output voltage for the uncompensated reference, and one to trim the value of the PTAT^2 voltage that is added in. A key advantage of the circuit configuration chosen is that it allows the PTAT^2 component to be adjusted independently from the basic reference. The first part of the trim procedure is to disconnect the PTAT^2 component (set I_T to zero), and trim the absolute output voltage. In the particular experimental device described

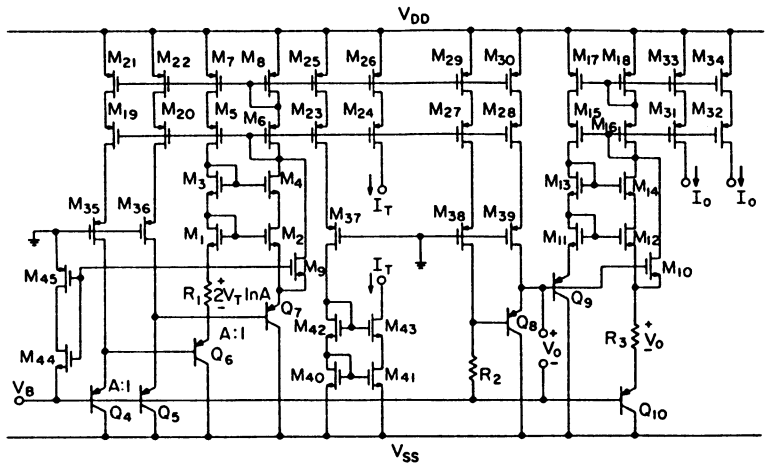


Fig. 5. PTAT and TI current generators for the bias currents I_T and I_o.

here, it is most convenient to do this with a combination of a capacitor array to adjust the C_2/C_1 ratio, and a resistor string to provide fine adjustment through adjustment of the bias current I_d. Even though two different physical trim arrays are involved, in effect one trim operation is performed and a total resolution of approximately 12 bits is achieved in the absolute value of the final output voltage. Next the PTAT current I_T is turned on and its ratio to I_o is adjusted with another resistor string to give a change in the output equal to the desired PTAT^2 compensation value. Both trimming operations are done at room temperature.

The currents I_T and I_o are generated using the circuit shown in Fig. 5. The bias current I_D is generated in the same manner as the TI current I_o. The stacked-cascode connection formed by transistors M_1 to M_8 causes the emitter currents of transistors Q_6 and Q_7 to be equal. The same scheme is also employed for the matching of the emitter currents of transistors Q_9 and Q_{10}. Transistors M_9 and M_{10} form a start-up circuit for this self-biased circuit. As indicated, transistors Q_4 and Q_6 have emitter areas larger by a factor of A than the remaining transistors. Therefore, the voltage developed across the resistor R_1 is PTAT and the current I_T through R_1 is also PTAT. The voltage V_o formed by the transistor Q_8 and the resistor R_2 is approximately temperature independent. The temperature stability of V_o is not critical because it affects only the PTAT^2 component of the output voltage (not PTAT). The TI voltage V_o is developed across R_3 and the current I_o through R_3 is also TI. Therefore, the current ratio I_T/I_o is easily trimmed by R_3.

In the presence of the mismatches of $M_1 - M_2$ and $M_{11} - M_{12}$ transistor pairs, the voltages across R_1 and R_3 deviate from the ideal values. If the gate–source voltage mismatches of MOS transistor pairs inclusive in the bias circuit of Fig. 5 are assumed to be V_{m1}, V_{m2}, and V_{m3}, respectively, and current mirrors and p-n-p transistors are assumed to be ideal, the three bias currents I_T, I_o, and I_D are given by

$$I_T = \frac{1}{R_1}(2V_T \ln A + V_{m1}), \tag{11}$$

$$I_o = \frac{1}{R_3}(V_o + V_{m2}) \tag{12}$$

and

$$I_D = \frac{1}{R_4}(V_o + V_{m3}). \tag{13}$$

The factor 2 in (11) results from cascading transistors $Q_4 - Q_6$ and $Q_5 - Q_7$ to reduce the error contribution of the mismatch voltage V_{m1}. For this application, cascading of two devices is enough for the PTAT current generation. Depending on applications, several transistors can be cascaded to reduce the offset error contribution in such a BGR implementation, discussed in Section II. Substituting (11), (12), and (13) into (9) and (10), and neglecting higher orders, we obtain

$$\begin{aligned} V_{\text{ref}} &= V_{BE}|_{\text{ideal}} + \frac{C_2}{C_1} V_T \ln A + 4\frac{C_2}{C_1}\frac{R_3}{R_1}\frac{\ln A}{V_o} V_T^2 + V_\epsilon \\ &= V_{BE}|_{\text{ideal}} + KV_T + FV_T^2 + V_\epsilon \end{aligned} \tag{14}$$

where

$$\begin{aligned} V_\epsilon = {} & \frac{\sigma}{100}\frac{1}{R}\frac{dR}{dT}\bigg|_{T_o}(V_{T_o} - V_T)T + 2\frac{C_2 R_3 V_T V_{BE}}{C_1 R_1 V_o^2} V_{m1} \\ & - 4\frac{C_2 R_3 V_T^2 \ln A}{C_1 R_1 V_o^2} V_{m2} + \frac{V_T}{V_o} V_{m3}. \end{aligned} \tag{15}$$

The constant σ represents the standard deviation (%) of the temperature coefficient of R_4. The error voltage V_ϵ is less than 1 mV at room temperature. Therefore, the temperature coefficient uncertainty resulting from the mismatch voltages is 50 times smaller than the uncertainty caused by the op amp offset of existing designs given by (4).

B. Offset-Cancelled Amplification

In order to remove dc offsets from the amplifier, it is divided into two stages as shown in Fig. 6. In the first offset-storage mode, all the MOS switches are closed to

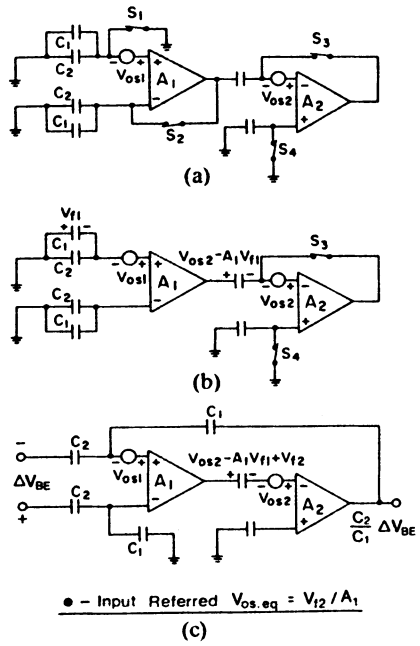


Fig. 6. Gain block which embodies offset-cancelled amplification. (a) First offset-storage mode. (b) Second offset-storage mode. (c) Amplification mode.

sample the offset voltages of the individual op amps. In the process of opening the MOS switches S_1 and S_2, the channel charges are injected into the op amp summing nodes to load the capacitors C_1 and C_2. The charge injection differential voltage V_{f1} due to the mismatch of switches S_1 and S_2 is sampled across C_1 and C_2 along with the offset voltage V_{os1}. In the second offset-storage mode, the first gain stage charges the coupling capacitor to compensate for the input differential voltage V_{f1}. After the switches S_3 and S_4 are opened, two stages are connected in a feedback amplification mode, and the amplification of ΔV_{BE} takes place by the capacitor ratio C_2/C_1. When referred to input, the feedthrough difference of the switches S_3 and S_4 is reduced by the open-loop gain of the first stage. Note the bottom plate of one capacitor C_1 should be connected to the diode voltage in the actual reference as shown in Fig. 4.

A single-pole folded-cascode CMOS op amp configuration for A_1 and A_2 is shown in Fig. 7. Two amplifiers are identical and designed to meet the following requirements:

1) moderate gain for each stage (100 to 300);
2) single dominant pole per stage;
3) inherent zero systematic offset voltage; and
4) capacitor driving capability (15 pF for one stage and 100 pF for two stages).

Transistors $M_{19}-M_{21}$ form a bias string for the amplifier. The replica bias circuit formed by transistors $M_{14}-M_{17}$ performs a level shift and a differential to single-ended conversion, and reduces the inherent systematic offset. In order to limit the gain of each stage, the lower part composed of transistors M_4-M_5 is not cascoded while the upper part M_6-M_9 is. The class-A source follower stages formed by transistors $M_{10}-M_{13}$ are added to meet the capacitor driving capability. In the offset-storage modes, each op amp is stabilized by connecting a frequency compensation capacitor between the high impedance node and the ac ground V_{ss}. When the two op amps are cascaded and the feedback loop closed around the composite amplifier, a Miller capacitance is switched in from the high impedance node of the second stage to the same node of the first stage to achieve a pole-splitting compensation.

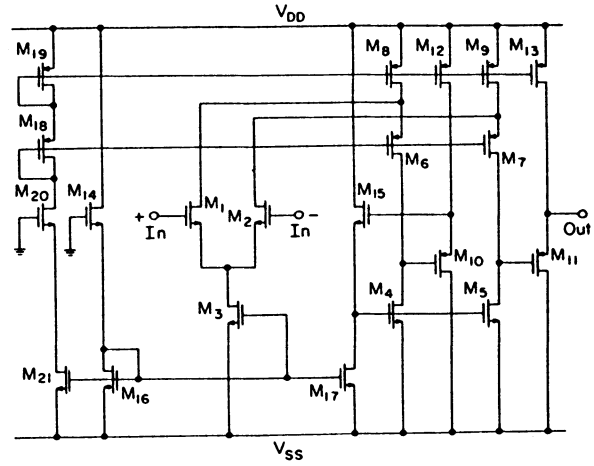


Fig. 7. Amplifier configuration for A_1 and A_2.

C. Base Current Cancellation

In the CMOS process used, the current gain of substrate p-n-p transistors is often limited and highly variable. Therefore, to compensate for the difference between the collector current and the emitter current, the base current has to be returned to the emitter as shown in Fig. 8(a). However, a simpler approach is to replicate the base current and allow it to flow into the emitter as shown in Fig. 8(b). The base current is cancelled with an accuracy of about 90 percent because the base currents typically match each other within 10 percent for the adjacent transistors on a single chip.

D. Base Resistance Cancellation

The effective series base resistance of the bipolar transistors consists of that produced by lateral flow in the base region under the emitter, and the extrinsic base resistance between the base contact and the active base. The former is bias dependent and difficult to predict, while the latter is more straightforward to predict given device geometry. The approach taken here is to include a lumped resistor R_{comp} made of the same n^--well diffusion material so as to achieve approximate tracking with temperature and process variations. The value of R_{comp} in Fig. 9 should be

$$R_{\text{comp}} = \left(\frac{1+\beta_1}{1+\beta_2} - \frac{1}{A}\right) r_{b1}. \tag{16}$$

For the process used here, the magnitude of the extrinsic compensation resistance is about one quarter of the measured intrinsic base resistance of p-n-p transistors with the same geometry.

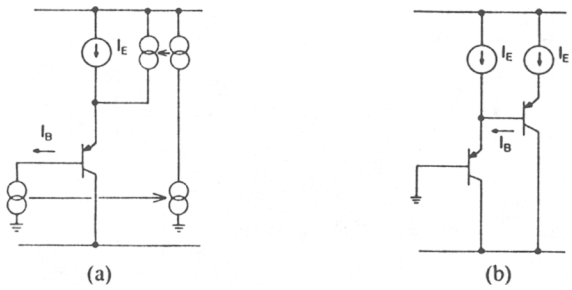


Fig. 8. Base current cancellation schemes. (a) I_B returning. (b) I_B replication.

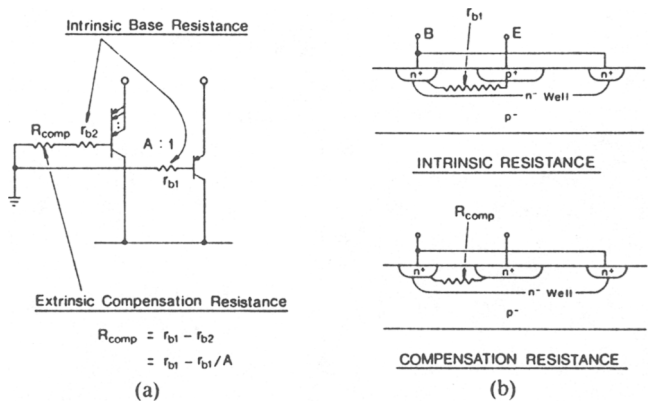


Fig. 9. Base resistance cancellation. (a) Extrinsic compensation resistance R_{comp}. (b) Difference between intrinsic base resistance and extrinsic compensation resistance.

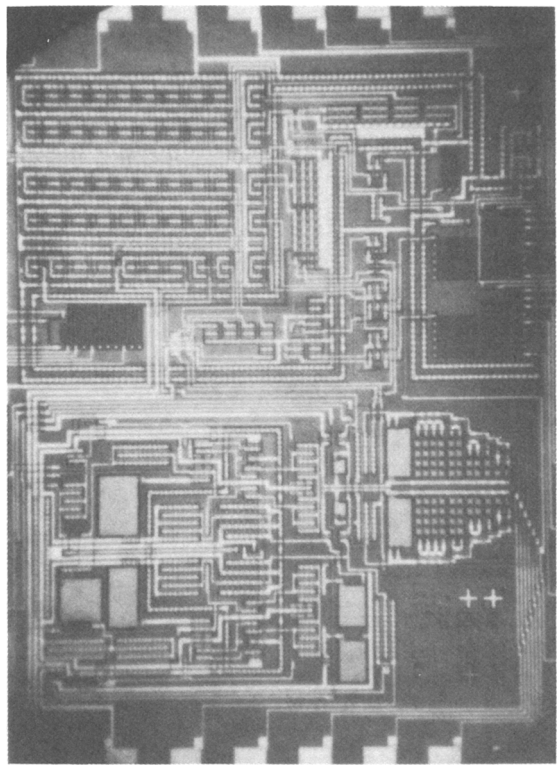


Fig. 10. Chip photo of the experimental prototype.

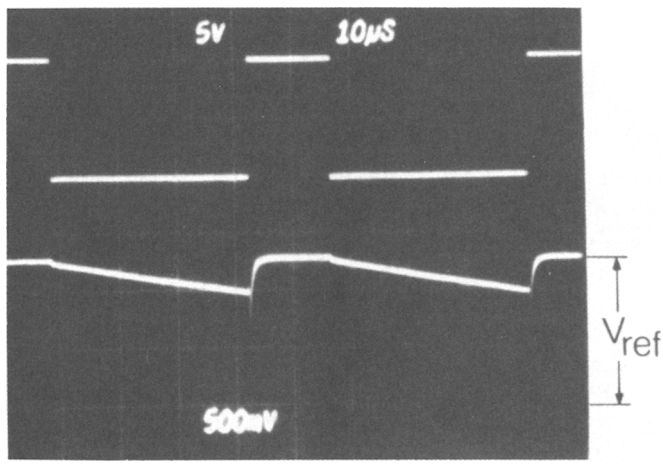


Fig. 11. Waveforms. (a) Output sync clock. (b) V_{ref}.

V. Experiment and Discussions

The experimental prototype circuit implementing the proposed reference was fabricated employing a self-aligned single-poly Si-gate CMOS process on a 20–30 $\Omega \cdot$cm boron-doped p-type $\langle 100 \rangle$ substrate. The gate oxide is 0.07 μm thick and the drawn minimum feature is 6 μm. Fig. 10 shows the microphotograph of the prototype chip and Fig. 11 shows its output waveform as well as output sync pulse.

Experimental data were gathered from seven representative samples from one wafer. Every experimental chip contains three types of reference voltages. The Type I reference has no curvature compensation, no base current cancellation, no base resistance cancellation, and no offset cancellation, and amplification is performed by a resistor ratio. The Type II reference which uses a capacitor-ratio amplification has the cancellations of offset, base current and base resistance, but no curvature compensation. However, the Type III reference which also uses a capacitor ratio amplification has all components of the Type II reference plus curvature compensation.

Following the procedure described in Section IV and the Appendix, one sample was adjusted to give a minimum temperature drift and the other six samples were trimmed at room temperature to an output voltage predetermined from the first sample. Statistical data from seven samples are summarized in Table I for three types of reference voltages. The optimum values of the PTAT^2 correction voltage, the first-order corrected and second-order corrected V_{ref}'s at 25 °C were found to be 61 mV, 1.256 and 1.192 V, respectively. Estimating from the measured data, the parameters V_{go1}, V_{go2}, and $4-n-\alpha$ necessary for specifying the prototype bandgap reference were 1.181, 1.158, and 2.623 V, respectively.

Note that the offset-cancelled amplification and compensation of r_b and β effects give a factor of 5 improvement in temperature stability and its deviation over the approach without any compensation. By curvature com-

TABLE I
STATISTICS OF MEASURED TEMPERATURE COEFFICIENTS OF 7 SAMPLES (ppm° C)

0 to 70 °C:		Mean	Standard Deviation	Minimum	Maximum
	Type I	105	43	42	167
	Type II	22.3	10.8	11.1	42
	Type III	13.1	7.1	5.6	25.7
−55 to 125 °C:					
	Type I	185	56.5	107	273
	Type II	35.1	18.8	17.6	66.7
	Type III	25.6	10.5	12.1	39.9

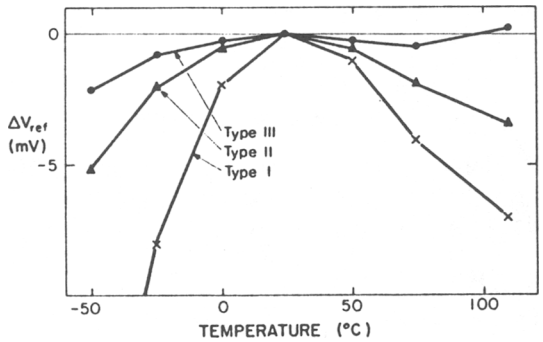


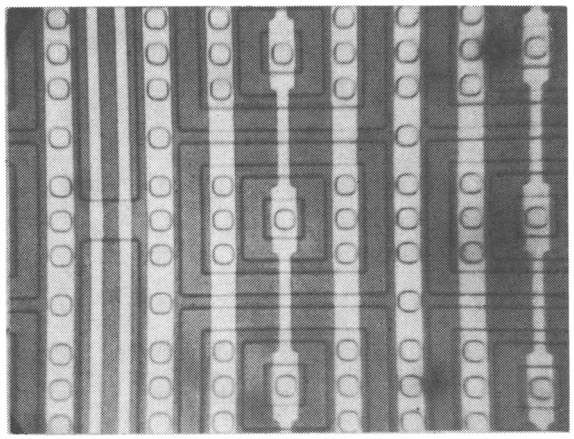
Fig. 12. Typical measured temperature variations of three types of references when they are optimally compensated.

TABLE II
PERFORMANCE SUMMARY

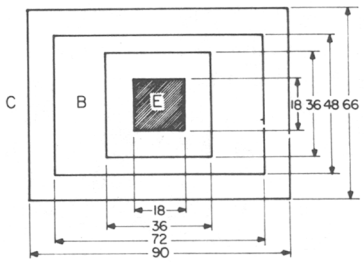
V_{ref}	1.192 V ± 1 mV at 25 °C
TC	See Table I
Power	12 mW with ± 5 V supply
Load	100 pF capacitor
Cycle time	5 μs
+ PSRR	50 dB (dc)
− PSRR	60 dB (dc)
Clock RR	75 dB
Output noise	400 μV (500 kHz)

pensation, a factor of 2 further improvement was obtained. Fig. 12 compares graphically those three types of optimally-compensated bandgap references. The experimental results are summarized in Table II. To improve the power supply rejection ratio (PSRR), the base of all p-n-p transistors should be biased at a constant voltage relative to the negative supply line. Otherwise, the base width modulation (Early effect) will limit the PSRR of the reference.

The critical aspect of the design is the substrate p-n-p transistor because the n^--well in CMOS processes usually has a relatively high resistivity. In order to minimize the intrinsic base spreading resistance, the base contact surrounds the emitter junction as shown in Fig. 13. The emitter junction and the base contact plug are separated by 9 μm. By the surrounding base, the intrinsic base resistance is reduced by a factor of 6 when compared to the parallel contact of emitter and base. For the process used here, the estimated r_b of this geometry is about 1.5 kΩ. With this geometry, the observed emitter area reduction due to the base crowding is insignificant over the 1 to 100 μA emitter



(a)



(b)

Fig. 13. Substrate p-n-p transistor. (a) A unit cell. (b) Drawn dimensions of a unit cell (μm).

current range. The current gain β is also relatively constant within this range. However, in the emitter current range over 100 μA, the base crowding effect is severe. Also, in the emitter current range below 1 μA, the current gain β decreases due to space charge recombination in the emitter–base junction. The unit cells of Fig. 13 were connected in parallel to obtain a multiple-emitter device. The current gain β is minimum at −55 °C and the average value at room temperature is 175. Temperature data for the p-n-p transistors and p^+ diffused resistors in the n^--well in the particular technology used here are listed in Table III. The temperature coefficients of diffused resistors match each other within 1.2 percent.

One potential problem in the use of the substrate p-n-p transistor is the fact that the dc collector current flows into the substrate. If this current gives rise to a large enough ohmic drop in the substrate, it could initiate latchup. To minimize the likelihood of this, each n^--well (base) was surrounded by the p^+ diffusion (collector) as shown in Fig. 13. No latchup was observed during experimental measurements even under transient conditions.

VI. CONCLUSION

A precision curvature-compensated switched-capacitor CMOS bandgap reference is reported whose monolithic prototype exhibits average temperature drifts of 13.1 and 25.6 ppm/ °C over the commercial and military tempera-

TABLE III
TEMPERATURE DATA OF DIFFUSED RESISTORS AND p-n-p TRANSISTORS
(−55 to 125°C)

Diffused Resistors:	Mean	Standard Deviation	Minimum	Maximum
Sheet resistance (Ω/square)	68.9	3.1	64.6	72.1
TC or R*	886	27.9	849	917
Ratio of $6R/R$**	5.952	0.02	5.9304	5.9794
TC of $6R/R$*	10.6	2.2	8.9	13.6
p-n-p Transistors:				
Current gain β at $I_E = 30\,\mu$A	175	83	91	273
TC of β*	6503	872	5471	7792

*TC unit is ppm/°C.
**R and $6R$ are composed of 40 and 240 squares of 6 μm-wide p^+ diffusion.

ture ranges, respectively, employing a straightforward room temperature trim procedure without thin-film resistor and laser trim. The design features second-order temperature compensation, simple room temperature trim up to 12-bit accuracy and complete cancellation of the offset and long-term offset drift of CMOS op amps. A reference voltage is obtained by systematically adding first-order and second-order correction voltages to the emitter-base potential of the substrate p-n-p transistor. Each correction voltage is individually trimmable to minimize temperature drift. The proposed reference is compatible with a standard digital CMOS process and is applicable to high-resolution monolithic CMOS data acquisition systems.

APPENDIX
BGR TEMPERATURE COMPENSATION TECHNIQUES

Employing (6) and neglecting higher orders, the forward-biased diode voltage is given by [10]

$$V_{BE} = V_g - V_T[(4-n-\alpha)\ln T - \ln EG] + HV_T - LV_T^2 \tag{17}$$

where V_g is the bandgap of silicon which is a function of base doping [13] and temperature [11], [12], n and α are parameters illustrated in (8), E and G are the parameters whose magnitude are insignificant in the temperature analysis [10], and H and L are defined from (6)

$$H = T_o \frac{1}{R}\frac{dR}{dT}\bigg|_{T_o} \quad \text{and} \tag{18}$$

$$L = \frac{T_o}{V_{T_o}}\frac{1}{R}\frac{dR}{dT}\bigg|_{T_o}. \tag{19}$$

That is, if the bias current variation and the silicon bandgap curvature discussed in Sections II and III are included, the V_{BE} in (17) is the actual diode voltage whose temperature variation is to be compensated to give a temperature stable reference voltage. In the next two subsections, first- and second-order temperature compensation techniques are discussed in more detail.

A. First-Order Temperature Compensation

If only the linear temperature variation of V_{BE} in (17) is compensated by adding the first-order correction voltage KV_T to V_{BE}, the reference output V_{ref} is

$$\begin{aligned} V_{\text{ref}} &= V_{BE} + KV_T \\ &= V_g - V_T(4-n-\alpha)\ln T \\ &\quad + (K + H + \ln EG)V_T - LV_T^2. \end{aligned} \tag{20}$$

By equating the derivative of V_{ref} at T_o to zero and eliminating the unknown constants, we obtain

$$\begin{aligned} V_{\text{ref}} = V_g - \frac{dV_g}{dT}\bigg|_{T_o} T + V_T(4-n-\alpha)\left(1+\ln\frac{T_o}{T}\right) \\ - LV_T^2\left(1-2\frac{T_o}{T}\right). \end{aligned} \tag{21}$$

The first two terms and the last term of (21) result from the nonlinearity of the Si bandgap over temperature and the bias current variation, respectively, and will not disappear until higher order temperature variations are compensated. Without these, (21) is identical to (8). Tsividis [12] recently explained the Si bandgap temperature dependence employing equations from Bludau *et al.*'s [11]. From (21), the nominal voltage at T_o is therefore

$$V_{\text{ref}}|_{T_o} = V_{go1} + V_{T_o}(4-n-\alpha) + LV_{T_o}^2 \tag{22}$$

where

$$V_{go1} = V_g(T_o) - \frac{dV_g}{dT}\bigg|_{T_o} T_o \approx 1.205\text{ V}. \tag{23}$$

As commonly called, V_{go1} is the linearly-extrapolated Si bandgap voltage at $T = 0$ K. Equation (22) indicates that the bias variation resulting from the temperature coefficient of diffused resistors causes the nominal voltage different from a theoretical value.

B. Second-Order Temperature Compensation

If the linear and the quadratic temperature variations of

V_{BE} in (17) are compensated as illustrated in Fig. 3 by adding both the first-order correction voltage KV_T and the second-order correction voltage FV_T^2 to V_{BE}, the reference output V_{ref} is

$$\begin{aligned} V_{\text{ref}} &= V_{BE} + KV_T + FV_T^2 \\ &= V_g - V_T(4-n-\alpha)\ln T \\ &\quad + (K + H + \ln EG)V_T + (F-L)V_T^2. \end{aligned} \tag{24}$$

By equating the first-order and the second-order derivatives of V_{ref} at T_o to zero and eliminating the unknown constants, we obtain

$$\begin{aligned} V_{\text{ref}} = V_g - \left.\frac{dV_g}{dT}\right|_{T_o} T - \frac{1}{2}\left.\frac{d^2V_g}{dT^2}\right|_{T_o} T^2\left(1-2\frac{T_o}{T}\right) \\ + V_T(4-n-\alpha)\left(\ln\frac{T_o}{T} + \frac{1}{2}\frac{T}{T_o}\right). \end{aligned} \tag{25}$$

Therefore, the nominal voltage at T_o is

$$V_{\text{ref}}|_{T_o} = V_{go2} + \frac{1}{2}V_{T_o}(4-n-\alpha) \tag{26}$$

where

$$V_{go2} = V_g(T_o) - \left.\frac{dV_g}{dT}\right|_{T_o} T_o + \frac{1}{2}\left.\frac{d^2V_g}{dT^2}\right|_{T_o} T_o^2 \approx 1.179\ \text{V}. \tag{27}$$

Now V_{go2} is the quadratically-extrapolated Si bandgap voltage at $T=0$ K. The voltage V_{go2} is closer to the Si bandgap at $T=0$ K than V_{go1} of (23). The theoretical value of $V_g(0\ \text{K})$ is approximately 1.179 V [12]. The bias current variation no longer affects (26) because it is compensated by the PTAT^2 correction voltage.

Only after the PTAT voltage is added, the intermediate voltage at T_o is

$$(V_{BE} + KV_T)|_{T_o} = V_{go2} + \frac{1}{2}\left.\frac{d^2V_g}{dT^2}\right|_{T_o} T_o^2 - LV_{T_o}^2. \tag{28}$$

Correspondingly, the magnitude of the PTAT^2 voltage FV_T^2 at T_o is obtained by subtracting (28) from (26):

$$FV_T^2|_{T_o} = \frac{1}{2}V_{T_o}(4-n-\alpha) - \frac{1}{2}\left.\frac{d^2V_g}{dT^2}\right|_{T_o} T_o^2 + LV_{T_o}^2. \tag{29}$$

This PTAT^2 correction voltage includes the inherent Si bandgap curvature as well as the bias current variation. Other error sources neglected can be included in (29) easily in the same manner as the bias current variation. No matter how many error sources are included, the final V_{ref} given by (25) is independent of these instabilities as far as their temperature variations are compensated properly.

REFERENCES

[1] R. J. Widlar, "New developments in IC voltage regulators," *IEEE J. Solid-State Circuits*, vol. SC-6, pp. 2–7, Feb. 1971.

[2] K. E. Kujik, "A precision reference voltage source," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 222–226, June 1973.

[3] A. P. Brokaw, "A simple three-terminal bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 288–393, Dec. 1974.

[4] C. R. Palmer and R. C. Dobkin, "A curvature corrected micropower voltage reference," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 1981, pp. 58–59.

[5] G. C. M. Meijer, P. C. Schmale, and K. van Zalinge, "A new curvature-corrected bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1139–1143, Dec. 1982.

[6] Y. P. Tsividis and R. W. Ulmer, "A CMOS voltage reference," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 774–778, Dec. 1978.

[7] E. A. Vittoz and O. Neyroud, "A low-voltage CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 573–577, June 1979.

[8] R. Gregorian, G. A. Wegner, and W. E. Nicholson, Jr., "An integrated single-chip PCM voice codec with filters," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 322–333, Aug. 1981.

[9] W. H. White, D. R. Lampe, F. C. Blaha, and I. A. Mack, "Characterization of surface channel CCD image arrays at low light levels," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 1–14, Feb. 1974.

[10] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*. New York: Wiley, 1977, pp. 256.

[11] W. Bludau, A. Onton, and W. Heinke, "Temperature dependence of the bandgap in Si," *J. Appl. Phys.*, vol. 45, pp. 1846–1848, 1974.

[12] Y. P. Tsividis, "Accurate analysis of temperature effects in $I_C - V_{BE}$ characteristics with application to bandgap reference source," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1076–1084, Dec. 1980.

[13] J. W. Slotboom and H. C. DeGraaf, "Measurements of bandgap narrowing in Si bipolar transistors," *Solid-State Electron.*, vol. 19, pp. 857–862, 1976.

CMOS Voltage References Using Lateral Bipolar Transistors

MARC G. R. DEGRAUWE, OSKAR N. LEUTHOLD, ERIC A. VITTOZ, MEMBER, IEEE, HENRI J. OGUEY, MEMBER, IEEE, AND ARTHUR DESCOMBES, MEMBER, IEEE

Abstract—**Two bandgap references are presented which make use of CMOS compatible lateral bipolar transistors. The circuits are designed to be insensitive to the low beta and alfa current gains of these devices. Their accuracy is not degraded by any amplifier offset.**

The first reference has an intrinsic low output impedance. Experimental results yield an output voltage which is constant within 2 mV, over the commercial temperature range (0–70°C), when all the circuits of the same batch are trimmed at a single temperature. The load regulation is 3.5 μV/μA and the Power Supply Rejection Ratio (PSRR) at 100 Hz is 60 dB.

Measurements on a second reference yield a PSRR of minimum 77 dB at 100 Hz. Temperature behavior is identical to the first circuit presented. This circuit requires a supply voltage of only 1.7 V.

I. Introduction

DURING THE LAST FEW YEARS, there has been an increasing trend to realize analog and digital circuits on the same chip. Bipolar technologies are more adequate to implement analog functions but CMOS technologies are more interesting if large digital parts have to be realized.

Voltage references are a key element of analog-to-digital converters. In the past, several CMOS compatible voltage references [1]–[5] have already been proposed but none of them achieves the precision of purely bipolar bandgap references. The circuits suffer from the weaknesses of CMOS technologies (large amplifier offset, poor matching of transistors, etc.) or are quite complex and do not deliver a continuous reference voltage [4].

In this paper, a family of more accurate voltage references is presented. The circuits which make use of lateral bipolar transistors eliminate the problems of the previously published CMOS voltage references.

CMOS compatible lateral bipolar transistors [6] are first briefly discussed and an alternative way for their realization is presented. In the next section, the basic principle of the references is studied in detail. Limitations of temperature stability due to technology variations are explained. A low output impedance reference is then presented with exhaustive experimental results. Finally a voltage reference with a very high Power Supply Rejection Ratio (PSRR) is discussed.

Manuscript received April 22, 1985; revised July 22, 1985. This work was supported by the "Fonds National Suisse pour la Recherche Scientifique, PN13."

M. G. R. Degrauwe, E. A. Vittoz, and H. J. Oguey are with Centre Suisse d'Electronique et de Microtechnique S.A. (CSEM), Recherche et Développement (formerly CEH) Maladière 71, 2000 Neuchâtel 7, Switzerland.

O. N. Leuthold was with Ebauches Electroniques S.A., MEM, 2074 Marin, Switzerland. He is now with Hughes Aircraft Co., Newport Beach, CA.

A. Descombes is with Ebauches Electroniques S.A., MEM, 2074 Marin, Switzerland.

II. Lateral Bipolar Transistors

It has been shown that a bipolar transistor, with a collector which is not tied to the substrate, is available in any CMOS technology [6]. This device is obtained by operating a MOS transistor in the lateral bipolar mode.

A cross section of a concentric n-channel transistor realized in a p-well CMOS process is shown in Fig. 1. By biasing the gate of the MOS transistor far below its threshold voltage, an accumulation layer is created below the gate. This prevents MOS transistor operation between the two concentric n^+ diffusions. By properly biasing the p-well (B) and the drains (E, C), a bipolar operating mode is obtained. In addition a second, unwanted, vertical bipolar transistor is also activated. In order to favor the lateral bipolar transistor, the gate length should be as small as possible and the perimeter-to-surface ratio of the emitter should be maximized. A symbolic representation of this five-terminal device is shown in Fig. 2.

An alternative way to realize this device is shown in Fig. 3. The normal thin oxide MOS transistor has been replaced by a parasitic field oxide transistor with an aluminum gate. If the threshold voltage of this transistor is high enough, this device operates already in the bipolar mode when gate and emitter are tied together. This structure thus has the advantage that one terminal can be eliminated and that no negative gate voltage need be created on chip.

III. Basic Principle and Theory

A classical approach to realize CMOS bandgap references is shown in Fig. 4 [9], [10]. Assuming an ideal amplifier, the output voltage of the circuit is given by

$$V_{\text{out}} = V_{BE_1} + \frac{R_2}{R_1}\left(V_{BE_1} - V_{BE_2}\right). \tag{1}$$

Since V_{BE_1} decreases approximately linearly with absolute temperature T and $(V_{BE_1} - V_{BE_2})$ increases with T, the

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1151–1157, Dec. 1985.

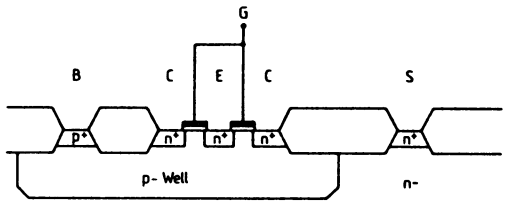


Fig. 1. Cross section of a lateral bipolar transistor.

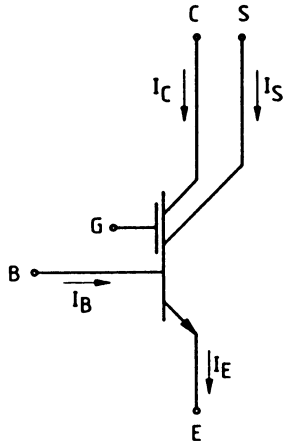


Fig. 2. Symbol for a lateral bipolar transistor.

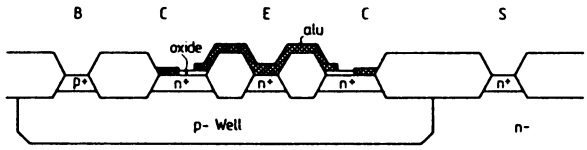


Fig. 3. Cross section of a lateral bipolar transistor formed with a parasitic MOS transistor.

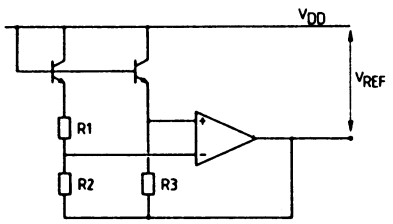


Fig. 4. Classical bandgap reference.

circuit can be made almost independent of temperature by choosing an appropriate ratio of R_2/R_1. The resulting output voltage at the reference temperature will be called the ideal reference voltage ($V_{ref}(T_r)$).

However the MOS amplifier is not ideal. It has a typical offset of a few millivolts. The output voltage will be given by (1) plus the additional term

$$\left(1+\frac{R_2}{R_1}\right)\cdot V_{os} \tag{2}$$

Since the ratio R_2/R_1 is about 10, the output voltage will have an unwanted component of about 20–50 mV. Since the offset voltage is not proportional to absolute temperature (PTAT), it cannot be fully compensated. Furthermore, it decorrelates the relationship which exists between output voltage and temperature behavior. Obtaining a very accurate voltage reference by trimming the output voltage at a single temperature becomes impossible.

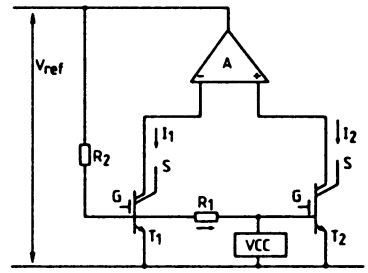


Fig. 5. Principle of voltage reference A-current comparator-V_{CC}-voltage controlled current source which draws a current proportional to I_1 and much larger than the base current of T_1.

To improve the accuracy of the voltage reference, the influence of the amplifier offset must be decreased. This can be achieved by using a chopper stabilized amplifier [4] or a stack of bipolar transistors [5]. These techniques lead, however, to circuits quite complex that are and/or require relatively large supply voltages.

A better solution is to use lateral bipolar transistors as shown in Fig. 5. The circuit consists of two bipolar transistors operated at different current densities, two resistances R_1 and R_2, a current comparator, and a voltage controlled current source. This current source draws a current much larger than the maximum base current of T_1 to achieve (1) independently of the current gain T_1. By realizing the current comparator by MOS transistors operating deep in strong inversion, the effect of the current offset can be neglected. Furthermore, any mismatch of bipolar pair T_1, T_2 from their nominal area ratio results in an output voltage component which is PTAT and can thus be compensated for.

The temperature behavior of the circuit will thus be as described by Tsividis [7] and is shortly repeated hereafter.

If the effective mobility for the minority carriers in the base can be represented with sufficient accuracy by

$$\mu(T) = C\cdot T^{-p} \tag{3}$$

with C and p appropriate constants, if the current through the bipolar transistors is given by

$$I_C = I_C(T_r)\left(\frac{T}{T_r}\right)^m \tag{4}$$

where

T_r = reference temperature

and if

$$\frac{R_2}{R_1}(V_{BE_1}-V_{BE_2})\bigg|_{T_r} = V_{G0_r}+(n-m)kT_r/q - V_{BE_1}(T_r) \tag{5}$$

where

V_{G0_r} = extrapolated bandgap voltage from T_r to 0 K

$n = 4-p$

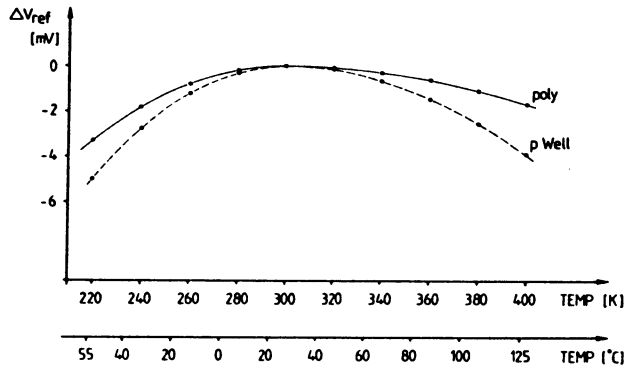


Fig. 6. Theoretical variation of the reference voltage.

then the output voltage of the circuit will be equal to

$$V_{ref}(T) = V_{ref}(T_r) + f_r(T) \tag{6}$$

where

$$V_{ref}(T_r) = V_{G0_r} + (n-m)\cdot\frac{k\cdot T_r}{q} \tag{7}$$

$$f_r(T) = V_G(T) - V_{G0_r} + \frac{T}{T_r}\left(V_{G0_r} - V_G(T_r)\right) + (n-m)\cdot\frac{k}{q}\cdot(T - T_r - T\ln(T/T_r)) \tag{8}$$

"$f_r(T)$" in (7) expresses the nonideal behavior of the voltage reference. The only way to decrease this nonideality is to choose an appropriate m.

Since the current in the bipolar transistors is fixed by PTAT voltage and a resistance, the coefficient m is determined by the temperature coefficient of this resistance.

If the temperature behavior of the resistance is modeled by

$$R(T) = R(T_r)\cdot\left(\frac{T}{T_r}\right)^a \tag{9}$$

then

$$m = 1 - a. \tag{10}$$

In Fig. 6 the formula (8) is evaluated for standard poly resistances ($a = 0.1$) and p-well resistances ($a = 2$). For the bandgap, the model given in [7] was used and from measurements the coefficient n was found to be 2.1. A bandgap reference realized with poly resistances will thus be more accurate than one realized with p-well resistors. For the commercial temperature range, the reference with poly resistance is stable within about 0.4 mV and with p-well resistances within 0.8 mV.

In practice all the circuits of a same batch will be trimmed at a reference temperature T_r to the same reference voltage given by (7). However, due to technology variations the ideal reference voltages of the circuits differ from each other. Measurements have shown that the temperature coefficient a of the poly resistances has a standard deviation of 0.05. From (6), (7), and (10) it is seen that the standard deviation of the ideal reference voltage $V_{ref}(T_r)$ will thus be about 1.25 mV at T_r. This means that at worst (3σ) a circuit will be trimmed 3.75 mV above or below its ideal reference voltage. This results, with respect to (8), in an additional PTAT temperature variation of about 0.9 mV (plus or minus) for the commercial temperature range.

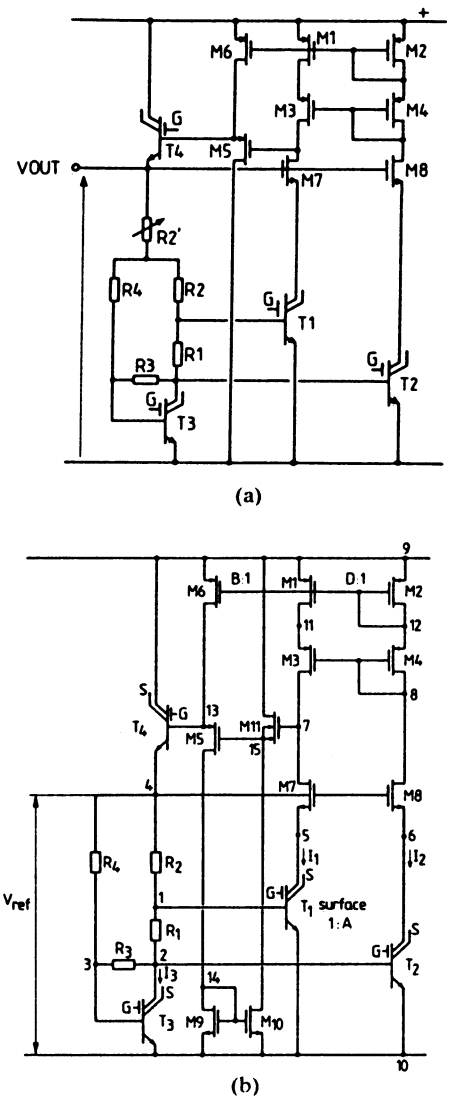


Fig. 7. (a) Circuit of low output impedance bandgap reference. (b) High-voltage supply variant.

IV. Low Output Impedance Voltage Reference

A straightforward implementation of the principle described above is shown in Fig. 7(a). The current comparator is realized as a cascode current mirror (M_1–M_4) followed by two source followers (M_5, T_4). Transistors M_1, M_2, and M_6 operate deep in strong inversion in order to minimize the sensitivity to threshold mismatch [11]. MOS follower M_5 and cascode transistors M_7, M_8 (biased at V_{ref}) avoid any variation of the current density ratio of T_1

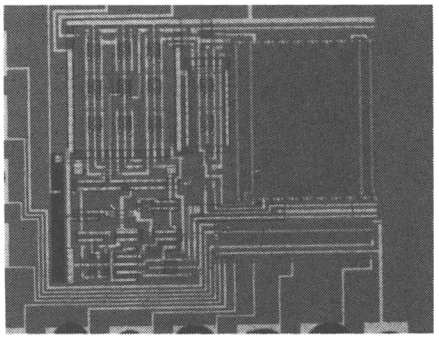


Fig. 8. Chip photograph of low output impedance bandgap reference (area = 0.42 mm^2).

and T_2. The bipolar source follower provides a low output impedance. Output conductance is the product of the transconductance of T_4 and the loop gain.

The circuit has been integrated (Fig. 8) in a low threshold 4-μm-Si gate p-well technology with standard polysilicon resistors of 50 $\Omega/\square$. In order to satisfy the condition given by (5), several taps are placed on the resistor. These taps can be short-circuited outside the chip. In the final version, fixed resistors will be used and the adjustment of the reference voltage will be done by adjusting the current mirror ratio of $M_1 - M_2$ [8]. The lateral bipolar transistors are realized with active transistors (polygate). Their gate voltage is biased at about -2 V.

The most important measurement results are shown in Fig. 9 and summarized in Table I.

The standard deviation of the reference voltage is 5.3 mV before trimming and 150 μV after trimming. The output voltage is constant within 2mV, over the commercial temperature range (0–70°C), when all the circuits of the same batch are trimmed at the same voltage at a single temperature. It can be noted that, at room temperature, some circuits have a positive temperature coefficient while others have a negative one. This means that their ideal reference voltage is, respectively, lower or higher than the voltage on which they are trimmed.

The load regulation is 3.6 μV/μA. The circuit can thus be loaded with approximatively 4 K without loss of temperature accuracy (reference voltage changes only 1 mV).

The PSRR (Fig. 9(b), (c)) of the positive rail is 60 dB at 100 Hz and the PSRR of the bipolar gate voltage is 75 dB at 100 Hz. Long but straightforward calculations show that the PSRR of the positive rail can be further improved by increasing the gain of the folded amplifier formed by T_1, T_2, $M_1 - M_4$, M_7, and M_8. In the actual design, this gain is about 4000. A gain of 8000 would increase the PSRR by 6 dB.

The $1/f$ noise is essentially due to transistors M_1 and M_2. In this design they are only 20 μm by 10 μm and can eventually be made larger to reduce the $1/f$ noise.

The dynamic behavior of the voltage reference is shown in Fig. 9(d). The settling time is smaller than 15 μs with a load of 250 pF. Further reductions can be obtained by

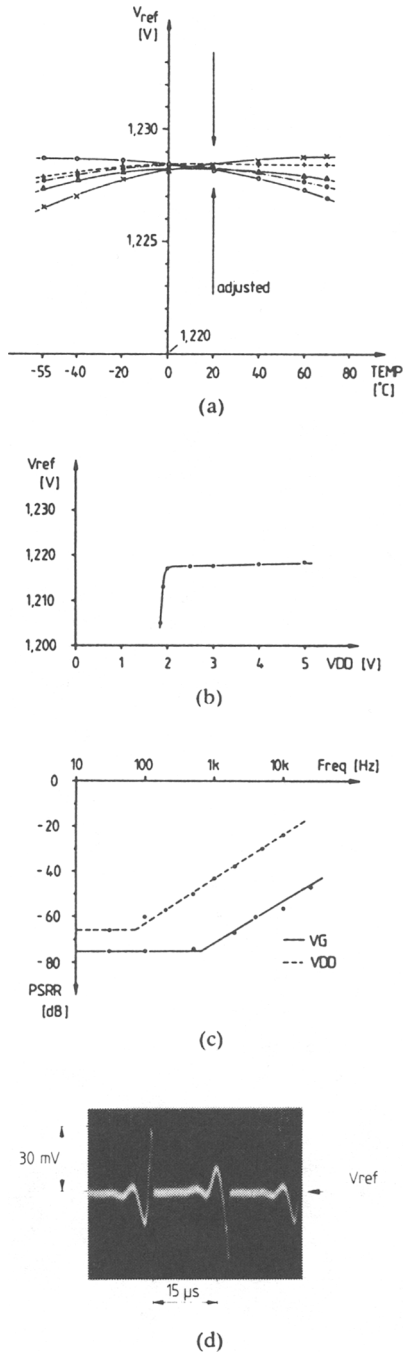


Fig. 9. (a) Temperature behavior of samples of different wafers. Output voltage is trimmed at a single temperature by short circuiting (8 bits) parts of a resistive divider; 90 percent of the measured circuits (30) were as precise as the samples shown. (b) Reference voltage as a function of supply voltage. (c) PSRR of voltage reference. (d) Dynamic behavior of voltage reference (pulsed load capacitance).

increasing the current (use smaller resistances), which is now only 70 μA for the whole reference circuit.

For high supply voltages, the threshold of transistor M_5 becomes large and can eventually push transistor M_7 out of saturation, this degrades circuit performances drastically. The variant scheme shown in Fig. 7(b) is less sensitive to this effect. This circuit has also been integrated and has the same characteristics as the circuit of Fig. 7(a). Only the minimal supply voltage has increased to 2.8 V.

TABLE I
MEASUREMENT RESULTS OF LOW OUTPUT IMPEDANCE VOLTAGE REFERENCE

Output voltage $\bar{x}$	1.2285 V
σ	150 μV
Minimal supply voltage	2.2 V
Supply current	79 μA
Noise spectra	
white	316 nV/$\sqrt{Hz}$
1/f (at 1 kHz)	560 nV/$\sqrt{Hz}$
RMS noise voltage (0.01 - 250 kHz)	162 μV
PSRR at 100 Hz	60 dB
Load regulation ($\Delta V_{out}/I_{out}$)	3.6 μV/μA
Chip area	0.42 mm^2

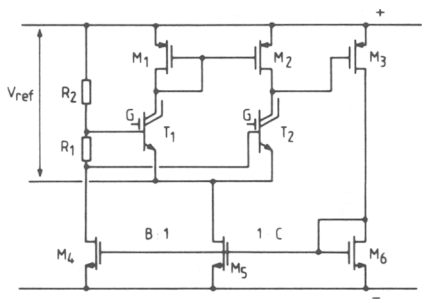


Fig. 10. High PSRR bandgap reference.

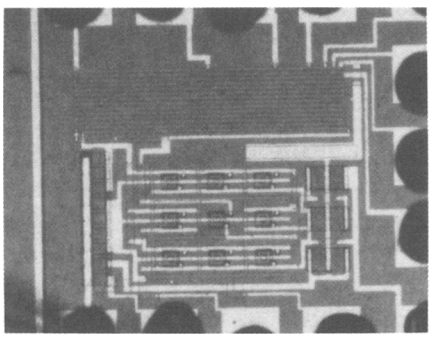


Fig. 11. Chip photograph of high PSRR bandgap reference.

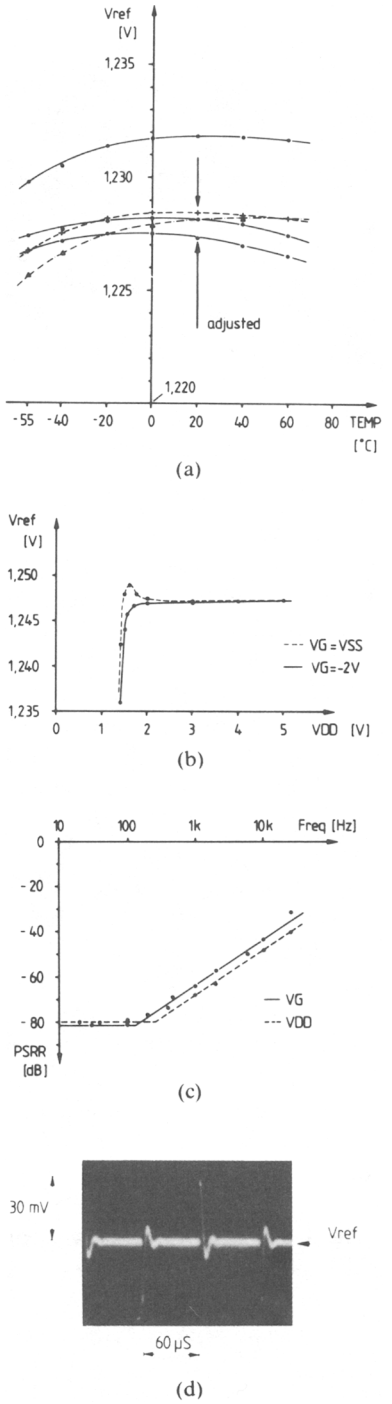


Fig. 12. (a) Temperature behavior of samples of different wafers. Output voltage is trimmed at a single temperature by short-circuiting (5 bits) parts of a resistive divider. (b) Reference voltage as a function of supply voltage. (c) PSRR of voltage reference. (d) Dynamic behavior of voltage reference (pulsed load capacitance).

V. HIGH PSRR VOLTAGE REFERENCE

The drawback of the above described voltage references is that polyresistances have to be used to obtain the mentioned PSRR. Indeed since resistors are fixed with respect to V_{SS}, the ratio of well resistors will be affected by bulk modulation of V_{DD} which will result in a degraded PSRR. Despite their poor temperature behavior, p-well resistances remain interesting to realize, on a small die area, micropower voltage references.

Fig. 10 presents a voltage reference which accepts p-well resistors without degrading PSRR. This circuit, which differs slightly from the principle shown in Fig. 4, is also insensitive to low alfa and beta of the bipolar transistors.

The circuit consists of a two-stage amplifier, with lateral bipolar input transistors, and two feedback resistors. The current through R_1 is designed to be much larger than the maximal base current of T_1. The current density of the transistors T_1 and T_2 is different which causes a PTAT offset of about 54 mV at room temperature. Due to the feedback mechanism, this voltage is multiplied with R_2/R_1.

TABLE II
MEASUREMENT RESULTS OF HIGH PSRR VOLTAGE REFERENCE

Output voltage $\bar{x}$	1.2281 V
σ	350 μV
Minimal supply voltage	1,7 V
Supply current	20 μA
Noise spectra	
white	500 nV/$\sqrt{Hz}$
1/f (at 1 kHz)	1 μV/$\sqrt{Hz}$
PSRR at 100 Hz	77 dB
Load regulation ($\Delta V_{out}/I_{out}$)	4.1 mV/μA
Chip area	0.18 mm^2

The reference voltage appears between the positive rail and the emitter of T_1. Better PSRR is achieved since the potentials of all critical nodes are constant with respect to the substrate. Changes of the supply voltage do not affect the value of resistances R_1 and R_2 and thus do not degrade the PSRR. They only affect the current mirror ratios B and C (see Fig. 10). Straightforward calculations show that the PSRR is inversely proportional to the length of transistors M_4-M_6.

The circuit has been integrated in two different technologies. The most important measurement results of the integration (Fig. 11) in the aforementioned technology are shown in Fig. 12 and summarized in Table II.

The standard deviation of the reference voltage is 7.2 mV before trimming and 350 μV after trimming. The temperature behavior is identical to that of the first circuit presented. In Fig. 12(a) the reference voltage is shown as a function of temperature. The temperature behavior of an untrimmed circuit is also given.

Fig. 12(b) shows the reference voltage as a function of the supply voltage. The full line is for an externally applied gate voltage (V_G) of -2 V and the dotted line is in case the gate of the bipolar transistors is connected to the negative rail.

The PSRR of the negative rail and that of the bipolar gate voltage are almost identical. Measurement yields a PSRR of 77 dB at 100 Hz.

The load regulation ($\Delta V_{ref}/I_{out}$) is 4.1 mV/μA, which is the inverse of the sum of the transconductances of transistors T_1 and T_2.

The dynamic behavior of the voltage reference is shown in Fig. 12(d). The settling time is smaller than 40 μs with a load of 250 pF.

The current consumption of the circuit is 20 μA and is independent of the α of the bipolar transistors. Since the current through T_1 depends on α, transconductance of T_1 as well as the settling time depend on α. In case this is not acceptable, one can insert a lateral bipolar transistor, with base and collector tied together, between resistor R_1 and the drain of M_4; however, power consumption will then depend on α.

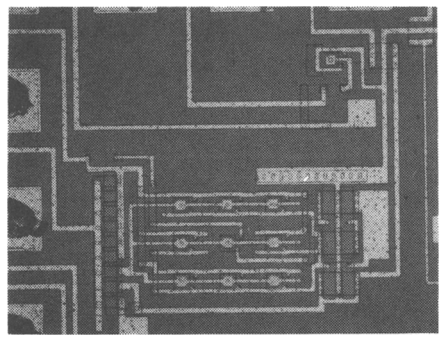


Fig. 13. Chip photograph of high PSRR bandgap reference which makes use of p-well resistors.

The circuit has also been integrated in a 4-μm p-well technology with threshold values of ±1 V. In order to be able to bias the transistors M_1 and M_2 deep in strong inversion, a lateral bipolar transistor was inserted between resistor R_2 and the positive supply voltage (base and collector connected to V_{DD}). Reference voltage appears between V_{DD} and the base of T_1.

The circuit (Fig. 13) was realized with p-well resistors and lateral bipolar transistors formed with parasitic transistors. The PSRR of this circuit is 87 dB (dc). Current consumption is 4.2 μA and the minimal required supply voltage is 2.2 V.

VI. CONCLUSIONS

In this paper, two new bandgap references are presented which make use of CMOS compatible lateral bipolar transistors. The circuits are designed to be insensitive to the possible low beta and alfa current gains of these devices. Their accuracy is not degraded by any amplifier offset.

The circuits have been integrated in three consecutive batches. The experimental results yield a reference voltage which is stable within 2 mV after trimming at a single temperature. The temperature accuracy is limited mainly by technology fluctuations, especially those of the temperature coefficient of resistors.

The first circuit can be used as a voltage regulator since it has a low output impedance. The second circuit permits the use of p-well resistances without any degradation of the PSRR. Both circuits use only a few components, which results in a moderate die area, and deliver a continuous reference voltage.

REFERENCES

[1] Y. Tsividis and R. Ulmer, "A CMOS voltage reference," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 774–778, Dec. 1978.
[2] E. Vittoz and O. Neyround, "A low-voltage CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 573–577, June 1979.
[3] H. Oguey and B. Gerber, "MOS voltage reference based on polysilicon gate work function difference," *IEEE J. Solid-State Circuits*,

vol. SC-15, pp. 264–269, June 1980.
[4] B. S. Song and P. Gray, "A precision curvature compensated CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 634–643, Dec. 1983.
[5] B. Ahuja, W. Baxter, and P. R. Gray, "A programmable CMOS dual channel interface processor," in *ISSCC Dig. Tech. Pap.*, Feb. 1984, pp. 232–233.
[6] E. Vittoz, "MOS transistors operated in the lateral bipolar mode and their application in CMOS technology," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 273–279, June 1983.
[7] Y. Tsividis, "Accurate analysis of temperature effects in $I_C - V_{BE}$ characteristics with application to bandgap reference sources," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1076–1084, Dec. 1980.
[8] H. Oguey, "Référence de tension et détecteur de tension de pile à faible consommation," in *Proc. 57th Swiss Congress of Chronometry* (Montreux, Switzerland), Oct. 1982, pp. 59–63.
[9] R. Ye and Y. Tsividis, "Bandgap voltage reference sources in CMOS technology," *Electron. Lett.*, vol. 18, no. 1, pp. 24–25, Jan. 1982.
[10] J. Michejda and S. K. Kim, "A precision CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 1014–1021, Dec. 1984.
[11] E. Vittoz, "The design of high performance analog circuits on digital CMOS chips," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 657–665, June 1985.

A Low Drift Fully Integrated MOSFET Operational Amplifier

ROBERT POUJOIS AND JOSEPH BOREL, MEMBER, IEEE

Abstract—**A fully integrated MOSFET amplifier with very low drift has been built using standard technology. Input offset voltages as low as 5 μV and drift values of this offset voltage less than 0.05 μV/°C are measured.**

INTRODUCTION

ANALOG amplifiers using standard MOS technology could be thought to have dc performances very different from the bipolar ones mainly as far as drift and offset voltages are concerned. Specific advantages of MOS devices can be used to avoid these limitations [1], [2], and results similar to those obtained in chopper amplifiers can be measured [3], [4]. The basic principle is to use memory capacitances to store the existing defects (either drift, noise in a given bandwidth, or offset voltages) before amplification and refresh these analog data as soon as possible.

The main MOSFET's limitations when these are used in analog amplifiers are the following.

1) Offset voltages as high as 50 mV are observed and drift voltages in the range of 50 μV/°C are measured.

2) Noise voltage levels of 250 μV peak are typical.

These values are related to the Si-SiO_2 interface properties and can be decreased slightly by improved SiO_2 technology.

Other limitations are in the electrical properties of analog amplifiers built in standard MOS technologies.

1) High values of voltage gain per stage are difficult to obtain and a value of 3 is typical. It follows that a large number of stages (five or six) must be used, giving additional phase shift.

2) Maximum gain-bandwidth products of 10 MHz can be achieved, and increased values can only be obtained with new technologies (self-registered gates, depletion enhancement technology, SOS technology, etc.).

Manuscript received January 13, 1978.

The authors are with the Laboratoire de Microelectronique Appliquee, C.E.A.-C.E.N.G., Grenoble, France.

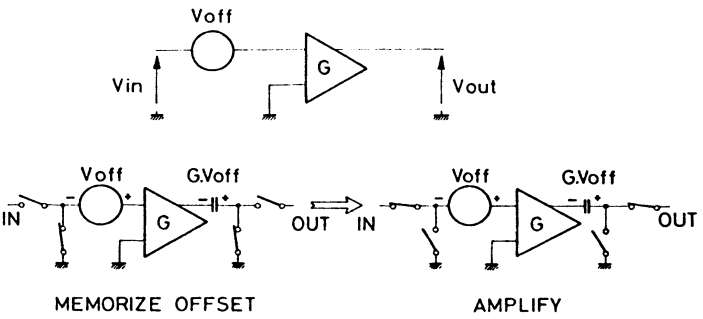


Fig. 1. Offset memorization principle.

On the other hand, MOSFET devices have specific advantages that can be used in analog amplifiers.

1) On chip compatibility with LSI logic circuits leads to complex arrays where digital and analog signals are processed.

2) Refresh memory capabilities allow one to store analog values of defects like offset voltages.

3) No offset current is seen due to the very high input impedance of the SiO_2 gate layer.

4) Fully integrated amplifiers without external components can be built.

5) Low cost of the MOS technology is a very attractive argument for standard products.

We present some results obtained on analog MOSFET amplifiers (chopper amplifiers) using the general considerations given above.

AMPLIFIERS WITH OFFSET VOLTAGE MEMORIZATION

The basic principle is to store the analog offset voltage in an MOS capacitance and to use this offset voltage to compensate for the amplifier input offset voltage. The circuit configuration is given Fig. 1 where we have represented the offset mem-

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-13, no. 4, pp. 499–503, Aug. 1978.

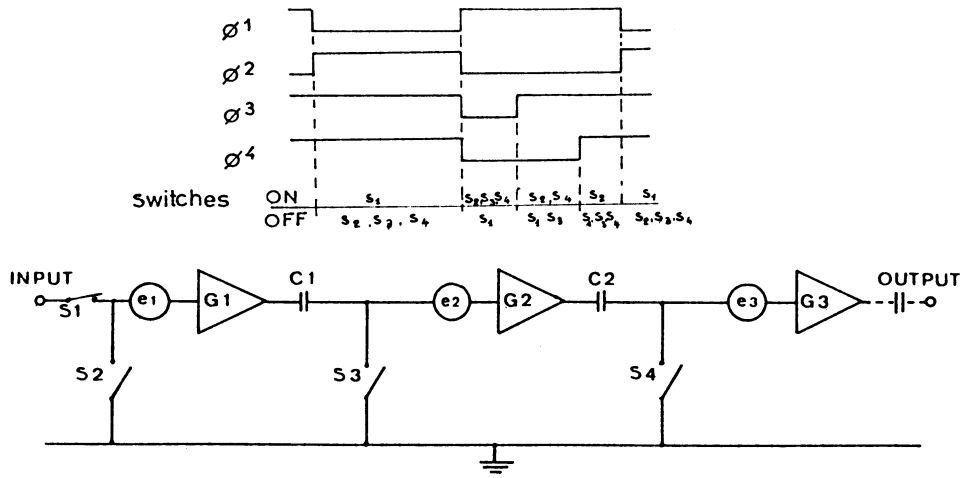


Fig. 2. Residual voltage successive memorization (RSM) amplifier.

orization principle. An actual amplifier has an offset voltage V_{off} due to the mismatch of the electrical parameters of the input devices.

For one elementary amplification stage, we use four switches and a storage capacitance as seen in the bottom part of Fig. 1.

During memorization time, input node and output node through the storage capacitance are grounded. In this capacitance a value $(G \cdot V_{off})$ is memorized. We must avoid saturation of the amplifier by the $(G \cdot V_{off})$ value at the output. So a limited value of the voltage gain per elementary stage is allowed $(G \cdot \max \leqslant 100)$. Several elementary stages in series are needed for a high voltage gain amplifier.

During the next sequence, the amplifier configuration is changed and the input and output nodes are connected in the signal path for amplification with a low residual input offset voltage. The sequence is repeated at the clock rate (16 kHz in our case), allowing us to use dc amplifiers (no large area connecting capacitances).

Storage capacitances must have high values compared to switch capacitances to avoid parasitic offset signals due to switching. A MOSFET switch presents a mean parasitic capacitance C_p of 0.1 pF between gate and channel and a memory capacitance C_m of 10 nF will be necessary if offset voltages ϵ as low as 100 μV are wanted. For a $V = 10$ V clock voltage on the gate of the switch transistor, we have

$$\epsilon = V \cdot \frac{C_p}{C_m} \simeq 10\ \text{V} \times \frac{0.1}{10^4} = 100\ \mu\text{V}.$$

A differential configuration allows us to compensate for the switching parasitics and leakage currents of the switches and of memory capacitances. Only the mismatch between these elements is to be considered, allowing the use of smaller values for the memory capacitances (compatible with integration).

Residual Voltage Successive Memorization (RSM) Amplifier

This technique uses integrated storage capacitances as low as 4 pF and memorizes parasitic pulses as well as circuit defects. The elementary stages are sequentially compensated

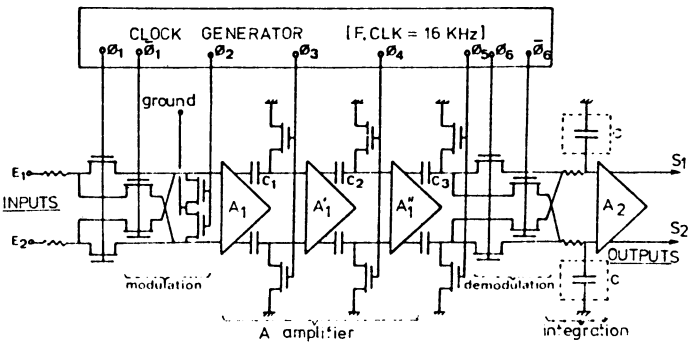


Fig. 3. Block diagram of the RSM amplifier.

starting from the input of the amplifier. Fig. 2 is a schematic view of the circuit with clock voltage waveforms: if switches 2, 3, and 4 are closed and switch 1 is opened, C_1 and C_2 are, respectively, charged with $G_1 \cdot e_1$ and $G_2 \cdot e_2$. These values must not saturate the amplifiers (relatively low values of G_1 and G_2).

If switch 3 is then opened, giving an extra contribution ϵ_2 to the offset voltage e_2, this can be compensated for in capacitance C_2. So the net input offset voltage comes only from the input offset voltage (and the parastic input clock pulse) of the last amplifier stage. The input offset voltage is then given by

$$V_{off} = (e_n + \epsilon_n)/(G_1 \times G_2 \cdots \times G_{n-1})$$

and can be decreased by increasing the overall voltage gain. e_n and ϵ_n are, respectively, the input offset voltages of the nth amplifier stage and parasitic clock pulse at the input of this amplifier stage. G_1-G_{n-1} are the voltage gains of the various stages. Using such considerations, an MOS integrated amplifier has been built with integrated clock generator and modulation-demodulation stages to compensate for the parasitic signals coming from switches 1 and 2 at the input. Standard p channel aluminum gate technology is used.

Fig. 3 is a block diagram of the RSM amplifier. It has five basic blocks.

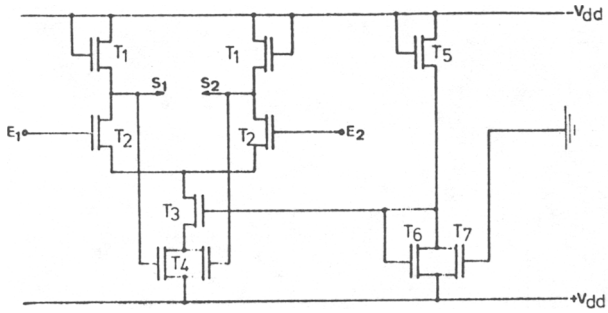


Fig. 4. Basic differential amplifier with compensated bias.

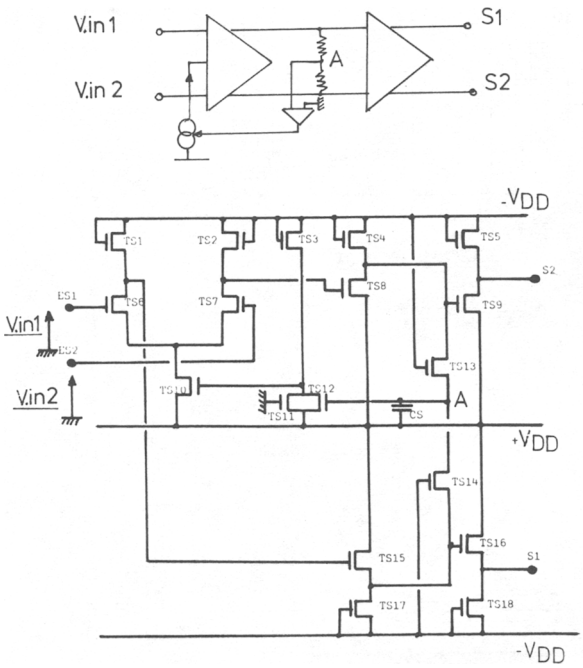


Fig. 5. Output amplifier stage diagram.

1) The amplifier A: it is a three-stage differential RSM amplifier. All signals are amplified differentially to compensate for the parasitics coming from the memory switches.

2) The modulation-demodulation system: it is built to eliminate very low frequency drift input signals (mainly from thermal origin) and parasitic pulses from ϕ_1, $\overline{\phi_1}$, and ϕ_2.

3) The clock generator working at a clock frequency of 16 kHz.

4) The integration circuit working with two external capacitances C.

5) The output stage A_2 with a voltage gain of 200.

The crossed modulation-demodulation system is able to amplify very low frequency signals without any drift (input signals are sequentially switched toward the two inputs of A). Moreover, amplifier A is working at the clock frequency and the $1/F$ noise spectrum is lowered. The chopper amplifier behaves like a high-pass filter for the noise (see Fig. 8, for example).

Fig. 4 is a diagram of one elementary stage of the amplifier (A_1, A_1', or A_1''). The main features are

1) common mode rejection through $T4$,

2) output level stabilization ($S1$ and $S2$) versus supply voltage variations ($T7$), threshold shifts ($T6$), and geometrical spreads of devices ($T5$).

Fig. 5 shows a block diagram and a complete configuration of amplifier A_2. A feedback loop is used to obtain a virtual ground in A, resulting in symmetrical dc output voltages.

Fig. 6 is a view of the RSM amplifier including clock generation and amplifier stages—chip size is 3.1 mm X 2.2 mm.

The next figures give the measured performances of the amplifier.

1) Fig. 7 is a plot of open loop voltage gain versus frequency. Gain-bandwidth products of 4 MHz are achieved.

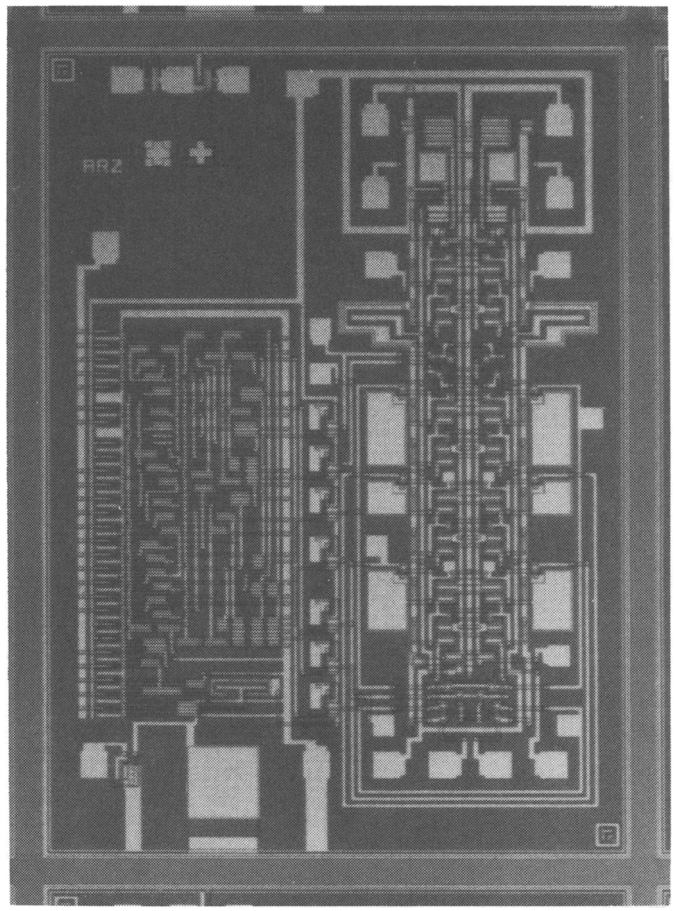


Fig. 6. Amplifier chip with clock generation.

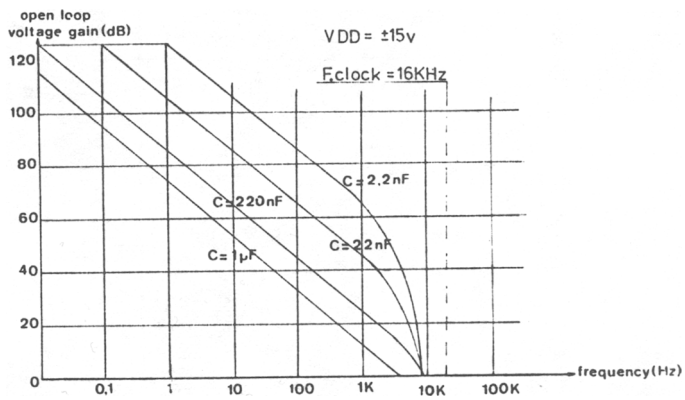


Fig. 7. Amplifier open loop frequency response.

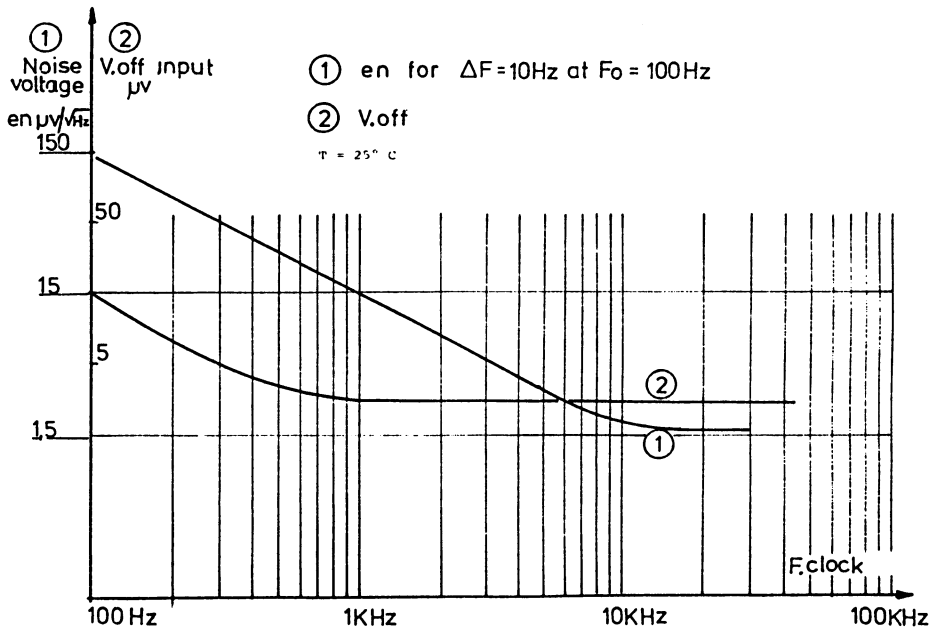


Fig. 8. Input offset voltage and input noise voltage versus frequency.

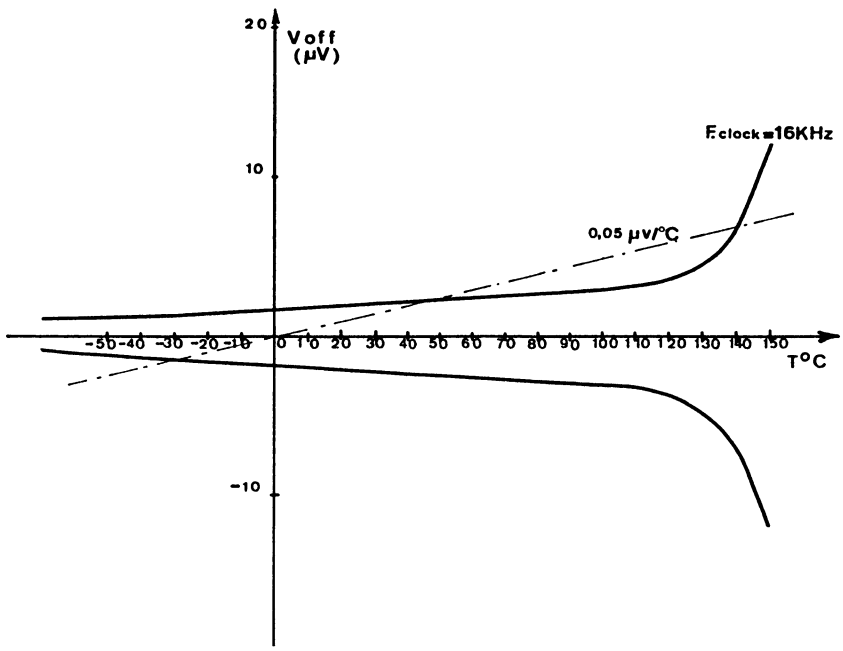


Fig. 9. Input offset voltage versus temperature.

Values of integration capacitances C are given as a parameter. Cutoff frequencies are limited below F clock/2 (Nyquist limit).

2) Fig. 8 is a plot of input offset voltage V_{off} and input noise voltage versus clock frequency. An offset voltage of 5 μV is reached and is independent of clock frequency in a wide range. Narrow-band noise voltage (ΔF = 10 Hz bandwidth) at central frequency F_0 = 100 Hz is a function of clock frequency showing evidence of $1/f$ noise contribution from the input devices.

3) Fig. 9 gives an example of the behavior of offset voltage versus temperature for a 16 kHz clock frequency. It is clearly seen that at higher temperatures leakage current compensation is not good, and higher clock frequencies should be necessary.

4) Fig. 10 is a typical spread of input offset voltages measured on the amplifier of Fig. 6. This corresponds to input offset voltages spread either for chips on the same wafer or from different wafers. Most of the input offset voltages are within 5 μV.

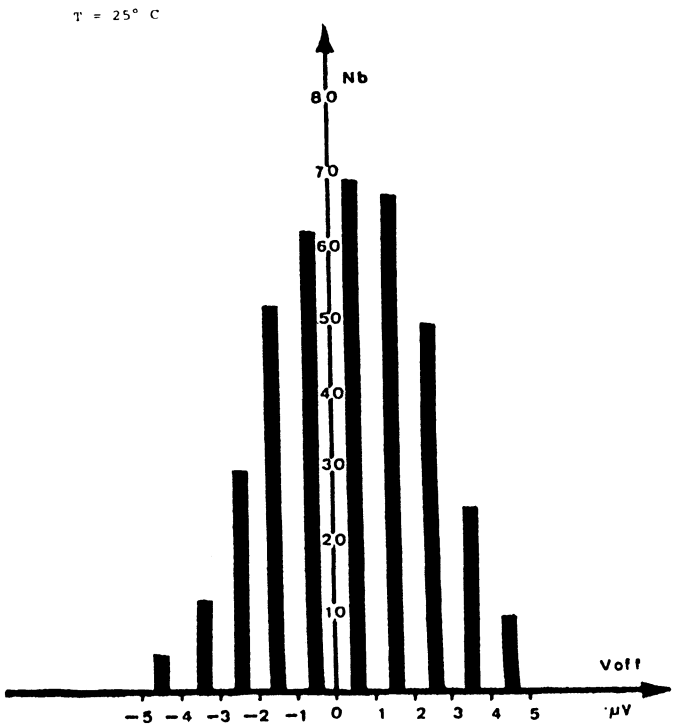


Fig. 10. Typical spread of input offset voltages on various chips coming from different wafers.

TABLE I
RSM AMPLIFIER CHARACTERISTICS (AMBIENT TEMPERATURE)

voltage gain	2 000 000
input offset voltage	<5 µV
drift voltage	<0.05 µV/°C
input noise voltage	2.5 $\mu V/\sqrt{Hz}$
supply voltage	±10 V to ±18 V
power consumption	240 mW (V_a = ±15 V)
common mode rejection ratio	120 dB
supply voltage rejection ratio	>120 dB
external components	2 capacitors

CONCLUSION

We have presented a new technique to compensate for offset voltages and drift voltages in analog amplifiers using standard MOSFET technology. Amplifiers with improved performances are obtained, and their association with digital MOSFET circuits is a powerful tool in custom integrated circuit design. These amplifiers can be used as elementary blocks in more complex analog or digital arrays or as standard products.

Some performance still needs to be improved for specific applications such as input noise level and maximum operating frequency.

In summary, we present in Table I measured results that have been obtained with the integrated amplifier of Fig. 6.

Let us only point out the very high value of supply voltage rejection ratio observed in the case of RSM amplifiers. This is due to the internal memorization of supply voltage shifts.

Best frequency responses are expected with depletion enhancement self-registered technologies or with SOS technology, and lower noise levels can be obtained when optimizing input stages.

ACKNOWLEDGMENT

Thanks are due to J. M. Ittel for making the measurements, to D. Barbier for valuable discussions, and to E. Mackowiak for a critical reading of the manuscript.

REFERENCES

[1] R. Poujois, B. Baylac, D. Barbier, and J. M. Ittel, "Low level MOS transistor amplifier using storage techniques," in *ISSCC 73*, pp. 152-153.

[2] R. Poujois, J. M. Ittel, and J. Borel, "A low drift MOSFET operational amplifier: A.R.Z.," presented at ESSCIRC, Toulouse, France, Sept. 1976.

[3] J. L. Villevieille, "Amplificateur opérationnel stabilisé par chopper," in *Texas Instruments Seminar Dig.*, 1974, p. 177.

[4] I. Hackel and H. Hagemann, "Construction of chopper amplifiers," *Elektron. Reinschan* (Germany), vol. 16, pp. 509-512, Nov. 1962.

Large Swing CMOS Power Amplifier

KEVIN E. BREHMER, MEMBER, IEEE, AND JAMES B. WIESER

Abstract—A CMOS class *AB* power amplifier is presented wherein supply-to-supply voltage swings across low impedance loads are efficiently and readily handled. The amplifier consists of a high gain input stage and a push–pull unity gain amplifier output stage. The amplifier dissipates only 7 mW of dc power and delivers 36 mW of ac power to a 300 Ω load, using standard power supplies of ±5.0 V. Lower impedance loads can be driven to higher power levels, providing the internal current limiting level is not exceeded.

I. Introduction

During the past few years, CMOS has emerged as an industry standard because of its low power dissipation. However, the implementation of analog functions in MOS has presented a challenge to many circuit designers. One area of MOS analog design where considerable work is taking place is the design of an efficient, large dynamic range power amplifier [1], [2].

Prior art MOS power amplifiers used output stage configurations that were subject to various limitations. Such limitations included the size of the output driver devices and the control of the dc bias current in these output drivers. In order to control the dc bias current in the output driver devices, previous designs used a source follower device and a controlled current sink as an output stage [1], [2]. This type of design has dynamic range limitations and requires large output driver device sizes. By replacing the source follower output driver with a bipolar emitter follower, transistor die area and dynamic range limitations can be reduced considerably. However, instabilities arise when using a bipolar emitter follower to drive high capacitive loads, thus limiting its application. Also, bipolar transistors are parasitic devices in a standard CMOS process which are not necessarily well controlled.

In order to obtain an efficient power amplifier with a large dynamic range capable of driving a low impedance load, a push–pull class *AB*, fully CMOS power amplifier is presented. This amplifier is designed to operate at voice-band frequencies for telecommunication and audio applications, and to be used in an inverting configuration so that users can design gain or attentuation into their systems.

Manuscript received July 11, 1983; revised July 26, 1983.
The authors are with the Telecom Group, National Semiconductor, Santa Clara, CA 95051.

II. Description of the Power Amplifier

The complete power amplifier consists of a high gain input stage driving a unity gain push–pull output stage. The input stage is comprised of a differential amplifier and a common source amplifier. Compensation of this stage is achieved by a Miller multiplied feedback capacitor.

The output stage includes two unity gain amplifiers in push–pull configuration. Each amplifier contains a differential input stage whose output controls the gate of the output driver device. The drain of the output driver device is directly fed back to the noninverting input of the differential stage to form a noninverting unity gain amplifier, and one half of the push–pull configuration.

In the event of an offset between the two push–pull amplifiers, a feedback circuit controls the dc bias current in the output drivers. Fig. 1 shows a pseudo block diagram form of the power amplifier and the associated feedback circuitry required to stabilize the dc bias current in the output driver devices.

A. Output Stage Analysis

From Fig. 1 we can see that amplifier *A*1 and transistor *M*6 form the unity gain amplifier for the positive half of the output voltage swing, and conversely amplifier *A*2 and transistor *M*6*A* form the negative half cycle circuit. For the sake of simplicity, only the circuit referring to the positive half output swing of the output stage will be discussed. The operation of the negative half circuit is an inverted mirror image of that of the positive half swing circuit. Components performing similar functions in each circuit are designated with an additional letter "A" for the negative half circuit.

Shown in Fig. 2 is a detailed schematic of the positive unity gain amplifier. The differential amplifier input stage of the unity gain amplifier has a large positive common mode range (CMR), which allows transistor *M*6 to source large amounts of current to the load, while still being of a reasonable physical size for incorporation into a monolithic circuit. Large current sourcing is provided by producing the highest possible gate drive on *M*6. The maximum gate-to-source voltage *M*6 can have while still keeping *M*1 and *M*2 in the saturation region is given by

$$V_{GS6_{max}} = -(V_{CC} - (V_{IN} - V_{GS1} + V_{DSAT1})) \qquad (1)$$

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-18, no. 6, pp. 624–629, Dec. 1983.

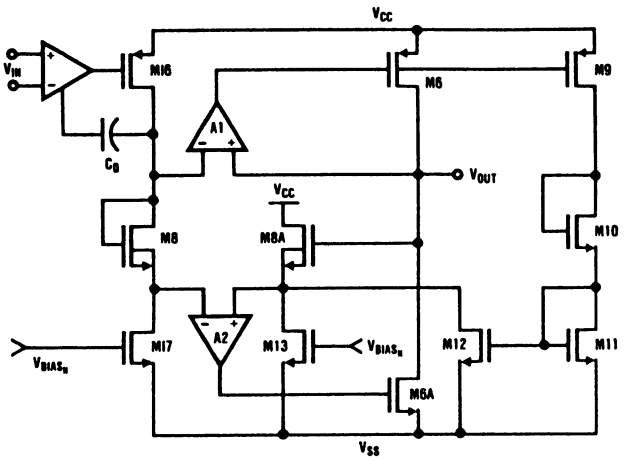


Fig. 1. Block diagram of power amplifier.

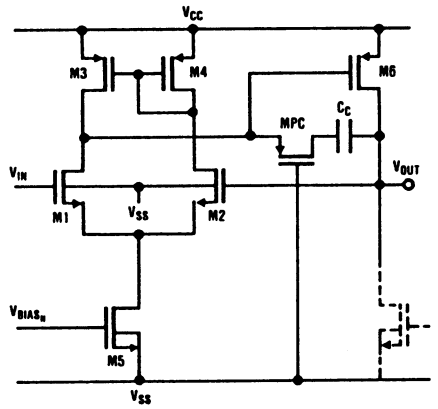


Fig. 2. Positive output stage unity gain amplifier.

It first appears that increasing V_{GS1} by increasing the overdrive voltage of $M1$ and $M2$ will increase the gate drive of $M6$. However, after substitution of the saturation equation (2) into the expression for V_{GS6max}, we see that this is not the case:

$$V_{DSAT} = V_{GS} - V_T \tag{2}$$

$$V_{GS6_{max}} = -(V_{CC} - V_{IN} + V_{T1}). \tag{3}$$

From (3), maximum gate drive on $M6$ is provided by increasing the threshold voltage (4) of $M1$ and $M2$. One method of increasing the threshold voltage of $M1$ and $M2$ is to modify V_{T0} by a threshold implant. Implanting devices $M1$ and $M2$ to produce a higher V_{T0} pushes the common source voltage lower, allowing the gate of $M6$ to drop more while still keeping $M1$ and $M2$ in saturation.

$$V_T = V_{T0} + \gamma\sqrt{V_{BS} - 2\varphi_F}. \tag{4}$$

Further enhancement of the positive CMR is obtained by connecting the substrate of transistors $M1$ and $M2$ to V_{SS}, thereby modulating the source–substrate voltage of these transistors. The effect of this substrate modulation on the threshold voltage of $M1$ and $M2$ can be seen in (4). As the output swing increases, the common source voltage also increases, however, not by the same amount, since the threshold voltage of $M1$ and $M2$ has increased due to substrate modulation. The increased threshold voltage of $M1$ and $M2$ tends to reduce the common source mode voltage of the differential pair thus driving the gate of $M6$ more negative while still keeping $M1$ and $M2$ in saturation.

The current in the output driver device $M6$ is typically controlled by the current mirror developed in the differential amplifier of the positive unity gain amplifier and matches the current set in the negative output driver device $M6A$ by the negative unity gain amplifier. If an offset occurs between amplifiers $A1$ and $A2$, the current balance between output drivers $M6$ and $M6A$ no longer exists and either massive amounts of current or no current at all will flow through these devices. The feedback loop, shown in Fig. 1, consisting of transistors $M8A-M13$, stabilizes the current through output drivers $M6$ and $M6A$ in the event of an offset between amplifiers $A1$ and $A2$. The feedback loop operates as follows. Assume that amplifier $A1$ has an offset such that transistor $M6$ begins to source excessive amounts of current. The excessive current is sensed by transistor $M9$ and is fed back to the source follower $M8A-M13$. The increase of current provided to transistor $M8A$ produces a greater voltage drop across the source follower $M8A-M13$ and more differential signal on the input of amplifier $A2$. The larger differential signal on amplifier $A2$ results in lower gate drive on output driver $M6A$, thereby reducing the current in the output drivers $M6-M6A$. The output voltage now has increased due to the fact that the positive swing amplifier $A1$ attempts to keep both of its inputs at the same potential. The complete power amplifier is in a feedback loop, wherein amplifier feedback drops the voltage of the negative input of amplifier $A1$ in attempting to keep the output of the complete power amplifier at 0 V in the dc bias condition. Transistor $M8$ transfers this voltage drop to the negative input of amplifier $A2$, thus balancing offset of amplifier $A2$. The offset that was initially introduced by amplifier $A1$ is absorbed by the source follower transistor $M8A$.

Because the output stage current feedback is not unity gain, some current variation in transistors $M6$ and $M6A$ occurs. Offsets between amplifiers $A1$ and $A2$ can produce a 2:1 variation in dc current over temperature and process variations. Equation (5) predicts the change in the output driver current assuming that V_{out} is at ground, and any offset between amplifiers $A1$ and $A2$ can be reflected as a difference between the inputs of amplifier $A1$. From this equation, V_{off1} must be zero for ΔI_0 to be zero.

$$\Delta I_0 = -gm_{6A}A_2\left[V_{off1} - \sqrt{\frac{2\beta_9\beta_{12}}{\beta_{8A}\beta_6\beta_{11}}}\left[\sqrt{I_{B1}\left(\frac{\beta_6\beta_{11}}{\beta_9\beta_{12}} + \frac{1}{2}\frac{\beta_5\beta_6}{\beta_7\beta_3}\right) + \Delta I_0} - \sqrt{I_{B1}\left(\frac{\beta_6\beta_{11}}{\beta_9\beta_{12}} + \frac{1}{2}\frac{\beta_5\beta_6}{\beta_7\beta_3}\right)}\right]\right] \tag{5}$$

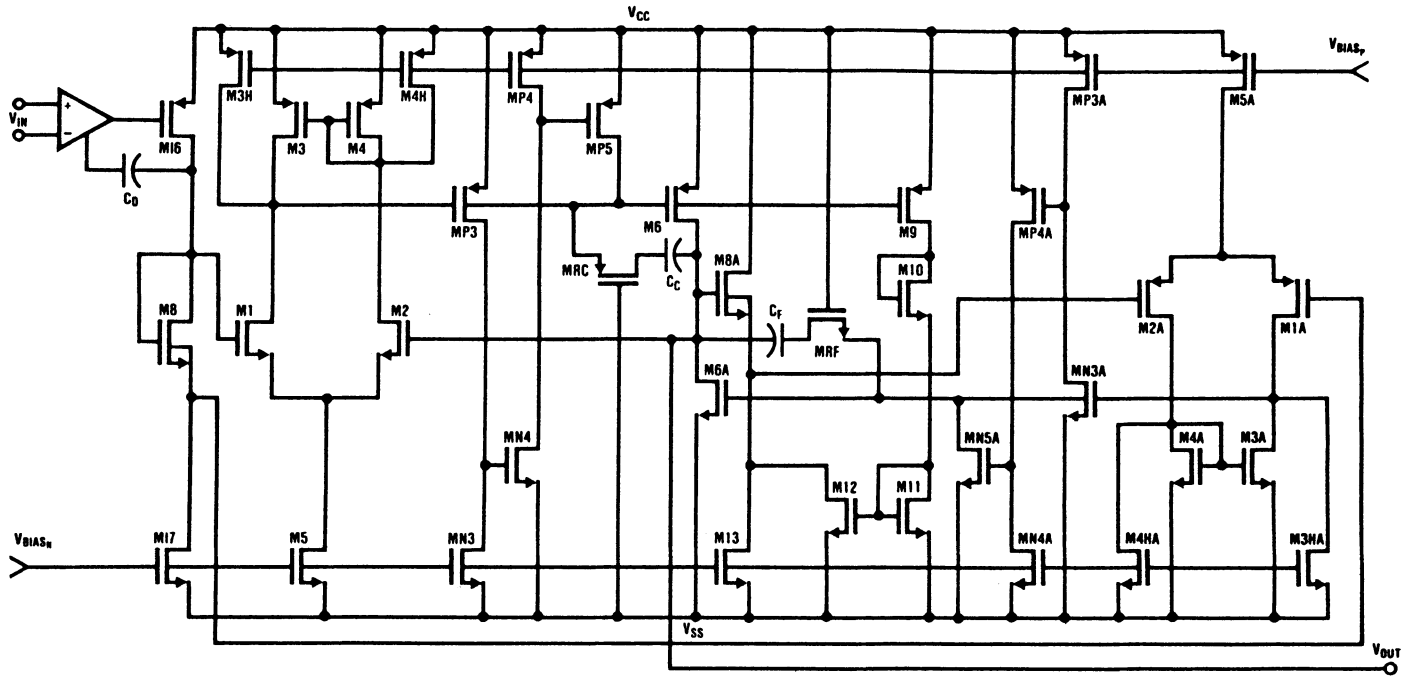


Fig. 3. Complete power amplifier schematic.

where

$$\beta = \frac{\mu_{\text{eff}} C_{\text{ox}}}{2}\left(\frac{W}{L}\right)_{\text{eff}} \text{ and } I_{B1} = I_{MI7}.$$

Since transistor $M6$ can supply large amounts of current, care must be taken to ensure that this transistor is off during the negative half cycle of the output voltage swing. For large negative swings, the drain of transistor $M5$ pulls to V_{SS}, turning off the current source that biases the differential amplifier $A1$. As the bias is turned off, the gate of transistor $M6$ floats and tends to pull towards V_{SS}, turning on transistor $M6$.

Shown in Fig. 3 is circuitry which ensures that transistor $M6$ remains off for large negative voltage swings. As transistor $M5$ turns off, transistors $M3H$ and $M4H$ pull up the drains of transistors $M3$ and $M4$, respectively. As a result, transistor $M6$ is turned off and any floating nodes in the differential amplifier are eliminated. Positive swing protection is provided for the negative half cycle circuit by transistors $M3HA$ and $M4HA$, which operate in a manner similar to that described above for the negative swing protection circuit. The swing protection circuit does, however, degrade the step response of the power amplifier since the unity gain amplifier not in operation is completely turned off.

Short circuit protection is also included in the design of the amplifier. From Fig. 3, we can see that transistor $MP3$ senses the output current through transistor $M6$, and in the event of excessively large output currents, the biased inverter formed by transistors $MP3$ and $MN3$ trips, thus enabling transistor $MP5$. Once transistor $MP5$ is enabled, the gate of transistor $M6$ is pulled up towards the positive supply V_{CC}, thus limiting the current transistor $M6$ source to approximately 60 mA.

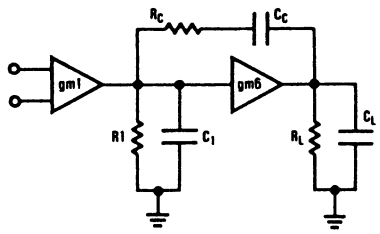


Fig. 4. RC compensation for the output stage unity gain amplifier.

B. AC Compensation of the Power Amplifier

AC stability of the complete power amplifier is achieved by providing a wide-band output stage and by using compensation at the input stage to produce the dominant pole. The dominant pole in the input stage is produced by the Miller multiplied capacitor C_D. The compensation of each unity gain amplifier in the output stage is achieved by a Miller multiplied capacitor and feedforward resistor. From Figs. 2 and 3, we can see that transistor MRC and capacitor C_C comprise the RC network for the positive unity gain amplifier and transistor MRF and C_F comprise the RC network for the negative unity gain amplifier. To simplify the calculation of the open loop transfer function of Fig. 2, a model as shown in Fig. 4 has been developed [1]. In this figure, the effective output impedance of the first stage is represented by resistor R_1 and capacitor C_1, while the effective output impedance of the second stage is represented by resistor R_L and capacitor C_L. From this model, it can be shown that the poles and zero of each unity gain amplifier before closing the feedback loop are

$$P_1 \simeq \frac{-1}{gm_2 R_L R_1 C_C}\text{; for high } R_L$$

$$\simeq \frac{-1}{gm_2 R_L R_1 C_C + R_1(C_1 + C_C)}\text{; for low } R_L$$

$$P_2 \simeq \frac{-gm_2 C_C}{C_L(C_1 + C_C) + C_1 C_C}\text{; for high impedance load}$$

$$\simeq \frac{-(gm_2 + g_2)C_C}{C_C(C_1 + C_C) + C_1 C_C}\text{; for low impedance load}$$

$$P_3 \simeq \frac{-1}{R_C C_1}$$

$$Z \simeq \frac{-1}{R_C C_C - \dfrac{C_C}{gm_2}}.$$

Note that the pole splitting between $P1$ and $P2$ is a function of the load resistance and load capacitance. For low resistive loads the open-loop gain of the positive swing amplifier drops, and $P1$ and $P2$ both move out in frequency. Since both $P1$ and $P2$ move out in frequency, phase degradation could occur depending on the location of $P3$; however, by bringing the zero shown above, into the left half plane, some phase shift that occurs can be cancelled. The zero placement also helps to cancel phase shift that will occur as a result of capacitive loading on the output. As the output capacitance increases, $P2$ decreases and phase margin degradation occurs; however, by careful placement of the zero, the reduction in the phase margin can be minimized.

The ac stability of the total output stage can be modeled, to a first order, as two independent amplifiers in parallel. Since the negative unity gain amplifier is an inverted mirror image of the positive unity gain amplifier with equal drive requirements, the dominant poles and zeros of each amplifier are approximately the same, and the complete output amplifier transfer function simplifies to that of a half cycle stage. The number and location of the poles and zeros of the complete output stage are identical to those in each of unity gain amplifiers; therefore, no additional compensation is required to stabilize the complete output stage.

The total amplifier circuit is capable of driving 300 Ω and 1000 pF to ground. The gain–bandwidth product is approximately 500 kHz and is limited by the output stage 1000 pF load requirement. The output stage bandwidth is approximately 1.0 MHz.

III. Experimental Results

The power amplifier presented was fabricated using National Semiconductor's proprietary P^2CMOS process. A die photo of the power amplifier is shown in Fig. 5. The total die area of the power amplifier is 1500 mils^2.

Table I shows a comparison between the simulated and measured results of the power amplifier and Table II shows the device sizes that correspond to Fig. 3. Since the ampli-

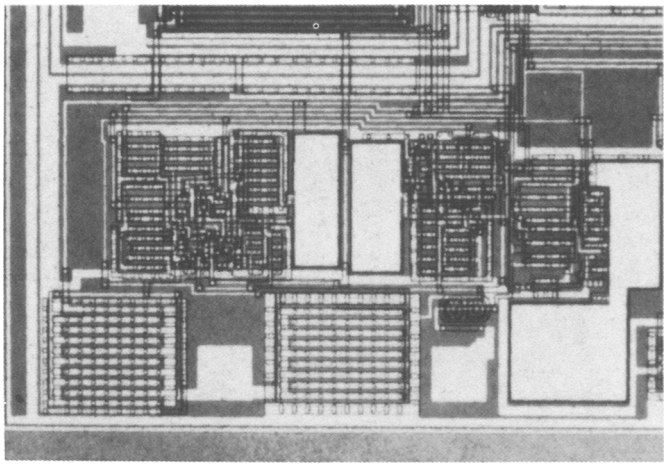


Fig. 5. Die photo of power amplifier.

TABLE I
Power Amplifier Performance

Parameter	Simulation	Measured Results
Power dissipation	7.0 mW	5.0 mW
A_{vol}	82 dB	83 dB
F_u	500 kHz	420 kHz
V_{offset}	0.4 mV	1 mV
PSRR+ (dc)	85 dB	86 dB
(1 kHz)	81 dB	80 dB
PSRR− (dc)	104 dB	106 dB
(1 kHz)	98 dB	98 dB
THD $V_{IN} = 3.3$ V_p $R_L = 300$ Ω	0.03%	0.13%(1 kHz)
$C_L = 1000$ pF	0.08%	0.32%(4 kHz)
$V_{IN} = 4.0$ V_p $R_L = 15$K Ω	0.05%	0.13%(1 kHz)
$C_L = 200$ pF	0.16%	0.20%(4 kHz)
T settling (0.1%)	3.0 μs	< 5.0 μs
Slew rate	0.8 V/μs	0.6 V/μs
$1/f$ noise at 1 kHz	N/A	130 nV/Hz
Broad-band noise	N/A	49 nV/Hz
Die area		1500 mils^2

TABLE II
Component Sizes
(μm,pF)

MI6	184/9	*M8A*	481/6
MI7	66/12	*M13*	66/12
M8	184/6	*M9*	27/6
M1, M2	36/10	*M10*	6/22
M3, M4	194/6	*M11*	14/6
M3H, M4H	16/12	*M12*	140/6
M5	145/12	*MP3*	8/6
M6	2647/6	*MN3*	244/6
MRC	48/10	*MP4*	43/12
CC	11.0	*MN4*	12/6
M1A, M2A	88/12	*MP5*	6/6
M3A, M4A	196/6	*MN3A*	6/6
M3HA, M4HA	10/12	*MP3A*	337/6
M5A	229/12	*MN4A*	24/12
M6A	2420/6	*MP4A*	20/12
MRF	25/12	*MN5A*	6/6
CF	10.0		

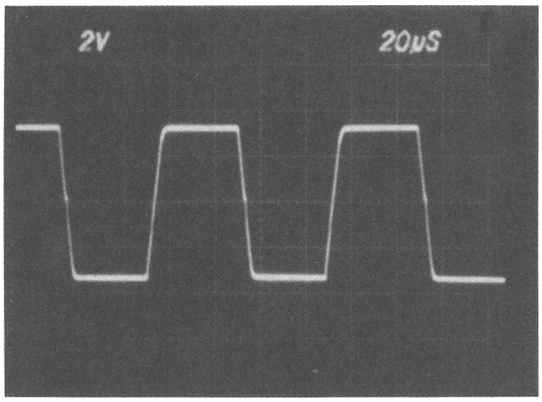


Fig. 6. Output time response for a ± 3.3 V_P input pulse.

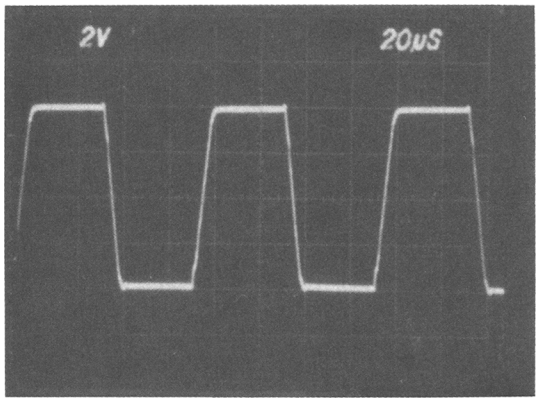


Fig. 7. Output time response for a ± 4.0 V_P input pulse.

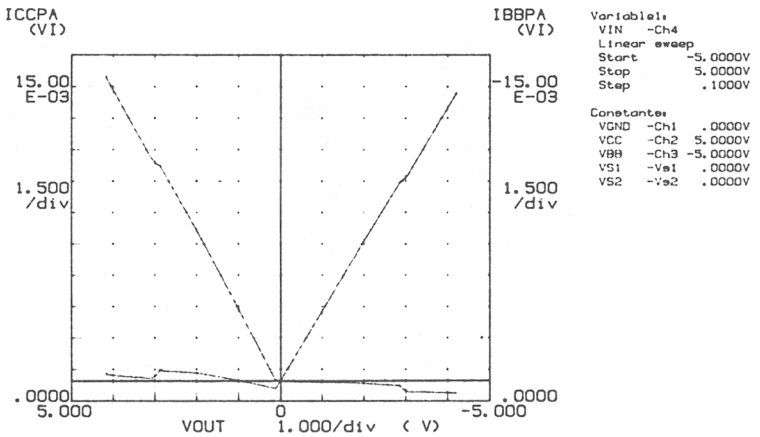


Fig. 8. Power supply current versus output voltage level for $R_L = 300\ \Omega$.

fier was designed to be used in an inverting configuration, the common-mode range and CMRR are not applicable op amp parameters.

Shown in Figs. 6 and 7 are step responses of ± 3.3 V_P and ± 4.0 V_P for loads of 300 Ω/1000 pF and 15K Ω/200 pF, respectively, using ± 4.75 V supplies. The slight cross-over distortion, as seen in Fig. 6, while slewing negatively, is a result of the delay caused by the offset feedback circuit when amplifier $A2$ is required to drive large amounts of current to the load. This distortion is negligible in THD measurements for telecommunication and audio applications.

Fig. 8 shows the efficiency of the power amplifier as the output voltage swings from rail-to-rail. For this case the power supplies used were ± 5.0 V. The slight discontinuity in the supply current curves at approximately ± 3.0 V is a result of the opposite unity gain amplifier turning off in the output stage. This discontinuity in supply current has no effect on the distortion seen in the amplifier. The dc operating current at 0 V is the operating current for two power amplifiers since the application in which the power amplifiers is to be used requires a differential signal. The dc operating current per amplifier is 500 μA. The ICCPA current in the positive supply is greater due to the fact that the high current $M6$ source is mirrored around to the source follower $M8A$ by the output driver current offset feedback circuit.

Amplifier offset voltages of typically 1 mV were measured. These low offsets were a result of a good process and careful layout. From these data and the fact that amplifier feedback will reduce any offset between the output unity gain amplifiers, the dc current variation in the output driver devices should be very small.

IV. Conclusion

Presented was a fully CMOS class AB power amplifier wherein supply-to-supply voltage swings across low impedance loads are efficiently and readily handled. The amplifier dissipates only 7 mW of dc power and can deliver 36 mW of ac power to a 300 Ω load using ± 5.0 V power

supplies. Other features of the design include typical offset voltages of 1 mV and THD of less than 0.4 percent.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of C. Laber to the design of the presented power amplifier.

REFERENCES

[1] W. C. Black, Jr., D. J. Allstot, and R. A. Reed, "A high performance low power CMOS channel filter," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 921-929, Dec. 1980.

[2] D. Senderowicz, D. Hodges, and P. Gray, "High-performance NMOS operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 760-766, Dec. 1978.

A High-Performance CMOS Power Amplifier

JOHN A. FISHER, MEMBER, IEEE

Abstract—A high-performance CMOS power amplifier consisting of a new input stage especially suited to power amplifier applications and a variation on a class *AB* output stage is presented which has been fabricated using a conventional silicon gate p-well process. The configuration results in several performance improvements over previously reported high-output current amplifiers without requiring process enhancements. Design details and experimental results are described.

I. Introduction

THE CONTINUING SEARCH for a better CMOS power amplifier seems to be leading to configurations which take advantage of a common source type output stage in order to achieve higher load current capability along with a higher output swing [6], [7]. By introducing a local feedback network around the common source transistor in the form of a simplified operational amplifier, a pseudo source follower is formed which offers substantially better voltage swing than that available from a conventional enhancement source follower. Unfortunately, the bandwidth of the amplifier must often be substantially reduced to maintain a stable frequency response due to an excessive amount of phase shift through this pseudo source-follower circuit.

Other recent developments include variations in amplifier compensation techniques which improve the amplifier's ability to reject noise on the power supplies. Usually this improvement comes at the expense of other performance parameters such as common-mode range, offset voltage, or output swing.

This paper presents a new input stage which exhibits excellent supply rejection properties without the previously mentioned disadvantages and allows a potentially large increase in gain over the classic two stage design. Also, an output stage variation is presented which exhibits an improved frequency response over previously reported large swing stages. Although the output swing of this stage is comparatively limited, it still provides up to ± 3 V into 200 Ω without requiring process enhancements.

Manuscript received January 4, 1985; revised July 25, 1985.
The author is with Siemens AG, WIS TE PE 23, Balanstrasse 73, 8000 Munich 80, West Germany.

II. Circuit Description

A. Core Amplifier

In modern CMOS system design, many analog and digital circuits are often included on the same die. This situation can result in system degradation in the form of power supply noise contamination of the analog signals unless there is sufficient rejection of this noise within the amplifiers in the circuit.

Unfortunately, the classic *RC* compensation technique exhibits very poor supply rejection at high frequencies.

One solution to this problem is to return the compensation capacitor (*Cc*) to a virtual ground, such as the source of a cascode transistor in the input stage [1], [2]. This type of connection works well for inverting gain configurations but the cascode transistors limit the common-mode range for voltage follower or positive gain applications. A separate bias string connected to the output of the first stage can also be used to generate a virtual ground [3], [4]; however, inexact current cancellation in this type of connection can degrade the offset voltage of the amplifier.

Another solution is to use a single gain stage arrangement where the load capacitance is used to compensate the amplifier [4]. This approach leads to a degraded output swing due to the necessity of using cascode transistors to increase the gain of the amplifier to a reasonable level.

The core amplifier used in this design is shown in Fig. 1.[1] It was felt that a three-stage configuration, consisting of a wide bandwidth input stage and two high gain stages, offered much more flexibility in optimizing the performance requirements. As such, the principle of returning *Cc* to a virtual ground can be used for improving the supply rejection of the amplifier without introducing the previously mentioned disadvantages. Also, the three-stage architecture allows for a significant increase in gain without sacrificing the stability of the amplifier.

[1] Since the submission of this paper, a similar transconductance amplifier has been reported by D. B. Ribner and M. A. Copeland, "Design techniques for a cascoded CMOS op-amps with improved PSRR and common-mode input range," *IEEE J. Solid State Circuits*, vol. SC-19, pp. 919–925, Dec. 1984.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1200–1205, Dec. 1985.

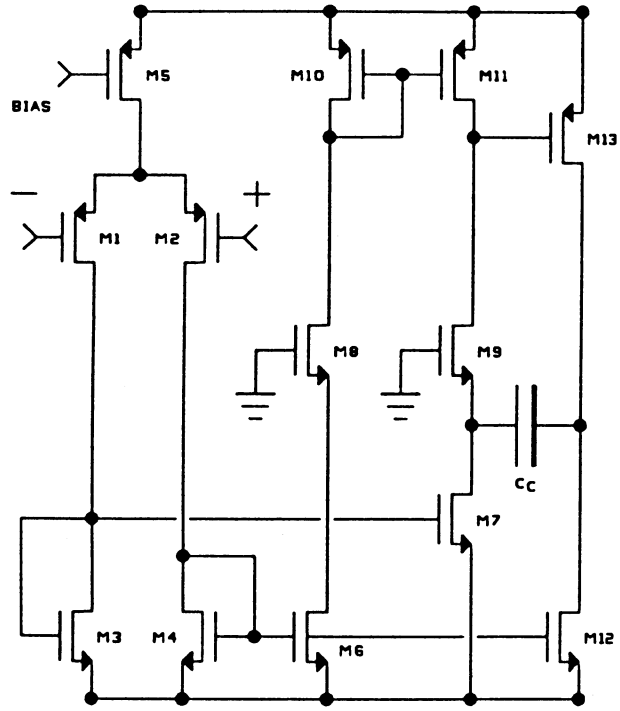


Fig. 1. Schematic diagram of the input stage.

The pole-zero structure of this type of compensation is different than that of the normal RC type. The transfer function of Fig. 1 is given in (1), where it should be understood that $gm1 = gm2$, $gm3 = gm4$, etc.:

$$A_V \approx \frac{a(1+Sb)}{1+Sc+S^2d+S^3e}$$

$$a \approx \frac{gm2\ gm6\ gm13}{gm4\ g_{ds}10\ g0}$$

$$b \approx \frac{Cc\ gm6\ gm13 + C1\ gm8\ gm12}{2\ gm6\ gm8\ gm13}$$

$$c \approx \frac{gm13\ Cc}{gds10\ g0}$$

$$d \approx \frac{C_1(Cc+C_0)}{g_{ds}10\ g0}$$

$$e \approx \frac{C_1CcC_0}{gm8\ g_{ds}10\ g0} \tag{1}$$

where

$$g0 = g_{ds}12 + g_{ds}13$$

$$C_0 = C_L + C_{db}12 + C_{db}13$$

$$C_1 = C_{gs}13 + C_{db}11 + C_{db}9 + C_{gd}9. \tag{2}$$

The zero of this transfer function occurs at

$$Z \approx \frac{-2\ gm6\ gm8\ gm13}{Cc\ gm6\ gm13 + C_1\ gm8\ gm12}. \tag{3}$$

While reducing $M12$ to a simple current source changes the zero of the amplifier slightly, the configuration shown generally results in less phase shift at high frequencies.

The poles in (1) are not all widely spaced and, although $P1$ is easily extracted at

$$P1 \approx \frac{-g_{ds}10\ g0}{gm13\ Cc} \tag{4}$$

$P2$ and $P3$ normally form complex conjugates at

$$P2, P3 \approx \frac{-gm8(Cc+C_0)}{2C_0Cc} \pm j\left[\frac{gm8\ gm13}{C_0C_1} - \left(\frac{gm8\ (C_0+Cc)}{2\ C_0Cc}\right)^2\right]^{1/2} \tag{5}$$

Generally, this type of structure does not lend itself to hand calculation of the appropriate values for stability. Suggested guidelines are to make $gm8$ large and $gm13 \gg gm6$. The unity gain bandwidth occurs at

$$BW \approx \frac{gm2\ gm6}{gm4\ Cc}. \tag{6}$$

B. Output Stage

Fig. 2 shows two previously reported output stages [4], [5] which have been merged to form the output stage in this design. The configuration shown in Fig. 2(a) exhibits very desirable frequency characteristics in the form of one pole and one zero at very high frequencies for normal loads. The voltage swing for the stage is limited to slightly more than a V_{gs} from either supply, which becomes quite significant for large output currents into low impedance loads.

The configuration shown in Fig. 2(b) has several problems associated with it. First, connecting the inputs of Ampl and Amp2 together results in a voltage between the gates of $M1$ and $M2$ that is offset dependent. This means that unless some method is devised to control the quiescent current through $M1$ and $M2$, the current through these transistors will vary widely with variations in V_{os_1} and V_{os_2}. Second, the common-mode range of Ampl and Amp2 must be equal to the desired swing in V_0 if the stage is to work properly. Third, careful consideration must be given to the frequency characteristics of Ampl and Amp2 if the overall amplifier is to be stable. This type of output stage exhibits a large amount of phase shift at high frequencies which typically has required limiting the bandwidth of the overall amplifier in order to insure stability.

Many of these problems are easily solved within the merged output stage shown in Fig. 3. By building a small offset voltage into Ampl and Amp2 as shown, transistors $M1$ and $M2$ are turned off in the quiescent state. The quiescent output current is therefore controlled by transistors $M3$–$M6$. The quiescent output current will be proportional to the current through $M3$ and $M4$ and is a function of the size ratio of $M5$ to $M3$ and $M6$ to $M4$. Under full load conditions in the negative direction, $M2$

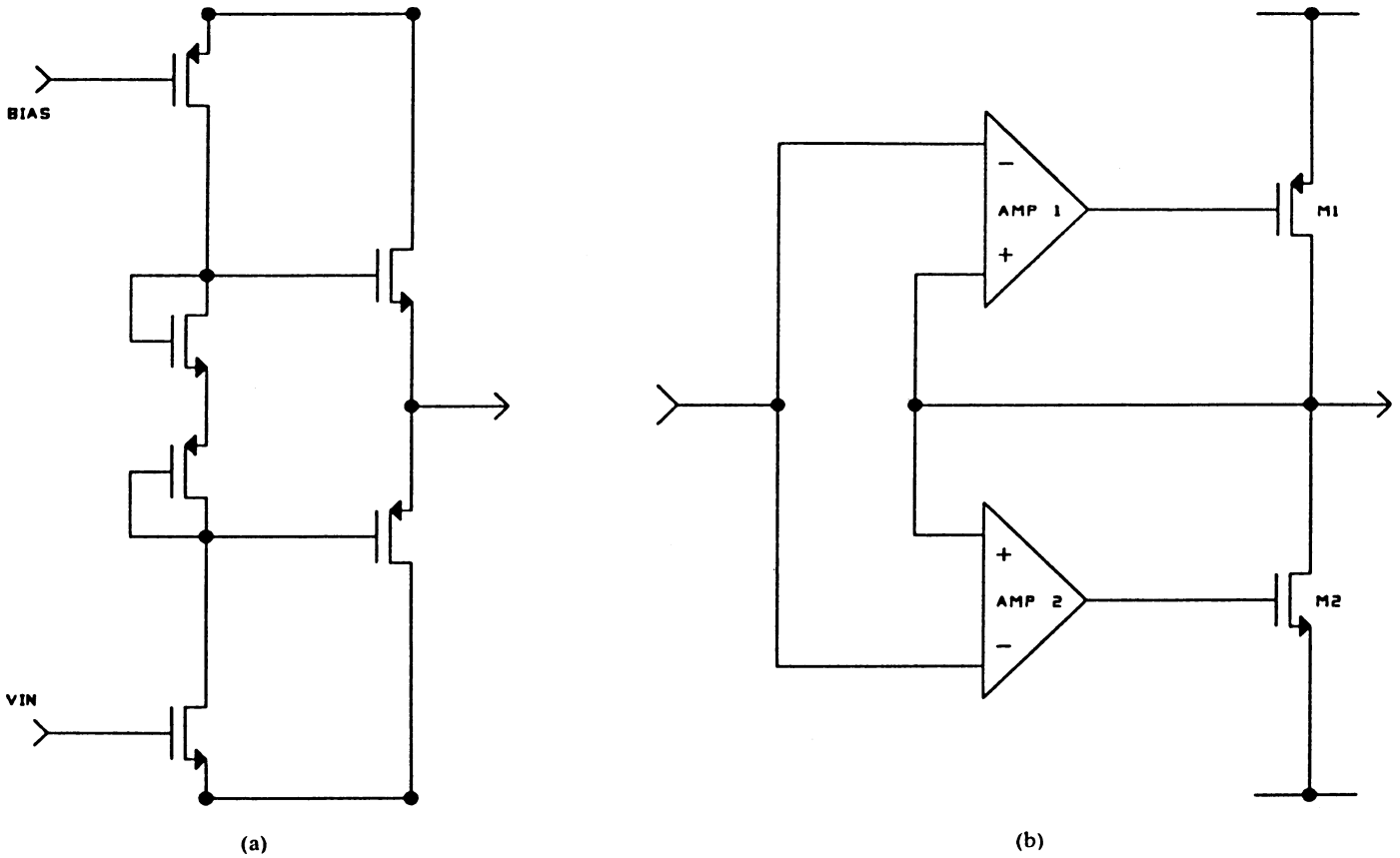


Fig. 2. Previous output stages. (a) Class AB source follower. (b) Pseudo source follower.

sinks approximately 95 percent of the required load current and $M6$ the remaining 5 percent. The actual current control mechanism through $M6$ takes place as Amp2 attempts to equalize the source voltages of $M4$ and $M6$, producing a similar V_{gs} across both transistors, and thus, ratioing their currents. A similar action takes place during the positive swing. This configuration results in a output swing which is still limited to within a V_{gs} of either supply. However, it is a relatively well controlled V_{gs}, dependent on the current through $M3$ and $M4$ rather than on a large output current through $M5$ or $M6$. The limiting factor on output swing in this design is the large threshold voltage of $M4$ and $M6$ due to back bias effects. Naturally, if a low threshold p-channel were available, the output swing could be improved considerably.

Although transistors $M5$ and $M6$ supply a fraction of the load current, their real usefulness lies in quiescent current control and in reducing the excess phase shift introduced by Ampl and Amp2 by providing a feed-forward path to the output at high frequencies. Ampl and Amp2 still require some minimal phase compensation in order to make them stable entities in the closed-loop mode, but the overall amplifier tends to adopt the frequency characteristics of $M5$ and $M6$ rather than that of the composite source followers. For example, during a fast input transient, it would be expected that the slow pseudo source-follower circuits would have a large amount of error associated with them. Observation, however, showed no visually discernible crossover point nor the threshold point of the pseudo source followers. Apparently these errors are corrected by the much faster $M3$–$M6$ source-follower combination. This combination also insures stable operation into a pure capacitive load.

More specifically about Ampl, Fig. 4 shows the structure used in this design. The dc requirements for Ampl are to be able to operate with its inputs near the positive supply while driving the gate of $M9$ to near the negative supply. Therefore, n-channel inputs are used which, with back bias effects, provide a common-mode range exceeding the positive supply. A second stage is used to maximize the gate drive to $M9$. During negative excursions of the output voltage, the gate–source voltage of $M1$ and $M2$ will tend to decrease the current through $M5$ as the negative common-mode range is exceeded. The effect of this current reduction is decreased gate drive to $M6$, which in turn drives $M9$ on. $M9$ corresponds to $M1$ in Fig. 3. As can be seen in Fig. 3, $M1$ should be prevented from turning on during the negative swing to maximize the amplifier's efficiency. Therefore, transistors $M13$–$M15$ have been added in Fig. 4 to force $M9$ off for output voltages more negative than the threshold of $M13$. As was also shown in Fig. 3, $M1$ and $M2$ are held off in the quiescent state by a small offset voltage built into Ampl and Amp2. Note that when the output of the amplifier is at ground, the source of $M3$ and $M4$ are likewise near ground. Thus, connecting the inputs of Ampl and Amp2 together as shown in Fig. 3 will

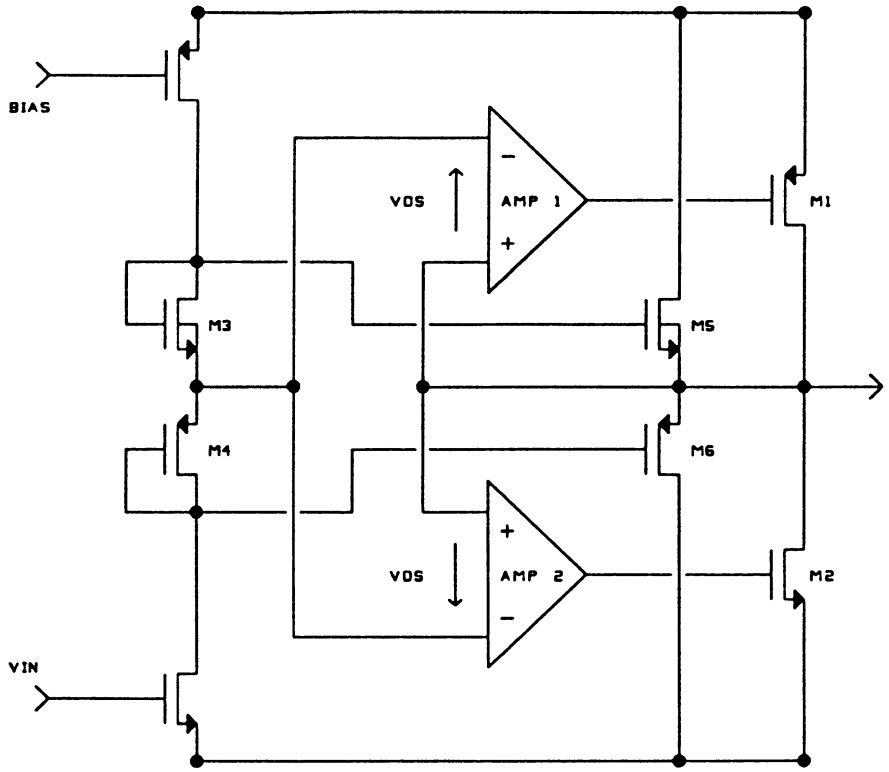


Fig. 3. Combined output stage.

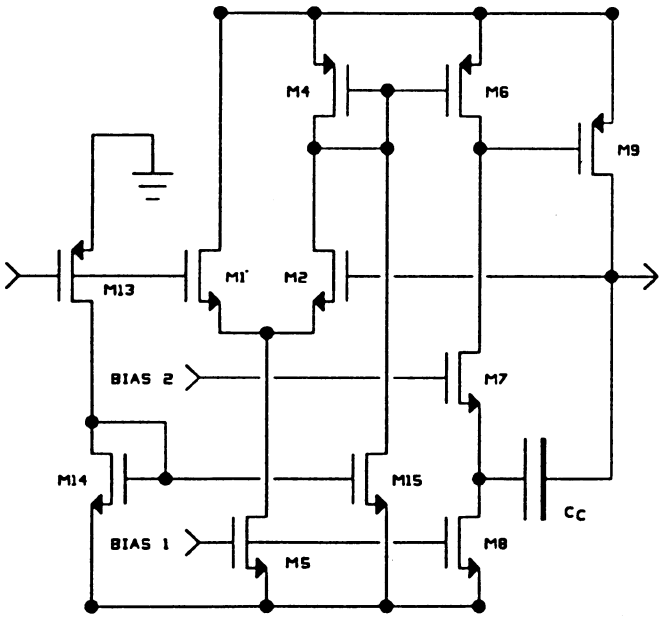


Fig. 4. Schematic diagram of the error amplifier.

result in positive gate drive to $M1$ and negative gate drive to $M2$ when the offsets of the error amplifiers are positive as labeled. These offsets do not affect the operation of the circuit for output voltages other than ground. In Fig. 4, the offset has been introduced by making $M2$ slightly wider than $M1$. A mirror image of this circuitry has been used within Amp2. Ampl in conjunction with $M9$ can be viewed as a three-stage amplifier in unity gain with pole splitting compensation between the source of $M7$ and the output. This type of compensation has a large advantage over the standard RC type when used in the output stage because the zero in the transfer function is completely independent of $gm9$, which varies widely with output swing into a resistive load. This independence generally simplifies stabilizing the amplifier for varying load conditions. The transfer function of Ampl with $M9$ is given in (7):

$$
\begin{aligned}
Av &\approx \frac{a(1+Sb)}{1+Sc+S^2d+S^3e} \\
a &\approx \frac{gm2\ gm6\ gm9}{2\ gm4\ g_{ds}6\ gL} \\
b &\approx \frac{Cc+C_{gs}7}{gm7+gm_{bs}7} \\
c &\approx \frac{CL+Cc\ \dfrac{gm9}{g_{ds}6}}{g1} \\
d &\approx \frac{C_1(CL+Cc)}{g_{ds}6\ gL} \\
e &\approx \frac{C_1\ Cc\ CL}{gm7\ g_{ds}6\ gL}
\end{aligned}
\tag{7}
$$

where

$$
C_1 = C_{gs}9 + C_{db}6 + C_{db}7 + C_{gd}7. \tag{8}
$$

As was the case for the input stage, $P2$ and $P3$ are not widely spaced and again normally form complex con-

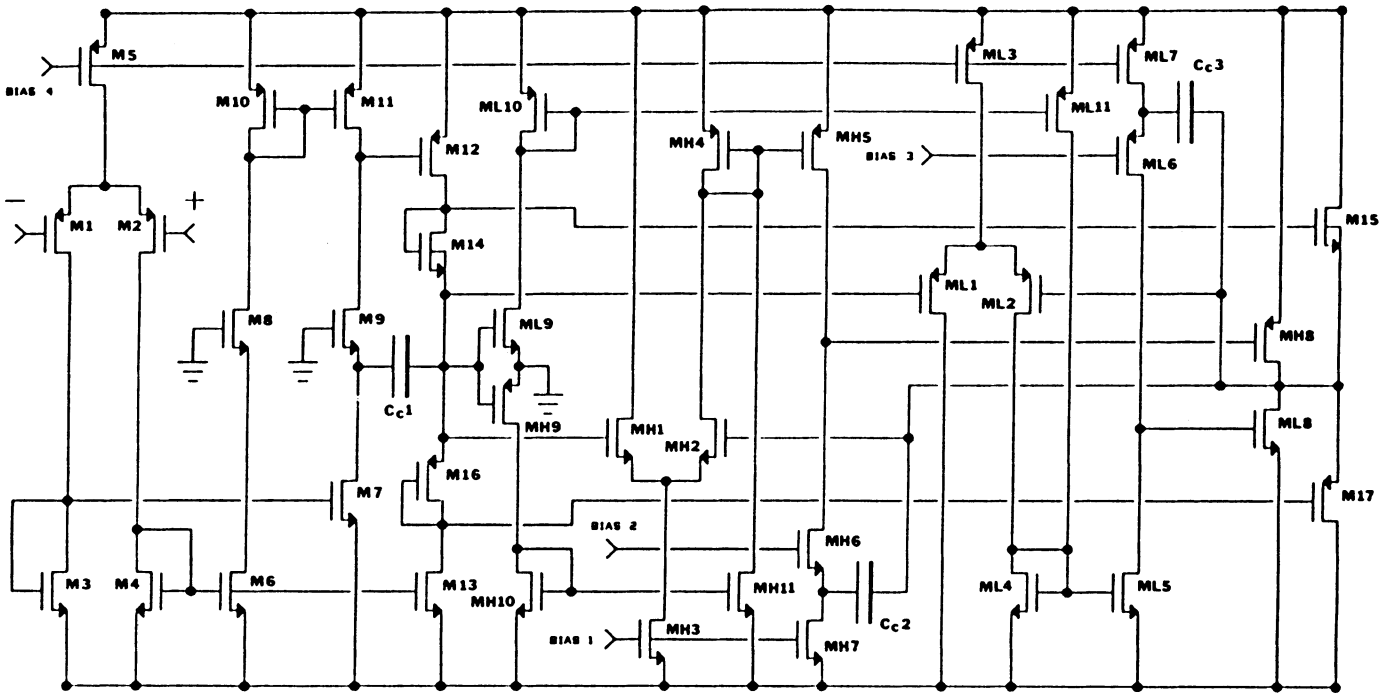


Fig. 5. Schematic diagram of the complete power amplifier.

jugates. The open-loop poles and zero of Ampl with $M9$ are given by

$$Z \approx -\frac{(gm7 + gm_{bs}7)}{Cc + C_{gs}7}$$

$$P1 \approx \frac{-gL}{CL + Cc\dfrac{gm9}{g_{ds}6}}$$

$$P2, P3 \approx \frac{-gm7\,(Cc + CL)}{2\,Cc\,CL} \pm j\left[\frac{gm7\,g_{ds}6\left(CL + Cc\dfrac{gm9}{g_{ds}6}\right)}{Cc\,CL\,C_1} - \left(\frac{gm7(Cc + CL)}{2\,CcCL}\right)^2\right]^{1/2}. \quad (9)$$

Fig. 5 shows a complete schematic of the power amplifier circuit and a device size listing is given in Table I.

TABLE I
COMPONENT SIZES

M1 400/15	MH1 48/10	ML1 48/6
M2 400/15	MH2 50/10	ML2 50/6
M3 150/10	MH3 500/15	ML3 300/15
M4 150/10	MH4 300/6	ML4 150/5
M5 100/15	MH5 300/6	ML5 100/5
M6 150/10	MH6 200/5	ML6 300/6
M7 150/10	MH7 250/15	ML7 100/15
M8 300/5	MH8 700/6	ML8 400/5
M9 300/5	MH9 15/6	ML9 5/5
M10 300/10	MH10 10/15	ML10 5/15
M11 300/10	MH11 20/15	ML11 15/15
M12 1200/10	Cc1 20 pf	
M13 600/10	Cc2 4 pf	
M14 200/5	Cc3 4pf	
M15 200/5		
M16 600/6		
M17 600/6		

III. Experimental Results

A die photograph of this amplifier is shown in Fig. 6. The amplifier was fabricated using a polysilicon gate CMOS process with an n+-implant to generate the capacitor bottom plate. The minimum geometry used in this circuit is 5 μm and the die area of the amplifier, excluding bonding pads, is 1000 mil^2. A summary of the amplifier characteristics is presented in Table II.

Fig. 7 shows the step response of the amplifier with a load of 200 Ω, 1000 pf using supplies of ± 5 V. Fig. 7(a) shows a large signal response of ±3.1 V and Fig. 7(b) shows a small signal response of ±20 mV.

A power amplifier has been described which provides a high degree of performance from a standard polysilicon gate CMOS process. Several new circuit configurations have been incorporated.

Acknowledgment

The author wishes to thank W. C. Black, Jr. for many useful discussions on amplifier design.

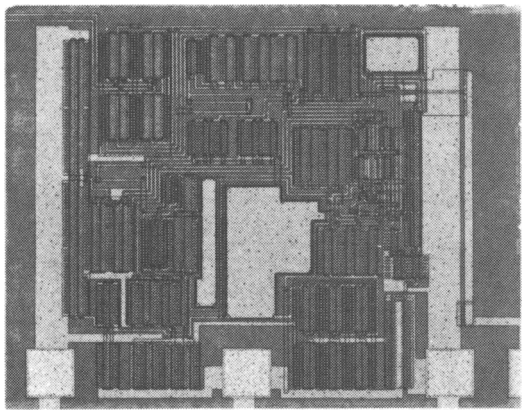
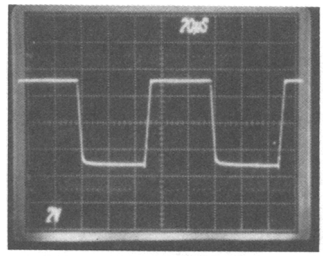


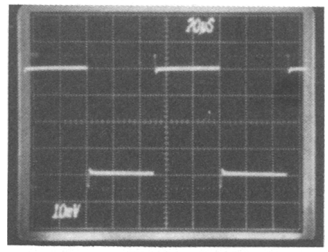
Fig. 6. Die photograph of the power amplifier.

TABLE II
POWER AMPLIFIER PERFORMANCE SUMMARY
(First Revision)

Parameter		Measured Results
Supplies		±5 V
Open-Loop Gain		93 dB
Bandwidth		1.2 MHz
Power Dissipation	$\bar{x}$	12.7 mW
	σ	1.76 mW
Output Swing (R_L=200Ω)		±3.1 V
PSRR+ at DC		93 dB
1 kHz		91 dB
10 kHz		76 dB
100 kHz		60 dB
PSRR- at DC		102 dB
1 kHz		89 dP
10 kHz		75 dB
100 kHz		53 dB
Slew Rate		1.5 V/μs
Input Common Mode Range		+3.3 V
		-5.5 V
Die Area		1000 mils²
Harmonic Distortion (3 kHz) V_{in} = 3 V_p R_L = 200Ω		
HD2		-73 dB
HD3		-78 dB



(a)



(b)

Fig. 7. Step response. (a) Large signal. (b) Small signal.

REFERENCES

[1] R. D. Jolly and R. H. McCharles, "A low-noise amplifier for switched capacitor filters," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1192–1194, Dec. 1982.

[2] D. J. Allstot and W. C. Black, Jr., "Technological design considerations for monolithic MOS switched-capacitor filtering systems," *Proc. IEEE*, vol. 71, pp. 967–986, Aug. 1983.

[3] B. K. Ahuja, "An improved frequency compensation technique for CMOS operational amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 629–633, Dec. 1983.

[4] P. R. Gray and R. G. Meyer, "MOS operational amplifier design—A tutorial overview," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 969–982, Dec. 1982.

[5] W. C. Black, Jr., D. J. Allstot, and R. A. Reed, "A high performance low power CMOS channel filter," *IEEE J. Solid State Circuits*, vol. SC-15, pp. 921–929, Dec. 1980.

[6] K. E. Brehmer and J. B. Wieser, "Large swing CMOS power amplifier," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 624–629, Dec. 1983.

[7] B. K. Ahuja, W. M. Baxter, and P. R. Gray, "A programmable CMOS dual channel interface processor," in *Dig. Tech. Pap. Int. Solid-State Circuits Conf.*, Feb. 1984, pp. 232–233.

An Improved Frequency Compensation Technique for CMOS Operational Amplifiers

BHUPENDRA K. AHUJA

Abstract—The commonly used two-stage CMOS operational amplifier suffers from two basic performance limitations due to the RC compensation network around the second gain stage. First, this frequency compensation technique provides stable operation for limited range of capacitive loads, and second, the power supply rejection shows severe degradation above the open-loop pole frequency. The technique described here provides stable operation for a much larger range of capacitive loads, as well as much improved V_{BB} power supply rejection over very wide bandwidths for the same basic op amp circuit. This paper presents mathematical analysis of this new technique in terms of its frequency and noise characteristics followed by its implementation in all n-well CMOS process. Experimental results show 70 dB negative power supply rejection at 100 kHz and an input noise density of 58 nV/$\sqrt{\text{Hz}}$ at 1 kHz.

I. Introduction

LINEAR CMOS techniques have achieved significant progress over the last five years to provide high-performance low-power analog building blocks like operational amplifiers (op amp), comparators, buffers, etc. These circuits have demonstrated comparable performance to their bipolar counterparts at much less silicon area and power dissipation, thus enabling single chip implementations of complex filtering functions, A/D and D/A conversions with quite stringent specification. Due to relatively simple circuit configurations and flexibility of design, CMOS technology has an edge over NMOS technology and is gaining rapid acceptance as the future technology for linear analog integrated circuits, especially in the telecommunication field [1], [2]. The most important building block in any analog IC is the op amp of which numerous implementations have been reported in both the technologies [3], [6].

The most commonly used op amp configuration in CMOS has two gain stages, the first one being the differential input stage with single-ended output, and the second one being either class A or class AB inverting output stage. Each stage typically is designed to have gain in the range of 40 to 100. Fig. 1(a) shows the circuit configuration while

Manuscript received July 11, 1983; revised August 23, 1983.
The author is with the Intel Corporation, Chandler, AZ 85224.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-18, no. 6, pp. 629-633, Dec. 1983.

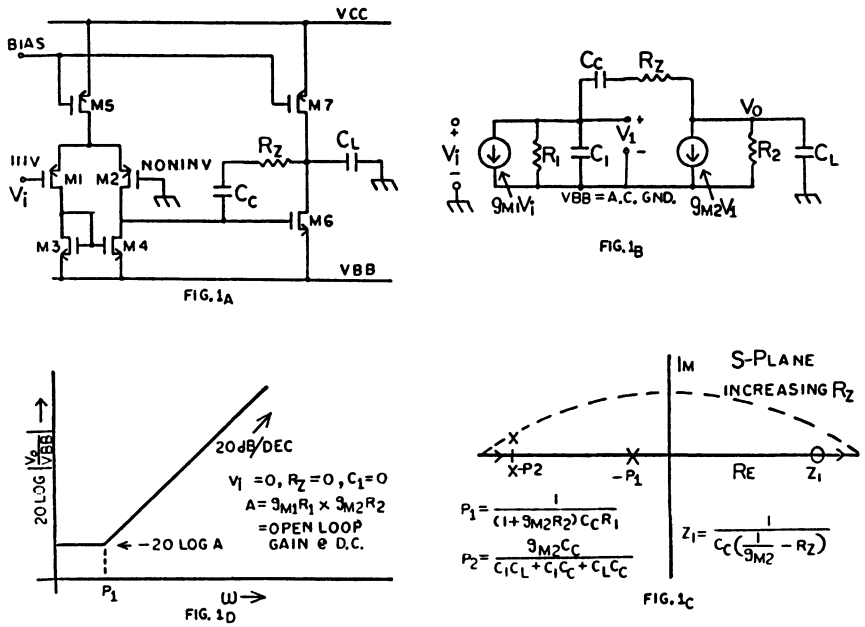


Fig. 1. (a) Commonly used two-stage operational amplifier. (b) Small signal equivalent model for the two-stage amplifier. (c) Pole-zero diagram of Fig. 1(b). (d) V_{BB} PSRR in unity gain configuration.

its first-order ac equivalent model is shown in Fig. 1(b). This configuration is most suitable for internal usage in the IC for driving capacitive loads only. Briefly, transistors $M1$ to $M5$ form the input differential stage and $M6$ and $M7$ form the output inverting gain stage. The series RC network across the second gain stage provides the frequency compensation for the op amp. This circuit, previously analyzed by many authors [5], [7], displays a dominant pole, two complex high frequency poles, and a zero which can be moved from the right half plane to the left half plane by increasing the compensating resistor value R_Z. This is pictorially shown in Fig. 1(c). Due to feedforward path with no inversion from the first stage output to the op amp output provided by the compensation capacitor at high frequencies, the op amp performance shows the following degradations:

1) The op amp stability is severely degraded for capacitive loads of the same order as compensation capacitor (C_L must be less than $g_{m2}C_c/g_{m1}$ to avoid second pole crossover of the unity gain frequency).

2) In case of p-channel MOS transistors for the input differential stage, the negative power supply displays a zero at the dominant pole frequency of the op amp in unity gain configuration. This results in serious performance degradation for sampled data systems which use high-frequency switching regulators to generate their power supplies. (In the case of n-channel MOS transistors for the input differential pair, it is the positive supply which shows similar degradation.) This is illustrated in Fig. 1(d).

The circuit technique described in this paper overcomes both of these limitations. This technique has been referenced earlier [7] as a private communication by Read and Weiser [8]. This paper provides analysis, implementation, and experimental results on the realization in an n-well CMOS process.

II. Improved Frequency Compensation Technique

The technique is based on removing the feed forward-path from the first stage output to the op amp output. The circuit shown in Fig. 1 has a current $C_c d(V_0 - V_1)/dt$ flowing into the first-stage output. If one can devise a circuit where only $C_c dV_0/dt$ current flows into the first-stage output, one would have eliminated the feedforward path while still producing a dominant pole due to the Miller effect. The only difference is that Miller capacitance is now A_2C_c rather than $(1 + A_2)C_c$ where A_2 is the second-stage voltage gain. Thus, the conceptual ac equivalent of such a circuit is shown in Fig. 2(a). Here the compensation capacitor is shown to be connected between the output node and a virtual ground (or ac ground), while the controlled current source having the same value as $C_c dV_0/dt$ charges the first-stage output. It can be shown that for such an arrangement, the open-loop gain of the op amp is given by

$$A = \frac{-A_1A_2}{1 + s(R_1C_1 + R_2C_L + R_2C_c + A_2R_1C_c) + s^2R_1R_2C_1(C_c + C_L)}$$

where $A_1 = g_{m1}R_1$ = dc gain of the first stage and

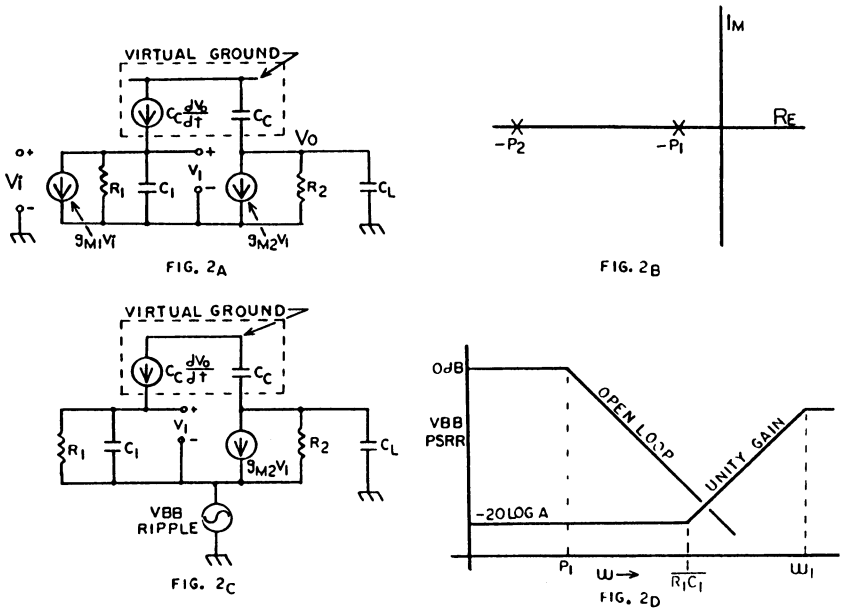


Fig. 2. (a) The new frequency compensation concept. (b) Resultant pole locations in s-plane. (c) Small signal model for V_{BB} PSRR analysis. (d) Expected V_{BB} PSRR frequency response of Fig. 1(a).

$$A_2 = g_{m2}R_2 = \text{dc gain of the second stage.} \tag{1}$$

Fig. 2(b) shows its pole-zero location. Notice that there is no finite zero in this circuit and that both the poles are real and are widely spaced.

$$P_1 \cong \frac{1}{(g_{m2}R_2)C_cR_1} \tag{2}$$

$$P_2 \cong \frac{g_{m2}C_c}{C_1(C_c+C_L)}. \tag{3}$$

Assuming the internal node capacitance C_1 being much smaller than the compensation capacitor C_c or the load capacitance C_L, the unity gain frequency W_1 is still given by g_{m1}/C_c. This results in

$$\frac{P_2}{W_1} = \frac{g_{m2}}{g_{m1}} \cdot \frac{C_c}{C_1} \cdot \frac{C_c}{(C_c+C_L)}. \tag{4}$$

Taking some typical design values of a two-stage amplifier as given by

$$g_{m2}/g_{m1} = 10,\ C_c = 5\text{ pF},\ C_1 = 0.5\text{ pF, and } P2/W_1 \geqslant 5,$$

the new compensation technique can drive up to 100 pF capacitive load as compared to 10 pF capability of the commonly used RC technique as shown in Fig. 1. Thus, the new technique offers an order of magnitude improvement in capacitive load capability for the same performance. The improvement factor is given by C_c/C_1, where C_1 can be reduced by careful layout and design of the first stage.

Another major performance improvement is found in the negative power supply rejection characteristics. Fig. 2(c) shows the model for computing the open-loop negative power supply rejection with grounded inputs. It can be shown that open-loop V_{BB} PSRR is given by

$$\frac{V_0}{V_{BB}} = \frac{1+sC_1R_1}{1+s[R_1C_1+R_2(C_c+C_L)+A_2R_1C_c]+s^2R_1R_2C_1(C_c+C_L)} \cong \frac{1+sC_1R_1}{(1+s/P_1)(1+s/P_2)} \tag{5}$$

which indicates that is has the same poles as the open-loop gain and a zero which is created by the parasitic capacitance at the first-stage output. Thus, in a unity gain configuration, the V_{BB} PSRR is given by

$$\frac{V_0}{V_{BB}} = \frac{1+sC_1R_1}{(1+s/P_1)(1+s/P_2)} \cdot \frac{1}{1+A_1A_2/(1+s/P_1)(1+s/P_2)} \cong \frac{(1+sC_1R_1)}{A_1A_2(1+s/W_1)}. \tag{6}$$

This implies a flat response at $-20\log A_1A_2$, until the parasitic zero frequency of the first stage where it starts to degrade at 6 dB/octave rate and becomes flat again at unity gain frequency W_1. This is illustrated in Fig. 2(d).

III. A Circuit Implementation and Experimental Results

Although the above described scheme can be applied to any MOS amplifier design, it lends a relatively simple

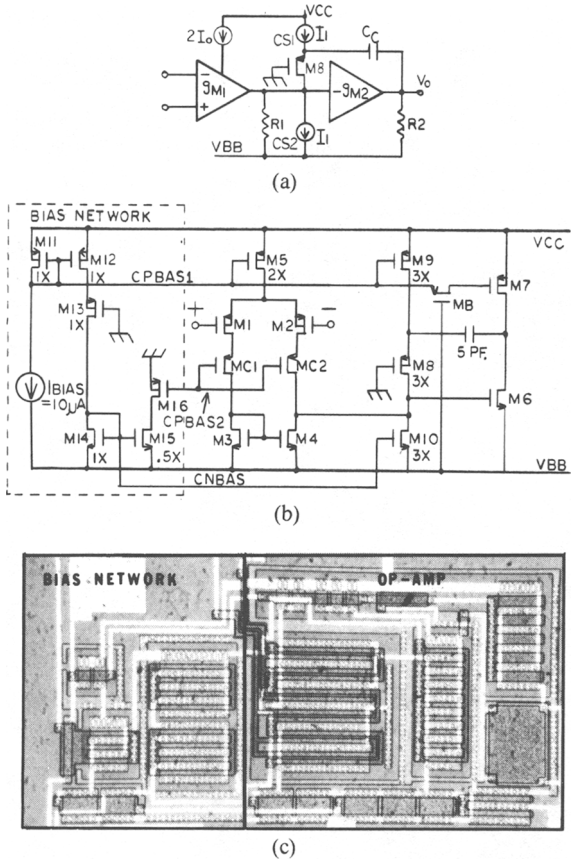


Fig. 3. (a) Implementation of the current transformer providing virtual ground. (b) Circuit schematic of the implemented amplifier. (c) Photomicrograph of the amplifier.

implementation in CMOS technology. Fig. 3(a) shows an implementation of the current transformer providing a virtual or ac ground to the compensation capacitor, while still able to dump $C_c dV_0/dt$ current into the second-stage input. The current source $CS1$ biases the source of $M8$ at a fixed dc potential above ground, thus providing the ac ground for the compensation capacitor. By matching the $CS2$ value to $CS1$, all displacement current $C_c dV_0/dt$ flows into or out of the first stage output.

Under large differential input conditions, the output can slew at a rate determined by the total input differential bias current $2I_0$, i.e.,

$$C_c \cdot \frac{dV_0}{dt} = \pm 2I_0. \tag{7}$$

In order to keep the current transformer biased during the slewing intervals, one must make I_1 greater than $2I_0$. Also, the size of M_8 and the value of I_1 should be large enough to keep V_{GS} of M_8 relatively constant under worst-case slewing conditions.

Fig. 3(b) shows a circuit schematic of the implemented amplifier. The input differential stage, formed by $M1$ to $M5$ transistors, uses cascode devices $MC1$ and $MC2$ to reduce supply capacitance from the negative power supply for switched-capacitor applications [5]. The current transformer is being realized by $M8$, $M9$, and $M10$. Due to its cascode configuration, this technique has been referred to as the "grounded gate cascode compensation" in [7]. The output stage is formed by $M6$ and $M7$. The transistor MB and the gate capacitance of the $M7$ transistor provide RC low-pass filtering of the high-frequency noise on the bias line $CPBAS1$. The associated bias circuit shown in the dotted box is shared among several such amplifiers, thus reducing power and area overhead cost due to this compensation technique. Fig. 3(c) shows the die photo of the amplifier. The amplifier has been designed in a 4 μm n-well CMOS process and occupies about a 165 mil^2 die area.

TABLE I
$V_{CC} = +5$ V, $V_{BB} = -5$ V, AND $T = 27°$C

Parameter	Measured Value
Open-Loop Gain	80 dB
Unity Gain Frequency	3.8 MHz
Phase Margin with $C_L = 15$ pF	70°
Input Common Mode Range	+4 to −2.5 V
CMRR at 1 kHz	−74 dB
Input noise density at	
1 kHz	58 nV/$\sqrt{\text{Hz}}$
100 kHz	8 nV/$\sqrt{\text{Hz}}$
C-msg Input Noise	−21 dBrnc
V_{CC} PSRR at	
1 kHz	−84.5 dB
10 kHZ	−84.5 dB
100 kHz	−73 dB
V_{BB} PSRR at	
1 kHz	−84 dB
10 kHz	−84 dB
100 kHz	−70 dB

The input referred noise of this amplifier is slightly worse than the one shown in Fig. 1(a) due to the noise contributions from transistors $M9$, $M10$, $M12$, and $M14$. However, these contributions can be reduced significantly by choosing large values of channel lengths of these devices with respect to the channel lengths of input transistors $M1$ and $M2$ [3], [7].

Some of the measured performance parameters are listed in Table I. The op amp exhibits open-loop gain of 80 dB, unity gain frequency of 3.8 MHz, and a phase margin of 70° with 15 pF load capacitance. The V_{CC} and V_{BB} PSRR at low frequencies are better than −80 dB due to the bias circuit design and the cascode transistors $MC1$ and $MC2$, respectively. The V_{BB} PSRR shows zero at about 60 kHz, which closely matches the simulated value of the parasitic zero frequency. The op amp displays an input referred noise density of 58 and 8 nV/$\sqrt{\text{Hz}}$ at 1 and 100 kHz frequencies, respectively.

Conclusions

An improved frequency compensation technique has been described with a brief review of the existing techniques. A CMOS implementation of the technique has also been presented with experimental results which show considerable high-frequency power supply rejection improvement over the existing techniques which would result in approximately −30 to −35 dB V_{BB} PSRR at 100 kHz.

Furthermore, the technique provides extended capacitive drive capability for the same size of the compensation capacitor.

ACKNOWLEDGMENT

The author would like to thank Dr. P. Gray for technical discussions on the noise analysis of this compensation technique. Also, the technical assistance in the performance evaluation by T. Barnes is greatly appreciated.

REFERENCES

[1] R. Gregorian and G. Amir, "A single chip speech synthesizer using a switched-capacitor multiplier," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 65–75, Feb. 1983.

[2] B. K. Ahuja *et al.*, "A single chip CMOS PCM codec with filters," in *ISSCC Dig. Tech. Papers*, pp. 242–243, Feb. 1981.

[3] P. R. Gray, "Basic MOS operational amplifier design—An overview," in *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980, pp. 28–49.

[4] D. Senderowicz, D. A. Hodges, and P. R. Gray, "A high performance NMOS operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 760–768, Dec. 1978.

[5] W. C. Black *et al.*, "A high performance low power CMOS channel filter," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 929–938, Dec. 1980.

[6] V. R. Saari, "Low power high drive CMOS operational amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 121–127, Feb. 1983.

[7] P. R. Gray and R. G. Meyer, "MOS operational amplifier design—A tutorial overview," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 969–982, Dec. 1982.

[8] R. Read and J. Wieser, as referred in [7].

Design Considerations for a High-Performance 3-μm CMOS Analog Standard-Cell Library

CARLOS A. LABER, MEMBER, IEEE, CHOWDHURY F. RAHIM, MEMBER, IEEE, STEPHEN F. DREYER, GREGORY T. UEHARA, MEMBER, IEEE, PETER T. KWOK, MEMBER, IEEE, AND PAUL R. GRAY, FELLOW, IEEE

Abstract—**Several design aspects of a high-performance analog cell library implemented in 3-μm CMOS are described, including an improved central biasing scheme, a new circuit for high-swing cascode biasing, an impact ionization shielding technique, and a family of operational transconductance amplifiers (OTA's) including a precision low offset-voltage amplifier utilizing lateral bipolar transistors.**

I. Introduction

THE application of standard-cell methodology to the design of digital integrated circuits has greatly reduced their design time and associated engineering costs. While the same potential advantages are available in the mixed analog–digital domain, the application of standard-cell-based design methodology in the analog domain is more difficult because of the much wider range of applications encountered, the wider variety of types of cell functions required, and the tendency of analog blocks to interact with one another in a variety of ways, some difficult to predict *a priori*. This variety of different performance levels and types of functions required makes the systematic application of a standard-cell-based methodology to high-performance mixed analog and digital designs a challenging task.

This paper will describe several circuit design approaches intended to alleviate some of the problems mentioned above, and to improve circuit robustness in the presence of wide process parameter variations found in different silicon foundries. The techniques have been applied to a general-purpose CMOS analog cell library, described in general terms at the end of the paper. This library consists of a variety of analog and digital standard cells intended for application in IC's for high-performance telecommunications systems, precision data-acquisition and instrumentation systems, and general analog processing. In Section II, the overall characteristics of the cell library are discussed. In Section III, an impact ionization shielding concept which allows the realization of 10-V circuitry using technologies which display significant impact ionization in the 10-V range of V_{ds} without encountering the degrading effects of impact ionization will be described. In Section IV, the approach taken to the optimum biasing of the active circuitry under the control of a central bias source is discussed, including an optimized central biasing scheme that incorporates circuitry to allow the accurate control of the bias points of the active circuitry in the presence of wide excursions in process parameters such as resistor sheet resistance. Also, an approach for optimum biasing high-swing cascode current sources so as to allow maximum possible signal swing in the active circuitry will be discussed.

In Section V, several basic analog cells are described including a low-input offset-voltage operational transconductance amplifier which utilizes lateral n-p-n bipolar transistors. Finally, a summary of the standard-cell library is given, as well as examples of actual integrated circuit implementations which make use of this library.

II. Library Objectives

The circuit design techniques described in this paper were dictated largely by the specific objectives the library was intended to address. A key objective of the cell library was to allow integration of telecommunication and data-acquisition subsystems with performance compatible with 12-bit linearity and 14-bit resolution in A/D interfaces and 90-dB dynamic range in analog signal paths, including switched-capacitor filters. This consideration dictated relatively high levels of performance in areas such as operational-amplifier input noise and input offset voltage, power supply rejection, and voltage reference drift. The requirement for high-linearity A/D interfaces was realized through the use of self-calibration techniques. The library also makes extensive use of differential circuitry in filters and amplifiers, so as to optimize power supply rejection and dynamic range. Single-ended structures are also utilized for less demanding applications.

A second objective was that the functions to be implemented include operating voltages from 4.5 to 11 V, split or single supplies, and temperature ranges of from -55 to

Manuscript received September 9, 1986; revised January 5, 1987.

C. A. Laber, C. F. Rahim, and S. F. Dreyer are with Micro Linear Corporation, San Jose, CA 95131.

G. T. Uehara was with Micro Linear Corporation, San Jose, CA 95131 He is now with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

P. T. Kwok was with Micro Linear Corporation, San Jose, CA 95131. He is now with Exar Corporation, Sunnyvale, CA 94088.

P. R. Gray is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

IEEE Log Number 8613379.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-22, no. 2, pp. 181–189, Apr. 1987.

+125C. They must be sufficiently insensitive to process variations such that the same circuit can be fabricated on different foundry fabrication lines with similar but not equal process parameters utilized. As discussed later, an important aspect in achieving this goal is the provision for accurate control of the active-circuitry bias currents in the presence of variations in supplies, temperature, and process. Good control of the bias points is also important in order to achieve an optimum speed, power, and swing trade-off for a given application, particularly when extensive use is made of folded cascode amplifiers as is the case here.

Finally, due to the wide spectrum of requirements on speed, power dissipation, dynamic range, complexity, and cost which are encountered in practice, a set of cells is required for each function which encompasses a spectrum of area/performance and power/performance trade-offs so as to economically address a wide range of applications. This dictates a design approach in which cells can be scaled through bias current modification and other means to easily achieve different levels of speed and dynamic range with near-optimum power dissipation and area.

III. Impact Ionization Shielding

Impact ionization is a severe problem in scaled MOS technologies operating at supplies voltages above 5 V. In a typical n-channel transistor, illustrated for reference in Fig. 1, as the drain–source voltage is increased the electric field strength at the drain end of the channel eventually becomes high enough to induce significant impact ionization currents originating in the drain depletion region. The magnitude of the peak field for a given bias point is a function of gate oxide thickness, drain junction depth, doping concentration in the substrate, the voltage between the drain terminal and the drain end of the channel region, and the gate–drain voltage. It is not a strong function of channel length and the magnitude of the impact ionization current is not dramatically reduced by simply making the channel length longer. For technologies with feature sizes in the 2–3-μm range which use a nongraded implanted arsenic source–drain region for the n-channel device, the condition at which the substrate current equals 1 percent of the drain current typically occurs at voltages between the drain and the drain end of the channel of between 4 and 9 V, assuming the device is biased in saturation with a V_{gs}-V_t of a few hundred millivolts. In p-channel devices the effect occurs at substantially higher field strengths.

Operation in the impact ionization mode has several undesirable effects from a circuit standpoint. One potentially catastrophic consequence is the inducement of latch-up due to the ohmic drops induced by the ionization current. Assuming this can be controlled by proper strapping, another negative consequence is the degradation of impedances at the output of cascode current sources, as used, for example, in folded cascode amplifiers which rely on high output impedances to function properly. Also, the presence of a significant amount of substrate current increases the noise of the device. A third serious potential consequence of continuous operation in the impact ionization mode is threshold shift. If the electric field is high enough, high-energy carriers can be created which can be trapped in the gate oxide, and the resulting long-term threshold shift can have the effect of debiasing the active circuitry if the threshold shift occurs in a device whose threshold voltage is important in setting up bias currents [2].

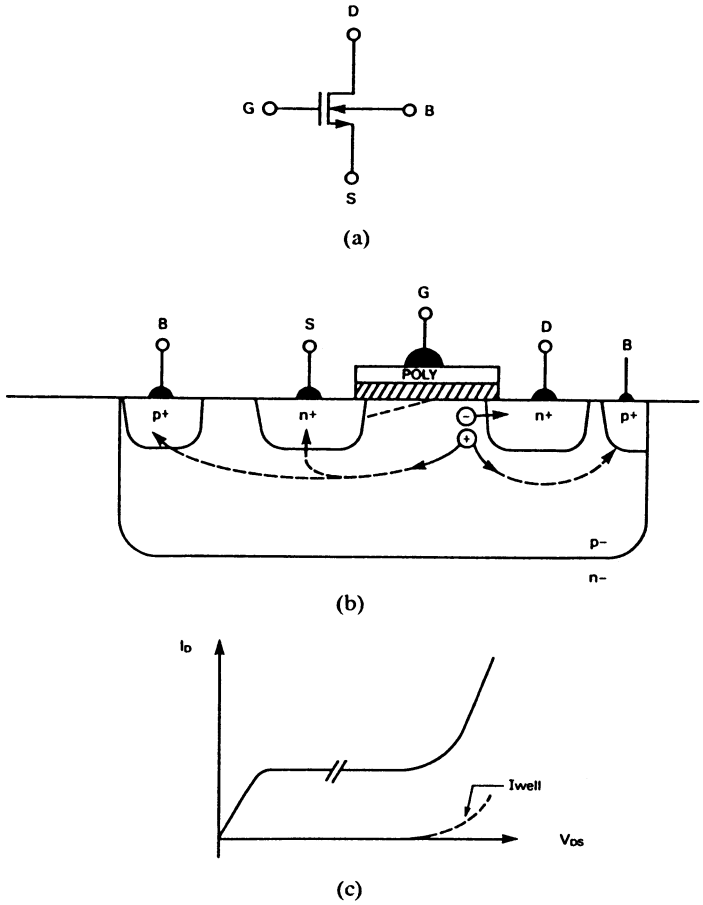


Fig. 1. (a) N-channel device. (b) N-channel device cross section showing impact ionization current flow. (c) $I-V$ characteristic observed as a result of impact ionization.

A number of process modifications have been proposed and implemented to alleviate this problem by lowering the impurity gradient in the drain junction using lightly doped drain (LDD) structures. These structures are effective at raising the voltage at which impact ionization becomes a problem. However, major modifications are required for a typical 2–3-μm digital CMOS technology to completely eliminate impact ionization from the n-channel transistor for values of V_{ds}-$V_{d\,sat}$ of 11 V, as required in 10-V$\pm$10-percent supply systems. Since in this case the objective was to realize A/D mixed functions operating on 10 V with near-standard foundry digital CMOS technology, a circuit solution to the impact ionization problem was preferable to extensive technology modification. A design philosophy was adopted in which all 10-V circuitry would be designed in such a way that no n-channel transistor would experience a drain–source voltage larger than one-half the power

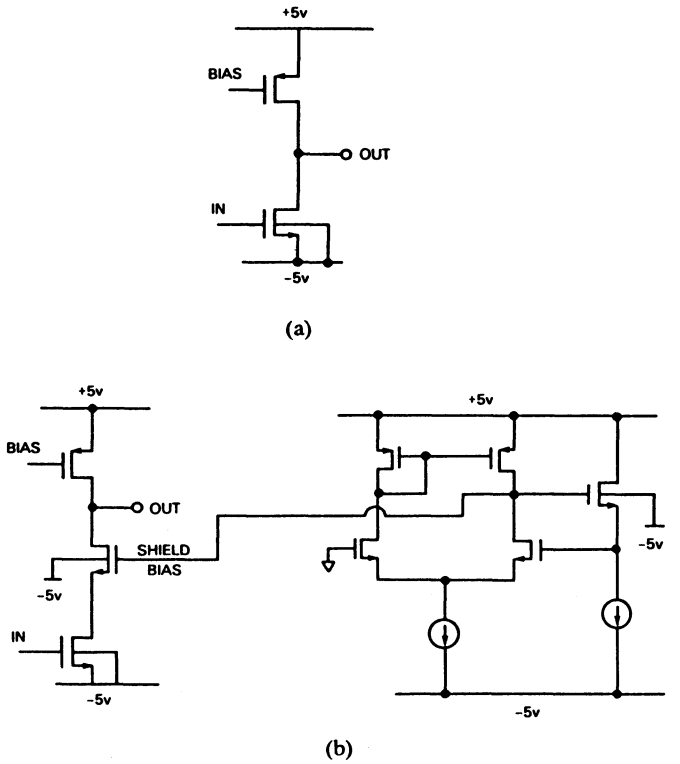


Fig. 2. (a) Conventional CMOS gain stage. (b) Same as (a) but with impact ionization shielding device and associated bias circuit.

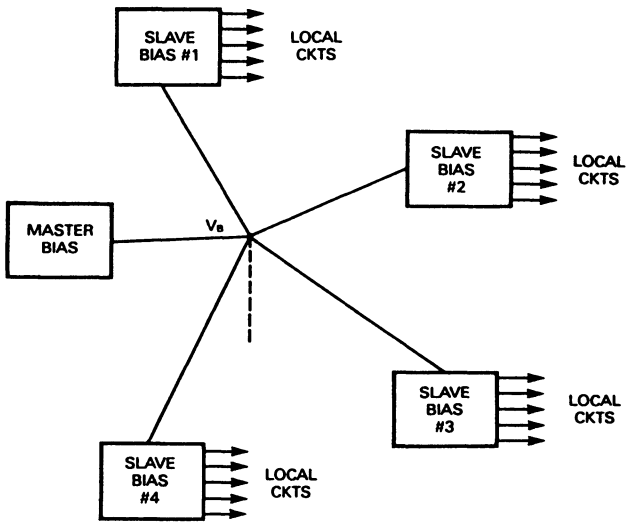


Fig. 3. Central biasing configuration with master bias supplying bias reference to a number of slave bias circuits placed around the chip.

supply voltage, except for the case of transmission gates carrying only transient displacement currents. This was achieved by inserting shielding devices in series with each common-source n-channel transistor.

In a conventional CMOS gain stage, illustrated in Fig. 2, the n-channel active device experiences impact ionization for positive output swings, as the output gets closer to the positive rail. One way to deal with this problem is to place a shield n-channel device in series with the driver, as shown in Fig. 2(b). For positive swings, the shield transistor goes into saturation, preventing either n-channel device from having a V_{ds} greater than approximately one-half of the supply. For negative swings, the shield transistor enters the deep triode region, without significantly affecting the performance of the circuit.

The shield transistor must be biased in such a way that its source resides approximately at ground potential. Grounding the gate of this transistor would be simplest, but because of the large body effect in some CMOS technologies, this would result in large V_{ds} drops across the shield transistor itself. Instead, a dedicated bias circuit is used to more precisely set the source of the shield transistor at ground. This is accomplished by the circuit on the right of Fig. 2(b), which contains a shield mirror transistor, whose source is forced by the feedback loop to a voltage which is approximately ground (or the midpoint potential for single supply systems). This bias scheme is independent of the body effect of the technology, because of the feedback action around the mirror device.

The shield device was incorporated in all 10-V circuitry in the library at an area cost of about 10 percent. The shield bias voltage is contained in the standard bias bus and is routed to all analog cells. In a total of ten products designed for 10-V operation and fabricated with the library, no performance or reliability problems attributable to impact ionization have been encountered.

IV. Cell Bias Technique

One critical element in achieving high-performance CMOS analog circuits is the accurate control of the bias currents of the different transistors employed in a given circuit topology. Excessive variation of bias currents with temperature and process tends to sacrifice speed at the low extreme of bias current, and power dissipation and output voltage swing at the high end. Furthermore, the variation of bias currents with supply voltage results in poor power supply rejection (PSR). Achieving the goal of minimizing the dependence of bias points on supply voltage, temperature, and process variations requires a bias circuit of some complexity, and one which is therefore uneconomical to implement on a per-cell basis. Therefore a central biasing scheme, as illustrated in Fig. 3, was adopted for this cell library. A central master bias circuit, described below, produces a voltage which is distributed around the chip. This voltage is used by slave bias cells, which generate the locally required bias voltages, in order to power the nearby circuitry.

The main advantage of this approach is the flexibility of the architecture, which allows different slave bias cells to work at different current levels, as required in certain applications. This scheme also helps in preventing crosstalk between critical circuits through the bias lines, as can occur when a bias line sharing approach is employed. The device which generates the output control master bias voltage is designed with a large $V_{d\,sat}$, so that small voltage gradients across the chip in the supply lines do not significantly affect the resulting current.

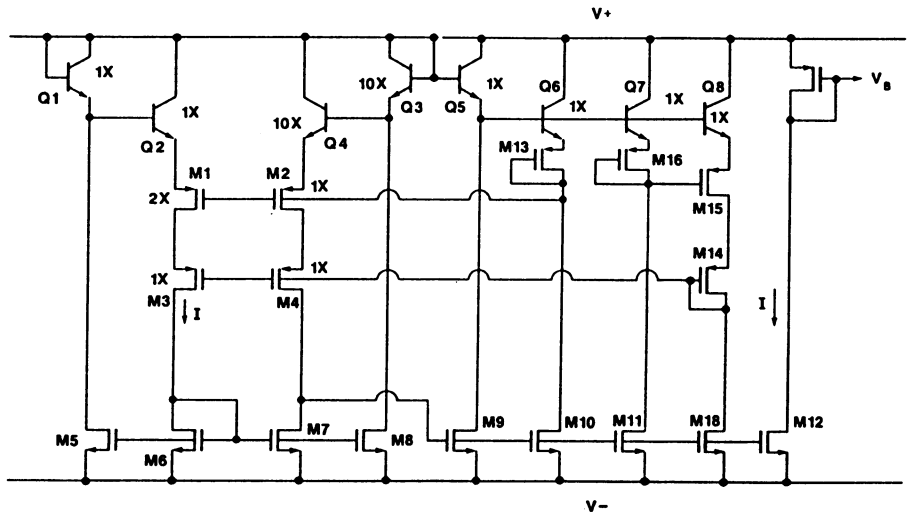


Fig. 4. Simplified schematic of the master bias circuit.

In critical circuits where tight limits on power dissipation or speed require a tighter control on bias current, the master bias cell allows the use of polysilicon fuse trimming, giving the desired result at the cost of trim pad area. Another advantage of central biasing is that full-chip power down can be accommodated, assuming that the slave bias cells and active circuits are properly designed to avoid floating-gate-induced parasitic paths in the power-down state.

A. Master Bias Implementation

A simplified schematic of the master bias cell is shown in Fig. 4. This is a self-biased circuit whose output current has been optimized for best temperature coefficient, absolute tolerance, and power supply rejection. The core of the circuit is the vertical bipolar transistors $Q1-Q4$. Since they are forced to conduct the same amount of current, by the bottom n-channel current mirrors $M5-M8$, a voltage difference is developed between the emitters of $Q2$ and $Q4$, which is approximately equal to $2(kT/q)\ln(10)$, or about 120 mV at room temperature. When this voltage is applied across p-channel devices $M1$ and $M2$, which are biased by the same gate voltage, it is easy to show that a current is generated whose value is given to a first order by

$$I = \frac{2\mu C_{ox}\left(\frac{W}{L}\right)_2\left[\left(\frac{kT}{q}\right)\ln 10\right]^2}{\left[1-\sqrt{\frac{(W/L)_2}{(W/L)_1}}\right]^2}.$$

Taking into account the effects of tolerance on oxide thickness and mobility, as well as the effects of offset voltages in the MOS and bipolar transistors, this results in an absolute room-temperature tolerance on bias current which is somewhat better than for the more typical $(kT/q)/R$ based current reference for most technologies, where R is a poly, source–drain diffusion, or well resistor. More importantly, the process dependence of the derived bias current results in a very well-defined value of the $V_{d\,\mathrm{sat}}$ of the MOS transistors in the active circuitry, making it possible to maintain high-voltage swings over wide variations in process parameters. Further, since the channel mobility of the MOS device has an approximate $T^{-3/2}$ temperature dependence, where T is the absolute temperature, the result is a bias current with a net residual $T^{+1/2}$ temperature coefficient (TC). This is a very desirable feature, since in amplifiers biased by this current the transconductance of saturated MOS device is to a first order, inversely proportional to the square root of absolute temperature. The net consequence is that the duration of the slewing portion of the transient response of those active analog circuits exhibits a small positive TC, whereas the duration of the small-signal settling transient portion has a small negative TC. As a result the overall settling time behavior of the active circuitry does not display severe degradation at temperature extremes.

The right-hand part of the circuit shown in Fig. 4 is responsible for the high-swing cascode biasing of transistors $M1-M4$, and also the negative feedback which insures proper setting of the voltage at the drain of $M7$, which is the only high-impedance node in the circuit. The output current is finally forced to flow through diode-connected p-channel transistor $M17$, which produces the output master bias voltage V_b. As mentioned above, $M17$ has a large $V_{d\,\mathrm{sat}}$ to absorb any difference in the threshold voltage between $M17$ and the slave bias mirror device, as well as V_{DD} drops across the chip. The high-swing cascode biasing scheme illustrated in Fig. 4 is the same as the one used in the slave bias cell, and is discussed below. Not shown in Fig. 4 are the start-up circuit which prevents the zero-current state at power up, the impact ionization shield devices and shield bias generator for 10-V operation, and cascode transistors on the n-channel current sources. Table I summarizes the main performance parameters of the master bias cell.

TABLE I
TYPICAL MASTER BIAS CIRCUIT PERFORMANCE (ROOM TEMPERATURE)

Parameter	value
Output bias current	13.1uA
Output current standard deviation	1.3UA
Output current temperature coefficient	+1700ppm/deg C
PSRR	1500ppm/volt
Supply range	4.5-11 volts
Power dissipation	2mW@10V

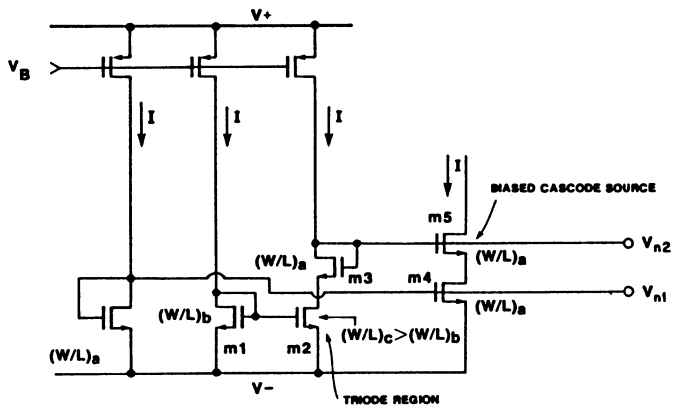


Fig. 5. Simplified schematic of the slave bias circuit. The bias lines V_{n1} and V_{n2} are used to supply the high-swing cascode current sources in the local active circuitry. A similar circuit is used to supply the p-channel current courses.

B. Slave Bias Implementation

As mentioned earlier, local slave bias cells are used in conjunction with each local group of circuits to produce the much smaller $V_{d\,sat}$'s required to bias the gates of transistor current sources within each active cell. The slave bias cell produces bias voltages for p- and n-channel current sources, and the proper voltages to optimally bias the cascode devices in series with these current sources as used in, for example, folded cascode operational amplifiers.

Within the various cells of the library, cascode current sources are used extensively to generate high-impedance current source loads. Because of the requirement that the cells operate on a single 5-V supply, it is important that these cascodes be biased for optimum voltage swing [3]. The generation of high-swing bias voltages can be accomplished in many ways. However, previous reported techniques [4] have suffered from a strong body-effect sensitivity and poor control of the absolute tolerance on the current-source current value.

A simplified schematic of the slave bias cell is shown in Fig. 5. The basic concept utilized in this circuit is to force device $M2$ into the triode region by making its aspect ratio greater than that of transistor $M1$, since both $M1$ and $M2$ conduct the same amount of current. The V_{ds} of $M2$ then takes a value, neglecting second-order effects, of

$$V_{ds2} = \left\{\sqrt{\frac{2I}{\mu C_{ox}(W/L)_1}}\right\}\left[1-\sqrt{1-\frac{(W/L)_1}{(W/L)_2}}\right]$$

or, in terms of the $V_{d\,sat}$ of $M1$

$$V_{ds2} = V_{d\,sat\,1}\left[1-\sqrt{1-\frac{(W/L)_1}{(W/L)_2}}\right].$$

Assuming that devices $M3$ and $M5$ are sized so that they have equal gate–source voltages, the drain–source voltage of $M4$ will be equal to the drain–source voltage of $M2$, given above. Thus by choosing device ratios, the V_{ds} of $M4$ can be chosen to be an arbitrary multiple, usually on the order of 1.5, of its $V_{d\,sat}$, independent of process parameters, by means of transistors $M1$ and $M2$. The margin of drain–source voltage over $V_{d\,sat}$ required to insure operation fully in the saturation region with accompanying output conductance is in the 200-mV range. Another feature of this circuit is that when the master bias described earlier is used to generate the input current I in Fig. 5, then $V_{d\,sat}$ and thus the V_{ds} of $M2$ become approximately independent of process variations, and only proportional to absolute temperature. The body-effect sensitivity is avoided because, to a first order, the body effect of $M3$ cancels that of $M5$, resulting in a voltage across $M4$ which is approximately independent of gamma.

V. BASIC ANALOG ACTIVE CELLS

Most low- and mid-frequency active circuitry is implemented using a family of folded cascode operational transconductance amplifiers (OTA's) which can drive internal capacitive loads and a family of class AB unity-gain buffers for driving on-chip resistive loads and for driving signals off-chip. The OTA's include an n-channel input single-ended amplifier, a p-channel input single-ended amplifier, a bipolar input single-ended cell, and a differential output amplifier. Simplified circuits for the NMOS and PMOS input amplifiers are shown in Fig. 6. The NMOS input OTA includes cascode devices in the input stage and has the input devices in a well so as to optimize power supply rejection in single-ended switched-capacitor filter applications. The bipolar input amplifier is discussed below.

The OTA configuration was chosen as the basic active element because of the flexibility of these circuits and the ease with which they can be combined into more complex blocks. For example, in applications requiring very high voltage gain, two OTA's can be cascaded and compensated with a single pole-splitting capacitor and nulling resistor between the two high-impedance nodes. For applications requiring higher transconductance or lower $1/f$ noise and thermal noise, OTA's can be connected directly in parallel, unlike two-stage operational amplifiers. This capability

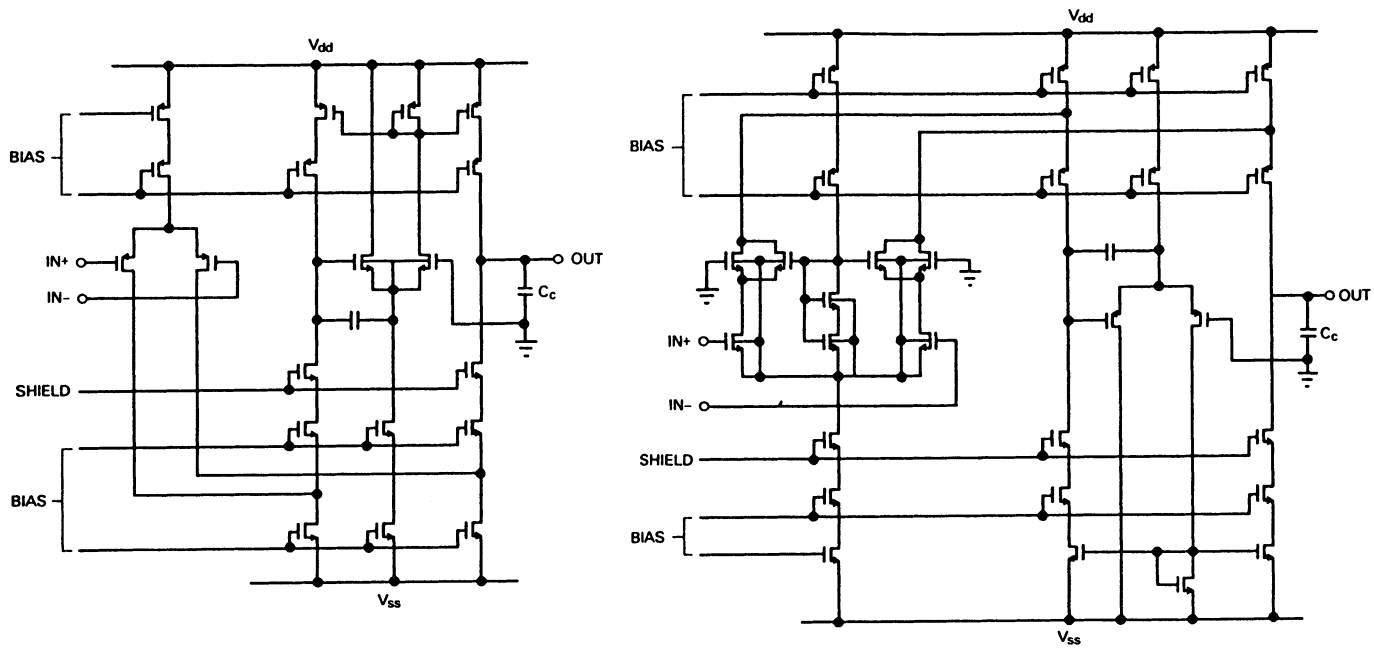


Fig. 6. Schematic diagram of the (a) PMOS and (b) NMOS input OTA's. The auxiliary differential pairs in the differential-to-single-ended converter balance the output stage voltages so as to achieve a small value of systematic offset voltage.

also permits the realization of a rail-to-rail common-mode capability by paralleling an NMOS input OTA with a PMOS input OTA. The large phase margin of the structure allows it to be used as the front-end gain stage in composite off-chip driver operational amplifiers in which a unity feedback path is closed from output to input. Another important advantage for the wide spectrum of requirements encountered in the custom environment is the fact that the compensation capacitor can be connected to whatever potential is being used as the signal reference, giving good high-frequency power supply rejection. Depending on the application, this can be either supply or a separate ground.

These elements also have great flexibility for use in conjunction with nonlinear feedback elements to perform such functions as peak detectors and comparators. In such applications the compensation capacitor can often be removed completely to maximize the speed of operation. In all applications, the bandwidth and transient response of the block can be improved at the cost of power dissipation and voltage swing by scaling up the bias current using the slave bias approach described earlier.

The MOS amplifiers described above display the broad input offset-voltage distribution typical of MOS differential amplifiers. While switched-capacitor offset cancellation techniques can be used to remove system offsets in many cases, the need often arises for continuous amplifiers with input dc offset voltages which are smaller than those achievable in MOS amplifiers, and also which display the PTAT drift characteristic found in bipolar amplifiers. The latter characteristic is important in bandgap references, for example. This is best satisfied with a bipolar input stage, and to satisfy this need a precision transconductance amplifier using a combination of lateral and vertical transistors was designed. This cell is then used as an element of more complex blocks such as references and instrumentation amplifiers that require low offset or PTAT drift.

This circuit, which is illustrated in simplified form in Fig. 7, uses the lateral transistor structure previously described by Degrauwe *et al.* [5]. In a p-well CMOS technology, this device is the lateral n-p-n transistor formed by the source and drain diffusions of an NMOS transistor in a p^- well. The well terminal becomes the base and the source and drain diffusions the emitter and the collector.

The drawback of this device is the inevitable presence of the parasitic n-p-n vertical transistor, which has a saturation current from 4 to 20 times larger than the lateral device. To eliminate the effects of this excess current, a biasing scheme was utilized as shown in Fig. 7, which insures that the lateral transistor maintains a well-controlled collector current, and hence transconductance, even in the presence of large variations in lateral-to-vertical current ratio. This is achieved by placing a dummy transistor Q_1 in a feedback loop, which forces an emitter current I_e, such that the lateral collector current of Q_1 is equal to the desired value I. The current I_e becomes the tail current of the input stage which biases devices Q_3–Q_5. Because of this and assuming that Q_1 and Q_3–Q_5 are matched, the lateral current of transistors Q_3–Q_5 is approximately given by I. Notice that this is independent of the lateral-to-vertical current ratio, and also that this causes the actual tail current of the input stage to vary over a 4- or 5-to-1 ratio while keeping the collector currents constant.

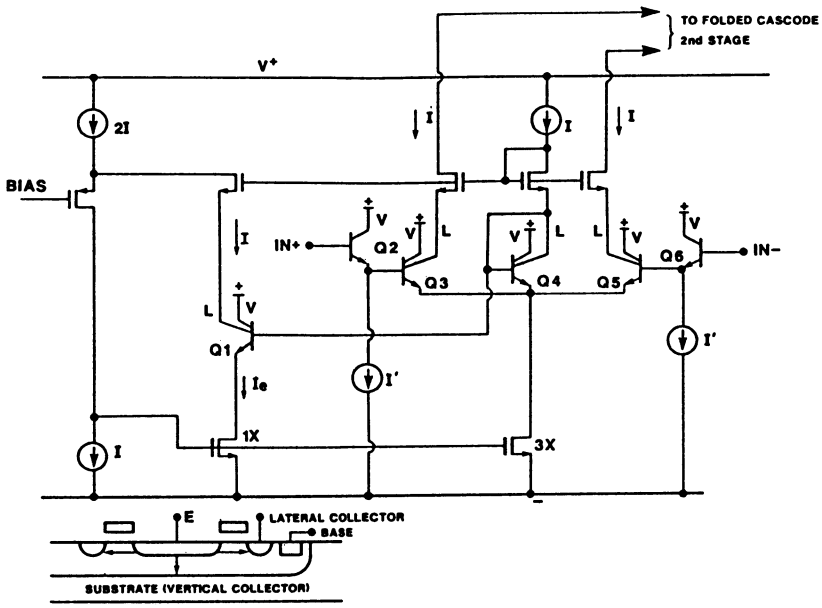


Fig. 7. Simplified schematic of the lateral bipolar input stage. The lateral transistor is shown in cross section at the bottom of the drawing. The gate overlying the lateral base region is tied to the negative power supply.

TABLE II
OTA TYPICAL PERFORMANCE (ROOM TEMPERATURE, ±5-V SUPPLIES)

Parameter	PMOS OTA	NMOS OTA	Bipolar OTA	units
Small-signal transconductance	80	90	200	uA/V
Max output current	10	10	10	uA
Open loop gain	40K	40K	50K	V/V
Unity-gain bandwidth C_L=3pF	2.0	1.8	2.0	Mhz
Output swing	0.6	0.6	0.6	Volts from supply
Input offset volt. (std deviation)	3	3	1	mV
Input bias current	neg	neg	2	na
Equiv. input noise (1khz)	75	75	20	$\frac{nV}{\sqrt{hz}}$

Another important factor in the design is that the vertical-to-lateral current ratio is a strong function of the collector–emitter voltage V_{ce} of the vertical and lateral devices because of the high Early effect of these transistors. For this reason, the dummy device Q_1 is biased off the common-mode point of the input stage, so as to keep the V_{ce}'s approximately the same as that of Q_3–Q_5. This is particularly important, for example, in voltage followers where the input common-mode voltage can be quite high. Transistor Q_4 is used to derive the bias voltages for the cascode n-channel devices, which are used to avoid the degradation in common-mode rejection ratio (CMRR) caused by the poor Early voltage of these bipolar devices. Transistors Q_2 and Q_6 are used to decrease the input bias current due to the low beta of the laterals (typically 50). The remainder of the amplifier is a folded-cascode second stage in which the p- and n-channel current source devices have been made large in geometry, interdigitated, and operated at relatively large $V_{d\,sat}$ in order to reduce their contribution to input offset voltage. This is also true of the NMOS current sources which bias the input transistors Q_2 and Q_6.

Over a large sample of units, the input offset voltage of this amplifier displays a mean value of 0.25 mV, and a standard deviation of 1 mV in a sample taken from many wafers. The typical input bias current is 2 nA. The performance parameters of all three OTA's are summarized in Table II.

For driving off-chip loads and resistive on-chip loads, four class *AB* unity-gain output stages are provided which are capable of driving from 10 kΩ down to 300 Ω with full-swing signals, and lower resistive loads with corresponding lower output swing. A complete power operational amplifier can be assembled by combining one of these output stages with one of the amplifiers mentioned above. The configuration used is the conventional composite-device common-source class *B* [4].

Additional building blocks include a one-pin crystal oscillator, a voltage control oscillator (VCO) capable of operating with a center frequency of up to 5 MHz, and two bandgap references with different area–performance trade-offs. The precision reference is oriented toward high-performance applications and is polysilicon fuse trimmable to 0.1 percent with a typical tempco of 20 ppm/C at room temperature. This cell is described in more detail elsewhere [7]. In addition, the library contains a large number of utility functions such as 5–10-V logic level shifters, analog switches, analog multiplexers, clock generators for switched capacitor filters, and so forth. Finally, these analog cells are supplemented with a conventional family of 75 digital cells which implement most common logic functions.

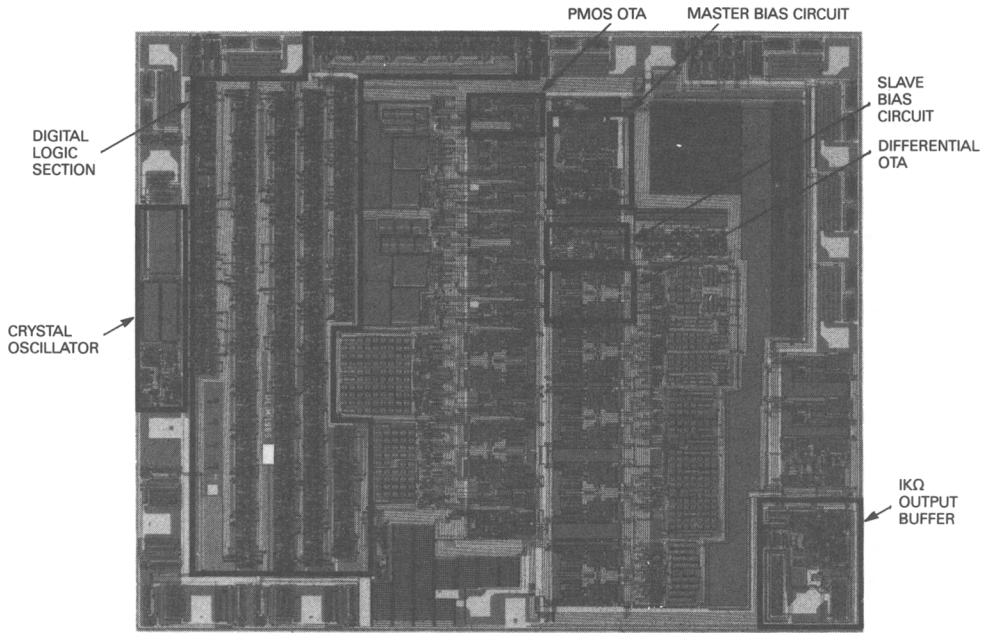


Fig. 8. Tone signaling chip example. The chip contains 20 operational amplifiers and 5000 gates of digital logic. The die size is 25 200 mils^2 (16.3 mm^2).

From a geometric viewpoint, internal cells containing active circuitry are laid out at constant height of 500 μm with standard placement of power supply, ground, and bias lines running along the top and bottom interior of the cells so that these lines are automatically routed when the cells are placed abutting each other. I/O cells such as analog output buffers are configured so that they can lie directly on the chip periphery and incorporate pads. Passive components and random transistor level circuitry are hand placed and interconnected in the areas above and below the assemblages of active cells, although the cell layouts are compatible with the ultimate use of analog place and route packages and block compilers. In a typical application the cells are combined into blocks as just described, and these blocks together with automatically placed and routed digital blocks are assembled and interconnected to compose the complete chip.

VI. Higher Level Analog Functional Blocks

More complex functions such as filters and A/D converters are made up from these basic cells together with additional passive and active circuitry. In order to address a wide spectrum of applications, a very large number of the higher level cells is required in order to adequately cover the spectrum of performance at economical die area and cost. Thus the number of cells in a practical library will continually expand over time as the spectrum of applications expand. Initially, the D/A conversion function is provided in this library by a conventional untrimmed current switched segmented 300-ns DAC cell at the 8-bit level. At the 12-bit level a self-calibrated cyclic D/A cell is utilized. The details of operation of this cell have been described elsewhere [7]. For the A/D function, the same algorithmic self-calibrating cell is used at the 12-bit level, giving a conversion time of 25 μs. This cell incorporates both the sample/hold function as well as a programmable gain function. For the 8-bit level, a two-step flash 1.5-μs cell is used with no trimming. These are not fixed-height internal cells but are placed as large blocks in the chip layout and manually interconnected.

Switched-capacitor filters are implemented using internal fixed-height operational amplifier, bias, and clock generator cells together with manual placement and interconnect of capacitor elements. Both single-ended and fully differential filter architectures can be accommodated with this approach.

A. Example

The application of the cell library described here is illustrated by the chip shown in Fig. 8. This integrated circuit is used for tone signaling on trunk circuits and contains switched-capacitor filters, peak detectors, timers, and so forth for performing precision voice-band tone detection. In addition it provides an audio speech-path filtering function. Various of the cells mentioned above are used on this chip and are indicated in the photograph. Most of the digital cells are concentrated on one side of the die and were manually placed and routed in this case, although automatic place and route has been subsequently used. The analog blocks implemented in the rest of this chip include continuous active *RC* filters, differential switched-capacitor filters, discrete-time gain blocks, tone detectors, and so forth. Cells included in this die are the master bias, slave bias, differential OTA, single-ended operational amplifiers with resistive drive capability, com-

parators, one-pin crystal oscillator, and a number of smaller cells.

A second example of the application of the cell library is the telephone trunk equalizer, discussed elsewhere in this issue [6], which illustrates the dynamic range capability of the library.

References

[1] C. Laber, C. Rahim, S. Dreyer, G. Uehara, P. Kwok, and P. R. Gray, "A high-performance 3μ CMOS analog standard cell library," in *Proc. IEEE 1986 Custom Integrated Circuits Conf.* (Rochester, NY), May 1986, pp. 21–24.

[2] C. Hu, "Hot electron effects in MOSFETS," in *IEDM Tech. Dig.*, 1983.

[3] T. Choi *et al.*, "High-frequency CMOS switched-capacitor filters for communications applications," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 652–665, Dec. 1983.

[4] B. K. Ahuja *et al.*, "A programmable CMOS dual interface processor for telecommunications applications," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 892–899, Dec. 1984.

[5] M. Degrauwe, E. Vittoz, and H. Oguey, "A family of CMOS compatible bandgap references," in *Dig. Tech. Papers, IEEE Int. Solid-State Circuits Conf.* (New York, NY), Feb. 1985.

[6] C. F. Rahim, C. A. Laber, B. L. Pickett, and F. J. Baechtold, "A high-performance custom standard-cell CMOS equalizer for telecommunications applications," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 2, pp. 174–180, Apr. 1987.

[7] M. Armstrong, H. Ohara, H. Ngho, and P. R. Gray, "A self-calibrating CMOS 13-bit self-calibrating analog interface processor," in *Dig. Tech. Papers, IEEE Int. Solid-State Circuits Conf.* (New York, NY), Feb. 1987.

A MOS Switched-Capacitor Instrumentation Amplifier

ROBERT C. YEN AND PAUL R. GRAY, FELLOW, IEEE

Abstract—**This paper describes a precision switched-capacitor sampled-data instrumentation amplifier using NMOS polysilicon gate technology. It is intended for use as a sample-and-hold amplifier for low level signals in data acquisition systems. The use of double correlated sampling technique achieves high power supply rejection, low dc offset, and low $1/f$ noise voltage. Matched circuit components in a differential configuration minimize errors from switch channel charge injection. Very high common mode rejection (120 dB) is obtained by a new sampling technique which prevents the common mode signal from entering the amplifier. This amplifier achieves 1 mV typical input offset voltage, greater than 95 dB PSRR, 0.15 percent gain accuracy, 0.01 percent gain linearity, and an rms input referred noise voltage of 30 μV/input sample.**

Manuscript received April 27, 1982; revised June 7, 1982.

This work was sponsored by the Xerox Corporation and the National Science Foundation under Grant ENG78-11397.

R. C. Yen was with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720. He is now with the Hewlett-Packard Laboratory, Hewlett-Packard, Inc., Palo Alto, CA 94304.

P. R. Gray is with the Department of Electrical Engineering and Computer Sciences and the Electronic Research Laboratory, University of California, Berkeley, CA 94720.

I. INTRODUCTION

ANALOG data acquisition systems often need to perform analog-to-digital conversion on signals of very small amplitude or signals superimposed on large common mode components. This problem is traditionally solved by using fixed gain differential amplifier implemented as a stand-alone component in bipolar technology [1], [2]. However, MOS technology is increasingly being utilized to implement monolithic data acquisition systems, either as a stand-alone component or as part of a control-oriented microcomputer or signal processor [3].

A typical data acquisition system generally consists of an input amplifier, sample-and-hold stage, and an A-to-D converter. The amplifier serves to increase the signal level prior to analog-to-digital conversion. Input offset voltage is a key aspect of amplifier performance since it can limit dc system accuracy. In some systems, it is possible to measure and subtract the dc offset, but the equivalent input noise voltage represents a fundamental limit on the resolution of the system.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-17, no. 6, pp. 1008–1013, Dec. 1982.

Also, gain accuracy and gain linearity are critical parameters for instrumentation applications. In some cases, a low-amplitude signal input is superimposed on large common mode components due to electrostatic or electromagnetic induction. This adds a requirement for high common mode rejection. The data acquisition circuit may reside on the same chip as the digital LSI processor; therefore, the ability to reject power supply noise is also very important.

Compared to bipolar devices, MOS transistors display smaller transconductance at a given drain current level. This makes it difficult to achieve large values of voltage gain in a single MOS amplifier stage, and also results in high dc offsets in source coupled pairs. Also, the MOS devices inherently displays much larger $1/f$ noise than bipolar devices.

This paper describes a switched-capacitor circuit technique for the implementation of the instrumentation amplifier and sample/hold function in MOS technology. Double correlated sampling [4] is used to reduce the circuit dc offset and low-frequency noise, and a balanced circuit configuration is used to achieve first-order cancellation of switch channel charge injection. A charge redistribution scheme is described in this paper which allows the circuit CMRR to be independent of the op amp CMRR, thus resulting in a very high overall common mode rejection ratio.

In Section II, a circuit approach to implement the sample-and-hold instrumentation amplifier is described. The prototype implementation of this circuit using NMOS technology is depicted in Section III. The switch channel charge injection problems are addressed in Section IV, and the fundamental noise limitation of kT/C noise is also discussed in Section V. Finally, in Section VI, the experimental results of this circuit fabricated using local oxidation NMOS polysilicon gate technology are presented.

II. Circuit Description

The MOS implementation of the differential double-correlated sampling amplifier is shown in Fig. 1. This circuit consists of a pair of sampling capacitors $C1$, $C2$; gain setting capacitors $C3$, $C4$; offset cancellation capacitors $C5$, $C6$; and two differential amplifiers $A1$ and $A2$ where amplifier $A1$ is a broad-band low-gain differential preamplifier and $A2$ is a high-gain differential operational amplifier. The input signal is sampled on to the sampling capacitors $C1$, $C2$, and subsequently transferred to the gain setting capacitors $C3$, $C4$ through a sequence of switching operations. The output voltage will be a replica of the input differential signal with a voltage gain defined by the capacitor ratio $C1/C3$ if capacitor $C1$ matches $C2$ and $C3$ matches $C4$. The circuit is fully differential so that all the switch charge injection and power supply variations are cancelled to the first order.

Operation of the circuit takes place in two phases as illustrated in Fig. 2. In the sample mode, the switches are closed as shown in Fig. 2(a). In this mode, the differential and common mode input voltages appear across both $C1$ and $C2$. The difference between the offset voltages $A1$ and $A2$ is impressed across $C5$ and $C6$. The instantaneous value of the $1/f$ noise of both amplifiers is also stored. A requirement on this amplifier, $A1$, is that its gain be low enough so that its output does not saturate on its own offset when the inputs are shorted.

The input signal is sampled and a transition to the hold mode is made when clock one goes negative, turning off the input sampling switches and feedback switches. Subsequently, the switches are closed as shown in Fig. 2(b), and the voltage difference between the two inputs is forced to zero. This causes a charge redistribution in capacitors $C1$, $C2$, $C3$, and $C4$, which results in an output voltage which is only proportional to the input difference voltage. Any common mode input voltage will not cause charge redistribution error, even if the capacitors do not match each other exactly. Another requirement of the amplifier $A1$ is that it must have a high enough bandwidth such that the loop stability is assured in this mode. Differential amplifiers $A1$ and $A2$ together must provide enough loop gain to achieve the desired closed-loop gain accuracy.

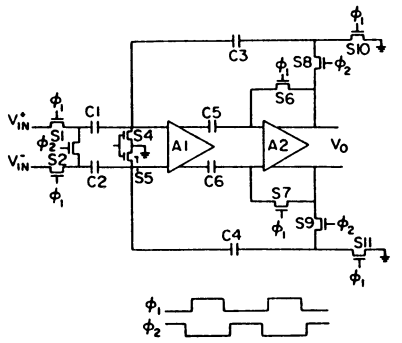


Fig. 1. Circuit schematic of differential double-correlated sampling instrumentation amplifier.

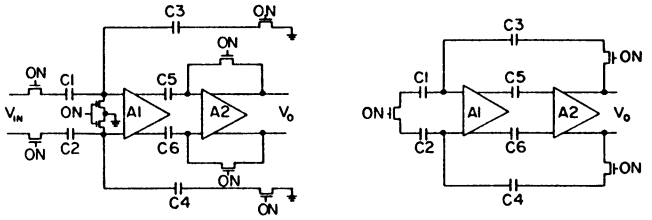


Fig. 2. Illustration of the switching sequence of the instrumentation amplifier.

This circuit has several advantages compared to other techniques. Because the amplifier does not experience any common mode shift, the overall common mode rejection of the circuit is independent of the common mode rejection of the operational amplifier. Because of the balanced nature of this circuit, switch charge injection and clock feedthrough are cancelled to the first order. Because of the equal and opposite voltage excursion on the capacitors, the capacitor nonlinearity is also cancelled to the first order. The sampling bandwidth of this circuit is determined by the RC time constant of the input switch and capacitor, which is usually much faster than the settling time of an operational amplifier. The gain is set by capacitor ratios, which has good initial accuracy [5], very good temperature stability, and is trimmable. Both the $1/f$ noise and the dc offsets are reduced by the use of double correlated sampling; and as a result of this fact and the balanced nature of the circuit, the power supply rejection is also very high.

The overall performance of the circuit is limited by the mismatch of charge injection from the input switches. In this switched-capacitor instrumentation amplifier circuit, cancellation of switch channel charge injection is guaranteed by the

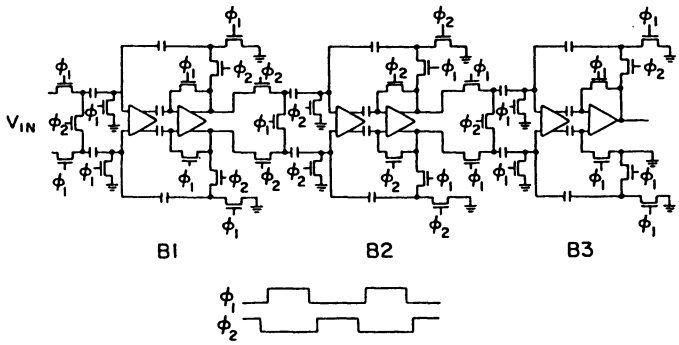


Fig. 3. Experimental implementation of a programmable single-ended output instrumentation amplifier.

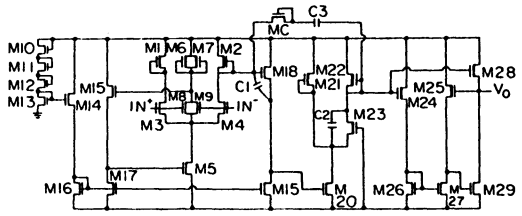


Fig. 4. Circuit schematic of single-ended NMOS operational amplifier.

symmetry of the circuit, and the offset becomes limited primarily by the mismatches of the switch charge injection. The mismatch in the switch channel charge is determined by mismatches in device parameters such as threshold voltage, channel geometry, and so forth. Experimental results to be presented later indicate that for the particular technology used here, the channel charge mismatch in an 8 μm MOS device is typically on the order of one percent.

III. Experimental Implementation of Switched-Capacitor Instrumentation Amplifier

Precision preamplifiers may be required to provide voltage gains from less than 10 to over 1000. The use of a single stage to obtain very large values of voltage gain requires operational amplifiers with very large open-loop gain, and also very large capacitor ratios. In NMOS technology, the voltage gain achievable in operational amplifiers is often limited, and as a result, it is more desirable to use a relatively small value of closed-loop gain. High values of overall closed-loop gain can be achieved by either cascading multiple stages, as shown in Fig. 3, or by using a single stage in a recirculating mode [6]. In the example described here, a fixed gain of ten is used. Another problem is the fact that many A-to-D converters require a single-ended input voltage. Thus, a single-ended output referenced to ground must be produced. This can be achieved as shown in Fig. 3 where the first stage of amplification is realized in a fully differential mode, and the last stage uses a single-ended output operational amplifier to generate a single-ended output voltage.

Operational Amplifier Design

The broad-band low-gain differential preamplifier used in this system is a single differential pair with enhancement load devices. The single-ended output operational amplifier shown in Fig. 4 is a conventional NMOS operational amplifier design [7]. Transistors $M1$-$M5$, $M20$-$M23$, and $M24$-$M29$ are the input, gain, and output stages, respectively. $C1$ and $C2$ are feedthrough capacitors, and $C3$ is the Miller compensation capacitor. Transistor MC is a depletion mode resistor for right half-plane zero compensation. The dc bias points of this op amp are set by the replica bias circuit composed of transistors $M6$-$M17$ [8]. The output stage has the ability to drive a large capacitive load without severely degrading the loop phase margin. This op amp realizes a voltage gain of about 1500 with output voltage swing of about 6.5 V (±7.5 V supply).

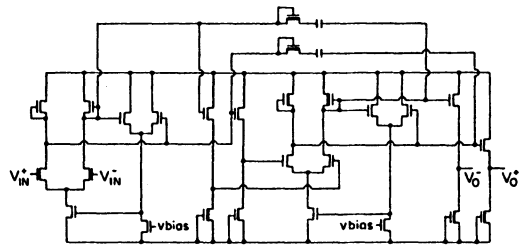


Fig. 5. Circuit schematic of differential NMOS operational amplifier.

A schematic of the fully differential operational amplifier is shown in Fig. 5. Two differential stages are used to achieve a voltage gain of 1500. Common mode feedback is used in these two stages to stabilize the dc bias condition. The output stages are simple source followers, and the capacitors and depletion device resistors are utilized for frequency compensation. Since this op amp is to be used only in the early stages of amplification where the signal swing is small, the output of this amplifier is required to develop a differential voltage of less than 1 V.

IV. MOS Switch-Induced Errors

MOS switches introduce a significant amount of error due to clock voltage feedthrough through the gate-source, gate-drain overlap capacitance and the channel charge stored in the MOSFET device. As shown in Fig. 6, the MOS switch is connected to a sampling capacitor which is charged to the input voltage level. When the MOS switch is turned off, the amount of channel charge injected into the sampling capacitor represents an error source as a result of the sudden release of the charge under the MOS gate. The amount of channel charge that flows into the sampling capacitor as opposed to the amount that flows back to the input terminal is a complex function of the gate voltage fall time, input impedance level, and the size of the sampling capacitor [9]. For a typical switch size of 8 X 8 μm and a sampling capacitor of 5 pF, for example, a 5 V gate overdrive will introduce an error on the order of 20 mV if half of the channel charge flows into the sampling capacitor.

One approach to the reduction of this type of error is the use of large external capacitors [10], [11]. However, the added complexity and the decreased circuit operating speed due to the large external capacitors would limit the usefulness of the circuit as a subsystem of a VLSI processor. A second approach is to use an on-chip capacitor with a dummy switch, as shown in Fig. 7(a), to cancel the charge from the main switch. Unfortunately, all of the channel charge in the dummy switch flows onto the sampling capacitor, while only a fraction of that from the main switch does. As mentioned above, this fraction is a function of gate waveform and source impedance. Another approach is to use dummy switches in combination

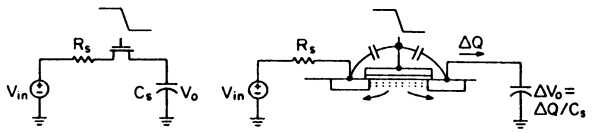


Fig. 6. Clock feedthrough and channel charge injection of an MOS switch.

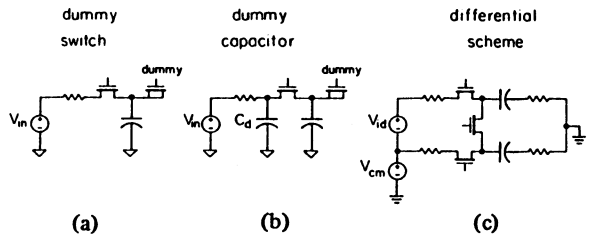


Fig. 7. Illustration of several switch channel charge cancellation techniques.

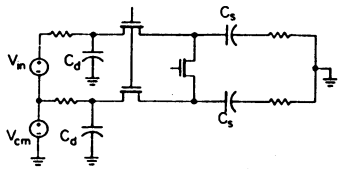


Fig. 8. Differential switch charge cancellation with dummy capacitors.

with dummy capacitors [12] as shown in Fig. 7(b). This configuration assures by symmetry that exactly one half of the channel charge will flow into the sampling capacitor. Thus, the dummy switch with one half the size of the input switch can guarantee the exact cancellation of the channel charge injection and the clock feedthrough problems. This technique works well when the source impedance, clock frequency, and fall time are all well controlled. However, when the signal is driven from an external source whose impedance level is low, the effect of the dummy capacitor on the channel charge cancellation is reduced and becomes a function of the source impedance.

The differential sampling configuration shown in Fig. 7(c) uses two matched switches and capacitors to sample the differential signal. The channel charge injection will introduce the same error voltage on these two sampling capacitors, thus giving no differential error in the sampled values. Although the differential input voltage introduces a difference of channel charge in the two matched switches, this error term is proportional to the input differential voltage, and can be considered as a gain error. One drawback in the configuration shown in Fig. 7(c) is that the channel charge cancellation relies on the matching of the two differential input impedances. However, the effects of source impedance mismatch can be reduced by the inclusion of additional dummy capacitors as shown in Fig. 8 [9].

V. Noise Performance

A key aspect of the performance of the circuit is the noise introduced into the signal path by the circuit, which in this case results from two principal sources, the $1/f$ noise in the operational amplifiers and the thermal noise in the channel resistance of the MOS switches making up the circuit.

An inherent aspect of the offset cancellation technique used is the fact that the noise contribution from the operational amplifier on any sample is the difference between the instantaneous input-referred noise voltage at the time when the input voltage is sampled and when the op amp output is sampled. For the case of the $1/f$ noise, since the noise energy is concentrated at low frequency, the successive samples of the input-referred noise are highly correlated. The noise spectrum which results from this correlated sampling process has been treated analytically elsewhere [13]. Assuming that all of the $1/f$ noise energy is concentrated below the sampling frequency, the spectrum of the noise added to the signal path has the form

$$\bar{\sigma}_{eq}^2(w) = S(w) * \{1 + |\mathrm{sinc}(wT/2)|^2 - 2 * \mathrm{sinc}(wT/2) * \cos(wT)\}$$

where $S(w)$ is the spectral density of the original input-referred $1/f$ noise and T is the sampling period. In the case of the amplifier described here where the sampling rate is in the 200 kHz range and the $1/f$ noise corner frequency is in the 10 kHz range, this reduction in noise contribution at low frequencies is enough such that the other noise source, kT/C noise, is dominant.

The fundamental limitation on the noise performance of the circuit results from thermal noise in the MOS switches. When the switches are on in the sampling mode, a low-pass filter is formed by the channel resistance of the switch and the sampling capacitor which band limits this noise. The resulting mean-square noise voltage appearing across the capacitor is kT/C. When the switch is opened, this noise is sampled, with the result that each sample of the incoming signal has a noise sample added to it. The amplitude distribution of these noise samples is Gaussian with a standard deviation of the square root of kT/C V. For example, for a 10 pF sampling capacitor, this standard deviation is 25 μV at room temperature.

A frequent application of an amplifier of this type would be in a signal processing system in which it samples an input signal periodically, after which the samples are processed to form an output sequence which is subsequently converted to analog form with a DAC and sample/hold and smoothed with a reconstruction filter which passes output spectral components up to one half the sampling rate. For this case, it can be shown that the contribution of the kT/C noise is equivalent to that of a continuous time white noise source at the input with a spectral density of 2 kT/fC where f is the sampling frequency. Again, assuming a 10 pF capacitor and a 200 kHz sampling rate, the equivalent input noise spectral density would be 65 nV/$\sqrt{\text{Hz}}$ or equivalent to the noise in a 250 kΩ resistor. The noise behavior of the amplifier is very similar to that of a switched-capacitor integrator with a 10 pF sampling capacitor. Increasing the sampling rate decreases the input noise density, but the total noise energy below the sampling frequency remains constant.

VI. Experimental Results

Experimental circuits for the fully differential stage and single-ended output stage were designed and fabricated using

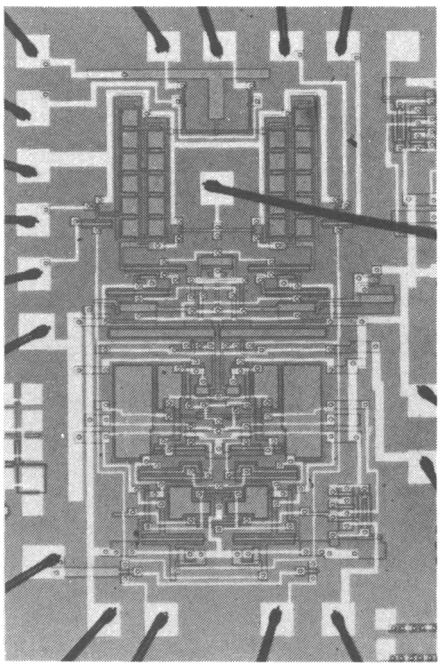


Fig. 9. Die photo of a fully differential instrumentation amplifier gain block.

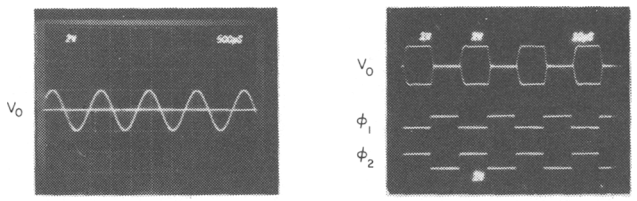


Fig. 10. Measured circuit output waveform.

a local oxidation polysilicon gate NMOS process with depletion load. The minimum geometry used in these circuits was 7 μm. The input sampling capacitors were 10 pF each, feedback capacitors were 1 pF, and the DCS capacitors were 3 pF.

A die photo of the switched-capacitor amplifier with differential output stage is shown in Fig. 9. The top part of the die photo contains the matched capacitor arrays; the bottom part contains the differential amplifier. The symmetrical layout of the circuit is crucial to the matching of circuit components. The die area of this circuit is 2500 mils^2.

In Fig. 10(a), the output waveform is shown with a 1 kHz sinusoidal input signal. The staircase-shaped waveform is the result of the sample-and-hold operation. Shown in Fig. 10(b) is the output waveform with a square-wave input signal. The output is reset to its own offset value during the sample mode when clock 1 is high, and generates an amplified input sample during the hold mode when clock 2 goes high.

The experimentally observed input offset voltage and common mode rejection ratio are shown in Fig. 11 for five typical samples of the circuit as a function of source resistance. For small values of source resistance, the average value of the input

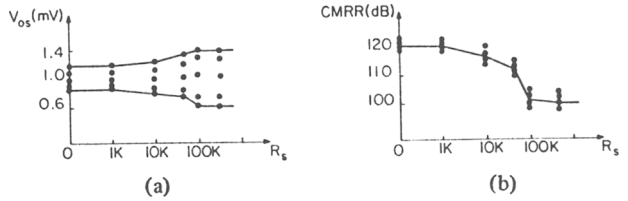


Fig. 11. Experimentally measured V_{os} and CMRR as a function of source resistance value.

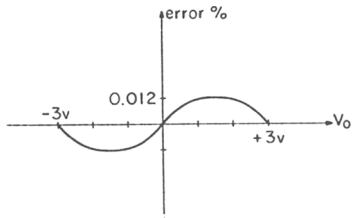


Fig. 12. Experimentally measured gain nonlinearity.

TABLE I
SUMMARY OF EXPERIMENTAL RESULTS; 25°C (15 SAMPLES)

Gain accuracy (G = 10)	
average value	0.15 percent
standard deviation	0.03 percent
Input offset voltage	
average value	1 mV
standard deviation	0.5 mV
Input offset drift	2 μV/C
Gain linearity	0.01 percent
Common mode rejection	
dc	120 dB
10 kHz	95 dB
Power supply rejection	
dc	>95 dB
1 kHz	>95 dB
Equivalent input noise (200 kHz sampling freq.)	65 nV/$\sqrt{\text{Hz}}$
Input voltage range	+4 V, -6 V
Power supply	+7.5 V, -7.5 V
Power dissipation	10 mW

offset voltage is 1 mV; this actually results from a layout-induced systematic offset in the amplifier offset cancellation circuit and not from charge injection in the input switches. The spread of this offset is about 500 μV, which increases at high values of source resistance because the percentage of the channel charge injected into the sampling capacitors increases [9]. The CMRR at low values of source resistance is about 120 dB. This is also degraded at high values of source resistance for the same reason as the offset voltage. The CMRR and PSRR of this circuit were measured using a spectrum analyzer. A large single frequency sinusoidal signal was applied to the common mode input and connected in series with the power supply, respectively, and the magnitude of the output component was measured.

Fig. 12 illustrates the typical gain nonlinearity observed for the device. This nonlinearity results primarily from the nonlinear open-loop characteristic of the operational amplifier. The peak deviation from linear behavior is about 0.012 percent of full scale at plus and minus 3 V output swing. Table I

shows a summary of the data measured at room temperature for power supply voltages of plus and minus 7.5 V.

VII. Conclusion

This paper has described one approach to the implementation of the instrumentation amplifier/sample-hold function in MOS technology. The performance levels achieved are generally somewhat inferior to recently reported bipolar instrumentation amplifiers. The significance of the results achieved is that the compatibility with MOS technology allows a higher level of integration in the analog data acquisition interface with the digital interfacing and/or processing circuitry without adding to the complexity of basic digital MOS technologies. The experimental amplifier described in this paper was implemented in NMOS technology, but it appears that a CMOS implementation could achieve significantly higher operating speed because of the higher available gain per stage and the resulting improvement in operational amplifier gain and bandwidth.

Acknowledgment

The authors would like to thank Prof. D. A. Hodges for his support on this project. The discussion with many graduate students in the Berkeley IC Group on the development of the IC process is also acknowledged.

References

[1] M. Timko and A. P. Brokaw, "An improved monolithic instrumentation amplifier," in *Dig. 1975 ISSCC*, Feb. 1975.
[2] C. T. Nelson, "A 0.01% linear instrumentation amplifier," in *Dig. 1980 ISSCC*, Feb. 1980.
[3] M. Townsend, M. E. Hoff, Jr., and R. E. Holm, "An NMOS microprocessor for analog signal processing," *IEEE J. Solid-state Circuits*, vol. SC-15, Feb. 1980.
[4] J. Stremler, *Introduction to Communication Systems*.
[5] J. L. McCreary, "Matching properties and voltage and temperature dependence of MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-16, Dec. 1981.
[6] R. H. McCharles and D. A. Hodges, "Charge circuits for analog LSI," *IEEE Trans. Circuits Syst.*, vol. CAS-25, July 1978.
[7] Y. P. Tsividis and D. L. Fraser, Jr., "A process insensitive NMOS operational amplifier," in *Dig. 1979 ISSCC*, Feb. 1979.
[8] P. R. Gray, D. A. Hodges, and R. W. Brodersen, Eds., *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980.
[9] R. C. Yen, "High performance MOS circuits," Ph.D. dissertation, Univ. California, Berkeley, Dec. 1982.
[10] R. Poujois and J. Borel, "A low drift fully integrated MOSFET operational amplifier," *IEEE J. Solid-State Circuits*, vol. SC-13, Aug. 1978.
[11] M. Coln, "Chopper stabilization of MOS operational amplifiers using feed-forward techniques," *IEEE J. Solid-State Circuits*, vol. SC-16, Dec. 1981.
[12] L. A. Bienstman and H. DeMan, "An 8-channel 8b μP compatible NMOS converter with programmable ranges," in *Dig. 1980 ISSCC*, Feb. 1980.
[13] R. W. Brodersen and S. P. Emmons, "Noise in buried channel charge coupled devices," *IEEE J. Solid-State Circuits*, vol. SC-11, Feb. 1976.

A Micropower CMOS-Instrumentation Amplifier

M. DEGRAUWE, E. VITTOZ, MEMBER, IEEE,
AND I. VERBAUWHEDE

Abstract—**A CMOS switched capacitor instrumentation amplifier is presented. Offset is reduced by an auto-zero technique and effects due to charge injection are attenuated by a special amplifier configuration. The circuit which is realized in a 4-μm double poly process has an offset (σ) of 370 μV, an rms input referred integrated noise ($0.5-f_c/2$) of 79 μV, and consumes only 21 μW ($f_c = 8$ kHz, $V_{DD} = 3$ V).**

I. Introduction

For the realization of intelligent sensors an instrumentation amplifier is very often required in order to detect small differential signals in the presence of a large common mode signal. Such amplifiers should have very low offset and $1/f$ noise, large CMRR, and especially for biomedical applications consume as less as possible current and be able to operate at a low supply voltage.

First an offset cancellation technique will be presented. Further the amplifier realization will be discussed and experimental results will be given.

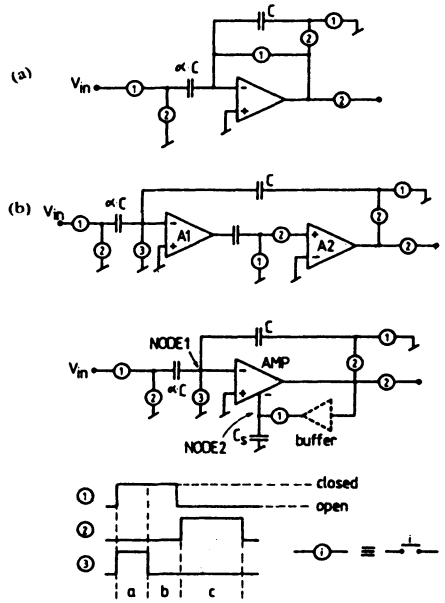


Fig. 1. Auto-zero techniques. (a) Information stored at the input. (b) Information stored after a first gain stage. (c) Information stored at a low sensitive auxiliary input.

II. Auto-Zero Realizations

Offset cancellation by means of auto-zero techniques can be implemented in different ways [1]. Up to now basically two approaches are used.

The first method (see Fig. 1(a)) consists of storing the offset information at the input of the amplifier. Ideally the stored information will be equal to the offset value. However due to charge injection of the switch S_1, this value will be significantly changed resulting in a degraded offset cancellation.

A second method (see Fig. 1(b)) [2] consists of storing the offset information at the output of the first amplifier stage. In practice the stored offset information will be about 100 times the offset value. Therefore, the charge injection of the switches S_1-S_2 will not significantly degrade the offset cancellation. However the need of a two-stage amplifier gives rise to potential stability problems, degraded noise, and PSRR performances [3].

In this correspondence an alternative auto-zero topology is presented (Fig. 1(c)) [5], [8].

AMP is an ordinary amplifier with gain A_1 (between node 1 and output) and offset V_{off1}. An auxiliary input (node 2) of reduced sensitivity (gain A_2) is added to the main amplifier. It is controlled by a compensation voltage V_2 stored in capacitor C_s. The buffer is added to speed up the compensation phase.

The compensation works as follows. During time slot "a", the input signal is sampled. In the same time the input terminals of the main amplifier are short-circuited. This amplifier will thus amplify its own offset. However, due to the feedback path through the auxiliary input, the output voltage will be stabilized and across the store capacitor C_s there will be a voltage equal to

$$V_2 = \frac{-A_1 V_{off1}}{1+A_2}. \tag{1}$$

At the beginning of time slot "b", the switches "3" are opened and a charge is injected at node 1. This causes an additional offset ΔV_1. The feedback mechanism is however still active and the voltage across C_s is now given by

$$V_2 = \frac{-A_1(V_{off1}+\Delta V_1)}{1+A_2}. \tag{2}$$

At the end of time slot "b" the switch "1" opens and causes charge injection (ΔV_2) at the node 2. The voltage at this node is now given by

$$V_2 = \frac{-A_1(V_{off1}+\Delta V_1)}{1+A_2} + \Delta V_2. \tag{3}$$

At the output node appears then a voltage equal to

$$V_{out} = -A_1 \cdot \left(\frac{(V_{off1}+\Delta V_1)}{1+A_2} + \Delta V_2 \cdot \frac{A_2}{A_1} \right) \tag{4}$$

which corresponds to an equivalent input which is A_1 times smaller.

Finally during the time slot "c" the charge stored on the capacitor αC (sample of the input signal) is transferred to the integration capacitor.

The equivalent offset of the whole amplifier contains thus two residual parts

the sum of the initial offset and the charge injection at node 1 which both are attenuated by a factor $(1+A_2)$; and

the charge injection at node 2 which is attenuated by a factor A_1/A_2.

The optimal value of the gain of the auxililary input is obtained by differentiation of (4)

$$A_2 = \left(\frac{(V_{off1}+\Delta V_1)A_1}{\Delta V_2} \right)^{1/2} - 1 \tag{5}$$

Manuscript received October 29, 1984; revised December 20, 1985. This work was partially supported by the "Fonds National Suisse pour la Recherche Scientifique, PN13."

M. Degrauwe and E. Vittoz are with Centre Suisse d'Electronique et de Microtechnique S.A. CSEM—Recherche & Développement—(formerly CEH) Maladière 71, 2000 Neuchâtel 7, Switzerland.

I. Verbauwhede was with CEH in 1983 on leave from Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3030 Heverlee, Belgium.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-20, no. 3, pp. 805–807, June 1985.

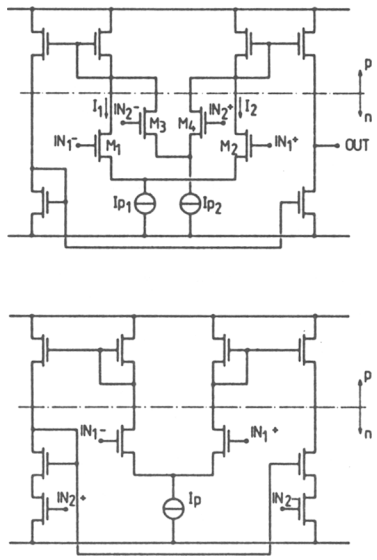


Fig. 2. Amplifier with auxiliary input. (a) Two parallel input stages. (b) Degenerated current mirror.

and results in a residual offset of

$$V_{off} = 2 \cdot \left(\frac{(V_{off1} + \Delta V_1) \cdot \Delta V_2}{A_1} \right)^{1/2} \quad (6)$$

The choice of A_1 and A_2 has to be so that the amplifier never saturate. Therefore the method will result in a smaller achievable residual offset for larger supply voltages. Further, it can be shown that the offset of the auxiliary amplifier can be neglected, provided it is of the same order of magnitude as the input offset.

From (5) it is seen that the optimum value of the auxiliary input depends on the initial offset of the main amplifier and on the clock injection on nodes 1 and 2. For small values of $(V_{off1} + \Delta V_1)$, the gain A_2 should be small and for large $(V_{off1} + \Delta V_1)$, the gain A_2 should be larger. For optimum compensation, the gain A_2 should thus not be constant. Recently it has been shown that a quadratic auxiliary input results in a better offset reduction [4].

If a fixed gain A_2 is used, this gain should be optimized for the largest expected values of the initial offset and of the clock feedthrough.

III. Circuit Configuration and Experimental Results

There are several ways to add an auxiliary input at a conventional amplifier. A first method consists of using two parallel input stages (Fig. 2(a)). The ratio of the gain A_1/A_2 will be determined by the ratio of the transconductances of the two input stages. The bias current I_{p2} can however not be chosen arbitrary small. Due to the offset voltage V_{off1}, the current I_1 and I_2 can deviate as much as 10 percent of their ideal value. Therefore, the current I_{p2} should be at least 10 percent of I_{p1}. Large A_1/A_2 can thus only be achieved by operating transistors $M_3 - M_4$ much deeper in strong inversion than transistors $M_1 - M_2$. This will however result in a large voltage drop across $M_3 - M_4$ which can be unacceptable for low-voltage battery operation.

An alternative method consists of degrading a current mirror ratio by inserting transistors operating in the linear region (Fig. 2(b)) [6], [7]. In this case the second input is realized with no additional current consumption. However, the gain A_1/A_2 will now depend on the supply voltage which will result in a reduced PSRR. For battery operation, the specifications for the PSRR can however be somewhat relaxed.

An SC instrumentation amplifier was developed according to the principles of Figs. 1(c) and 2(b). The circuit was realized in a differential way in order to further improve the performances [2].

The amplifier realization is shown in Fig. 3 (sampling and integrating capacitors are not shown). The main amplifier $M_1 - M_{15}$ is a differential transconductance amplifier whose input stage has a low gain. The secondary inputs are realized by adding four transistors to the main amplifier

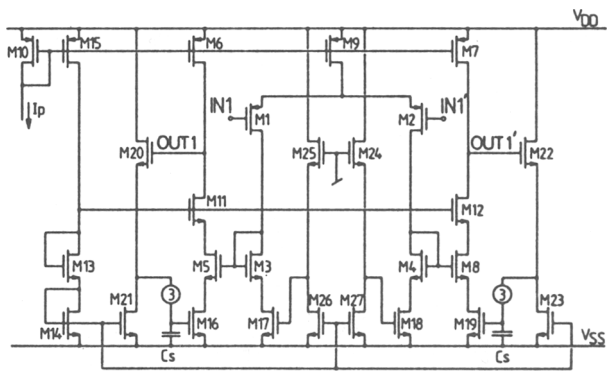


Fig. 3. Realized amplifier.

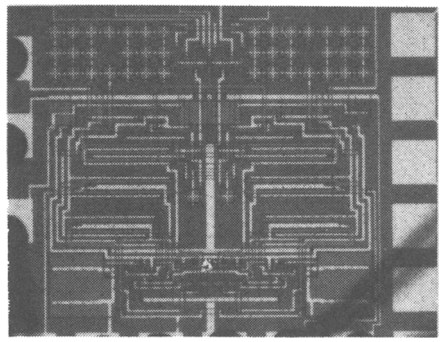


Fig. 4. Chip photograph.

TABLE I
Measurement Results

Gain ($= \propto C_1/C_2$)	20 dB
Max. clock frequency (C_L = 22 pF)	8 kHz
Offset (at 8 kHz) $\bar{x}$	90 μV
Offset (at 8 kHz) σ	370 μV
Equivalent input noise (0.5 Hz-4 kHz)	79 μV
No 1/f noise was observed above 0.5 Hz (= under limit of measurement equipment)	
CMRR	≥95 dB
Current consumption	7 μA
$PSRR^-$ (at DC)	54 dB
$PSRR^+$ (at DC)	66 dB

($M_{16} - M_{19}$). Those transistors who operate in the linear region can modulate the current ratio of two current mirrors. The buffer stages are simple source followers ($M_{20} - M_{23}$).

The circuit has been realized in a 4-μm double poly CMOS process. A chip photograph is shown in Fig. 4. The total chip area is 640 μm × 700 μm ($\simeq 0.45$ mm^2).

The most important measurement results are given in Table I. The standard deviation of the offset is 370 μV. Further reduction of the offset can be obtained by combining the offset cancellation technique of Fig. 1(a) and (c).

Recently the presented circuit has been redesigned in a 3-μm technology [9]. Measurement results of this circuit will be reported very soon [10].

IV. Conclusions

A micropower instrumentation amplifier has been presented which has a typical offset of 370 μV and consumes only 21 μW. The performances

are achieved by the use of a new offset cancellation technique in which the effects of charge injection are attenuated. For larger supply voltages ($\cong 10$ V) residual offsets of less than 50 μV are obtainable with the presented technique.

Acknowledgment

The authors wish to thank Dr. H. Oguey for useful discussions.

References

[1] R. Poujois and J. Borel "Low-level MOS transistor amplifier using storage techniques," in *ISSCC Dig. Tech. Papers*, 1973, pp. 152–153.

[2] R. C. Yen and P. R. Gray, "A MOS switched capacitor instrumentation amplifier," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1008–1013, Dec. 1982.

[3] B. J. Hosticka, W. Brockherde, and M. Wrede, "Effects of the architecture on noise performance of CMOS operational amplifiers," in *Proc. ECCTD*, 1983.

[4] E. Vittoz, "Dynamic analog techniques," "Advanced Summer Course on 'Design of MOS-VLSI Circuits for Telecommunications,'" L'Aquila, Italy, June 18–29, 1984.

[5] E. Vittoz and H. Oguey, Swiss Patent Application No. 2179/84, filed on April 5, 1984.

[6] H. Oguey, CEH Rep. HO/81, pp. 56–57, Mar. 1981.

[7] A. Acovic, "Amplificateur à offset compensé," Masters work of the EPFL, Lausanne, July 1983, (in French).

[8] M. Degrauwe, E. Vittoz, and I. Verbauwhede, "A Micropower CMOS instrumentation amplifier," in *Proc. ESSCIRC* '84 (Edinburgh, Scotland), September 1984, pp. 31–34.

[9] I. Verbauwhede, "Design and integration of a low-power CMOS SC-instrumentation amplifier," M.S. thesis, K.U. Leuven, June 1984 (in Dutch).

[10] P. Van Peteghem, I. Verbauwhede, and W. Sansen, "A micropower high performance S.C. building block for integrated low level signal processing," submitted for publication in the *J. Solid-State Circuits*.

A Precision Variable-Supply CMOS Comparator

DAVID J. ALLSTOT, MEMBER, IEEE

Abstract—Several new techniques are presented for the design of precision CMOS voltage comparator circuits which operate over a wide range of supply voltages. Since most monolithic A/D converter systems contain an on-chip voltage reference, techniques have been developed to replicate the reference voltage in order to provide stable supply-independent dc bias voltages, and controlled internal voltage swings for the comparator. These techniques are necessary in order to eliminate harmful bootstrapping effects which can potentially occur in all ac-coupled MOS analog circuits. An actively controlled biasing scheme has been developed to allow for differentially autozeroing the comparator for applications in differential A/D converter systems. A general approach for selecting the gain in ac-coupled gain stages is also presented. The comparator circuit has been implemented in a standard metal-gate CMOS process. The measured comparator resolution is less than 1 mV, and the allowable supply voltages range from 3.5 to 10 V.

I. Introduction

MOS switched capacitor analog circuits [1] have evolved rapidly during the past several years, with particular emphasis on industrial and telephony applications. As the power supplies in these applications are both fixed and relatively large (typically 10 V), depletion-load NMOS and CMOS have emerged as viable technologies, although CMOS is gaining wider acceptance because of its superior analog performance capabilities. This technology trend will continue as these techniques are used in new applications such as consumer and medical electronics where operation with both small supply voltages and large supply voltage compliances are required for battery-powered systems. Additionally, more process- and supply-independent design techniques will be needed for the analog–digital interfaces in the lower voltage digital VLSI technologies.

In this paper, some of these issues are addressed with new design techniques for a precision, variable-supply CMOS comparator. This particular comparator is used in the switched capacitor A/D section [2]-[3] of a through-the-lens autofocusing system for SLR cameras [4]. It is designed to operate with total supply voltages ranging from 3.5 to 10 V, with a nominal quiescent current of 500 μA. It is capable of resolving a 1 mV peak input signal in about 12 μs, although for this application, it was only required to resolve a 4 mV peak signal with a typical switching time of about 3 μs with a 10 V power supply. Differential structures are used extensively to minimize the effects of power supply coupling and to reduce the input-referred offset voltage due to switch feedthrough effects. Replication of the on-chip reference voltage as well as controlled biasing techniques are used to generate stable, supply-independent dc bias voltages for the comparator.

II. Selection of Comparator Topology

The most basic issue in selecting a comparator topology is to determine the minimum required gain. In this design, in order to obtain a 10 V logic swing with a 500 μV peak input signal, a minimum gain of 20 000 is necessary, which obviously requires multiple gain stages. There are several key issues involved in selecting both the number of gain stages, as well as the partitioning of the total gain relative to the placement of the interstage coupling capacitor(s). These considerations are illustrated in the following examples.

An obvious comparator topology is a conventional two-stage differential amplifier as shown in Fig. 1(a). It is initially configured in unity gain through a p-channel reset switch with the

Manuscript received July 16, 1982; revised August 9, 1982.
The author is with Nova Monolithics, Inc., Carrollton, TX 75006.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-17, no. 6, pp. 1080–1087, Dec. 1982.

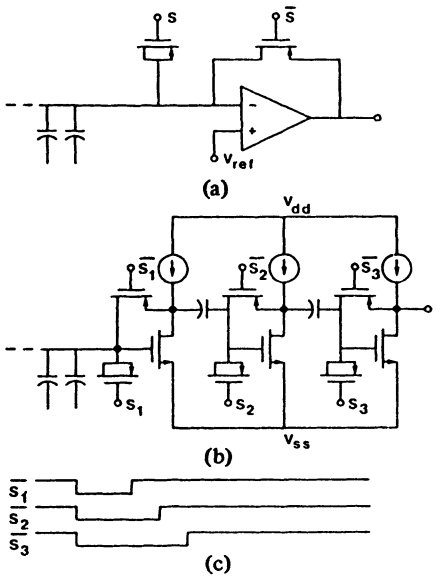


Fig. 1. Possible comparator topologies with offset cancellation. (a) A conventional two-stage differential amplifier which is periodically connected in unity gain. (b) A cascade of n self-biased inverters which are periodically reset using (c) n overlapping clock phases.

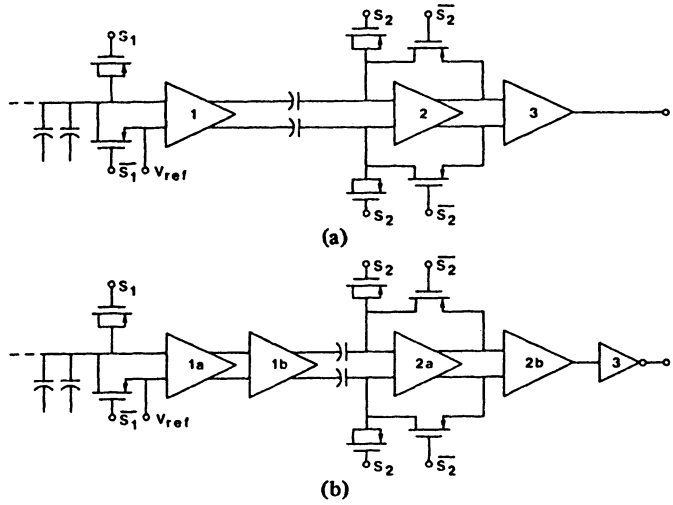


Fig. 2. Differential comparator topology. (a) A simplified block diagram of a three-stage design using only a single set of coupling capacitors as a result of the differentially reset second stage. (b) A more detailed block diagram of the actual implementation showing the partitioning of the five amplifiers relative to the coupling/reset network.

intent of storing the input-referred offset voltage on the common top plate of the A/D capacitor array where it is effectively subtracted during the A/D conversion process. This technique works well for removing the input offset voltage contribution of the amplifier itself. However, the clock-feedthrough [2]-[3] and channel charge-pumping [5] effects of the reset switch contribute a *residual* offset charge which is not completely eliminated, even with the use of a charge cancellation device as shown in the figure. Furthermore, the accuracy of the charge cancellation is strongly dependent on the supply voltage (clock swing) which is not constant in this application. Thus, in attempting to reduce the input offset voltage due to the residual feedthrough charge $V_{os} = Q_{ft}/C_{\text{array}}$, the array capacitance may be increased. Unfortunately, to maintain an equivalent time constant for charging the larger capacitor array, the size (W/L) of the reset and cancellation switches must also be increased, which results in a larger residual feedthrough offset charge, and hence, little or no effective decrease in the input-referred feedthrough offset voltage. Furthermore, the two-stage amplifier is usually pole-split frequency compensated for unity gain, which results in both limited switching speed because of slewing the compensation capacitance and poor high-frequency PSRR due to V_{dd} variations coupling directly to the amplifier output through the compensation capacitor.

Another commonly used comparator configuration is a cascade of ac coupled self-biased inverters [6] as shown in Fig. 1(b) where, for n gain stages, n overlapping reset signals are required [Fig. 1(c)]. By staggering the reset signals in time, only the feedthrough offset associated with the last clock phase contributes directly to the input offset voltage of the comparator as

$$V_{os} = \frac{V_{ft_3}}{A_1 A_2}. \tag{1}$$

All other offset terms are stored on the coupling capacitors. Obviously, the input-referred offset voltage can be made very small if the dc gain preceding the last stage, $A_1 A_2$, is made large. Unfortunately, V_{ft_1} and V_{ft_2} are indirectly very important since the dc output voltages of the first two stages are $A_1 V_{ft_1}$ and $A_2 V_{ft_2}$ (A_i is the dc gain of the ith stage), respectively, which can result in saturated gain stages for even modest dc gains considering the relatively large changes in the V_{ft}'s due to processing and power supply variations. Hence, the gain per stage must be kept small to avoid saturation, which results in a large number of gain stages for precision applications. Due to the use of single-ended amplifiers, another major disadvantage of the inverter cascade is its poor power-supply rejection which can limit the minimum resolvable signal to several millivolts in systems where a large amount of digital circuitry is contributing to power supply noise.

In order to circumvent these problems, as well as new problems which arise as a result of the variable-supply requirement (to be described later), a fully differential topology of the type shown in Fig. 2(a) is desirable. As in the single-ended case, the value of A_1 must be chosen sufficiently small to ensure that V_{ft_1} does not saturate the first stage. However, because of the differential outputs, A_1 can be chosen twice as large as in the single-ended case for the same degree of output saturation. The input-referred offset voltage for this circuit is given by

$$V_{os} = \frac{\Delta V_{ft_2}}{A_1} + \frac{V_{os_3}}{A_1 A_2} \tag{2}$$

where ΔV_{ft_2} is the *difference* in clock feedthrough voltages at the second stage inputs. In other words, the common-mode feedthrough charge terms are eliminated, with only the differential feedthrough referred to the input of the comparator. Although V_{ft_2} is large and varies widely with supply voltage and process changes, ΔV_{ft_2} is much smaller due to the matching and tracking between the V_{ft}'s for the two sides. Furthermore, the degree of saturation at the output of the second stage depends on $A_2 \Delta V_{ft_2}$, implying that for the same degree

of output saturation, the second-stage gain can be significantly increased relative to the single-ended design. Hence, by using the differential resetting and charge cancellation techniques, a precision design (small V_{os}) is obtained with only a single stage of ac coupling. Because the common-mode feedthrough terms are eliminated, the coupling capacitance can be reduced, which is beneficial in higher speed designs. (It should be noted that although the common-mode feedthrough terms do not contribute to the input offset voltage of the comparator, charge cancellation devices should still be included to ensure that the changes in the common-mode dc bias voltages as a result of the switching transients are small.) A more detailed block diagram showing that the three stages actually consist of five amplifiers is shown in Fig. 2(b).

The next section focuses on dc-biasing ac-coupled comparator circuits which is, in general, a very critical issue with variable power supplies.

III. DC Biasing Considerations

DC biasing of the second gain stage will first be considered in order to explain a rather subtle bootstrapping effect which can occur in *any* comparator with ac coupling, but which is much more likely to occur when the power supplies vary widely as in this application. Fig. 3 shows a simplified schematic of the second gain stage including the resetting and ac coupling networks. For reasons to be explained later, this stage is designed so that the dc bias voltage between V_{dd} and the n-channel inputs during the reset time approximately equals the reference voltage V_{ref}. (Note that this could be achieved by simply shorting the gates of the n-channel inputs directly to V_{ref}. However, in order to remove the second-stage offset, this approach requires another set of coupling capacitors with yet another relatively low gain stage.) When the switches are open, C_{in} represents the input capacitance of the second gain stage with C_s as the ac coupling capacitance. Thus, for a voltage change of ΔV_x at node Ⓧ, the voltage change at node Ⓨ is

$$\Delta V_y = \Delta V_x \, C_s/(C_s + C_{in}). \tag{3}$$

Usually, by design, $C_s \gg C_{in}$, and therefore, $\Delta V_y \cong \Delta V_x$. Assuming p-channel reset switches, there exists a parasitic diode (Fig. 3) from node Ⓨ to V_{dd}, namely, the drain-bulk diode of the p-channel switch. For a large positive voltage swing $\Delta V_y > V_{ref}$, node Ⓨ is bootstrapped above V_{dd}, and the diode becomes forward biased, resulting in a parasitic substrate current I_{sub} which has the deleterious effect of removing charge from C_s. Since this effect occurs for a range of intermediate to maximum input signals, it results in a gain change (nonlinearity) in the A/D converter transfer characteristic. Hence, another basic requirement in dc biasing the comparator is to ensure that the bias points and positive internal signal swings are well controlled to prevent this bootstrapping effect from occurring.

In biasing the comparator (as well as other analog circuitry), significant advantage can be taken of the on-chip voltage reference. For example, Fig. 4 shows a simplified schematic of the ΔV_t reference [7]-[8] with resistive trim circuitry as used in this design. By forming the reference voltage relative to V_{dd},

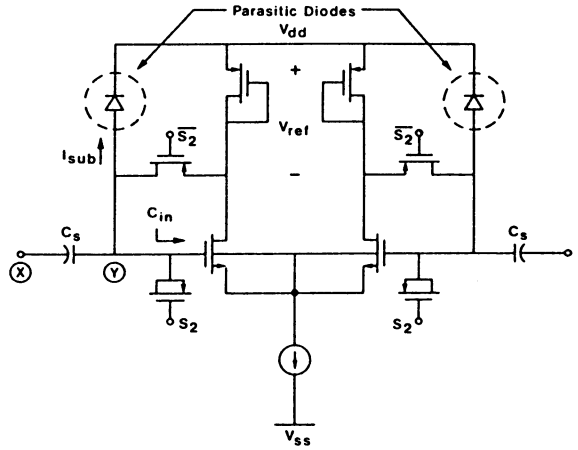


Fig. 3. A simplified schematic of the capacitive network coupling into the second gain stage (a_{2a}) showing the critical parasitic diodes which can conduct current if node Ⓨ is bootstrapped above V_{dd}.

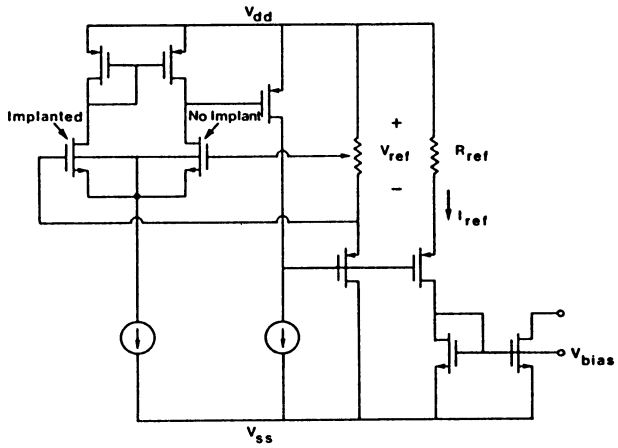


Fig. 4. A simplified schematic of the delta-V_t voltage reference with resistor trim network. An additional p-channel source follower replicates V_{ref} across R_{ref} to generate the bias voltage for other analog circuitry including the comparator.

with careful layout, and by also referencing the input signal(s) to V_{dd}, the substrate noise appears as a common-mode signal and is substantially rejected. In this design, the reference voltage was trimmed to $V_{ref} = 2.048$ V. The measured temperature coefficient of V_{ref} was less than 100 ppm/°C, and the measured dc PSRR was better than 55 dB [9]. In order to exploit these characteristics, circuit techniques were developed which replicated V_{ref} to produce stable supply-independent dc bias voltages for the comparator. Thus, a p-channel source follower was added to the basic reference to replicate V_{ref} across the p-well resistor R_{ref}, resulting in a current $I_{ref} = V_{ref}/R_{ref}$, which subsequently flows into the n-channel current mirror to generate the basic bias voltage V_{bias}. (Note that for biasing analog circuitry which is remote from the reference, it is preferable to supply a current bias since it is less susceptible to the substrate noise coupled through the interconnect stray capacitance.) Although V_{ref} is constant, I_{ref} varies by ±50 percent due to changes in the p-well resistivity over processing and temperature (6400 ppm/°C).

A schematic of the two differential gain stages [labeled 1a and 1b in Fig. 2(b)] preceding the coupling capacitors is shown in

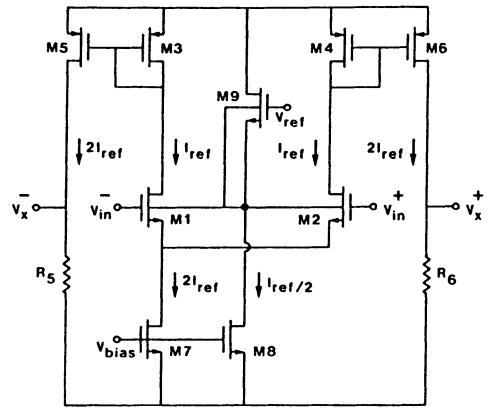


Fig. 5. Comparator input stage consisting of a cascade of two differential amplifiers with the resistive loads used in the second gain stage to replicate the reference voltage. This technique produces supply-independent dc bias voltages and controlled supply-independent output swings for the critical stages of the comparator.

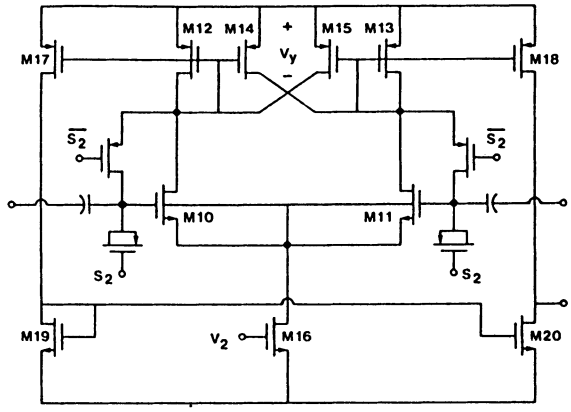


Fig. 6. Comparator second stage consisting of the ac coupling network and a cascade of two differential amplifiers. Positive feedback is used to provide increased gain. The bias voltage $V_2 \neq V_{bias}$ is derived from the bias control amplifier.

Fig. 5 with all bias currents indicated relative to I_{ref}. Rather than conventional active loads, p-well resistive loads are used in this design to achieve replication of V_{ref}.[1] In particular,

$$V_{R_5} = V_{R_6} = 2I_{ref}R_5 = 2\,V_{ref}(R_5/R_{ref}) \tag{4}$$

which generates a very reproducible fraction of V_{ref} since it is determined by a resistor *ratio*. In addition to providing a known bias voltage, the maximum *positive* output swing of the first stage is now well controlled at a value of

$$\Delta V_x(\max) = 2I_{ref}R_5 = 2\,V_{ref}(R_5/R_{ref}) \tag{5}$$

where (R_5/R_{ref}) is chosen to eliminate bootstrapping, i.e., ΔV_x (max) = V_{ref} in this design. (Negative output swings are of no direct concern since bootstrapping can occur only on large positive output swings.)

Note that to a first order, the first-stage bias voltage and output swing are independent of V_{dd}. However, the voltage coefficient of the p-well resistors introduces a second-order V_{dd} dependence since the voltage across R_{ref} is constant (V_{ref}) with respect to the substrate V_{dd}, while the voltage across R_5 relative to the substrate increases directly with V_{dd}. Experimentally, V_{R_5} varied by about 100 mV as V_{dd} was changed from 4 to 8 V.

The source follower $M8$-$M9$ produces a replicated dc bias voltage ($V_{GS_1} = V_{GS_2} = V_{GS_9}$) for the p-well of the input pair which is independent of the common source voltage. This simple technique provides a speed improvement for single-ended comparator applications since the large p-well-to-substrate capacitance is not charged or discharged as a function of the common-mode input signal component.

Fig. 6 shows a schematic of the second gain stage with V_2 as the bias voltage for the current source device $M16$. p-channel reset switches are used to differentially store the offset voltages onto the coupling capacitors, with the load voltages assumed equal to V_y during the reset time. Several important observations can be made regarding the ac-coupled differential stage.

[1] A potential disadvantage of this technique is the poor rejection of noise on the negative power supply. However, because of the differential topology used in this design, the V_{ss} noise appears as a common-mode signal to the second stage, and is substantially eliminated.

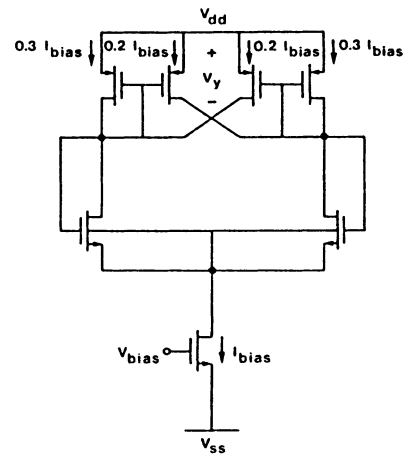


Fig. 7. A model of the second stage (a_{2a}) of the comparator during the reset interval with $V_2 = V_{bias}$.

Normally, in order to minimize the attenuation through the capacitive coupling network, n-channel cascode devices would be used in series with the input transistors $M10$-$M11$ to reduce the input capacitance due to the Miller-multiplied C_{gd}. Unfortunately, with small supply voltages, cascodes cannot be used. Next, assume that $V_2 = V_{bias}$, and that the p-channel reset switches are turned on and modeled as short circuits, as shown in Fig. 7. I_{bias} is nominally 150 μA with ±50 percent variations. Thus, over the worst case processing variations (Table I), the p-channel load bias voltage given by

$$V_y = |V_{tp}| + \left[\frac{2(0.3I_{bias})}{\mu_p C_{ox}(W/L)}\right]^{1/2} \tag{6}$$

ranges from 1.3 to 2.9 V, as shown in the graph of Fig. 8. In the minimum case, V_y (min) = 1.3 V, substrate injection can occur since the first stage positive output swing is ΔV_x (max) = V_{ref} = 2.048 V. In the maximum case, V_y (max) = 2.9 V, the p-channel reset switches may not be sufficiently turned on due to the small amount of available gate drive V_{GS} (min) = V_{dd} (min) - V_y (max). If necessary, full CMOS transmission gates can be used as reset switches to eliminate this problem at the expense of greater uncertainty in the clock feedthrough charge cancellation. The approach taken in this design overcomes

TABLE I
WORST CASE PARAMETER VARIATIONS FOR DEVICES WITH DRAWN CHANNEL LENGTHS OF 12 μm; $T = 25°C$

	N-Channel		P-Channel	
	min.	max.	min.	max.
V_t (Volts)	0.5	1.0	-0.5	-1.0
uCox (uA/volts^2)	14.4	24.0	5.9	9.7
T_{ox} (Angstroms)	800	1000	800	1000
Gamma (volts)$^{\frac{1}{2}}$	1.6	2.0	0.6	0.8
Lambda (V^{-1}) L=12μ	0.0031	0.0063	0.0104	0.0209

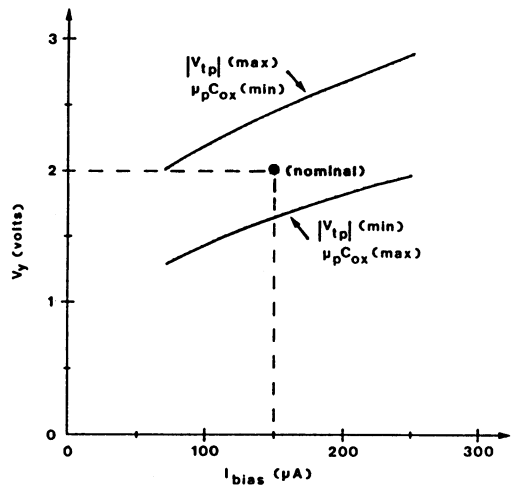


Fig. 8. Variations in the second-stage reset voltage as a function of processing and bias current variations for $V_2 = V_{bias}$.

both problems by combining the V_{ref} replica biasing techniques within an active feedback network to provide a solution whereby $V_y \approx V_{ref}$ independently of processing and operating variations. This so-called bias control amplifier shown in Fig. 9 consists of two major parts: 1) the replica bias string $M21$-$M23$ which nominally operates at the same current densities (same V_{GS}'s) as $M12$-$M10$-$M16$ in the second stage (Fig. 6), with $I_{rep} = 0.4 I_{bias}$ to conserve power; and 2) negative feedback around the differential amplifier forces $V_z \approx V_{ref}$ by varying the current I_{rep}, which results in a similar variation in I_{bias}, and hence, by replication, $V_y \approx V_z \approx V_{ref}$.

In simple terms, the controlled biasing of the gain stage(s) merely trades off the variation in the reset voltage V_y that existed previously (Fig. 8) for an increased variation in the bias current as shown in Fig. 10. With $V_2 = V_{bias}$, the bias current varied from 75 to 250 μA, while with the new controlled biasing, the current varies from 60 to 300 μA. This also results in a slight increase in the gain variance of the comparator (Table II).

Four nonideal effects contribute to a deviation of V_y from V_{ref}: 1) the common-mode reset/cancellation switch feed-

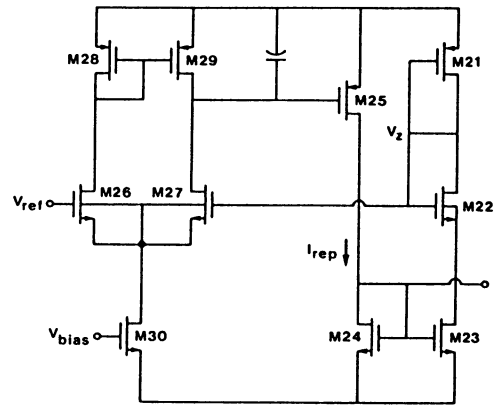


Fig. 9. Bias control amplifier. Negative feedback adjusts I_{rep} so that $V_z \cong V_{ref}$. By replication, the second-stage reset voltage V_y is approximately equal to V_{ref}.

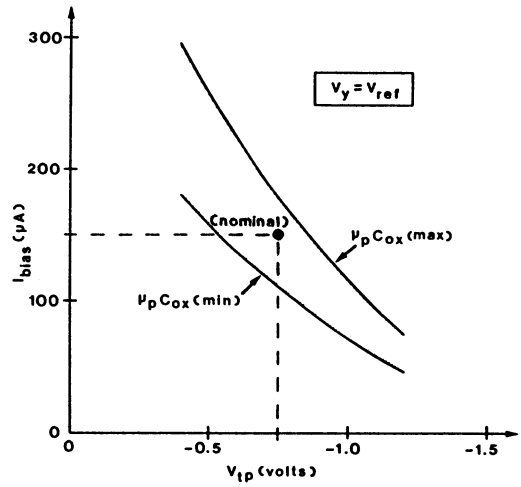


Fig. 10. Variations in the second-stage bias current as a function of processing variations. Controlled biasing is used so that the second-stage reset voltage $V_y \cong V_{ref}$.

TABLE II
CALCULATED DC GAIN VARIATIONS FOR EACH OF THE COMPARATOR STAGES

	a_{1a}	a_{1b}	a_{2a}	a_{2b}	a_3	a_{tot}
min	1.52	2.62	7.74	40.26	20.13	2.50×10^4
nom	1.96	3.68	9.98	87.13	43.57	2.73×10^5
max	2.52	5.81	12.83	231.04	115.52	5.01×10^6

through (in metal-gate CMOS technology, alignment-insensitive switch layouts should be used to provide the best possible matching between the V_{ft}'s [see Fig. 14 below stage 2]); 2) the input-referred offset voltage of the bias control amplifier (typically 20-50 mV); 3) the closed-loop gain error of the control amplifier (typically 0.05 V_{ref}); and 4) the matching of the replicated transistor strings (typically 1 percent mismatch). The worst case deviation in V_y from V_{ref} is thus about 200 mV. The amount of bias offset that can be tolerated is dependent on the attenuation through the capacitive coupling network $\Delta V_y / \Delta V_x = C_s/(C_s + C_{in})$ where, by design, $\Delta V_x \approx V_{ref}$. Assuming a gain of 0.9 through the capacitive divider and $V_{ref} = 2.048$ V, then ΔV_y is approximately 1.8 V, which partially compensates for the 200 mV maximum offset in the bias

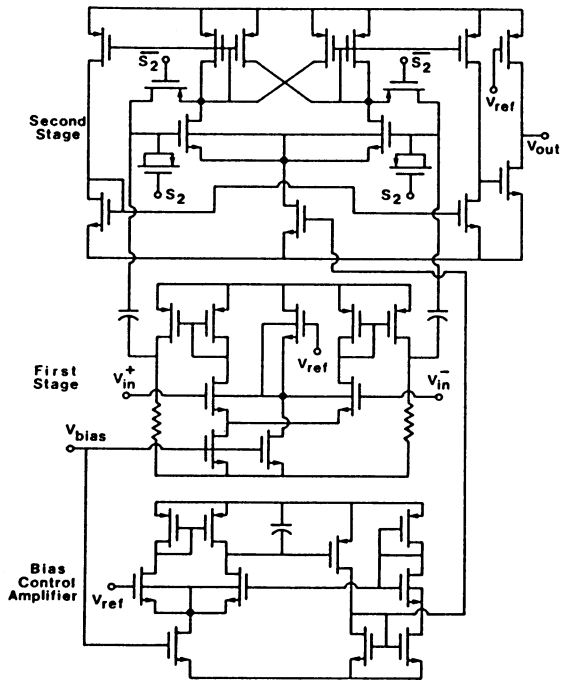


Fig. 11. A complete schematic of the comparator.

voltage V_y ensuring that no bootstrapping can occur. Fig. 11 shows a complete schematic of the comparator circuit.

IV. Gain/Bandwidth Considerations

The low frequency differential gain of the first comparator stage (Fig. 5) is approximately given by

$$a_1 = a_{1a}a_{1b} \approx \frac{g_{m1}}{g_{m3}} \cdot g_{m5}R_5 \tag{7}$$

which can be written as

$$a_1 \approx \sqrt{\frac{\mu_n}{\mu_p}}\sqrt{\frac{(W/L)_1}{(W/L)_3}} \cdot \frac{2I_5R_5}{(V_{GS} - V_{tp})_5} \tag{8}$$

where I_5 ($2I_{ref}$) is the dc bias current through $M5$. The first two terms in (8) represent the gain of an n-channel common-source amplifier with a p-channel diode load. It is similar in form to the gain of a single-channel enhancement inverter [10] with increased gain based on the square root of the n-channel to p-channel mobility ratio, but with more gain variance due to the independence of μ_n and μ_p. The third term represents the gain of the resistively loaded common-source amplifier where $I_5R_5 = V_{ref}$ is the dc bias voltage across the load resistor and $(V_{GS} - V_{tp})_5$ is the effective turn-on voltage of $M5$. (It is interesting to note that this gain term is similar to that of a resistively loaded common-emitter bipolar amplifier with kT/q replaced by $(V_{GS} - V_{tp})_5/2$.) The bias current levels must be chosen to provide an adequate large signal slew rate (note that there is no slewing in resistively loaded stages) and small-signal bandwidth. These are standard calculations which will not be considered further with respect to the first comparator stage.

Having completed the design of the first stage, the next issue is to design the ac coupling network. The size of the coupling capacitor(s) is usually selected based on feedthrough matching

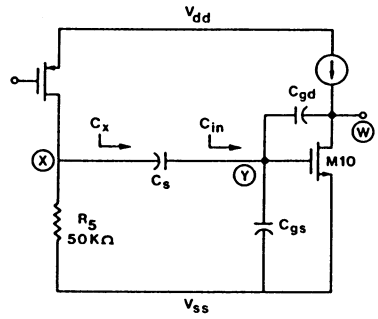


Fig. 12. A differential half circuit of the first-stage output and the series capacitor coupling into the input of the second gain stage.

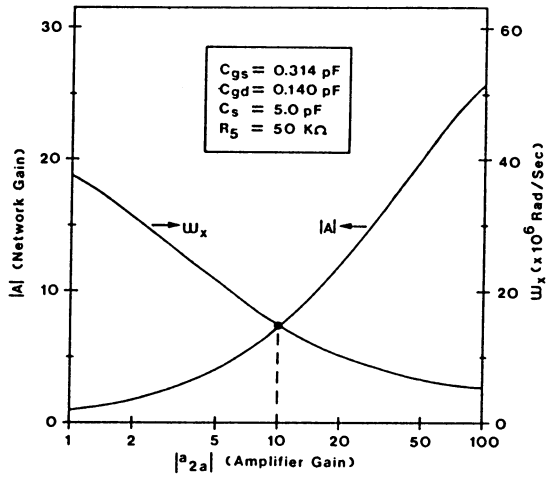


Fig. 13. The effective gain (A) of the second stage including the capacitive divider and the small-signal bandwidth (ω_x) of the first-stage output, both as a function of the gain (a_{2a}) of the second-stage input amplifier.

considerations. C_s = 5 pF was used in this design. At this point, the designer is faced with a conflicting set of requirements. Fig. 12 shows a differential half circuit of the coupling network between the first-stage output and the second-stage input. On the one hand, it is desirable to make C_x as small as possible to maximize the small signal bandwidth at node Ⓧ which is given by

$$\omega_x = 1/(R_5C_x) = (1 + C_s/C_{in})/R_5C_s \tag{9}$$

where $C_{in} = C_{stray} + C_{gs_{10}} + C_{gb_{10}} + (1 - a_{2a})\,C_{gd_{10}}$. Assuming that $C_s \gg C_{in}$, this implies that a_{2a} should be small to minimize the Miller-multiplied $C_{gd_{10}}$ term. On the other hand, the gain from node Ⓧ to node Ⓦ is given by

$$A = a_{2a} \cdot C_s/(C_s + C_{in}) \tag{10}$$

which implies that a_{2a} should be large. Fortunately, a near-optimum solution to this problem can be obtained graphically as shown in Fig. 13 where (9) and (10) are plotted versus the low frequency gain of the stage being ac coupled into a_{2a}. For low gain values, the bandwidth approaches $(R_5C_{in})^{-1}$, and for high gain values, it approaches $(R_5C_s)^{-1}$. For high a_{2a} values, A approaches $-C_s/C_{gd_{10}}$ since the input becomes a virtual ground with $C_{gd_{10}}$ as the feedback capacitor. The crossover point where a_{2a} = 10 was chosen for this design.

One approach for obtaining the gain of 10 is to again use an

n-channel differential pair with p-channel diode loads as used in the first stage. However, based on (8), it is difficult to achieve this relatively large gain with a reasonable aspect ratio. Therefore, it was decided to use a controlled amount of positive feedback to effectively increase the driver device transconductance [11]. The gain of the positive feedback cell of Fig. 6 is given by

$$a_{2a} = \frac{g_{m10}}{g_{m12}}(1 - \alpha)^{-1} \tag{11}$$

where $\alpha = (W/L)_{14}/(W/L)_{12}$. In this design, $\alpha = \frac{2}{3}$, giving a factor of three gain increase. Care must be exercised in selecting the value of α. If $\alpha = 1$, the stage becomes a positive feedback latch. If $\alpha < 1$, linear amplification with increased gain is obtained as in this design. For $\alpha > 1$, the stage becomes a Schmitt trigger circuit, with the amount of hysteresis dependent on the value of α. For linear amplification, $\alpha_{nom} = 0.9$ is a practical maximum because with mismatches and processing variations, the effective α can approach or even exceed one which leads to undesirable comparator hysteresis. The remaining comparator stages shown in Fig. 11 are conventional and will not be considered further.

V. Experimental Results

Fig. 14 shows a die microphotograph of the comparator circuit which was fabricated using a p-well metal-gate CMOS process with 8 μm minimum feature sizes. The die size is 560 μm (22.0 mils) by 950 μm (37.4 mils).

The most basic, and often the first, test for a comparator circuit is the qualitative "good or bad" test based on the performance of the complete A/D converter system. For example, Fig. 15(a) shows the measured dc transfer characteristic of an A/D converter which was judged to be bad due to the rather gross nonlinearity as indicated. This problem was traced to the comparator circuit which was different from the design described in this paper. More specifically, internal comparator nodes were being bootstrapped above V_{dd} as described earlier. This resulted in the gain change in the transfer characteristic (nonlinearity) due to the flow of parasitic substrate current which removed charge from the interstage coupling capacitors for large internal positive swings. Fig. 15(b) shows the measured transfer characteristic for the same A/D converter system using the comparator design described herein. The use of the controlled supply-independent biasing techniques has obviously eliminated the bootstrapping effect since the performance is now good. Furthermore, while the old design exhibited increased nonlinearity for larger supply voltages (larger positive voltage swings), the present transfer characteristic maintained linearity over the entire range of supply voltages.

Due to probe and instrument parasitics, direct comparator measurements are difficult to obtain without interfering with the performance of the circuit. Fortunately, many of the important parameters can be inferred quantitatively from the A/D transfer characteristic. For example, by increasing the conversion frequency, the available response time for the comparator is decreased; eventually, there will be insufficient time for the comparator to respond to an LSB input. As a result, the A/D digital output code will remain at all zeros for small input levels. The conversion frequency at the onset of the nonlinearity in the transfer characteristic indicates the switching speed. Using this method, the comparator's switching time was determined to be about 3 μs for a 4 mV LSB and about 12 μs for a 1 mV LSB.

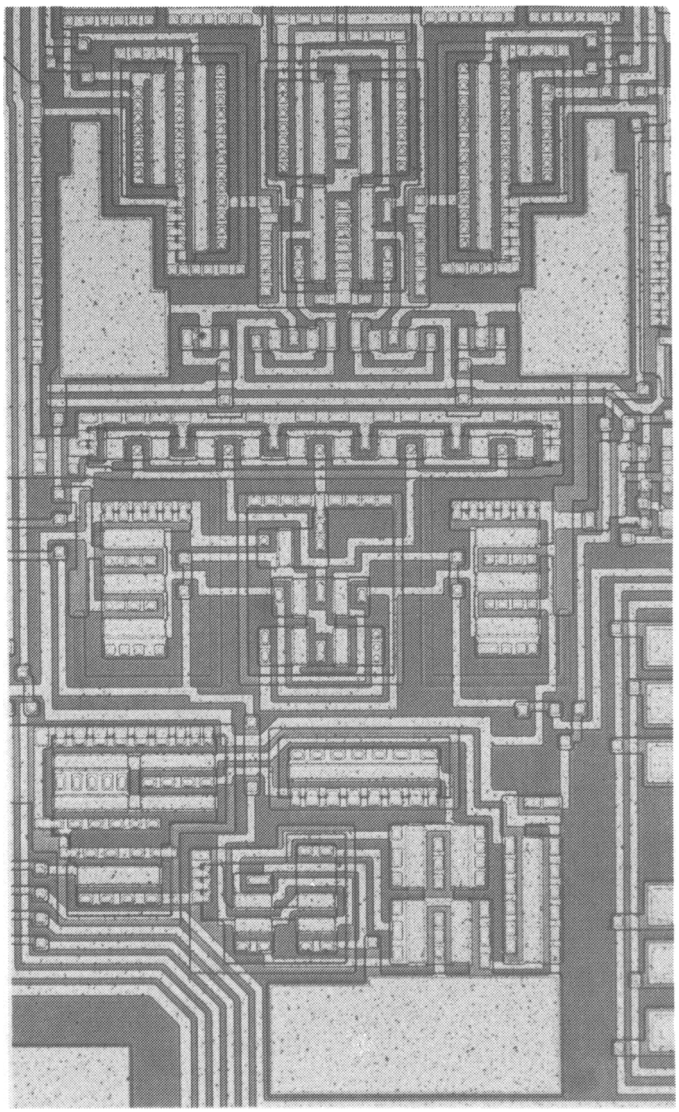


Fig. 14. Die microphotograph of the metal-gate CMOS comparator circuit. The layout orientation is identical to the schematic orientation of Fig. 11.

Although not performed here, this technique can also be used to infer the equivalent input noise of the comparator. As the LSB input level is decreased, the probability of error in the A/D digital code is increased. By taking a large number of transfer characteristics, the error probability versus input level can be determined, and from these data, assuming a Gaussian distribution, the equivalent rms input noise can be determined. The calculated rms input noise for this comparator is nominally about 100 μV.

VI. Conclusions

Several new design techniques for variable-supply MOS analog circuits have been described in terms of a precision CMOS comparator design. In A/D conversion systems that contain an on-chip voltage reference, techniques have been developed to replicate the reference voltage to produce stable supply-independent dc bias voltages for other analog circuitry including

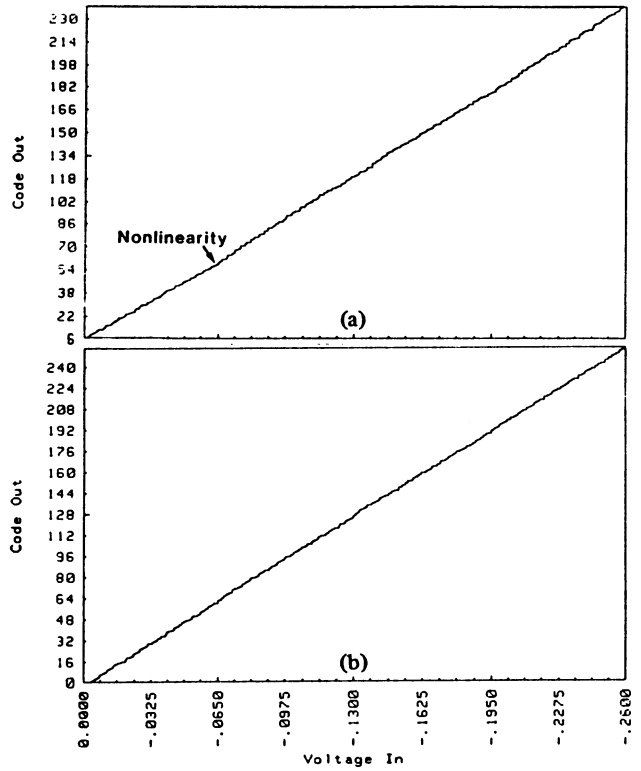


Fig. 15. Measured A/D dc transfer characteristics. (a) Using a previous comparator design, the A/D characteristic exhibited nonlinearity due to internal nodes of the comparator being bootstrapped above the supply V_{dd}. (b) Using the comparator described herein, the same A/D converter system maintained linearity over the entire range of supply voltages. V_{ref} was untrimmed in both cases.

the comparator. Controlled biasing techniques have also been developed which allow differential resetting in the comparator, which is beneficial in minimizing the input-referred offset voltage due to clock feedthrough and channel charge-pumping effects. These controlled biasing techniques eliminate the possibility of internal comparator nodes being bootstrapped above the power supply, which results in A/D nonlinearity if it occurs. A generalized technique for designing ac-coupled gain stages was also presented.

Acknowledgment

The author gratefully acknowledges the essential contributions and suggestions of Dr. W. C. Black, Jr., and J. R. Hellums, J. R. Ireland, and M. B. Terry. J. Baechler and V. Allstot are also acknowledged for helping prepare the final manuscript.

References

[1] P. R. Gray, D. A. Hodges, and R. W. Brodersen, *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980.

[2] J. L. McCreary and P. R. Gray, "All-MOS charge redistribution analog-to-digital conversion techniques–Part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, Dec. 1975.

[3] R. E. Suarez, P. R. Gray, and D. A. Hodges, "All-MOS charge redistribution analog-to-digital conversion techniques–Part II," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 379–385, Dec. 1975.

[4] F. Guteri, "Cameras that 'think,'" *IEEE Spectrum*, pp. 32–37, June 1982.

[5] J. S. Brugler and P. G. A. Jespers, "Charge pumping in MOS devices," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 297–302, Mar. 1969.

[6] A. R. Hamade, "A single chip all-MOS A/D converter," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 785–791, Dec. 1978.

[7] M. E. Hoff, Jr., J. Huggins, and B. M. Warren, "An NMOS telephone codec for transmission and switching applications," *IEEE J. Solid-State Circuits*, vol. SC-14, Feb. 1979.

[8] R. A. Blauschild, P. A. Tucci, R. S. Muller, and R. G. Meyer, "A new NMOS temperature-stable voltage reference," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 767–774, Dec. 1978.

[9] M. B. Terry, private communication.

[10] Y. P. Tsividis, "Design considerations in single-channel MOS analog integrated circuits–A tutorial," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 383–391, June 1978.

[11] K. B. Ohri and M. J. Callahan, Jr., "Integrated PCM codec," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 38–46, Feb. 1979.

A 750MS/s NMOS Latched Comparator

David C. Soo, Alexander M. Voshchenkov, Gen M. Chin,
Vance D. Archer, Maureen Lau, Mark Morris
AT&T Bell Laboratories
Holmdel, NJ

*Bruce A. Wooley**
Stanford University
Stanford, CA

Ping K. Ko, Robert G. Meyer***
University of California
Berkeley, CA

THIS PAPER WILL DESCRIBE THE DESIGN of analog circuits integrated in a polysilicon gate NMOS technology with 1μ effective channel length devices[1]. In particular, a latched comparator with 4b input resolution at 750MS/s and a wideband amplifier with 10dB of voltage gain over a bandwidth of 1.25GHz, when driving 130fF of on-chip capacitance, will be reported. The comparator was designed primarily for application to high-speed flash A/D conversion, but is is also suitable for use in integrated broadband fiber optic receivers. Its circuit configuration differs from previous designs[2,3] in that negative feedback is used in the preamplifier section to trade gain for bandwidth.

Preamplification in the comparator has been accomplished by differential implementation of the wideband amplifier topology shown in single-ended form in Figure 1. This amplifier design is suitable for a wide variety of both small and large-signal on-chip applications within larger NMOS circuits. For driving off-chip loads, a simple broadband output buffer, also shown in Figure 1, was added.

In Figure 1, transistor M_1 forms a transconductance input stage, while M_2 and the feedback transistor M_3 form a transresistance stage. M_4 and M_5 are source followers used to buffer the transconductance stage output, and M_6-M_{10} are depletion mode current sources that provide dc biasing for the amplifier. The compensation capacitor, C_c is used to optimize the complex pole response of the amplifier to suit the particular application.

Active, rather than resistive, feedback is used in the circuit in Figure 1, because it provides higher loop gain, and the amplifier gain is, therefore, less sensitive to variations in device transconductance and output resistance. If the effects of M_4 and M_5 are neglected, the dc gain and the per-stage gain-bandwidth product of the amplifier are approximated by

$$G = \frac{g_{m1}}{g_{m2}}\left[1 + \frac{1}{g_{m2}r_{o2}}\left(1 + \frac{1}{g_{m2}r_{o1}}\right)\right] - 1 \qquad (1)$$

$$\sqrt{G} \times BW = \left[\frac{g_{m1}g_{m2}}{C_1C_2}\left(\frac{1}{1 + \frac{C_c}{C_1} + \frac{C_c}{C_2}}\right)\right]^{1/2} \qquad (2)$$

As is apparent from (*1*), the dc gain is proportional to the ratio of two well-matched transconductances and is therefore relatively insensitive to variations in temperature and processing. As seen from (*2*), the per-stage gain bandwidth product of the amplifier approaches the transistor ω_T ($=g_m/C_s$) as C_c approaches zero.

To test the small-signal response of the preamplifier, the single-ended version was fabricated together with a broadband output buffer capable of driving a 50Ω load. The effects of the output buffer were subtracted from the test system measurements to determine the preamplifier response. The buffer has an input capacitance of 130fF and a voltage gain of 0.75 when driving a 50Ω load. The frequency response measured for the preamplifier is shown in Figure 2.

The amplifier configuration described has been used for preamplification in a 750MS/s NMOS latched comparator. This comparator differs from conventional designs in that wideband controlled gain, rather than open-loop high-gain, preamplification, is combined with a sensitive latch in a 2-phase clock architecture. By this means the large signal digital outputs are obtained via regeneration in the latch rather than gain in the preamplifier, and an increase in comparator speed is thus achieved.

Figure 3 is the schematic of the latched comparator. It is essentially a differential wideband amplifier (M_1-M_{23}) driving a positive feedback latch (M_{22}-M_{30}). M_1 and M_2, together with the input capacitance of M_3 and M_4, form an analog sample-and-hold circuit with a -3dB bandwidth above 1GHz. Common mode charge injection and clock coupling from M_1 and M_2 are rejected by the differential pair M_3 and M_4. ϕ_1 and ϕ_2 are complementary overlapping clocks. At the falling edge of ϕ_1, the input signal is sampled onto the amplifier input capacitance and held for the entire time clock ϕ_2 is high. While ϕ_2 is high, M_{21} enables the negative feedback loop and the wideband amplifier establishes a differential voltage at its output which represents the initial latch voltage, V_i. When ϕ_1 goes high again (and ϕ_2 falls) M_{30} enables the latch. Regenerative feedback then drives the latch output to one of the two stable states as governed by the initial imbalance V_i. The large signal voltage developed across the outputs during ϕ_1 will be referred to as the final latch voltage, V_f.

The length of time ϕ_1 must be kept high (denoted as τ_1) is determined by the regeneration in the latch. Neglecting effects

---

*Bruce A. Wooley and Ping K. Ko were formerly with AT&T Bell Laboratories.

**Research supported in part by the U.S. Army Research Office under Grant DAAG29-84-K-0043.

[1] Ko, P.K., et. al., "SIGMOS-A Silicon Gigabit/sec NMOS Technology", *IEDM Technical Digest*, p. 751-753; 1983.

[2] Suzuki, M., et. al., "A Bipolar Monolithic Multigigabit/s Decision Circuit", *IEEE J. Solid State Circuit*, Vol. SC-19, p. 462-467; Aug., 1984.

[3] Meignant, D., et. al., "A High Performance 1.8GHz Strobed Comparator for A/D Converter", *IEEE GaAs IC Symposium Technical Digest*, p. 66-69; 1983.

Reprinted from *1985 IEEE Int. Solid-State Circuits Conf.*, pp. 146–147, 327, Feb. 1985.

of source followers M_{28} and M_{29}, and assuming linear operation, τ_1 is approximated by

$$\tau_1 \approx \tau_T \left(\frac{A_l}{A_l - 1}\right) \ln \frac{V_f}{V_i} \qquad (3)$$

where $\tau_T = C_l/g_{m21,22}$, $A_l{}^2$ is the positive feedback loop gain and C_l is the capacitive loading at the output nodes. Depending on the technology used, τ_T can be made to approach the transistor transit time, C_g/g_m. Regeneration thus offers the fastest possible generation of a large signal from a small one, assuming $A_l \gg 1$.

The length of time ϕ_2 is high, τ_2, is governed by the acquisi- for the output voltage to change from V_f to the minimum initial latch voltage, $V_i(\min)$, when the input of the amplifier is driven only by a voltage, V_{lsb}, equal to the minimum resolvable voltage expected of the comparator; that is, $V_i(\min) = A_p V_{lsb}$, where A_p is the preamplifier gain. Qualitatively, τ_2 increases with increases in both A_p and, to a lesser extent, the size of the compensation capacitor, C_c.

Circuit simulation was used to determine the optimum choices for A_p and C_c. Since τ_2 decreases and τ_1 increases with decreasing A_p, A_p is chosen so that $\tau_1 = \tau_2$. This choice also simplifies the design requirements for generating two clock phases. After A_p is chosen, C_c is then adjusted so that overshoot at the output is always less than the minimum initial latch voltage, $V_i(\min)$. In the present design A_p is about 3.

The comparator was integrated together with clock drivers and output buffers that facilitate monitoring the outputs with 50Ω test equipment.

Measured dc offset voltage is on the order of 25mV and the comparator input common mode range is +1V to −2V. For a full-scale input voltage of 2.5V, the comparator has an equivalent input resolution of 4b at a sampling rate of 750MS/s. Figure 4 shows the output waveform at 750MS/s when a periodic pattern (101100111000...) is applied to the input. Figure 5 is the eye diagram obtained when a 1/2b density pseudorandom pattern is applied to the input. In Figure 6, the measured input resolution of the comparator for an error rate of 10^{-9} is plotted as a function of the sampling rate.

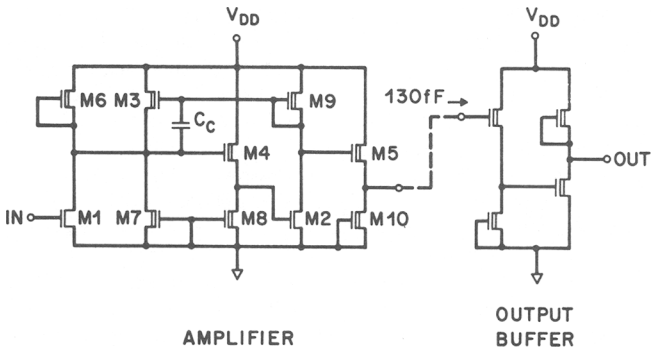


FIGURE 1—Wideband feedback amplifier.

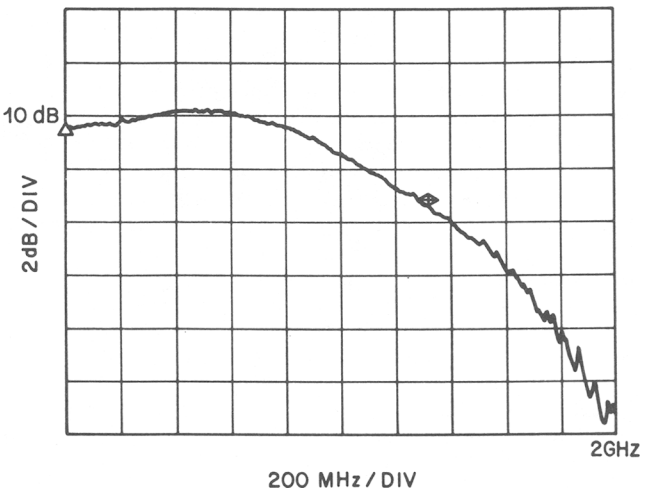


FIGURE 2—Frequency response of wideband amplifier.

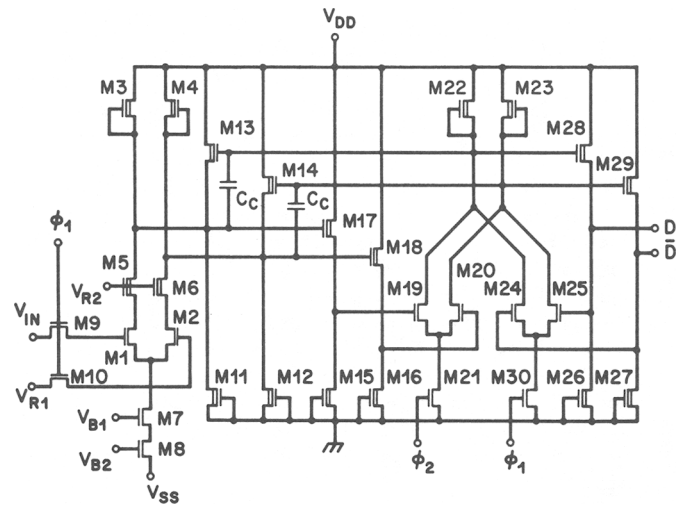


FIGURE 3—Latched comparator.

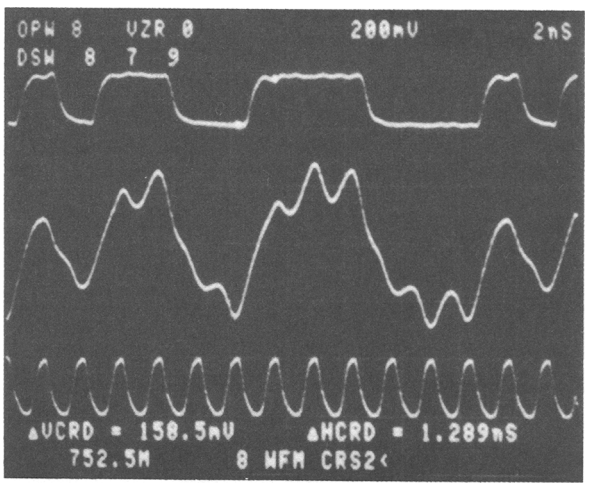


FIGURE 4—Output waveform of comparator.

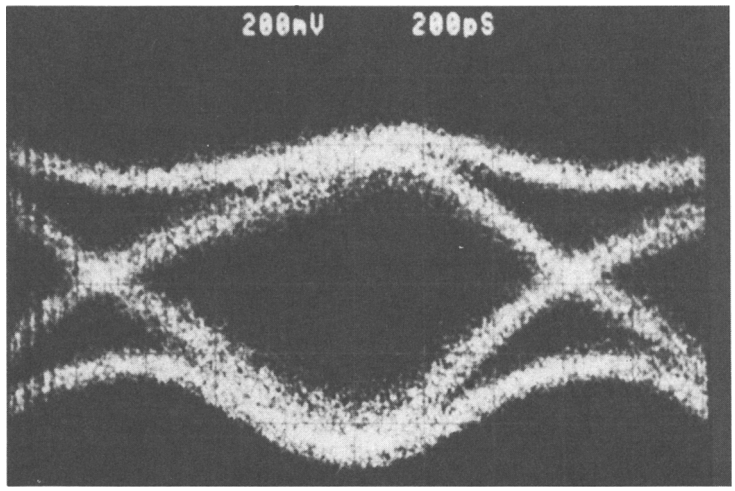


FIGURE 5—Eye diagram of comparator output (return to zero).

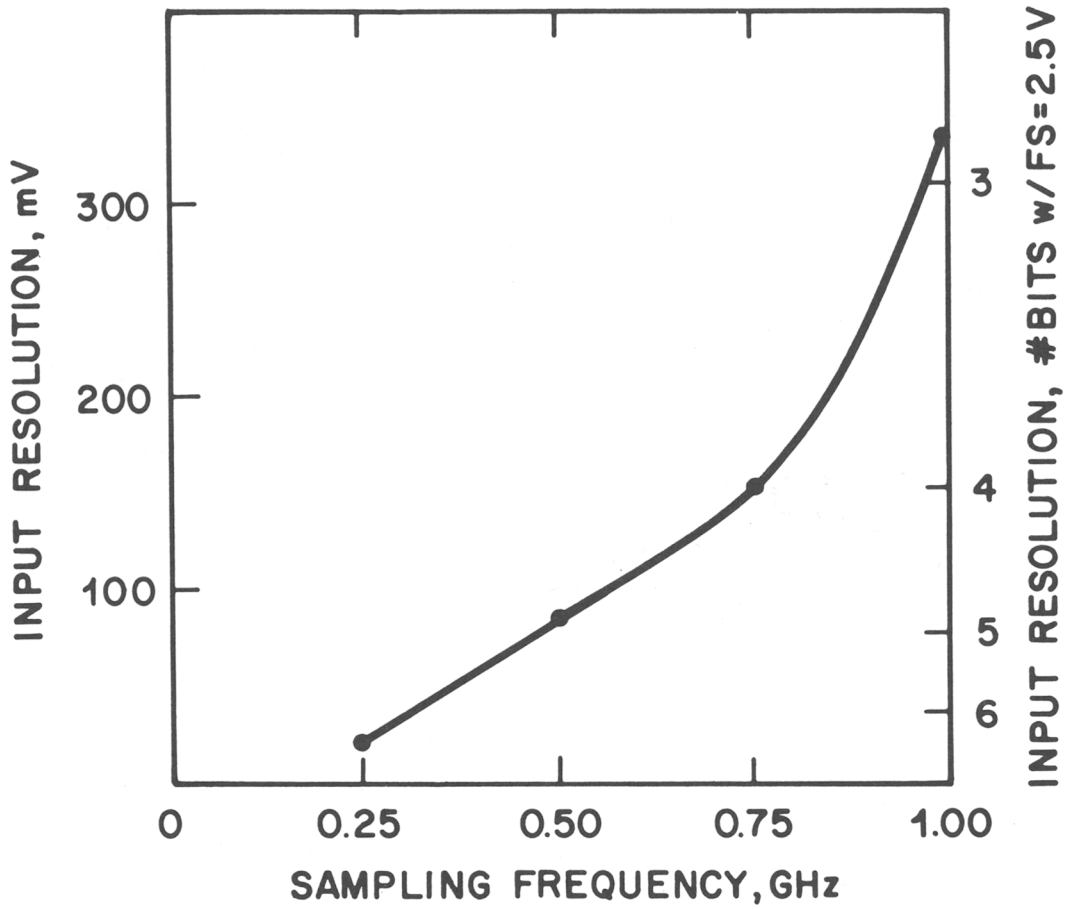


FIGURE 6—Input resolution of comparator.

Impact of Scaling on MOS Analog Performance

STEPHEN WONG AND C. ANDRE T. SALAMA, MEMBER, IEEE

Abstract—A first-order analysis of the impact of scaling on MOS analog performance under moderate scaling conditions is presented in this paper. Assuming a polysilicon gate ion-implanted MOS technology, quasi-constant voltage (QCV) scaling is shown to be the optimal scaling law, offering the best overall analog performance and resulting in an increase in functional density, gain-bandwidth product with a moderate degradation in gain, and signal-to-noise ratio. The first-order analysis agrees fairly well with computer simulation. A typical case study shows that under moderate scaling conditions, CMOS can generally offer a higher voltage gain when compared to depletion load NMOS and is the preferred technology for scaled analog implementations.

I. Introduction

With the ever increasing complexity of very large scale integrated circuits, it has become highly desirable to integrate analog and digital circuits on a single silicon chip. The primary VLSI focus has been on those technologies which permit high density integration of analog and digital circuits on a single chip. The most useful MOS technologies in this context are NMOS and CMOS [1].

The increasing complexity of digital MOS VLSI was spurred by the concept of device scaling [2]. The density of digital circuits has increased continuously and the area consumed by the analog portion of a digital-analog circuit has become relatively larger. This fact has provided a strong impetus for scaling the analog portion of the circuit as well as the digital one.

The scaling of analog circuits can benefit from the considerable amount of information available on the scaling of digital MOS IC's. Various scaling laws have been proposed and are summarized in Table I [2], [3]. The first of these is constant field (CE) scaling [2] which was introduced as a means of alleviating the difficulties arising from two-dimensional parasitic effects associated with short channels in MOS transistors. Basically, CE scaling involves a reduction of voltages, currents, and lateral and vertical dimensions by a factor λ and an increase in the substrate doping concentration by the same factor λ ($\lambda \geq 1$). The linear reduction of voltage associated with CE scaling generally results in a sizeable deterioration of the signal-to-noise ratio as well as a lack of TTL interface compatibility. To avoid this difficulty, the current trend is towards nonconstant field scaling in the form of constant voltage (CV) or quasi-constant voltage (QCV) scaling [3]. Constant voltage scaling implies a fixed nonscaled power supply (compatible with TTL) and a slower than λ (approximately $\sqrt{\lambda}$) scaling of gate-oxide thickness to reduce oxide field and improve reliability and yield. Quasi-constant voltage scaling implies a slower than λ (taken as $\sqrt{\lambda}$ for convenience sake) decrease in voltage while the other factors scale as in the CE case.

TABLE I
Scaling Laws

	CE	QCV	CV
Voltages	λ^{-1}	$\lambda^{-1/2}$	1
Lateral dimensions	λ^{-1}	λ^{-1}	λ^{-1}
Vertical dimensions	λ^{-1}	λ^{-1}	$\lambda^{-1/2}$
Doping concentration	λ	λ	λ

QCV scaling was found to provide optimum drive current capability in digital VLSI MOS circuits [3]. Although this factor is an important criterion in digital applications, parameters such as gain, bandwidth, and signal-to-noise ratio are more suitable criteria to evaluate the performance of analog circuits.

One of the objectives of this paper is to investigate the impact of the scaling laws discussed above on MOS analog component performance. Only moderate scaling factors are considered and no attempt is made to approach the ultimate scaling limits being considered for digital circuits [4]. Present day MOS analog circuits have minimum channel lengths of about 8 μm. The scaling factors considered here will be limited to the range $\lambda = 1$ to $\lambda = 4$. A scaling of $\lambda = 4$ would reduce the minimum channel length to 2 μm and the circuit area by approximately a factor of 16 while still keeping second-order effects within bounds [5], [6]. Any further scaling would lead (as will be seen from the following discussion) to unacceptable degradation in analog performance. Another objective of this paper is to establish the optimum technology for scaled analog MOS circuits and to evaluate the performance of a scaled MOS analog op amp as a typical example of analog circuit implementation.

Many intricate technology related problems are expected during device scaling. Presently, innovative processing techniques are being applied to reduce such problems as device mismatch, junction depth control, and reliability of scaled down oxide layers. It is neither the aim nor the scope of this paper to suggest solutions to these related problems, but merely to investigate the effects on analog performance when scaling is achieved.

II. MOSFET Modeling

MOS analog integrated circuits[1] consist for the vast majority of MOSFET's used as drivers, active loads for analog switches,

Manuscript received December 4, 1981; revised April 23, 1982. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

The authors are with the Department of Electrical Engineering, University of Toronto, Toronto, Ont., Canada M5S 1A4.

[1]Ion-implanted silicon gate technology is assumed throughout the discussion.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-18, no. 1, pp. 106–114, Feb. 1983.

as well as MOS capacitors, resistors (diffused or polysilicon), and interconnections.

Since the MOSFET's are the most critical components in any analog implementation, the discussion will focus mainly on those MOSFET parameters which have a direct influence on analog performance.

A. I-V Characteristics

In order to account for second-order effects which may arise due to scaling, a set of analytical equations must be used. Modeling of small geometry (short and narrow channel) devices have been treated extensively by many authors [7]-[10]. These models generally include the following phenomena: 1) threshold dependence on two-dimensional charge sharing effects [11], [12], 2) mobility-degradation due to increased normal and lateral fields [13], and 3) velocity saturation of carriers causing premature saturation of current and lower output conductance [14], [15].

When moderate scaling effects are considered, two additional factors must be considered. These are 1) mobility degradation due to impurity scattering when the effective substrate doping exceeds 5×10^{16} cm^{-3} [16], and 2) gain reduction via feedback through the parasitic drain and source resistances.

The saturation region of the MOSFET plays an important role in determining analog performance parameters such as transconductance, output resistance, and voltage gain. The set of equations selected for the analysis are based on a semi-empirical model of the MOSFET operation in this particular region of operation.

The threshold voltage V_T of the device can be expressed as

$$V_T = V_{TO} - \Delta V_{TS} + \Delta V_{TN} \tag{1}$$

where V_{TO} is the threshold voltage for a large geometry device and ΔV_{TS} and ΔV_{TN} are the corrections for short and narrow channel effects, respectively [7]. V_{TO} is given by

$$V_{TO} = \phi_{MS} + 2\phi_F - \frac{Q_{SS} + Q_B}{C_0} \pm \frac{qN_iD}{C_0} \tag{2}$$

where ϕ_{MS} is the work function difference between gate and bulk material, ϕ_F is the Fermi potential given by

$$\phi_F = \frac{kT}{q} \ln \frac{N_B}{n_i} \tag{3}$$

where N_B is the bulk doping concentration under the channel and Q_{SS}, Q_B are the oxide and bulk charges, respectively. The last term in (2) represents the effect of the ion implant used to adjust the threshold voltage. N_iD is the effective implant dose and D is the implant depth which is assumed to be very shallow.

Mobile carriers in the channel of a scaled MOST normally experience scattering effects due to the electric fields and increased impurity levels. This leads to a significant degradation of the channel mobility μ which is usually expressed as

$$\mu = \frac{\mu_0}{\delta_1 \delta_2} \tag{4}$$

where μ_0 is the zero field mobility and δ_1, δ_2 are second-order parameters describing the degradating effects of high field and high doping concentration on mobility, respectively. Above saturation, the parameter δ_1 is defined as

$$\delta_1 = 1 + \frac{\theta}{t_0}(V_G - V_T) + \frac{V_{D,\text{sat}}}{E_c L} \tag{5}$$

where θ is a constant (typically 3×10^{-7} cm/V), E_c is the critical field for velocity saturation (2×10^4 V/cm for NMOS and 2×10^5 V/cm for PMOS), $V_{D,\text{sat}}$ is the saturation voltage and is taken as $(V_G - V_T)$ for a first-order analysis. The parameter δ_2 accounts for impurity scattering effects and is defined as

$$\delta_2 = \left(1 + \frac{N_B}{5 \times 10^{16}}\right)^{1/2}. \tag{6}$$

The gain constant β of the transistor is defined as

$$\beta = \frac{\mu C_0}{\delta_3} \frac{Z}{L} \tag{7}$$

where Z is the channel width and δ_3 accounts for the feedback degradation due to the source resistance R_s

$$\delta_3 = 1 + \beta R_s (V_G - V_T). \tag{8}$$

R_s is proportional to the length but inversely proportional to the width, depth, and doping of the source junction. The effects of the inversion layer capacitance and contact resistance on β are neglected here. These effects are only relevant for channel lengths below 1.5 μm [6].

The saturation current of the device can be expressed as

$$I_{D,\text{sat}} = \frac{\mu C_0 Z}{2L} \frac{(V_G - V_T)^2}{\delta_3 \delta_4} \tag{9}$$

where δ_4 is a second-order parameter which takes into account the body effect on the channel. δ_4 is defined as

$$\delta_4 = 1 + \frac{\gamma}{2(2\phi_F - V_B)^{1/2}} \tag{10}$$

where

$$\gamma = \frac{(2\epsilon_s q N_B)^{1/2}}{C_0} \tag{11}$$

where ϵ_s is the dielectric constant of silicon.

The transconductance in the saturation region can be expressed as

$$g_m = \frac{\mu C_0 Z}{L} \frac{(V_G - V_T)}{\delta_3 \delta_4}. \tag{12}$$

B. Output Conductance

The output conductance of the device in the saturation region is critical in determining the voltage gain of inverters as well as the output impedance of current source. Three important phenomena affect the channel conductance of the saturated enhancement mode MOSFET. These are 1) classical pinchoff [17], 2) velocity saturation [14], [15], and 3) feedback caused by drain induced barrier lowering [18], [19].

For moderate channel lengths, the first two phenomena dominate. The current in the device beyond saturation is

increased by channel length modulation and is given by

$$I_D = I_{D,\text{sat}} \frac{L}{L - \Delta L}. \tag{13}$$

For long channel lengths, ΔL is the reverse bias depletion width formed between the drain and the channel and is defined as

$$\Delta L = K_1 [V_D - V_{D,\text{sat}}]^{1/2} \tag{14}$$

where

$$K_1 = [2\epsilon_s/qN_B]^{1/2}. \tag{15}$$

As the internal field in a scaled device increases as a result of nonconstant field scaling, velocity saturation of mobile carriers, and the deterioration of the gradual channel approximation make it necessary to modify the drain depletion length ΔL as follows [14], [15]:

$$\Delta L = \left[\left(\frac{E_p K_1^2}{2}\right)^2 + K_1^2(V_D - V_{D,\text{sat}})\right]^{1/2} - \frac{E_p K_1^2}{2} \tag{16}$$

where E_p is the lateral field at the drain during channel pinch-off. Out of convenience, E_p is usually equated to the critical field for velocity saturation E_c [20]. A more realistic value of E_p has been derived by Rossel [15] to ensure the continuity of current and conductance in the triode and saturation regions. His expression for E_p can be approximated by

$$E_p \simeq \left[\frac{qN_B(V_G - V_T)^2}{2\epsilon_s L \delta_3 \delta_4}\right]^{1/3}. \tag{17}$$

For moderate E_p, (16) can be rewritten as (see Appendix I)

$$\Delta L = K_1 (V_D - V_{D,\text{sat}})^{1/2} \delta_5 \tag{18}$$

where δ_5 is a correction factor for channel shortening given by

$$\delta_5 = 1 - \frac{E_p K_1}{2(V_D - V_{D,\text{sat}})^{1/2}}. \tag{19}$$

Differentiating (13), using the value of ΔL given by (18), yields the output conductance

$$g_{ds} = \frac{dI_D}{dV_D} = \frac{I_{D,\text{sat}} K_1}{2L(V_D - V_{D,\text{sat}})^{1/2} \delta_6} \tag{20}$$

where δ_6 is defined as

$$\delta_6 = \left(1 - \frac{\Delta L}{L}\right)^2. \tag{21}$$

From (20), g_{ds} is proportional to $I_{D,\text{sat}}$, as observed experimentally in moderately scaled devices.

For very short channel lengths, extensive drain induced barrier lowering (DIBL) occurs resulting in a large dependence of threshold voltage on V_D as observed by Masuda [21]. As a direct consequence of this effect, g_{ds} will be directly proportional to $(V_G - V_T)$ and inversely proportional to L^3 (as shown in Appendix II). This strong dependence of g_{ds} on L cannot be tolerated in analog applications and in general the DIBL regime must be avoided.

C. Subthreshold Current

The exponential dependence of the subthreshold current plays an important role in analog applications. For $V_G < V_T$ and $V_D >> kT/q$, the weak inversion current can be expressed as [2]

$$I_{D,\text{sat}} \simeq \frac{\mu C_0 Z}{L}\left(\frac{kT}{q}\right)^2 \exp\left[\frac{q(V_G - V_{\text{on}})}{nkT}\right] \tag{22}$$

where V_{on} is the turn on voltage at a surface potential $\phi_s = 1.5\phi_F$, C_d is the surface depletion region capacitance defined as

$$C_d = \left[\frac{\epsilon_s q N_B}{2\left(\phi_s + \frac{kT}{q}\right) - V_B}\right]^{1/2} \tag{23}$$

and n is given by

$$n = 1 + \frac{C_d}{C_0}. \tag{24}$$

Due to the exponential nature of the subthreshold current, it does not scale linearly with λ. A commonly used figure of merit to describe subthreshold behavior is the slope S_s of the $\log I_D$ versus V_G curve. S_s can be expressed as

$$S_s = \frac{d \log I_{D,\text{sat}}}{dV_G} = \frac{q}{2.3\, nkT}. \tag{25}$$

In an analog switch, for instance, S_s is a significant factor in determining the on-off current ratio of the device.

D. Noise

The noise in MOS devices, working at low frequencies, is high due to the dominant contribution of $1/f$ noise [22]. The rms equivalent gate noise voltage, in this case, can be expressed as

$$V_{ng} = \left(\frac{a_n}{ZL}\frac{q^2}{C_0^2}\frac{\Delta f}{f}\right)^{1/2} \tag{26}$$

where a_n is a constant dependent on the interface trap density at the Si-SiO_2 interface, Δf is the bandwidth and f is the frequency.

III. Effect of Scaling on MOS Parameters

In this section, the effect of scaling on the MOS device parameters previously discussed in first investigated. The effect of scaling on MOS circuit components is then discussed. The fact that scaling will limit the accuracy and the ability to match components is not considered in detail here since it is technology dependent.

A. Subthreshold Current

In order to avoid large subthreshold currents (and the onset of substantial DIBL), the long channel index M suggested by

[2] Assuming the surface state density at the Si-SiO_2 interface to be negligible.

Brews *et al.* [23] can be used. This index is defined as

$$M = \frac{A[(x_j t_0)(W_S + W_D)^2]^{1/3}}{L} \tag{27}$$

where A is a constant and W_S, W_D are the widths of source and drain depletion regions, respectively, and are given by

$$W_S = \left[\frac{2\epsilon_s}{qN_B}(2\phi_F - V_B)\right]^{1/2} \tag{28}$$

$$W_D = \left[\frac{2\epsilon_s}{qN_B}(2\phi_F - V_B + V_D)\right]^{1/2}. \tag{29}$$

For x_j, W_S, W_D, and L in microns and t_0 in Å, the value of A is $0.41(\text{Å})^{1/3}$ for n-channel devices. Upon scaling, a value of $M \leqslant 1$ will guarantee that long channel subthreshold behavior is maintained. When $M > 1$ undesirable short channel effects are expected.

B. Threshold Voltage

For successful scaling, the threshold voltage must scale with the other voltages. However, due to the nonscalable term $(\phi_{MS} + 2\phi_F)$ in (2), it is unlikely that V_T can be scaled properly without compensation. In general, scaling of this term produces a V_T which is too large for CE scaling and too small for QCV scaling. The problem is more critical in the PMOS case (assuming an n^+ polysilicon gate technology), where $(\phi_{MS} + 2\phi_F)$ is more significant. While some adjustment can be made by varying V_B in NMOS technology, this is not possible in CMOS.

Fortunately, ion implantation can be used in both NMOS and CMOS to adjust the threshold voltage. In a typical device which uses a shallow implant as a threshold adjust, it is common to obtain cancellation between the implant term and $(\phi_{MS} + 2\phi_F)$, whenever necessary[3], so that V_T becomes approximately

$$V_T \simeq \frac{-Q_B - Q_{SS}}{C_0} = \frac{\pm\sqrt{4\epsilon_s qN_B\phi_F} - Q_{SS}}{C_0} \tag{30}$$

(positive sign applies for NMOS while the negative sign applies for PMOS). If $Q_{SS} \ll Q_B$, the above equation will generally yield the desired threshold voltages.

With (30), proper scaling for V_T is easily achieved under the CV and QCV laws, resulting in a threshold voltage scaled by ≈ 1 and $\approx \lambda^{-1/2}$, respectively. To ensure that (30) remains valid, one requires that the condition

$$\left|\frac{qN_i D}{C_0}\right| \approx |\phi_{MS} + 2\phi_F| \tag{31}$$

holds under scaling. Since it is desirable that N_i scales with λ, this implies that the implant depth D must scale with $\lambda^{-1/2}$ in the CV case, and remain fixed for QCV and CE scaling. However, adjustment of D may be necessary to scale V_T with $1/\lambda$ (instead of $1/\sqrt{\lambda}$) under the CE law, and to compensate for the second-order short channel and narrow width effects.

[3]For instance, in an n^+ polysilicon gate p-well CMOS process, the PMOS threshold voltage is adjusted in this manner.

TABLE II
DEVICE SCALING

	CE	QCV	CV	Equation
$I_{D,\text{sat}}$	λ^{-1}	1	$\lambda^{1/2}$	9
g_m	1	$\lambda^{1/2}$	$\lambda^{1/2}$	12
$g_{ds(\text{enh})}$	1	$\lambda^{3/4}$	λ	20
$g_{ds(\text{dep})}$	1	$\lambda^{1/4}$	$\lambda^{1/2}$	34
$\frac{C_d}{C_0}$	1	$\lambda^{-1/4}$	1	23
S_s	1	>1	1	25
V_{ng}	1	1	$\lambda^{1/2}$	26
M	$\lambda^{-1/3}$	$\lambda^{-1/6}$	$\lambda^{1/3}$	27

Nevertheless, proper scaling of V_T seems feasible using ion implantation and has been assumed in the following discussion.

C. First-Order Parameter Scaling

In the model already described, the δ's represent second-order effects which do not scale linearly with λ. Their magnitudes are strong functions of the initial unscaled device and the scaling laws. To obtain a first-order estimate of the effect of scaling on device parameters as well as to keep the analysis independent of the unscaled conditions, the parameters δ are first assumed to be unity. This implies an ideal long channel MOSFET as the unscaled device. Justification for this assumption, and a comparison of first-order scaling estimates with computer simulated results on analog performance, taking into consideration the effect of δ's, are given in Section V.

The first-order scaling for some relevant device parameters are listed in Table II. Limitations can be expected from the parameters S_s and V_{ng}, which do not scale-down as desired.

D. Capacitor Scaling

In MOS technology, capacitors are realized between metal and heavily doped silicon or between two layers of heavily doped polysilicon. Thermally grown silicon dioxide is used as the dielectric. Scaling of capacitors is straightforward. A direct scale down of the surface area and dielectric thickness lead to a reduction in total capacitance. However, difficulties arise in absolute accuracy and matching of component values as the effects of misalignment and fringing become relatively more important. Electron quantization noise may also become a problem as capacitors become smaller.

E. Resistor Scaling

In MOS technology, resistors are fabricated using either diffused layers or polysilicon layers deposited on oxide. The latter alternative is preferred since it produces minimal parasitic capacitance. Scaling, again, is relatively simple; however accuracy and matching as well as conduction mechanisms [24] in the polysilicon become serious limits as the resistors become smaller.

F. Interconnection Scaling

Three items are of prime concern in interconnection scaling. These are electromigration, in the very thin, narrow aluminum

conductors, contact resistance in the very shallow junctions, and polysilicon interconnection sheet resistance. For the scaling conditions under consideration here, the first two items do not present serious constraints. As far as the third item is concerned, refractory metal silicides, which offer low sheet resistance, are a good alternative to polysilicon. In addition, it is reasonable to expect that, at least some of the longest interconnections will not scale-down at all, resulting in adverse effects on parasitic capacitance and line driving capabilities.

IV. Effect of Scaling on Analog Building Block Performance

This section investigates the scaling tendencies of some basic analog building blocks with the objective of determining the optimal scaling law for analog applications. Two technologies are considered: CMOS and NMOS with depletion load.

A. Gain Stage Performance

The most useful figure of merit in evaluating performance of gain stages is the voltage gain. In a CMOS stage, this gain is given by

$$A_{V,\text{CMOS}} = -\frac{g_m}{(g_{ds,n} + g_{ds,p})} \tag{32}$$

where $g_{ds,n}$ and $g_{ds,p}$ are the n- and p-channel device output conductances respectively, as defined by (20). In NMOS technology, where a depletion device is used as load, the bulk threshold feedback effect normally dominates over channel length modulation effects in determining $g_{ds,\text{dep}}$ [20]. The voltage gain, in this case, is given by

$$A_{V,\text{NMOS}} = -\frac{g_m}{g_{ds,\text{dep}}} \tag{33}$$

where

$$g_{ds,\text{dep}} = \frac{\gamma I_{D,\text{sat}}}{|V_{T,\text{dep}}|\,(V_{DD} - V_0 + 2\phi_F)} \tag{34}$$

where $V_{T,\text{dep}}$ is the threshold voltage of the depletion mode device, V_0 is the quiescent output voltage, and V_{DD} is the supply voltage.

B. Op Amp Performance

Since the operational amplifier is one of the basic building blocks in analog circuitry, an evaluation of its performance subject to scaling would be useful. For such an analysis, the following performance indices must be considered.

Voltage Gain A_v—In a typical two-stage op amp which uses a Miller capacitor for compensation, the voltage gain is

$$A_V \propto \left(\frac{g_m}{g_{ds}}\right)^2. \tag{35}$$

Unity Gain Bandwidth (GBW)—The unity bandwidth is mainly determined by the transconductance of the first stage g_{m1}, and the compensation capacitor C_c and can be expressed as

$$\text{GBW} = \frac{g_{m1}}{C_c}. \tag{36}$$

TABLE III
Scaling of Analog Circuits

	CE	QCV	CV	Equation
Gain Stages				
$A_{V,\text{CMOS}}$	1	$\lambda^{-1/4}$	$\lambda^{-1/2}$	32
$A_{V,\text{NMOS}}$	1	$\lambda^{1/4}$	1	33
Op Amp				
$A_{V\text{tot,CMOS}}$	1	$\lambda^{-1/2}$	λ^{-1}	35
$A_{V\text{tot,NMOS}}$	1	$\lambda^{1/2}$	1	35
Power/Area	1	$\lambda^{3/2}$	$\lambda^{5/2}$	39
C_c	λ^{-1}	λ^{-1}	$\lambda^{-3/2}$	–
GBW	λ	$\lambda^{3/2}$	λ^2	36
S_R/V_{DD}	λ	$\lambda^{3/2}$	λ^2	38
S/N	λ^{-1}	$\lambda^{-1/2}$	$\lambda^{-1/2}$	37

The GBW is normally determined by the second parasitic pole of the amplifier which can be approximately expressed as

$$\text{GBW} = \frac{g_{m2}}{C_1 + C_2} \tag{37}$$

where g_{m2} is the effective transconductance of the second stage, and C_1 and C_2 are the input (gate) and load capacitances associated with the second stage [25]. From (36) and (37), it appears that C_c will be proportional to $(C_1 + C_2)$. If one assumes C_2 to be the input (gate) capacitance of the subsequent stage, and that the overlap gate capacitance is small, then $(C_1 + C_2)$, and therefore C_c, will be directly proportional to AC_0, where A is the active gate area of an individual transistor. This implies that the area of C_c will scale with A and that the ratio of capacitor to op amp areas will remain constant under scaling.

Signal-to-Noise Ratio (S/N)—From (26), the signal to noise ratio for low-frequency analog applications can be expressed as

$$S/N = \frac{V_{sg}}{V_{ng}} \propto \frac{V_{sg}\sqrt{ZL}}{t_0} \tag{38}$$

where the input signal amplitude V_{sg} is assumed to scale with other voltages. This implies that the signal-to-noise ratio is dependent on device geometry.

Slew Rate S_R—The slew rate normalized to the supply voltage V_{DD} can be used to provide an indication of large signal op amp response

$$\frac{S_R}{V_{DD}} = \frac{I_{D,\text{sat}}}{V_{DD}C_c}. \tag{39}$$

Power Density—The power density is a useful figure of merit in determining the maximum packing density of op amps per unit area.

$$\text{power density} \propto \frac{I_{D,\text{sat}}V_{DD}}{ZL}. \tag{40}$$

C. Optimum Scaling Law

The three scaling laws presented in the introduction were applied to the first-order scaling of the basic gain stages and the op amp parameters. The results are listed in Table III.

Other than an increase in speed resulting from reduction of C_c, the majority of the analog performance parameters remain

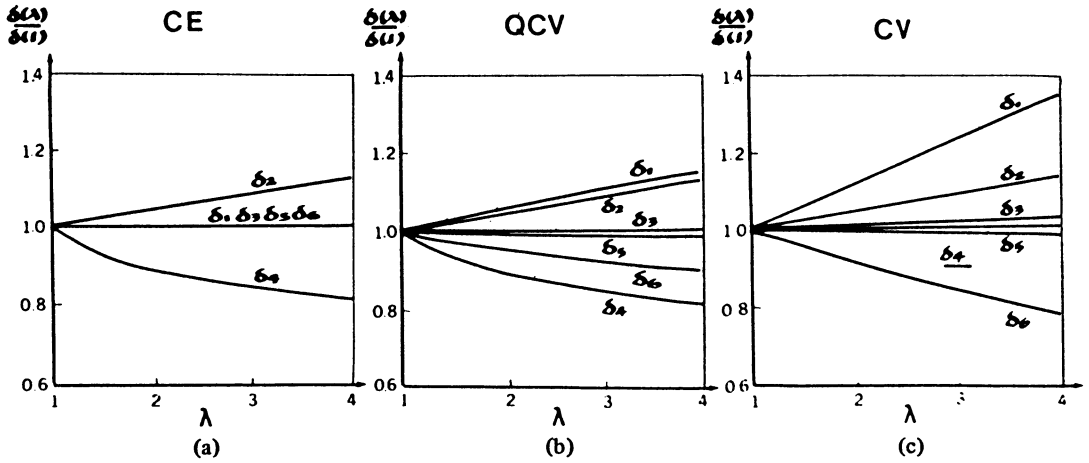


Fig. 1. Effect of scaling on δ's. (a) CE. (b) QCV. (c) CV.

invariant to the application of CE scaling. The exception being that the signal-to-noise ratio is severely degraded.

Although both nonconstant field scaling laws offer improvements in speed and frequency response, CV scaling appears unacceptable for analog applications because it results in the largest voltage gain degradation, the highest power dissipation per unit area without a significant gain in signal-to-noise performance. In addition (referring to (27) and Table II), the parameter M increases with CV scaling, resulting in short channel subthreshold effects becoming dominant at low values of λ.

QCV scaling offers an acceptable compromise, yielding improved speed and signal-to-noise ratio over the CE case, while exhibiting a higher gain and lower power dissipation than CV scaling.

V. Case Study

The scaling analysis performed in the previous sections neglected all second-order effects by assuming an ideal long channel MOSFET as the unscaled device. This allows simple first-order scaling factors to be computed without involving the nonlinear δ terms. In this section, computer simulation involving the δ's is performed using 7 μm channel length MOSFET's as the initial unscaled devices. The simulation focuses mainly on the QCV law in light of the results of the previous section. These results are compared with the first-order scaling theory of Tables II and III. Scaling simulation is achieved by manually scaling all voltages, dimensions, and doping concentrations in accordance with the QCV law and then inputing them into a simulation program which accommodates our model. A representative case study of such a simulation is presented in the following paragraphs.

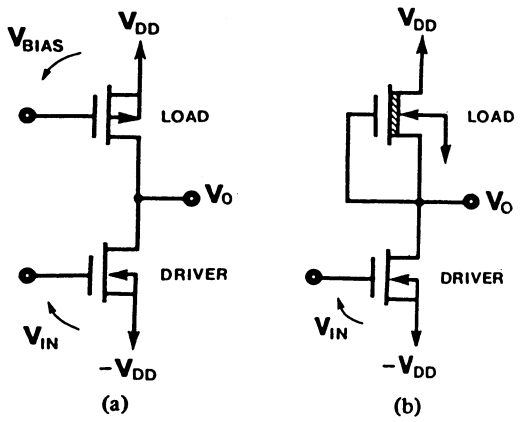
A. Effect of Scaling on the δ Parameters

As an example, consider a typical n-channel device having the following initial unscaled characteristics: $Z = 70\ \mu m$, $L = 7\ \mu m$, $t_0 = 800$ Å, $x_j = 1.2\ \mu m$, $N_A = 5 \times 10^{15}\ cm^{-3}$, $E_p = 1.1 \times 10^4$ V/cm, $V_D = 5$ V, $(V_G - V_T) = 1.5$ V, $V_B = 0$, $R_S = 20\ \Omega$. In this case the value of M is 0.788, implying that the device is still in its long channel mode of operation. From Table II and (27), it is seen that M does not increase for CE or QCV scaling, but becomes greater than 1 for $\lambda \geq 2$ in the case of CV scaling. Using the unscaled device characteristics defined above, the δ values for $\lambda = 1$ are $\delta_1(1) = 1.163$, $\delta_2(1) = 1.049$, $\delta_3(1) = 1.008$, $\delta_4(1) = 1.590$, $\delta_5(1) = 0.855$, and $\delta_6(1) = 0.785$. The effect of scaling on the $\delta(\lambda)$ parameters were computed and are plotted in Fig. 1(a), (b), and (c) for CE, QCV, and CV scalings. The graphs show that in most cases the $\delta(\lambda)/\delta(1)$ ratio remains very near unity (i.e., the δ's do not change drastically with scaling). For δ_3 and δ_5, this is true for all three scaling laws, implying that parasitic resistance and velocity saturation effects are still negligible in the scaling range under consideration. Under worst case conditions, the $\delta(\lambda)/\delta(1)$ factors do not deviate by more than 20 percent from unity for CE and QCV scaling and by more than 40 percent for CV scaling. If CV scaling is restricted to $\lambda \leqslant 2$ (which as discussed above is required to prevent the onset of DIBL), the maximum deviation is 20 percent. These results imply that the second-order effects described by the δ terms can be ignored in establishing the general trend of scaling on device parameters as was done in Sections III and IV. The error caused by ignoring the δ's may sometimes be significant but not dominant.

B. Gain Stage Simulation

The performance of the NMOS and CMOS gain stages, shown in Fig. 2, were investigated using computer simulation. To provide a fair comparison, the two stages were designed to have identical operating conditions and Z/L ratios. To achieve a zero quiescent output voltage in the NMOS stage with quiescent input voltage $V_{in} = V_{DD}/2$ (typical operating conditions in a gain stage), one requires the ratio $(Z/L)_{driver}$: $(Z/L)_{load}$ to be 2.8. In the CMOS stage, the same current and aspect ratios are maintained if V_{bias} is set at $-0.9\ V_{DD}$.

Scaling was carried out starting with a set of unscaled devices with characteristics listed in Table IV. The simulated gain for both gain stages as a function of the scaling factor λ, is shown in Fig. 3. Also shown in the figure are the corresponding gains observed from first-order scaling theory (Table IV). The maximum discrepancy between simulation and first-order theory is of the order of 20 percent. The unscaled gain of the CMOS stage remains higher than that of the NMOS stage. However, the gain in the NMOS case increases with increasing λ while the opposite is true in the CMOS case. The two gains approach each other at $\lambda \simeq 4$. The relative increase



$V_{T,enh}$	0.2 V_{DD}
$V_{T,dep}$	−0.7 V_{DD}
V_{IN}	0.5 V_{DD}
V_{BIAS}	−0.9 V_{DD}
$(Z/L)_{DRIVER}$	2.8
$(Z/L)_{LOAD}$	1
$V_{DD}(\lambda=1)$	5V

(c)

Fig. 2. Gain stages. (a) CMOS. (b) NMOS with depletion load. (c) Specifications.

TABLE IV
NMOS AND CMOS UNSCALED DEVICE CHARACTERISTICS

	NMOS Stage		CMOS Stage	
	Driver	Load	Driver	Load
Type	Enhancement	Depletion	n-channel	p-channel
V_T(V)	1	−3.5[a]	1	−1[a]
N_B(cm^{-3})	5×10^{15}	5×10^{15}	5×10^{15}	1.5×10^{15}
μ_0(cm^2/V · s)	600	600	600	300
E_c(V/cm)	2×10^4	2×10^4	2×10^4	2×10^5
X_j(μm)	1.2	1.2	2.2	1.2
L(μm)	7	7	7	7

[a]Threshold voltages adjusted by means of shallow implants.

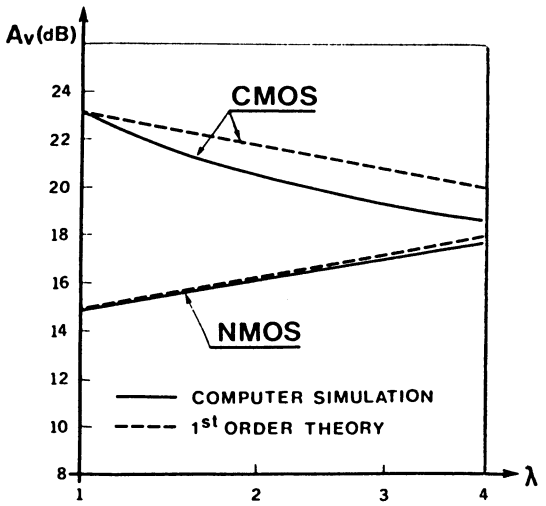
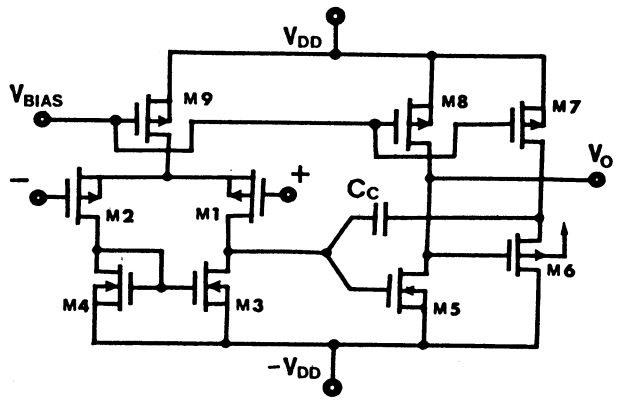


Fig. 3. Inverter voltage gain as a function of λ.

in the gain of the NMOS stage can be attributed to the fact that $g_{ds,dep}$ due to body effect increases at a slower rate ($\lambda^{1/4}$) than that due to channel length modulation ($\lambda^{3/4}$). However, this cannot be regarded as an advantage, since the latter mechanism present in the depletion NMOS would eventually overtake that due to body effect. From the above results, it appears that in the low $\lambda(\lambda < 4)$ range, the CMOS stage can offer a higher gain than that of its NMOS equivalent. This result is not specific to the particular set of device parameters chosen, but was found to be true in the majority of the cases investigated.



(a)

	Z (μm)	L (μm)
M1	100	7
M2	100	7
M3	100	7
M4	100	7
M5	400	7
M6	400	5
M7	100	7
M8	200	7
M9	100	7
V_{DD}	5V	
V_{BIAS}	2V	
C_C	10pF	

(b)

Fig. 4. CMOS operational amplifier. (a) Circuit diagram. (b) Unscaled specifications.

C. Op Amp Simulation

In light of the conclusion that, in general, scaled CMOS offers a gain advantage over scaled NMOS gain stages, a CMOS op amp was selected for simulation. The unscaled op amp and its important layout characteristics are shown in Fig. 4. The

TABLE V
SCALING RESULTS OF OP AMP

	Voltage Gain Av		GBW (MHz)		Power Dissipation (mW)		R_{out} (kΩ)	
	Simulation	First-Order Theory	Simulation	First-Order Theory	Simulation	First-Order Theory	Simulation	First-Order Theory
1	903	903	5	5	16.1	16.1	28	28
2	589	639	12.6	14.1	11.3	11.4	15.7	16.6
4	357	452	31.6	40	8.3	8.2	8.7	9.9
Scaling Factor		$\lambda^{-1/2}$		$\lambda^{3/2}$		$\lambda^{-1/2}$		$\lambda^{3/4}$

circuit configuration is typical of op amps presently used in telecommunication applications. It consists of two gain stages with a total gain of about 60 dB. A buffer stage is used in the feedback compensation loop to eliminate an undesirable zero at g_m/C_c [25]. The op amp has a useful feature associated with the fact that all the transistor channel lengths are nearly equal which guarantees equal V_T for all the transistors, even under scaled conditions.

Table V lists the results of the simulation as a function of λ. Also shown in the table are the results obtained from first-order scaling theory. Agreement between the simulation and first-order theory appears reasonable, confirming the validity of using the first-order theory to estimate trends in analog scaling.

VI. CONCLUSION

A theoretical analysis of the impact of scaling on analog component and circuit performance has been presented. Among the three scaling laws considered, QCV appears to be the optimum for analog scaling. Its application results in small area, high speed and moderate degradation in gain, power density, and signal-to-noise ratio. The selection of QCV is compatible with Chatterjee's [3] choice of the same scaling law for digital applications. Thus it appears feasible to scale both the analog and digital portions of a circuit using the same scaling law.

A typical case study comparing the performance of NMOS and CMOS gain stages under moderate scaling conditions shows that CMOS offers the optimum gain configuration for scaled analog implementations. A comparison between computer simulation (of gain stages and a CMOS op amp) and first-order scaling theory shows that second-order effects produce significant but nondominant errors in evaluating the performance of analog components under scaled conditions. Thus, first-order theory can be used to estimate scaling tendencies in analog applications.

APPENDIX I

Considering (16), if one lets

$$x = \frac{E_p K_1}{2(V_D - V_{D\,sat})^{1/2}}, \tag{A1}$$

this equation becomes

$$\Delta L = K_1 (V_D - V_{D\,sat})^{1/2} \{(1 + x^2)^{1/2} - x\}. \tag{A2}$$

If x is small, implying a moderate E_p, (A2) can be simplified to

$$\Delta L \simeq K_1 (V_D - V_{D\,sat})^{1/2} \left(1 - x + \frac{x^2}{2}\right). \tag{A3}$$

Differentiating (A3) yields

$$\begin{aligned}\frac{\delta \Delta L}{\delta V_D} &= \frac{1}{2} \frac{K_1}{(V_D - V_{D\,sat})^{1/2}} \left(1 - x + \frac{x^2}{2}\right) \\ &\quad + K_1 (V_D - V_{D\,sat})^{1/2} (x - 1) \frac{\delta x}{\delta V_D} \\ &= \frac{1}{2} \frac{K_1}{(V_D - V_{D\,sat})^{1/2}} \left\{1 - x + \frac{x^2}{2} + x\right\} \\ &= \frac{1}{2} \frac{K_1}{(V_D - V_{D\,sat})^{1/2}} \left\{1 - \frac{x^2}{2}\right\}. \end{aligned} \tag{A4}$$

Substituting (A3) and (A4) into

$$g_{ds} = \frac{I_{D,\,sat}}{L\left(1 - \dfrac{\Delta L}{L}\right)^2} \frac{\delta \Delta L}{\delta V_D} \tag{A5}$$

and neglecting all x^2 terms yields

$$g_{ds} = \frac{I_{D,\,sat} K_1}{2L(V_D - V_{D\,sat})^{1/2} [1 - K_1 (V_D - V_{D\,sat})^{1/2} (1 - x)]^2} \tag{A6}$$

which is (20). Note also that by neglecting x^2, (A3) becomes equivalent to (18).

APPENDIX II

Consider the ideal current equation

$$I_{D\,sat} = \frac{\beta}{2} (V_G - V_T)^2. \tag{B1}$$

When DIBL is present, the threshold voltage V_T is a function of V_{DS}, according to Masuda [21], the dependence is

$$V_T = V_{T_0} - \eta (V_{DS} - \phi) \tag{B2}$$

where

$$\eta = \frac{\eta_0 (x_j, N_B)}{L^3}. \tag{B3}$$

V_{T_0}, η_0, and ϕ are constants dependent on the technology. By substituting V_T into (B1) and differentiating with respect

to V_{DS}, one obtains

$$g_{ds} = \frac{\delta I_{D\,\mathrm{sat}}}{\delta V_{DS}} \simeq \frac{\beta\eta}{L^3}(V_G - V_T) \simeq \frac{g_m\eta}{L^3}. \tag{B4}$$

Therefore, as L becomes smaller, this effect will become a dominant factor in determining g_{ds}.

References

[1] C. A. T. Salama, "VLSI technology for telecommunication IC's," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 253-250, 1981.

[2] R. H. Dennard *et al.*, "Design of ion implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 256-266, 1974.

[3] P. K. Chatterjee, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," *Electron. Dev. Lett.*, vol. EDL-1, pp. 220-223, Oct. 1980.

[4] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics—I-MOS technology," *Solid-State Electron.*, vol. 15, pp. 819-829, 1972.

[5] E. Demoulin, "Process statistics of submicron MOSFET's," in *Tech. Dig., IEDM Conf.*, Washington, DC, 1979, pp. 34-37.

[6] Y. A. El-Mansy, "On scaling MOS devices for VLSI," in *Proc. IEEE Int. Conf. Circuits Comput.*, 1980, pp. 457-460.

[7] G. Merckel, "A simple model of the threshold voltage of short channel and narrow channel MOSFET's," *Solid-State Electron.*, vol. 23, pp. 1207-1213, 1980.

[8] P. P. Wang, "Device characteristics of short channel and narrow width MOSFET's," *IEEE Trans. Electron. Devices*, vol. ED-25, pp. 779-786, July 1978.

[9] Y. A. El-Mansy *et al.*, "A simple 2-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron. Devices*, vol. ED-24, pp. 254-262, Mar. 1977.

[10] M. H. White *et al.*, "High-accuracy MOS models for computer-aided design," *IEEE Trans. Electron. Devices*, vol. ED-27, pp. 899-906, 1980.

[11] H. S. Lee, "An analysis of the threshold voltage for short channel IGFET's," *Solid-State Electron.*, vol. 16, pp. 1407-1417, 1973.

[12] L. D. Yau, "A simple theory to predict the threshold voltage of short channel IGFET's," *Solid-State Electron.*, vol. 17, pp. 1059-1063, 1974.

[13] O. Leistiko, "Electron and hole mobilities in inversion layers on thermally oxidized silicon surfaces," *IEEE Trans. Electron. Devices*, vol. ED-12, pp. 248-254, 1965.

[14] F. M. Klaassen and W. C. J. de Groot, "Modeling of scaled down MOS transistors," *Solid-State Electron.*, vol. 23, pp. 237-242, 1980.

[15] G. Baum and H. Beneking, "Drift velocity saturation in MOS transistors," *IEEE Trans. Electron. Devices*, vol. ED-17, pp. 481-482, 1970.

[16] C. Hilsum, "Simple empirical relationship between mobility and carrier concentration," *Electron. Lett.*, vol. 10, no. 13, pp. 259-260, Jan. 1974.

[17] V. G. K. Reddi and C. T. Sah, "Source to drain resistance beyond pinchoff in MOS transistors," *IEEE Trans. Electron. Devices*, vol. ED-12, pp. 139-141, 1965.

[18] T. Poorter and J. H. Satter, "A dc model for an MOS transistor in the saturation region," *Solid-State Electron.*, vol. 23, pp. 765-772, 1979.

[19] R. R. Troutman, "VLSI limitations from drain induced barrier lowering," *IEEE Trans. Electron. Devices*, vol. ED-26, pp. 461-468, Apr. 1979.

[20] B. Hoefflinger *et al.*, "Model and performance of hot electron MOS transistors for VLSI," *IEEE J. Solid-State Circuits*, vol. 14, pp. 435-442, 1979.

[21] H. Masuda *et al.*, "Characteristics and limitation of scaled down MOSFET's due to two-dimensional field effect," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 980-986, 1979.

[22] H. Katto *et al.*, "MOSFET's with reduced low frequency $1/f$ noise," *Oyo Buturi* (Japan), vol. 44, pp. 243-248, 1975.

[23] J. R. Brews, "Generalized guide for MOSFET miniaturization," in *Tech. Dig., IEDM Conf.*, 1979, pp. 10-13.

[24] N. C. C. Lu *et al.*, "A new conduction model for polycrystalline silicon films," *Electron. Dev. Lett.*, vol. EDL-2, pp. 95-98, 1981.

[25] P. R. Gray, "Basic MOS operational amplifier design—An overview," in *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980, pp. 28-49.

[26] Y. P. Tsividis, "Design consideration in single channel MOS analog integrated circuits—A tutorial," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 383-391, June 1978.

Matching Properties, and Voltage and Temperature Dependence of MOS Capacitors

JAMES L. McCREARY, MEMBER, IEEE

Abstract—**The matching properties of MOS capacitors are modeled and compared with measured data. A weighted-capacitor array design approach is described. Voltage and temperature dependence of MOS capacitors are analyzed, modeled, and compared with measured data. It is shown that to a first-order heavily doped polysilicon accumulates and depletes similar to crystallyine silicon.**

List of Symbols

A/P	Area-to-perimeter ratio
BWC	Binary weighted capacitor
C_{ox}	Silicon dioxide capacitance per unit area, k_{ox}/t_{ox} (F/cm^2)
C_s	Semiconductor space charge capacitance per unit area (F/cm^2)
C_t	Total MOS capacitance per unit area (F/cm^2)
E_f	Fermi level (eV)
ϵ_0	Permittivity of free space
ϵ_{ox}	Dielectric constant of silicon dioxide
ϵ_{si}	Dielectric constant of silicon
k	Boltzmann's constant
L_{si}	Coefficient of linear thermal expansion of silicon, 2.8 ppm/C
M[WCD]	Mean value of the WCD
N_d	Donor impurity concentration
Phis	Surface electrostatic potential relative to semiconductor bulk
q	Electronic charge
Q_s	Semiconductor space charge per unit area
SD	Standard deviation
SD[x]	Standard deviation of the measured values of x
SD%[x]	Percent of full scale standard deviation of x
sgn[x]	The sign of x; or x divided by the absolute value of x
t_{ox}	Silicon dioxide thickness
T	Absolute temperature
T_{CC}	Temperature coefficient of capacitance (ppm/°C)
T_{CCj}	T_{CC} for a junction capacitor (ppm/°C)
U_f	Dimensionless Fermi potential, E_f/kT
U_s	Dimensionless surface potential, $q\,Phis/kT$
V_{CC}	Voltage coefficient of capacitance (ppm/V)
V	Voltage applied at the top plate of MOS capacitor with substrate or lower power held at ground
V_b	Barrier potential
V_{fb}	Flat-band voltage
V_j	Reverse-bias junction voltage
WCD	Worst-case deviation
%FS	Percent of full scale

Manuscript received May 5, 1981; revised June 6, 1981.
The author is with the Intel Corporation, Santa Clara, CA 95051.

I. Introduction

IN MOS capacitor circuits that perform A/D conversion [1]-[3], precision analog gain and attenuation [4], [5], and filtering [6], [7], the performance and also the cost effectiveness depend upon the accuracy of the capacitor ratio matching. This paper describes a systematic approach to capacitor array design that uses a statistical model for capacitor ratio errors. The voltage coefficient and temperature coefficient of MOS capacitors are also discussed. Measured data are compared with first-order calculations. The overall objective of the paper is not only to report this investigation, but also to establish a comprehensive design approach for MOS capacitor array circuits.

In the past, capacitor ratio errors have been modeled using a statistical analysis which treats these errors as random variables [8], [9]. More recently, Yee *et al.* [10] have done an analysis of capacitor errors in a two-stage capacitor array. All of these models have assumed that the errors were random, uncorrelated, and normally distributed. These assumptions are supported by data shown in this paper. However, little has been published regarding a systematic design approach. In addition, all previously reported capacitor matching data have been based upon limited data gathered using measurement techniques of limited accuracy. The analysis which follows is supported by data from more than 32 000 capacitor arrays taken from a total of 16 groups of wafers processed in five different technologies. Measurements were made using a 16-bit accurate, self-calibrating technique previously described [11]. Since the database is large enough, useful empirically deduced relationships are identified for aid in design.

II. Matching Properties of MOS Capacitors

Array Designs with Constant Area-to-Perimeter Ratio

Attention is now focused upon arrays with constant A/P ratio. Here, the ratio of the area of the capacitor to the total perimeter of the plate defining the capacitor is constant for each capacitor in the array. For purposes of this discussion, a capacitor array is defined as a set of capacitors with one common node connection. Consider an array of two capacitors C_m and C_k, made of m and k identical unit capacitors $C1$, respectively. Let SD[$C1$] be the SD (the standard deviation) in

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-16, no. 6, pp. 608-616, Dec. 1981.

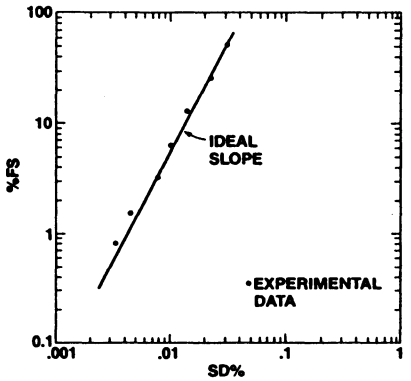


Fig. 1. Capacitor ratio in %FS versus the standard deviation of ratio errors for a 128-unit BWC array.

the distribution of errors in $C1$ from its ideal value for a particular process. Then, assuming that the errors are uncorrelated, random and normally distributed, the following equations can be derived:

$$SD[C_m] = [m]^{1/2}\, SD[C1], \tag{1}$$

and

$$SD[C_k] = [k/m]^{1/2}\, SD[C_m]. \tag{2}$$

A more useful parameter, however, is SD% which is the SD of the error in capacitor ratio (rather than capacitor value) expressed as percent of full scale, %FS:

$$SD\%[C_k, k+m] = 100\, SD[C_k]/(C_k + C_m) = [k/m]^{1/2}\, SD\%[C_m, k+m] \tag{3}$$

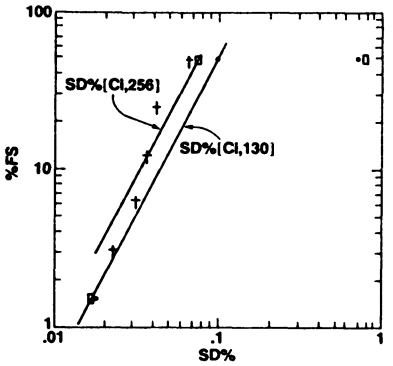
where $k + m$ is the total number of units in the array. For a BWC array, this equation predicts that the slope of log(capacitor ratio) versus log(SD%) is 2 as shown by the line on the graph of Fig. 1 where the capacitor ratio is expressed as %FS. This is true since capacitor ratio is directly proportional to k while SD% is proportional to the square root of k. This figure also shows measured data points for a certain BWC array design. These closely follow the predicted slope. For this array the unit plate size was approximately 72 μm X 72 μm and consisted of 128 total units and was fabricated in a technology denoted as CMOS-1. The A/P ratio was 18 μm.

Dependence upon Total Area for a Constant A/P Ratio

Consider now a 130-unit capacitor array having four capacitors: $C64A$, $C64B$, $C1A$, and $C1B$. Both $C64A$ and $C64B$ are each composed of 64 unit capacitors $C1$. $C1A$ and $C1B$ are each composed of a 1 unit capacitor. The unit capacitor size under study was 25 μm X 25 μm corresponding to an A/P ratio of 6.25 μm. When taken in parallel $C1A + C1B$ can also be considered as a single two-unit capacitor $C2$. From measurements on this array SD%[$C64A$, 130] was found to be 0.10%FS. Then the following equation can be written for $C2$:

$$SD\%[C2, 130] = [2/64]^{1/2}\, SD\%[C64A, 130] = 0.017\ \%FS \tag{4}$$

which agreed closely with the measured value of 0.018 %FS.



□ CALCULATED
• MEASURED FROM 130 UNIT ARRAY
† MEASURED FROM 256 UNIT ARRAY

Fig. 2. Measured and calculated values of capacitor ratio versus ratio error for a 130-unit array and a 256-unit array.

This is illustrated graphically in Fig. 2 along with a line labeled SD%[C_i, 130] representing the ideal behavior.

We can also calculate the matching of $C1A$ to $C1B$ by assuming that these are the only capacitors in a two-unit array. Using (2) and (3), it can be shown that in the general case the SD% of capacitor C_i composed of i units of $C1$ in an array of j units of $C1$ can be expressed in terms of the SD% of capacitor C_m having m units of $C1$ in a different array of k units of $C1$:

$$SD\%[C_i, j] = (k/j)\, [i/m]^{1/2}\, SD\%[C_m, k]. \tag{5}$$

Using this equation, the SD% can be calculated for each capacitor in the two-unit array based upon the SD% value measured for capacitor $C64A$:

$$SD\%[C1, 2] = (130/2)\, [1/64]^{1/2} \cdot SD\%[C64A, 130] = 0.80\ \%FS. \tag{6}$$

The measured value of the same parameter was found to be 0.64 %FS and is shown in Fig. 2 along with the calculated value.

Now assume that it is desirable to reduce the SD% and that this is done by increasing the total area such that the entire array now has 256 units of the same capacitor $C1$. The new SD% of the largest capacitor $C128$ is given by

$$SD\%[C128, 256] = (130/256)\, [128/64]^{1/2} \cdot SD\%[C64A, 130] = 0.072\ \%FS. \tag{7}$$

The measured value of the above parameter for the 256-unit array was found to be 0.068 %FS. Only limited data (provided by C. Laber) were available for this particular array. The measured data points for all capacitors in this array are also shown in Fig. 2 along with a line labeled SD%[C_i, 256] which denotes the ideal behavior of the array as predicted by the single data point for SD%[$C64A$, 130].

Capacitor Ratio Error Dependence upon A/P Ratio

Now consider the effect of different A/P ratios upon capacitor matching errors. One would expect that capacitor plate edge resolution limitations would cause larger matching errors for smaller plate sizes. It is obvious that at some point, photo-

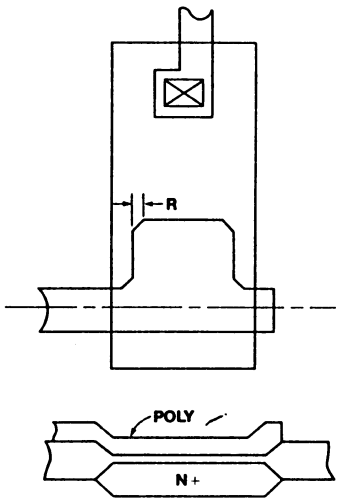


Fig. 3. Unit capacitor layout type-A for poly-to-Si capacitor showing slanted corners and poly interconnect.

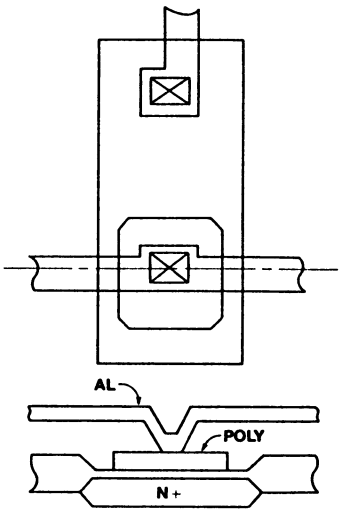


Fig. 4. Unit capacitor layout type-B for poly-to-Si capacitor using slanted corners and Al interconnect.

lithography limitations result in edge location uncertainties which cause significant capacitor area variations. Hence a larger A/P ratio would seem to be desirable. In fact, a single square or even a single circle would be a preferable geometry over a multiple-plate capacitor. However the largest capacitor cannot be a single plate since then the capacitor ratios would not be process insensitive (see [8] for a detailed discussion). Previous approaches have involved unit plate sizes ranging from the smallest capacitor [12] to intermediate values [8]. Other approaches have used voltage division between the main BWC array and an additional BWC array [13], [10]. These techniques have the advantage of keeping the A/P ratio larger than would otherwise occur for a single array and also reduce the total area required. The obvious extension of this approach would be a C-$2C$ ladder array in which two is the largest ratio between capacitors. However, this and the previous approach introduces an additional matching constraint upon the large voltage dependent parasitic capacitors in the array in order to maintain the required voltage divider precision. In this case absolute linearity may be difficult to achieve. Furthermore, it can be shown that although a second array or resistor string improves the overall resolution when added to a larger array, it does not substantially affect the integral linearity or WCD expressed as %FS. However, the linearity and WCD expressed in LSB (least significant bits) are degraded by approximately the same factor that the resolution is increased. This is a consequence of increasing the resolution without a corresponding increase in component matching accuracy.

How small the unit plate size may be before the perimeter effects become important is a function of the particular process technology. For very large unit plate sizes, the perimeter effects may be neglected. Assume that this is the case for the 128-unit array with A/P = 18 μm, previously shown in Fig. 1. Using this assumption, we now calculate the error SD%[$C64$, 130] for the 130-unit array having A/P = 6.25 μm. For square geometries, a general equation may now be written which normalizes for differences in the unit plate A/P ratio

$$\mathrm{SD\%}[C_i, j] = (A_k/A_j)^{1/2} \frac{k}{j} (i/m)^{1/2} \mathrm{SD\%}[C_m, k]. \quad (8)$$

Here A_k and A_j represent the unit plate areas of the k-unit and j-unit arrays, respectively. Using this equation, the error for $C64$ in the smaller 130-unit array may be calculated from measured data on the larger array. We calculate SD%[$C64$, 130] to be 0.085 %FS. This is slightly less than the 0.1 %FS measured value shown in Fig. 2. The difference is probably attributable to the nonnegligible perimeter effects in the 130-unit array.

Ratio Error Dependence upon Layout Design

Two experimental arrays, A and B, were identical in die area except that A used the layout shown in Fig. 3 while B used that of Fig. 4. The ratio error SD%[$C64$, 128] for the largest capacitor in each array was measured. For type A, this value was 0.026 %FS, while for type B, the value was 0.037 %FS. The reason that type A provides better matching is not obvious, however, this may be due to the larger A/P ratio (for the same die area) that is possible with type A layout. Also the process of etching the metal is not well controlled leading to variations in the capacitance of the plate-to-plate metal interconnect.

Other layout techniques involve the use of common centroid geometry [8] which offers some improvement for long-range oxide gradients. Short-range gradients often encountered with poly-to-poly capacitors are not cancelled by this technique.

Ratio Error Dependence upon Process Technology and Capacitor Structures

For a given array design, capacitor ratio errors are a function of the capacitor structure and may also depend upon the process technology. A particular 10-bit BWC array using the unit plate layout of Fig. 3 was fabricated in five different technologies. This particular design used a 72 μm X 72 μm unit plate size for the large array capacitors $C8$, $C16$, $C32$, $\cdots$ $C512$. However, the smaller capacitors $C4$, $C2$, $C1A$, and $C1B$ were approximately square geometries. The goal of this technique was to find an optimal compromise between a small unit plate size to minimize process sensitivity and a large unit plate size to increase A/P (and hence improve matching). Each design used identical geometries. The first technology was a metal-gate NMOS process denoted as NMOS-MG. This design has been

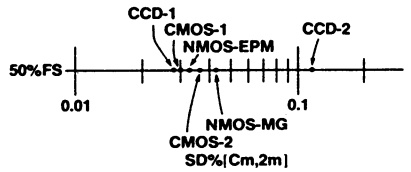


Fig. 5. Measured standard deviation of ratio error for a 10-bit BWC array fabricated in five different technologies. Only the error for the largest capacitor in each array is shown (the 50 %FS point).

previously reported [8]. The unit capacitor structure was an Al top plate with an n^+ silicon lower plate. The second technology was a two-level polysilicon CCD process used to fabricate capacitor arrays having two different structures. The first structure CCD-1 used a poly2 (upper level polysilicon) to n^+ silicon unit capacitor structure. The second CCD-2 was a poly2-to-poly1 capacitor structure similar to Fig. 3 except that the n^+ lower plate was replaced by poly1. The third technology, was an older poly gate CMOS process CMOS-1. Here the unit capacitor structure was identical to CCD-1. The fourth technology was an advanced, high-speed CMOS technology CMOS-2, which also used the same poly-to-silicon capacitor structure as CCD-1. The fifth technology was a modified NMOS EPROM process denoted as NMOS-EPM. This used a poly2-to-poly1 capacitor structure similar to CCD-2.

Fig. 5 shows the measured ratio-error data for each of the five arrays. Only the data point SD%$[C_m, 2m]$ at 50 %FS for the largest capacitor in the array C_m is shown in the figure. It is evident that the poly-to-silicon capacitor structures CMOS-1, CMOS-2, and CCD-1 have the best matching properties in the technologies examined. However, the matching properties of the poly2-to-poly1 structure is good for NMOS-EPM but extremely poor for CCD-2. The reason for this is not known, however, the CCD-2 structure was observed to have an extremely rough poly1 surface that may have contributed to the high ratio-error and also to an abnormally high oxide defect density. On the other hand, an EPROM process requires a high-quality, uniform poly-to-poly oxide. Lastly, the NMOS-MG structure matches worse than do any of the other except for CCD-2.

Total Yield Analysis

An important factor in capacitor array design is yield to a given linearity. For purposes of this discussion, the total capacitor array yield can be expressed as $Y_t = Y_f Y_r$. The functional yield Y_f is dependent primarily upon oxide defect density D and total capacitor oxide area A such that $Y_f = f(A, D)$. Several functional yield models have appeared in the literature [14], [15]. In a well-monitored process, the parameter D is known or else can be determined from test patterns.

The ratio-matching yield Y_r can be calculated from the required linearity and measured data. Let z be the maximum allowable nonlinearity in %FS. Let M[WCD] and SD%[WCD] be the mean and SD% of the WCD expressed as %FS. Y_r can be calculated directly by integrating the area under the Gaussian probability density function (having the standard deviation SD%[WCD]), centered at M[WCD], from $-z$ to $+z$. The following empirical relationship has been observed from measured data:

$$\text{SD\%[WCD]}/\text{SD\%}[C_m, 2m] = 1.5 \text{ to } 1.0. \tag{9}$$

For the 10-bit BWC array previously discussed, typical values of Y_f and Y_r were 95 and 80 percent (for WCD less than 0.05 %FS), respectively.

III. Voltage Dependence of MOS Capacitors with Degenerately Doped Silicon Plates

Voltage Coefficient of Capacitance

It is evident from the literature [16], [17] that the equations which describe the CV behavior of an MOS capacitor may be cumbersome, especially for structures involving degenerately doped surfaces. Usually, however, simplicity of description is desired. For those cases, which are often circuit design oriented, it is sufficient to specify the nominal capacitance value and the rate of change of capacitance over some voltage interval. For this reason, the voltage coefficient of capacitance V_{CC} is used: $V_{CC} = 1/C\ (dC/dV)$. This is the rate of fractional change in C per unit voltage at some dc voltage V. For large N_d and small V, C tends to become a linear function of applied voltage. The total MOS capacitance is given by the series combination of oxide and space charge capacitance:

$$C_t = \frac{1}{(1/C_{ox} + 1/C_s)}. \tag{10}$$

Equation (10) is valid for an MOS capacitor with one semiconductor surface, while the other surface is assumed to have a negligible space charge region (a metal). Using Maxwell-Boltzmann (MB) occupation statistics, the space charge capacitance is given by the well-known equation [18]

$$C_s = q/kT\, dQ_s/dU_s$$
$$= \text{sgn}[U_s] \left[\frac{\epsilon_{si} q^2 N_d}{2kT}\right]^{1/2} [e^{U_s} - 1]\ [e^{U_s} - 1 - U_s]^{-1/2}. \tag{11}$$

The CV model using this equation for C_s will be referred to as the MB CV model. V_{CC} may be expressed as

$$V_{CC} = \frac{C_{ox}^2}{\epsilon_{si} q N_d} \frac{e^{2U_s} - 2U_s e^{U_s} - 1}{(e^{U_s} - 1)^3}. \tag{12}$$

Taking the limit as U_s approaches zero (V approaches V_{fb}),

$$V_{CC}(V_{fb}) = \frac{C_{ox}^2}{3q\epsilon_{si} N_d}. \tag{13}$$

Since V_{fb} is small, (13) is nearly identical to the value at $V = 0$ for large values of N_d.

The values of V_{CC} as a function of doping are plotted in Fig. 6 [curve labeled $V_{CC}(0\text{ V})$]. It can be seen from (13) that V_{CC} can be reduced by increasing N_d, and increasing t_{ox}. To first order, V_{CC} is independent of temperature in this analysis. In addition $V_{CC}(-10)$ and $V_{CC}(+10)$, the voltage coefficients at -10 and $+10$ V, respectively, were also computed from (12). The V_{CC} curves for these two voltages are also plotted in Fig. 6.

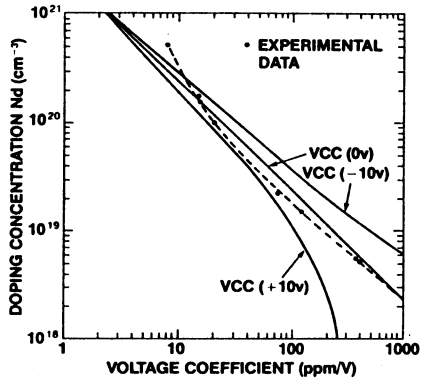


Fig. 6. Plot of calculated surface doping concentration versus voltage coefficient for 0, −10, and +10 V. A dashed line is an approximate fit to the experimental data points.

Experimental Procedure and Results For the Al-to-Si Capacitor

Several n-type Si wafers received a range of phosphorus predepositions from medium to heavy concentrations. Dry 1000 Å thermal oxides were then grown. Half of each wafer received a poly deposition followed by a second phosphorus doping. Then, Al was evaporated and capacitor top plates were patterned in the Al. This was followed by a poly etch and final anneal in hydrogen at 450°C. It has been assumed that this process results in negligible interface charge and fast surface state densities. However, some studies have suggested that this may not be true if the surface concentration becomes too large [19]. In this manner, both poly-to-Si and Al-to-Si capacitors were fabricated on the same Si substrate.

C-V measurements were performed at 1.5 MHz on all samples using a modified PAR model 410 *C-V* plotting system. Capacitance bridge measurements were also made as a function of dc voltage to confirm the plotted data. The samples were then grooved and spreading resistance data gathered from which the doping profiles in the poly and in the Si were determined [20]. Using the measured values of oxide thickness together with doping information, the *CV* curves were then computed for the MB *CV* model described in this paper. Allowing for the uncertainty in the exact position of the origin, the measured curve was then translated onto the same axes as the calculated curve to allow for comparison of slopes (voltage coefficients) rather than absolute values.

The measured values for V_{CC} obtained near the origin for all of the samples are plotted in Fig. 6. The experimental value for $N_d = 5 \times 10^{20}/\text{cm}^3$ was obtained from the literature [9]. One surprising result is the relatively good agreement between the MB model and the measured data. This is unexpected since it is well known that the MB model is valid only for the nondegenerate case (N_d less than 10^{19}). For the case of degenerately doped Si, the formal analysis becomes extremely complex [21].

Analysis of Poly-to-Si and Poly-to-Poly MOS Capacitor Structures

It will be shown in this paper that heavily doped poly accumulates and depletes in the same manner as heavily doped single crystalline Si. Consequently, the *CV* model for crystalline Si can be applied to the poly surface. In the case of a poly-to-poly or poly-to-Si capacitor, one expects that since one surface accumulates while the other depletes and vice versa that the voltage dependence tends to cancel if both surfaces are equally doped. A direct solution for *CV* dependence of this type of capacitor involves a simultaneous solution for both surface potentials and spaces charges such that the electric field in the oxide is identical at both interfaces. Referring to Fig. 7, an assumption is now made that simplifies the analysis. The electrostatic surface potential $Phis_2$ is considered negligible when computing $Phis_1$ and vice versa. This is a reasonable approximation when at least one surface is heavily doped. The device model can then be reduced to two space charge capacitances C_{s1} and C_{s2} in series with C_{ox} as shown in Fig. 7. Now the *CV* curves for a poly-to-Si device having different surface dopings can be computed in a direct manner:

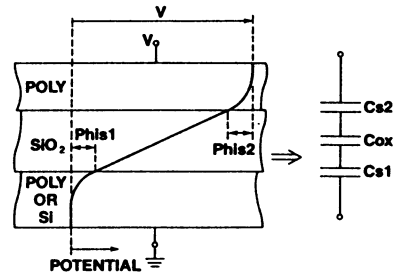


Fig. 7. Model for poly-to-Si and poly-to-poly capacitor showing two space charge capacitances.

$$1/C_t = 1/C_{ox} + 1/C_{s1} + 1/C_{s2}. \tag{14}$$

The analytical *CV* curves are obtained using (11) and (14). Based on this simple device model, the voltage coefficient for poly-to-Si and poly-to-poly capacitors is also approximately given by $V_{CC} = V_{CC1} - V_{CC2}$, where the plate numbering convention used is that shown in Fig. 7. The individual plate voltage coefficients are determined in the usual manner using Fig. 6.

Experimental Results for the Poly-to-Si Capacitor

Comparisons of some analytical and experimental results are shown in Figs. 8 and 9. The devices selected for display in these figures include one having both plates heavily doped, device *D*1 (Fig. 8). Fig. 9 illustrates a device *D*2 having a much smaller N_d in the poly than in the Si, while the opposite situation exists for the device *D*3. Agreement between the calculated and experimental data is not as good for lightly doped poly surfaces (doping less than $10^{19}/\text{cm}^3$). This may be associated with the error in spreading resistance measurements on thin lightly doped films or may be due to the onset of grain boundary effects.

It is interesting to note that for the device *D*2 having a poly plate that is more lightly doped than the crystalline silicon (Fig. 9) the poly plate provides the dominant space charge capacitance and hence the voltage dependence. In this case, the total capacitance decreases with increasing positive voltage since the poly surface is depleting and the Si surface is undergoing negligible accumulation. Hence the fundamental assertion that a heavily doped poly surface accumulates and depletes similar to that to crystalline Si is confirmed.

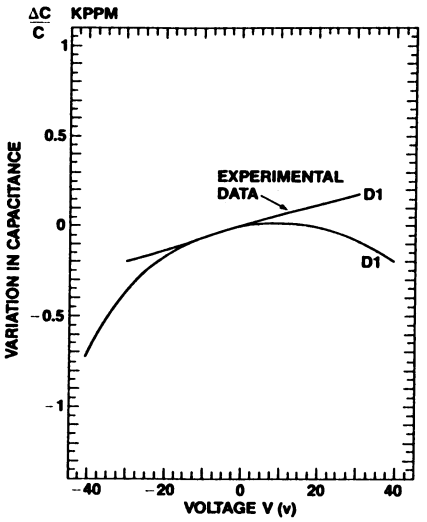


Fig. 8. Calculated and measured CV curves for a poly-to-Si capacitor $D1$ with N_d(poly) - 1.5 X 10^{20}/cm^3 and N_d(Si) = 1.1 X 10^{20}/cm^3.

If both the poly and Si were identically doped, the total capacitance variation would be symmetrical about the flat-band voltage, resulting in a nearly zero voltage coefficient. The MOS capacitor $D1$ in Fig. 8 exhibits a voltage coefficient of approximately 7 ppm/V.

Nonlinearities in Capacitor Arrays Used for A/D and D/A Conversion

When capacitor arrays are used in A/D and D/A conversion, a charge transfer error may be caused by the capacitor voltage dependence. This error may be expressed as a coding nonlinearity in terms of V_{CC}. Fig. 10 illustrates the final charge distribution between two capacitors $C1$ and $C2$. Here the V_{CC} polarities shown are chosen to minimize parasitic capacitance and leakage current at node X. The figure shows that one plate of $C1$ is raised to the full-scale reference voltage V_r. The nonlinearity in the voltage V_x due to V_{CC} can now be calculated by setting the charge on $C1$ equal to that on $C2$ and integrating the capacitance over the voltage transition. By using simplifying approximations, we get

$$V_x = V_r \frac{C1}{C1 + C2} - V_r^2 \frac{V_{CC}}{2} \frac{C1C2}{(C1 + C2)^2}. \quad (15)$$

The error voltage at node X can now be expressed as a nonlinearity in %FS:

$$\text{nonlinearity due to } V_{CC} = -50\, V_{CC} V_r N(1 - N)\ \%\text{FS} \quad (16)$$

where $N = C1/(C1 + C2)$. N corresponds to the fraction of full-scale signal being coded. It is seen that the maximum nonlinearity occurs at one-half full scale ($N = 1/2$) and has the value

$$\text{maximum nonlinearity due to } V_{CC} = -50\, V_{CC} V_r/4\ \%\text{FS}. \quad (17)$$

This is identical to that calculated for a capacitor A/D converter in which a comparator is connected to node X and a successive approximation algorithm is used to drive node X from $-V_{in}$ back to zero.

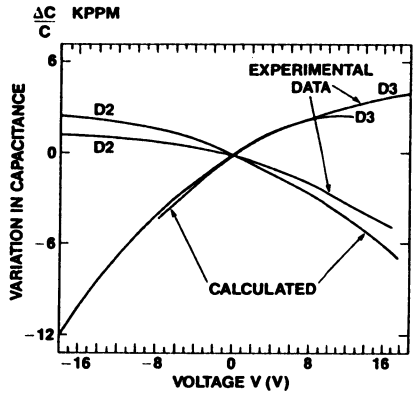


Fig. 9. Calculated and measured CV curves for poly-to-Si capacitor $D2$ with N_d(poly) = 9 X 10^{18}/cm^3 and N_d(Si) = 1.65 X 10^{20}/cm^3, and curves for poly-to-Si capacitor $D3$ with N_d(poly) = 1.2 X 10^{20}/cm^3 and N_d(Si) = 5.5 X 10^{18}/cm^3.

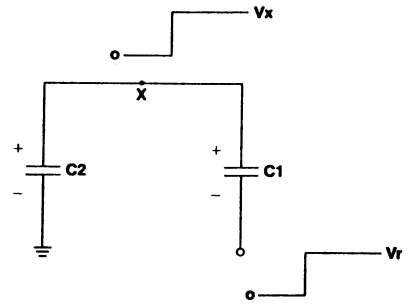


Fig. 10. Simplified representation of capacitor array charge redistribution during A/D or D/A conversion.

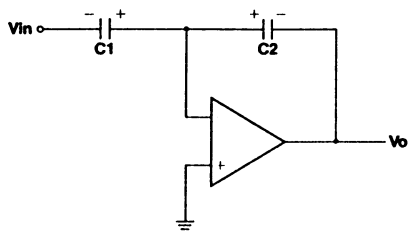


Fig. 11. Simplified schematic of precision analog amplifier, attenuator, or D/A converter.

Harmonic Distortion in Capacitor Array Circuits

Voltage coefficient of capacitance can also lead to harmonic distortion. Consider the precision analog amplifier, attenuator, or D/A converter shown in Fig. 11. Here again the polarity of the capacitor V_{CC} is chosen as shown due to considerations for leakage current and parasitic capacitance. By equating the charges transferred on each capacitor, we get an approximate equation:

$$V_0 = -\frac{C1}{C2} V_{in} + \frac{C1}{C2} \frac{V_{CC}}{2} \left(1 + \frac{C1}{C2}\right) V_{in}^2. \quad (18)$$

Letting $V_{in} = A \sin wt$, the second-harmonic distortion HD2 may be calculated:

$$\text{HD2} = (1 + G)\, V_{CC} \frac{A}{4} \quad (19)$$

in which $G = C1/C2$ is the nominal gain of the circuit and is

greater than zero. As seen from the equation above, distortion increases with gain and voltage amplitude.

IV. Temperature Coefficient of MOS Capacitors

The temperature coefficient of capacitance is defined as

$$T_{CC} = \frac{1}{AC_t}\frac{d}{dT}(AC_t) \tag{20}$$

in which C_t is the total MOS capacitance per unit area and A is the plate area of the device. For this study the MOS capacitor was the same Al-to-Si structure discussed in Section III.

The temperature coefficient of capacitance T_{CC} represents the fractional rate of change of total capacitance per unit temperature. Using the fact that C_s is much greater than C_{ox} for these structures, the value of T_{CC} may be computed and reduced to the following:

$$T_{CC} = \left[\frac{1}{A}\frac{dA}{dT} - \frac{1}{t_{ox}}\frac{dt_{ox}}{dT}\right] + \frac{C_{ox}}{C_s^2}\frac{dC_s}{dT} + \frac{1}{\epsilon_{ox}}\frac{d\epsilon_{ox}}{dT}$$
$$= T_{CC}(th) + T_{CC}(sc) + T_{CC}(\epsilon_{ox}). \tag{21}$$

In this form, T_{CC} is resolved into three components. The first term represents the change in capacitance for a plate area A and dielectric thickness t_{ox} due to thermal expansion. The second term corresponds to the temperature dependence of space charge capacitance. The third term represents the temperature dependence of the dielectric constant of the oxide k_{ox}. In this section, the first two components of T_{CC} will be discussed resulting in an expression which allows experimental evaluation of the third term.

Evaluation of the thermal expansion component $T_{CC}(th)$ requires an analysis of stresses and strains. This requires knowledge of coefficients of linear thermal expansion, values of elastic moduli and Poisson's ratios of the materials [22]. An exact solution has been performed by Thurston [23]; however, a reasonably good approximation can be made which simplifies the analysis and results in less than 5 percent error. The approximation involves neglecting the term $(1/t_{ox})\,dt_{ox}/dT$ and also the assumption that the thickness of the substrate (200 μm) is so large compared with that of the dielectric (0.1 μm) and the top plate (0.6 μm) that the Si expands freely carrying both thin films along with it. Hence for a plate of any geometry:

$$\frac{1}{A}\frac{dA}{dT} = 2L_{si} = 5.6 \text{ ppm/°C (from Appendix B).} \tag{22}$$

The space charge component $T_{CC}(sc)$ requires evaluation of the first derivative of C_s with respect to T which is difficult to evaluate for an arbitrary applied voltage since U_s is an intricate function of T. However the analysis becomes simplified if $U_s = 0$ (the flat-band condition). This is approximately achieved if the applied voltage is held at zero during the temperature excursion. This is valid for heavily doped Si substrates where U_s is small for applied voltages within several volts of flat band. Equation (11) for C_s in Section III may be simplified by using the series expansion of e^{U_s} for small U_s:

$$\frac{e^{U_s} - 1}{(e^{U_s} - 1 - U_s)^{1/2}} = (2)^{1/2}. \tag{23}$$

Therefore, at flat band, the space charge capacitance is

$$C_s(fb) = \left[\frac{\epsilon_{si} N_d q^2}{kT}\right]^{1/2} \tag{24}$$

Differentiating $C_s(fb)$ with respect to T and allowing k_{si} to be a function of T, it follows that

$$T_{CC}(sc) = \frac{C_{ox}}{C_s^2}\frac{dC_s}{dT} = -\frac{C_{ox}k}{2q(\epsilon_{si}N_d kT)^{1/2}} \cdot \left[1 - T\left(\frac{1}{\epsilon_{si}}\frac{d\epsilon_{si}}{dT}\right)\right] \tag{25}$$

which is valid in the vicinity of flat band. The term $(1/\epsilon_{si})\,d\epsilon_{si}/dT$ has been experimentally evaluated in Appendix A and its value was found to be 252 ppm/°C.

Calculated values of $T_{CC}(sc)$ are plotted in Fig. 12 as the curve labeled "0V." Since $T_{CC}(sc)$ is actually negative, the horizontal axis is the absolute value of $T_{CC}(sc)$. As illustrated, for large donor concentration, the value of $T_{CC}(sc)$ becomes small. $T_{CC}(sc)$ is also plotted for different applied voltages (with the Si substrate at ground). These are labeled in Fig. 12. For negative voltages and small N_d, $T_{CC}(sc)$ becomes a strong function of voltage, changes sign when strong surface depletion occurs, and then becomes very large.

It is suggested by Fig. 12 that $T_{CC}(\epsilon_{ox})$ could be evaluated for a device having N_d as large as possible in order to minimize the temperature dependence associated with the space charge relative to that for the oxide dielectric constant. The literature contains little information regarding measured values of $T_{CC}(\epsilon_{ox})$. Furthermore the value appears dependent upon the process associated with the oxide. For this reason, the value of $T_{CC}(\epsilon_{ox})$ is determined by the experimental technique just mentioned. This is described in Appendix B and the value of $T_{CC}(\epsilon_{ox})$ was found to be 21 ppm/°C, which is similar to (although not equal to) a previously reported value of 15 ppm/°C [24] for deposited oxides. Using (B2) and previous assumptions, it is now possible to compute T_{CC} for values of N_d from 10^{18} to 10^{21}/cm^3 for $V = 0$. This was done and the results are plotted in Fig. 13. Measured data points are also plotted along with a dashed line fitted to these points. As seen in this figure, there is generally good agreement between the measured and calculated data. This is better than would have been expected from the nondegenerate MB CV theory. The intersection of measured data and calculated data at $N_d = 1.6 \times 10^{20}$/cm^3 is of course due to the evaluation of $T_{CC}(\epsilon_{ox})$ at that point.

V. Conclusion

A comprehensive technique for designing MOS capacitor arrays has been discussed. This included a method of calculating capacitor ratio errors and subsequent total yield. Data illustrating the sensitivity of ratio matching to capacitor layout, structures, and technology were also presented.

This paper has compared measured and calculated voltage coefficients of MOS capacitors as a function of surface concen-

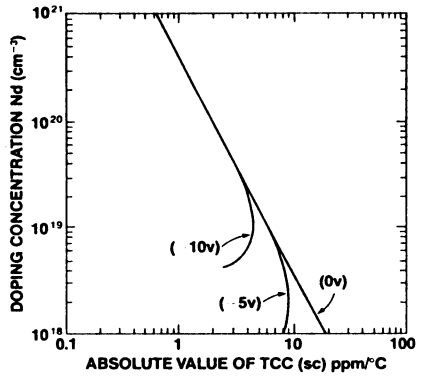


Fig. 12. Calculated curves of surface doping concentration versus absolute value of space charge capacitance temperature coefficient of and Al-to-Si capacitor for different applied voltages (0, -5, and -10 V). For +10 V, the calculated curve is close to that for 0 V.

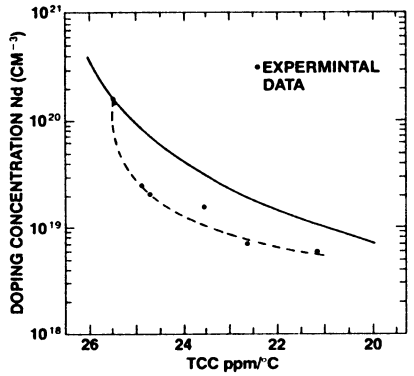


Fig. 13. Calculated and measured curves for surface doping concentration versus temperature coefficient of total capacitance of an Al-to-Si capacitor. The dashed line represents an approximate fit to the measured data.

tration. It has been shown that heavily doped polysilicon accumulates and depletes in a manner similar to that for single crystalline silicon.

It has also been demonstrated that the temperature dependence of space charge capacitance, thermal expansion, and temperature dependence of dielectric constant are the major components of T_{CC}. Of these, it has been found that the dielectric constant term dominates for the case of high substrate doping and low applied voltages.

For poly-to-Si capacitors with heavily doped plates, additional experiments have shown that the T_{CC} of these structures is also predicted by the same model. This is expected since the T_{CC} is due largely to expansion and to the dielectric properties as has been demonstrated.

Appendix A

The graded-junction capacitance per unit area is given by

$$C_j = K_j \epsilon_{si}^{2/3} (V_b - V_j)^{-1/3} \tag{A1}$$

in which K_j is a device constant and V_b is the barrier potential. Hence for a junction capacitor, of total capacitance AC_j, the temperature coefficient is

$$T_{CCj} = \frac{1}{A}\frac{dA}{dT} + \frac{2}{3}\frac{1}{\epsilon_{si}}\frac{d\epsilon_{si}}{dT} - \frac{1}{3}\frac{1}{(V_b - V_j)}\frac{dV_b}{dT}. \tag{A2}$$

The junction potential barrier is given by

$$V_b = \frac{KT}{q} \ln\left[\frac{N_a N_d}{n_i^2}\right] \tag{A3}$$

where N_a is the acceptor impurity concentration. The intrinsic carrier concentration n_i and silicon bandgap energy E_g are approximately given by [21]

$$n_i = 3.8 \times 10^{16} T^{3/2} \exp[-E_g/2kT] \tag{A4}$$

and

$$E_g = 1.2 - (2.83 \times 10^{-4})\, T. \tag{A5}$$

Finally, the desired result is obtained after the appropriate substitutions:

$$\frac{1}{\epsilon_{si}}\frac{d\epsilon_{si}}{dT} = \frac{3}{2}\left[T_{CCj} - 2L_{si} + \frac{1}{3}\frac{1}{(V_b - V_j)}\frac{dV_b}{dT}\right] \tag{A6}$$

Hence, to evaluate $(1/\epsilon_{si})\, d\epsilon_{si}/dT$, it is only necessary to measure the temperature coefficient of the junction capacitor and calculate those terms containing N_a, N_d, and L_{si}. This experiment has been performed and appears in the literature [8]. From that experiment, a value of $T_{CCj} = 230$ ppm/°C was measured. Using this value and data from that experiment we get

$$\frac{1}{\epsilon_{si}}\frac{d\epsilon_{si}}{dT} = 252 \text{ ppm/°C}. \tag{A7}$$

Appendix B

The following experiment was performed on an Al-to-Si capacitor. The surface concentration N_d was measured to be $1.65 \times 10^{20}/\text{cm}^3$ using a spreading resistance technique. C_{ox} for the 1000 Å thermal oxide was 3.4×10^{-4} F/cm^2.

The thermal expansion term $T_{CC}(th)$ becomes

$$2L_{si} = 5.6 \text{ ppm/°C}. \tag{B1}$$

From (25), Section IV, $T_{CC}(sc)$ is calculated to be -1.5 ppm/°C. The temperature coefficient T_{CC} was then measured. For ten devices, the average value of T_{CC} was 25.5 ppm/°C with a standard deviation less than 0.4 ppm/°C. Using these data in (21), Section IV, the effective value of $T_{CC}(\epsilon_{ox})$ becomes

$$T_{CC}(\epsilon_{ox}) = \frac{1}{\epsilon_{ox}}\frac{d\epsilon_{ox}}{dT} = 21 \text{ ppm/°C}. \tag{B2}$$

References

[1] R. Suarez and D. A. Hodges, "All-MOS charge redistribution A/D conversion techniques: Part II," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 379–383, Dec. 1975.

[2] J. L. McCreary and P. R. Gray, "All-MOS charge redistribution A/D conversion techniques: Part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, Dec. 1975.

[3] Y. P. Tsividis *et al.*, "A segmented mu-255 law PCM encoder utilizing NMOS technology," *IEEE J. Solid-State Circuits*, vol. SC-11, pp. 740–753, Dec. 1976.

[4] R. McCharles and D. A. Hodges, "Charge transfer circuits for analog LSI," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 490–497, July 1978.

[5] G. L. Baldwin and J. L. McCreary, "A CMOS digitally-controlled analog attenuator for voice band signals," in *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, AZ, Apr. 1977, p. 519.

[6] B. J. Hosticka, R. W. Brodersen, and P. R. Gray, "MOS sampled

data recursive filters using state variable techniques," in *Proc. Int. Symp. Circuits Syst.*, Phoenix, AZ, Apr. 1977, pp. 525-529.
[7] J. T. Caves *et al.*, "Sampled analog filtering using switched capacitors as resistor equivalents," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 592-599, Dec. 1977.
[8] J. L. McCreary, Ph.D. dissertation, Univ. California, Berkeley, 1975.
[9] R. Suarez, Ph.D. dissertation, Univ. California, Berkeley, 1975.
[10] Y. S. Lee, L. M. Terman, and L. G. Heller, "A two-stage weighted capacitor network for D/A and A/D conversion," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 778-780, Aug. 1979.
[11] J. L. McCreary and D. A. Sealer, "Precision capacitor ratio measurement technique for integrated circuit capacitor arrays," *IEEE Trans. Instrum. Meas.*, vol. IM-28, pp. 11-17, Mar. 1979.
[12] G. Smarandouiu and G. F. Landsberg, "A two-chip CMOS codec," in *Dig. Int. Solid-State Circuits Conf.*, Feb. 1978, pp. 180-181.
[13] K. Chri and M. Callahan, "Integrated PCM codec," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 38-46, Feb. 1979.
[14] R. M. Warner, *IEEE J. Solid-State Circuits*, p. 86, June 1957.
[15] R. Stapper, *IBM J. Res. Dev.*, vol. 20, no. 3, p. 228, 1976.
[16] A. S. Grove, *Physics and Technology of Semiconductor Devices.* New York: Wiley 1967.
[17] R. Seiwatz and M. Green, "Space charge calculations for semiconductors," *J. Appl. Phys.*, vol. 29, p. 1034, July 1958.
[18] A. Many *et al.*, *Semiconductor Surfaces.* Oxford, England: Pergamon, 1962.
[19] J. Snel, "Insulating films on semiconductors," in *Proc. Inst. Phys. Conf.*, Series 50, 1979, p. 119.
[20] *Spreading Resistance Symposium*, Nat. Bureau Standards Spec. Publ. 400-10, Dec. 1974.
[21] S. M. Sze, *Physics of Semiconductor Devices.* New York: Van Nostrand Reinhold, 1972.
[22] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity.* Reading, MA: Addison-Wesley, 1959.
[23] R. N. Thurston, *Physical Acoustics*, vol. 1A. New York: Academic, 1964.
[24] A. B. Grebene, *Analog Integrated Circuit Design.* New York: Van Nostrand Reinhold, 1972, p. 96.

Random Error Effects in Matched MOS Capacitors and Current Sources

JYN-BANG SHYU, GABOR C. TEMES, FELLOW, IEEE, AND FRANCOIS KRUMMENACHER

Abstract—**Explicit formulas are derived using statistical methods for the random errors affecting capacitance and current ratios in MOS integrated circuits. They give the dependence of each error source on the physical dimensions, the standard deviations of the fabrication parameters, the bias conditions, etc. Experimental results, obtained for both matched capacitors and matched current sources using a 3.5 μm NMOS technology, confirmed the theoretical predictions. Random effects represent the ultimate limitation on the achievable accuracy of switched-capacitor filters, D/A converters, and other MOS analog integrated circuits. The results indicate that a 9-bit matching accuracy can be obtained for capacitors and an 8-bit accuracy for MOS current sources without difficulty if the systematic error sources are reduced using proper design and layout techniques.**

Manuscript received June 8, 1984; revised July 30, 1984. This work was supported by the National Science Foundation under Grants ECS 81-05166 and ECS 83-15221 and by the Xerox Corporation and the University of California under MICRO Grant 104.

J.-B. Shyu was with the Department of Electrical Engineering, University of California, Los Angeles, CA 90024. He is now with American Microsystems, Inc., Santa Clara, CA 95051.

G. C. Temes is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90024.

F. Krummenacher is with the Department d'Electricité, Laboratorie d'Electronique Générale (LEG), Ecole Polytechnique Fedérâle De Lausanne, CH-1007 Lausanne, Switzerland.

I. INTRODUCTION

THE metal–oxide–semiconductor (MOS) technology employed in the large-scale integrated (LSI) fabrication of digital circuits has been also used recently to realize

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, pp. 948–955, Dec. 1984.

analog circuits [1]. Unlike for digital integrated circuits, the performance of analog MOS integrated circuits depends heavily upon the element matching accuracy [2], [3]. The key elements are usually the capacitors, but in some applications [4], [5], the matching accuracy of MOS transistors used as current sources is also critical. Therefore, the matching properties of both MOS capacitors and transistors are investigated in this paper. It is a continuation of our earlier work [6] in which we examined the local random errors of MOS capacitors only.

The elements of MOS integrated circuits are inherently subject to errors from two sources. One is the *systematic error*, which affects adjacent elements with identical geometries similarly. It can thus be reduced by proper matching techniques. The other is the *random error*, which differs from element to element, and therefore cannot be corrected by improved matching techniques. It hence represents the ultimate limitation on the achievable accuracy.

A statistical model proposed recently [6] is used to analyze the random errors which are due to *both* the *local* and *global* variations of the linear dimensions and parameters of the circuit elements. These random variations are modeled as random processes, with a zero mean and with stationary characteristics. This is more relevant to the realistic properties of MOS capacitor and transistor matching than the earlier model [6].

Two random capacitance error mechanisms are examined first. One is the random *edge effect* which is due to the local and global random variations of the ideally straight edges of the capacitor, and the other is the random *oxide effect* which is due to the local and global random fluctuations of the oxide thickness and permittivity in the capacitor. Multiple-plate capacitor effects are also considered. Then, four random error mechanisms for the transistor current are examined. First, the *random edge effect* is considered. Second, the *random surface-charge effect* due to both local and global variations of surface-state and ion-implanted charges is analyzed. Third, the *random oxide effect* is studied. Finally, the *random mobility effect* due to both local and global variations of the carrier channel mobility is examined.

In addition to the theoretical predictions, experimental results are presented, based on tests performed on both types of devices, to confirm the theoretical results. A large number of experimental test chips with "on-chip" measurement capability were fabricated for both devices, using 3.5 μm silicon-gate NMOS technology. The experimental results indicate that a 9-bit capacitor matching accuracy and an 8-bit transistor current matching accuracy can be easily achieved. The test results are in good agreement with the theoretical predictions.

II. Random Errors in MOS Capacitors

A. Single-Plate Capacitors

In the fabrication of an integrated circuit pattern, the edges of the lines and devices cannot be exactly located due to the uncertainty in the locations of the particle beams and mask dimensions [7]. The position of an edge is thus affected by a certain amount of "noise" so that an ideally straight line appears wavy. The edge variation includes a local jagged edge variation and a global distorted edge variation. The *local* edge variation is usually caused by the granular nature of the aluminum (or polysilicon) edge which is due to the evaporation onto a heated wafer, and also by the jagged edges of the developed photoresist which are caused, e.g., by light interference patterns [8]. The *global* edge variation may be caused by large-scale edge distortion which is due to the fact that the etchant solution may become saturated with the etched material in some areas of the chip [8] or by quantization effects if digital methods are used in the fabrication process, etc. The *local* jagged edge variation is similar to a wide-band thermal noise, containing high-frequency components in its spectrum. It has a very narrow autocorrelation range in terms of displacement [6]. The *global* distorted edge variation is like a narrow-band flicker noise containing only low-frequency components in its spectrum; it has a rather wide autocorrelation range in terms of displacement. Both of these variations introduce a random deviation of the area, and hence the capacitance, of the device.

In addition to the randomly varying edges, the uncertainty in the oxide thickness t permittivity ϵ also causes random errors in a capacitor. Therefore, e.g., two capacitors which have the same area will not have the same capacitance. As for the edge variations, the oxide variations also include *local oxide variations* and *global oxide variations*.

The local oxide variations may be due to the granularity of polysilicon, surface defects of crystalline silicon, etc. The global variations may be caused by slow variation of surface flatness, wafer warping, variations of oxide growth rate, etc.

In [6], the random capacitance errors due to *local* edge and oxide effects were derived. Following somewhat similar arguments, it is possible to find the corresponding errors due to *global* effects as well. Assuming that all of these effects are independent, the combined relative capacitance error can be expressed as

$$\frac{\Delta C}{C} = \sqrt{K_{le}C^{-3/2} + K_{ge}C^{-1} + K_{lo}C^{-1} + K_{go}} \qquad (1)$$

where K_{le} is the local edge effect factor, K_{ge} is the global edge effect factor, K_{lo} is the local oxide effect factor, and K_{go} is the global oxide effect factor. Somewhat pessimistic estimates can be derived for these factors [11], as follows:

$$K_{le} \simeq 8d_e\sigma_{le}^2\left(\frac{\epsilon}{t}\right)^{3/2}, \quad K_{ge} \simeq \frac{7\epsilon}{t}\sigma_{ge}^2$$

$$K_{lo} \simeq 4d_o^2\frac{\epsilon}{t}\left(\frac{\sigma_{l\epsilon}^2}{\epsilon^2} + \frac{\sigma_{lt}^2}{t^2}\right)$$

$$K_{go} \simeq \frac{\sigma_{g\epsilon}^2}{\epsilon^2} + \frac{\sigma_{gt}^2}{t^2}. \qquad (2)$$

In (2), d_e is the correlation radius and σ_{le} is the standard deviation of the local edge variation [6], [11], ϵ is the

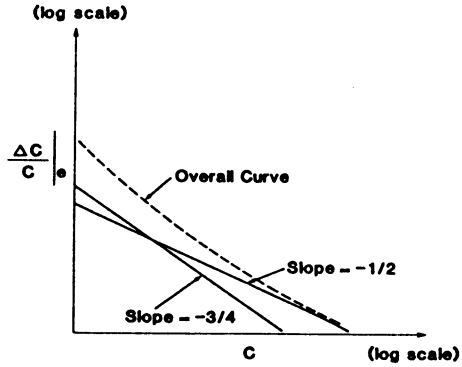


Fig. 1. The relative capacitance error versus capacitance due to local and global edge effects.

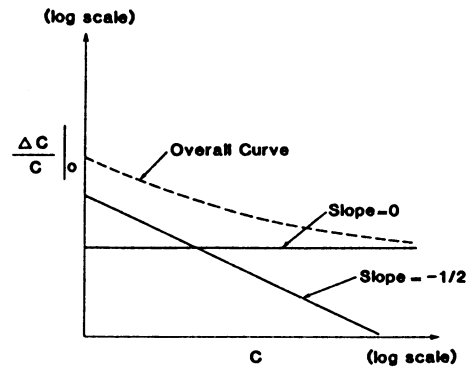


Fig. 2. The relative capacitance error versus capacitance due to local and global oxide effects.

permittivity and t is the thickness of the oxide layer, σ_{ge} is the standard deviation of the global edge variation, $\sigma_{l\epsilon}$ and σ_{lt} are the standard deviations of ϵ and t for local effects, and d_o is the local oxide correlation radius. Finally, $\sigma_{g\epsilon}$ and σ_{gt} are the standard deviations of ϵ and t for global effects.

Figs. 1 and 2 illustrate on a log–log scale the relations between the relative capacitance error and the value of C due to the various terms in (1). For very small C, the local edge effect ($\Delta C/C \propto C^{-3/4}$) dominates; for very large C, the global oxide error dominates. The other two error terms can be appreciable for intermediate values of C. The factors given in (2), and hence the relative importance of the four effects, are technology dependent.

B. Multiple-Plate Capacitors

In order to avoid process-caused systematic errors, MOS capacitors are often realized as the parallel combinations of several smaller "unit capacitors." Therefore, the random errors affecting such multiple-plate capacitors also are discussed. For an n-plate capacitor $C = nC_i$, the random capacitance error due to the local edge effect can be obtained from [6, eq. (25)] as $\sigma_{nC_i} = n^{1/4}\sigma_C = n^{1/2}\sigma_{C_i}$. Similarly, the random capacitance error due to the global edge effect can be found. The result is $\sigma_{nC_i} = n^{1/2}\sigma_C = n\sigma_{C_i}$. Here, the errors of the unit capacitors were assumed to be fully correlated since the *global* edge variation is considered. As far as both of the local and global oxide effects are concerned, there is no physical difference between the random errors of the single- and multiple-plate realizations. If both edge and oxide effects are considered for the multiple-plate capacitor, the overall relative capacitance error is

$$\frac{\sigma_{nC_i}}{C} = \sqrt{\frac{n^{1/2}K_{le}}{C^{3/2}} + \frac{K_{lo}}{C} + \frac{nK_{ge}}{C} + K_{go}} \qquad (3a)$$

or

$$\frac{\sigma_{nC_i}}{nC_i} = \sqrt{\frac{K_{le}}{nC_i^{3/2}} + \frac{K_{lo}}{nC_i} + \frac{K_{ge}}{C_i} + K_{go}}. \qquad (3b)$$

Equation (3a) shows that, *for a given C*, the more partitions we make, the worse the matching becomes. Equation (3b) shows, however, that *for a given unit capacitance* C_i, the matching can be improved by paralleling more unit capacitors since this increases C.

Consider now a realistic example of a single-plate capacitor. Assume that the edge length L of the unit capacitor is 50 μm, $t_o = 700$ Å, and d_e and d_o are about 1 μm. From (1), assuming also that $\sigma_{l\epsilon} = \sigma_{g\epsilon} = 0$, $\sigma_{le} = \sigma_{ge} = 0.2$ μm, and $\sigma_{lt} = \sigma_{gt} = 10$ Å, the various terms in (1) turn out to be $K_{le}^{1/2}C^{-3/4} \simeq 0.16$ percent, $K_{ge}^{1/2}C^{-1/2} \simeq 1.06$ percent, $K_{lo}^{1/2}C^{-1/2} \simeq 0.069$ percent, and $K_{go}^{1/2} \simeq 1.74$ percent. $\Delta C/C$ is thus dominated by the terms containing K_{ge} and K_{go}.

As this example shows, for a typical MOS process, the *global edge and oxide variations represent the crucial limitations on the achievable matching accuracy*. Of course, the relative random errors in MOS capacitors which are made with a different process could be significantly different. However, the parameters in (4) give us an insight into these random error sources and their effects for a typical situation. The dominance of the global effects shows why steps aimed at their elimination (common centroid geometries [8], guard rings, etc.) are particularly important for high-accuracy matching.

In most applications, it is the *ratio* α of two capacitances C_1 and C_2, rather than their individual values, which is of interest. Assuming that $C_1 = nC_i$ and $C_2 = mC_i$ are both multiplate capacitors, it can readily be shown [11] that the relative rms error of α is

$$\frac{\sigma_\alpha}{\alpha} = \sqrt{\left(\frac{\sigma_{nC_i}}{C_1}\right)^2 + \left(\frac{\sigma_{mC_i}}{C_2}\right)^2}$$

$$= \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\left(\frac{K_{le}}{C_i^{3/2}} + \frac{K_{lo}}{C_i}\right) + \frac{K_{ge}}{C_i} + 4K_{go}}. \qquad (4)$$

III. Experimental Results on Capacitance Matching

A large number of experimental capacitor test chips with on-chip capacitance measurement capability was fabricated using the 3.5 μm silicon gate NMOS technology of the Xerox Microelectronics Center. A total of 42 test struc-

TABLE I
LAYOUT STRATEGIES OF MATCHED MOS CAPACITORS IN CAPACITOR TEST CHIPS

	Capacitor Ratio	Unit Capacitor Size (μm^2)	Layout Strategy	Number of Test Structure
Group 1	1:1 2:1 4:1 8:1 16:1	25×25 38×38 50×50 75×75 100×100	Precision unit capacitor layout technique: multiple-plate realization, common-centroid geometry, guard ring, corner cutting, mask alignment tabs.	25
Group 2	1:1 2:1 4:1 8:1	25×25 50×50 100×100	Multiple-plate realization used only.	12
Group 3	1:1 4:1 2:2	25×25 50×50 100×100	Same as Group 1, except that corner cutting was not used.	5

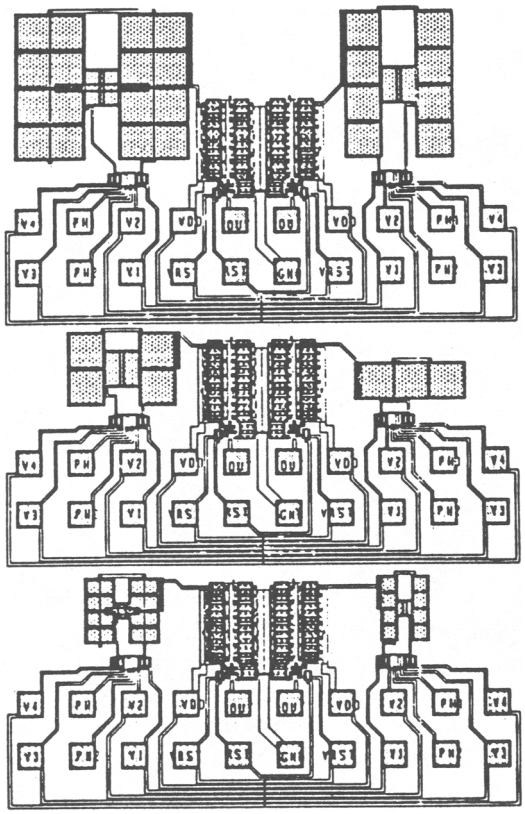


Fig. 3. Parts of the layout of Group 1 in a capacitor test chip.

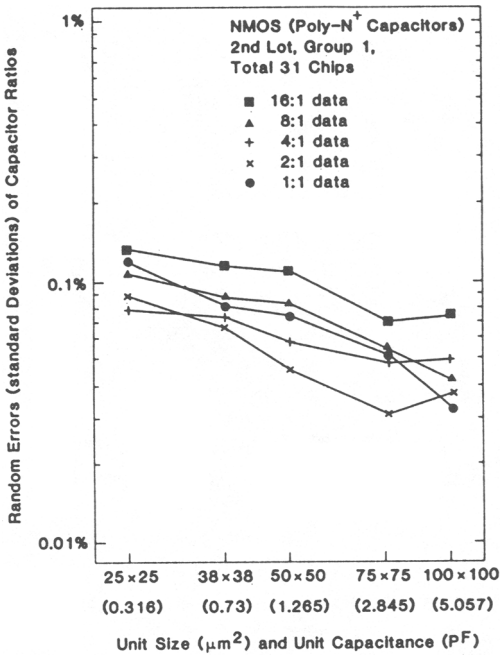


Fig. 4. Random errors of MOS capacitor ratios for Group 1.

tures, which included different capacitor ratios with different unit capacitor sizes, were designed using three different layout strategies and were implemented on 6×6 mm^2 monolithic chips. The properties of the three groups, according to their layout strategies, are summarized in Table I. Part of the layout of a capacitor test chip (Group I capacitors) is shown in Fig. 3. The test circuit was a stray-insensitive buffered switched-capacitor circuit which transforms the ratio of two capacitors to a voltage ratio. After the circuit is zeroed, the variable voltage gives the capacitance ratio [11]. Since the circuit is highly sensitive and immune to parasitics, its accuracy is much better than 0.1 percent. The experimental test data are plotted in two charts. One (Fig. 4) shows the random errors for Group 1 of the test structures, and the other (Fig. 5) shows the random errors for Groups 2 and 3. Each data point gives the measured standard deviation of the capacitance ratio for pairs of capacitors, with nominal ratios 16:1, 8:1, etc.

The conclusions which can be drawn from the experimental results are as follows.

1) Since the slopes of the relative random error versus C plots on a log–log scale are close to $-1/4$ and the capacitance ratio errors are slowly varying functions of the capacitance ratios, the dominant random error sources for this process are clearly the global edge effect and the global oxide effects. This means that in (10b), K_{ge} and K_{go} are the dominant factors. The theoretical predictions made in connection with (12) have thus been verified.

2) Group 1 provided major improvements over Group 2 for large capacitor ratios and small unit capacitor sizes. Hence, a more careful layout strategy indeed gives a better capacitor matching.

3) The optimum geometries for this process are between 50×50 μm^2 and 75×75 μm^2 because the crossovers between systematic edge and oxide effects for all three groups of test data (not shown) are within this range. For area

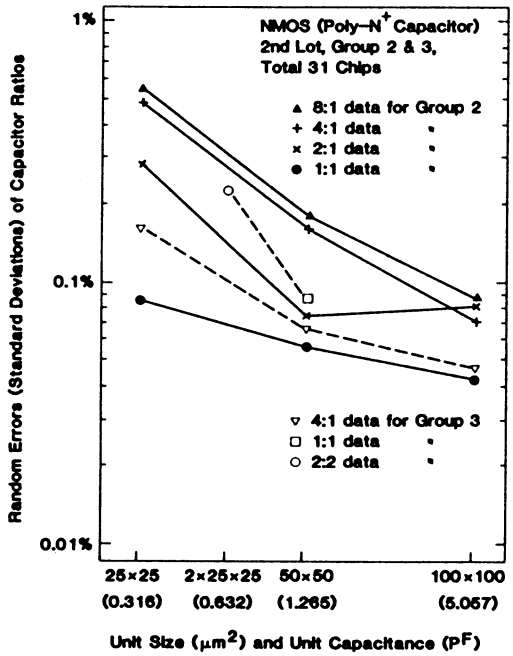


Fig. 5. Random errors of capacitor ratios for Groups 2 and 3.

efficiency, the 50×50 μm^2 unit capacitor size can be used to achieve a 9-bit capacitor ratio accuracy since the random errors at this geometry are around 0.1 percent. This accuracy is adequate for most precision switched-capacitor circuits or A/D and D/A converters.

IV. Random Errors in MOS Transistors

Consider next a $W \times L$ μm^2 polysilicon gate n-channel MOS transistor where the width W and the length L are so large that first-order models are valid. The simple square-law drain current relation of a MOSFET operating in saturation is then [9]

$$I_o = \frac{\overline{\mu C_{ox}}}{2} \frac{W}{L} (V_{GS} - \bar{V}_T)^2 \tag{5}$$

where I_o is the nominal drain current, $\bar{\mu}$ is the nominal effective electron mobility, $\bar{C}_{ox} \triangleq \epsilon / \bar{t}$ is the nominal gate capacitance per unit area, W is the nominal channel width, and L is the nominal channel length, while $\bar{V}_T$ is the nominal threshold voltage (including the body effect).[1]

A. Random Errors Due to Edge Effects

Because the gate area of a MOSFET is determined by several different masks (i.e., the poly mask determines the channel length, while the diffusion mask determines the channel width), we should analyze the edge effects due to random length variation and random width variation independently. As for MOS capacitors, the edge effects also include the local and global random variations. First we consider random length variations only. A MOSFET can then be modeled as a parallel combination of many differential-width "strip transistors" with equal widths dW and slightly different lengths $L + l(z)$ as shown in Fig. 6. Here, $l(z)$ represents the random length variation and is assumed to be a zero-mean stationary random process. On the basis of this model, the relative current error, due to *both* local and global length variations, can be found as

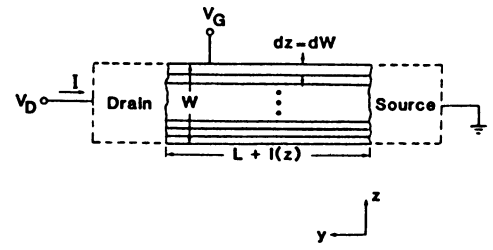


Fig. 6. Top view of an MOS transistor with random length variations.

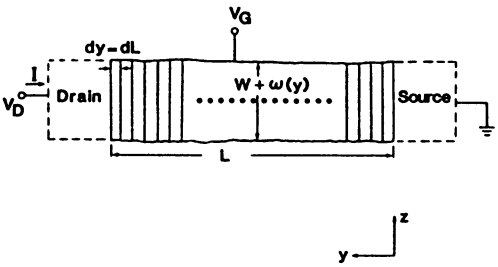


Fig. 7. Top view of an MOS transistor with random width variations.

$$\frac{\Delta I}{I_o} = \frac{1}{L} \sqrt{B_{gL} + \frac{B_{lL}}{W}} . \tag{6}$$

Here, $B_{gL} = \sigma_{gl}^2$ where σ_{gL} is the standard deviation of L due to global variations and $B_{lL} = 2 d_e \sigma_{lL}^2$ where d_e is the correlation radius of the local length variation and σ_{lL} is the standard deviation of L due to local variations.

Next, the channel width variations are considered. The transistor can then physically be modeled as a series combination of many wide "strip transistors" with equal differential lengths dL and slightly different widths $W + \omega(y)$ as shown in Fig. 7. On the basis of this model [9], the relative current error due to both local and global width variations can be obtained as

$$\frac{\Delta I}{I_o} = \frac{1}{W} \sqrt{B_{gW} + \frac{B_{lW}}{L}} \tag{7}$$

where $B_{gW} \triangleq \sigma_{gW}^2$ is the global-width-effect factor and $B_{lW} \triangleq 2 d_e \sigma_{lW}^2$ is the local-width-effect factor.

If *both* length and width random variations are taken into account, the relative current error due to edge effect is found

$$\left. \frac{\Delta I}{I_o} \right|_e = \sqrt{\left(\frac{B_{gL}}{L^2} + \frac{B_{gW}}{W^2} \right) + \left(\frac{B_{lL}}{L} + \frac{B_{lW}}{W} \right) \frac{1}{LW}} . \tag{8}$$

Equation (8) shows that for a given current I_o, the current matching can be improved by increasing the device

[1] Note that the random variations of substrate doping can also have significant effects on the matching, e.g., for PMOS transistors in a CMOS process. Our simple drain–current relation is unsuitable for an analysis of this effect, and a more elaborate device model is required.

size. The matching is not affected by the body effect if the edge effect is the dominant factor.

B. Random Errors Due to Surface-State-Charge and Ion-Implanted-Charge Effects

In a typical MOS process, ion-implanted charges are usually employed to shift the threshold voltage [10]; also, some inherent interface charges always exist at and inside the interface of the silicon oxide and substrate. The surface-state charge density Q_{ss} and the ion-implanted charge density Q_{ii} are randomly distributed over the whole channel area. Hence, the deviations of these charge densities from their mean values cause a deviation on the nominal threshold voltage $\bar{V}_T$ and thus also a current error. The nominal threshold voltage in the presence of body effect is given by [9], [14]

$$\bar{V}_T = \phi_{MS} + 2|\phi_F| + \frac{\sqrt{2q\epsilon_{Si}N_a(2|\phi_F| + V_{BS})}}{\bar{C}_{ox}} - \frac{\bar{Q}_{ss}}{\bar{C}_{ox}} + \frac{\bar{Q}_{ii}}{\bar{C}_{ox}} \tag{9}$$

where the meaning of the symbols is as follows.

ϕ_{MS}: Potential difference of the work functions of the gate and substrate.

ϕ_F: Built-in potential in the bulk of the substrate.

$Q_D \triangleq \sqrt{2q\epsilon_{Si}N_a(2|\phi_F| + |V_{BS}|)}$: Space-charge density per unit area.

Q_{ss}: Surface-state charge density per unit area.

Q_{ii}: Ion-implanted charge density per unit area.

$C_{ox} \triangleq \epsilon/t$: Gate capacitance per unit area.

The bar above a symbol indicates the nominal value. Hence, the nominal drain current, from (5), becomes

$$I_o = \beta\left[K_1 + \frac{1}{\bar{C}_{ox}}(\bar{Q}_{ss} - \bar{Q}_{ii})\right]^2 \tag{10}$$

where $\beta \triangleq (\bar{\mu}\bar{C}_{ox}/2)(W/L)$ is the gain factor and $K_1 \triangleq V_{GS} - \phi_{MS} - 2|\phi_F| - (Q_D/\bar{C}_{ox})$ is a constant.

We then describe the random variations of Q_{ss} and Q_{ii} as functions of y and z as follows:

$$\begin{aligned} Q_{ss}(y,z) &= \bar{Q}_{ss} + q_{ss}(y,z) \\ Q_{ii}(y,z) &= \bar{Q}_{ii} + q_{ii}(y,z) \end{aligned} \tag{11}$$

where $q_{ss}(y,z)$ and $q_{ii}(y,z)$ are assumed to be zero-mean stationary two-dimensional random processes. We can express $\bar{Q}_{ii}$ and $\bar{Q}_{ss}$ in (10) using (11) and then integrate it over the whole channel area. After some fairly complicated calculations [11], the relative current error due to *both* local and global variations of Q_{ss} and Q_{ii} is found to be

$$\left.\frac{\Delta I}{I_o}\right|_q = \frac{1}{\bar{C}_{ox}(V_{GS} - \bar{V}_T)}\sqrt{(B_{gss} + B_{gii}) + \frac{1}{WL}(B_{lss} + B_{lii})} \tag{12}$$

where $B_{gss} \triangleq 4\sigma_{gq}^2$ and $B_{gii} \triangleq 4\sigma_{gq_{ii}}^2$ represent the global surface-state charge and ion-implanted charge-effect factors, while $B_{1ss} \triangleq 16d_q^2\sigma_{1q_{ss}}^2$ and $B_{1ii} \triangleq 16d_q^2\sigma_{1q_{ii}}^2$ are the local surface-state charge and ion-implanted charge effect factors, respectively. The subscript "q" here and in the following means the random surface-charge effect. Equation (12) shows that a larger $\bar{C}_{ox}$ and V_{GS} and smaller body effect in V_T results in a better current matching for a given transistor size, but at the cost of a poorer dynamic range due to the larger V_{GS}.

A practical application of our results is to the random input offset voltage of a differential input amplifier. This is found to be [12]

$$\Delta V_{GS} = \frac{\Delta I}{g_{mo}} \tag{13}$$

where $g_{mo} \triangleq \bar{\mu}\bar{C}_{ox}(W/L)(V_{GS} - \bar{V}_T)$ is the nominal transconductance, while ΔI represents the random current error. Hence, from (12) and (13), the random input offset voltage due to random surface-charge effects is

$$\Delta V_{GS}|_q = \frac{1}{2\bar{C}_{ox}}\sqrt{(B_{gss} + B_{gii}) + \frac{1}{WL}(B_{lss} + B_{lii})}\,. \tag{14}$$

C. Random Errors Due to Oxide Effects

Consider now the random fluctuations of the oxide thickness t. The gain factor as well as the threshold voltage in (5) are affected by the oxide thickness variation, and hence so is the drain current of the device. From (5) and (9), the nominal drain current can be expressed as

$$I_o = \frac{K_2}{\bar{t}}(K_3 - \bar{t}K_4)^2 \tag{15}$$

where $K_2 \triangleq \beta\bar{t}$ and $K_3 \triangleq V_{GS} - \phi_{MS} - 2|\phi_F|$ and $K_4 \triangleq (Q_D - \bar{Q}_{ss} + \bar{Q}_{ii})/\epsilon$. Next, t is considered to be composed of a mean $\bar{t}$ and a random variation $\Delta t(y,z)$ across the gate area, i.e., $t(y,z) = \bar{t} + \Delta t(y,z)$.

Performing a statistical analysis, the relative current error due to both local and global variations of oxide thickness is found to be

$$\left.\frac{\Delta I}{I_o}\right|_o = \frac{1}{\bar{t}}\left[1 + \frac{2Q_T}{\bar{C}_{ox}(V_{GS} - \bar{V}_T)}\right]\sqrt{B_{go} + \frac{B_{lo}}{WL}} \tag{16}$$

where $Q_T \triangleq Q_D - \bar{Q}_{ss} + \bar{Q}_{ii}$ represents the surface charge, and $B_{go} \triangleq \sigma_{g\Delta t}^2$ and $B_{lo} = 4d_o^2\sigma_{l\Delta t}^2$ are the global oxide and local oxide-effect factors, respectively. (The subscript "o" refers to the oxide effect.) An interesting aspect of (16) is that the first term in the square brackets represents the oxide effect acting through the gain factor of the device, while the second term represents the oxide effects due to the threshold voltage of the device. If the first term is dominant because, e.g., a large $V_{GS} - \bar{V}_T$ is used, the matching accuracy will be improved by increasing $\bar{t}$; however, it should be recalled that as (12) showed, the relative error due to random surface-charge effects increases with increasing $\bar{t}$. Therefore, the relation between $\bar{t}$ and the matching accuracy due to both oxide and surface-charge effects is not obvious; it depends upon the process, i.e., the actual values of B_{gss}, $B_{go}, \cdots$, etc.

Our results can be applied again to a differential input amplifier. The random input offset voltage variation due to oxide effects acting through the threshold voltage is found to be

$$\Delta V_{GS}|_o = \frac{Q_T}{\epsilon}\sqrt{B_{go} + \frac{B_{lo}}{WL}}. \tag{17}$$

If both oxide and surface-charge effects are taken into account using (14) and (17), then the random input offset voltage due mainly to the threshold voltage mismatch can be expressed as

$$\Delta V_{GS}|_{V_T} = \sqrt{\frac{B_{gss}+B_{gii}}{4\bar{C}_{ox}^2} + \frac{Q_T^2 B_{go}}{\epsilon^2} + \frac{1}{WL}\left(\frac{B_{lss}+B_{lii}}{4\bar{C}_{ox}^2} + \frac{Q_T^2 B_{lo}}{\epsilon^2}\right)}. \tag{18}$$

Hence, ΔV_{GS} is *not* a function of $V_{GS} - \bar{V}_T$. This is in agreement with the argument in [12].

D. Random Errors Due to Channel Mobility Effects

Due to the random impurity scattering and lattice scattering mechanisms [13], the effective channel mobility of the carriers varies randomly over the entire channel area of the device. The effective channel mobility μ can be regarded as composed of a mean value $\bar{\mu}$ and a random variation $\Delta\mu(y, z)$, i.e., $\mu(y, z) \triangleq \bar{\mu} + \Delta\mu(y, z)$. Hence, the deviation of I can be found to be

$$\Delta I = K_9 \int_o^W \int_o^L \Delta\mu(y, z)\, dy\, dz. \tag{19}$$

Performing a statistical analysis [11], the relative current error due to both local and global variations of μ turns out to be

$$\left.\frac{\Delta I}{I_o}\right|_\mu = \frac{1}{\bar{\mu}}\sqrt{B_{g\mu} + \frac{B_{l\mu}}{WL}} \tag{20}$$

where $B_{g\mu} \triangleq \sigma_{g\Delta\mu}^2$ and $B_{l\mu} = 4d_\mu^2\sigma_{l\Delta\mu}^2$ represent the global and local channel mobility-effect factors, respectively.

In conclusion, combining the relative current errors due to all four random effects, the overall relative current error from (8), (12), (16), and (20) can be expressed as

$$\frac{\Delta I}{I_o} = \left\{\left[\left(\frac{B_{gL}}{L^2} + \frac{B_{gW}}{W^2}\right) + \frac{B_{gss}+B_{gii}}{\bar{C}_{ox}^2(V_{GS}-\bar{V}_T)^2} + \frac{B_{go}}{\bar{t}^2}\left(1 + \frac{2Q_T}{\bar{C}_{ox}(V_{GS}-\bar{V}_T)}\right)^2 + \frac{B_{g\mu}}{\bar{\mu}^2}\right] + \frac{1}{WL}\left[\left(\frac{B_{lL}}{L} + \frac{B_{lW}}{W}\right) + \frac{B_{lss}+B_{lii}}{\bar{C}_{ox}^2(V_{GS}-V_T)^2} + \frac{B_{lo}}{\bar{t}^2}\left(1 + \frac{2Q_T}{\bar{C}_{ox}(V_{GS}-\bar{V}_T)}\right)^2 + \frac{B_{l\mu}}{\bar{\mu}^2}\right]\right\}^{1/2}. \tag{21}$$

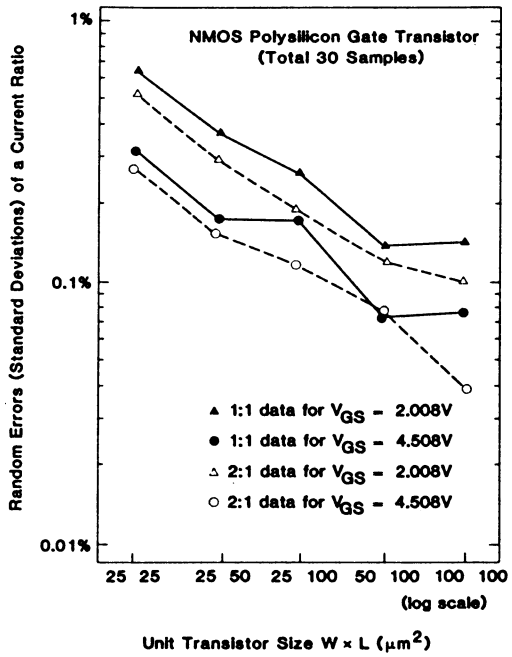


Fig. 8. Random errors of the current matching.

Here, the terms in the first pair of square brackets in (21) represent the *global* random effects, while those in the second pair of square brackets represent the *local* random effects. Since WL enters in the denominators of all terms in the second brackets, all local random effects can be improved by using a larger transistor size.

V. Experimental Results on Transistor Current Matching

An experiment aimed at verifying the MOSFET's predicted current matching properties was carried out using the same 3.5 μm NMOS technology as for capacitor matching at the Xerox Microelectronics Center. Hence, the results obtained previously for capacitance matching regarding edge effects, oxide effects, and optimum geometry could be utilized in the analysis of the transistor current matching. A total of 300 test transistors with five unit transistor sizes ($W \times L = 25\times25$, 25×50, 25×100, 50×100, and 100×100 μm^2) were fabricated and tested using off-chip current meters on five test chips. These large transistor sizes guaranteed that the short-channel and narrow-width effects were not significant. The nominal threshold voltage for this technology was $\bar{V}_T = 0.9$ V. Two different gate-to-source voltages ($V_{GS} = 2$ and 4.5 V) were used in order to be able to distinguish the current errors due to the threshold voltage effects. The nominal current levels for $V_{GS} = 2$ V were 5.06, 10.13, and 20.25 μA, while for $V_{GS} = 4.5$ V, they were 53.69, 107.39, and 214.79 μA. Matched pairs of MOSFET's were designed for two current ratios (1:1 and 2:1) using careful layout techniques. The experimental data obtained are shown in Fig. 8.

The following conclusion can be drawn from our experimental results.

1) The global random edge variations were the dominant effects for smaller unit transistor sizes (25×25 to 25×100

μm^2), while the global random oxide effects dominated for larger unit transistor sizes (50×100 to 100×100 μm^2). Furthermore, the *global* random surface-charge effects were also important over the entire range of unit transistor sizes. This follows because the relative current errors for $V_{GS} = 2$ V were always larger than for $V_{GS} = 4.5$ V. Also, the *local* random surface-charge effects were influential for the smaller unit transistor sizes. This follows because the improvement of random errors in (21) due to a larger V_{GS} is seen to be reduced by increasing the unit transistor size. This means that the random surface-charge effects are functions of W and L. From (21), we realize that the *local* random surface-charge variations also appear to have a large effect for smaller unit transistor sizes.

2) The optimum geometry for this process is about 50×100 μm^2 ($W \times L$). To reduce power dissipation, the smaller gate-to-source voltage (i.e., $V_{GS} = 2$ V) can be used to obtain an 8-bit current ratio accuracy. This is adequate for many precision applications.

3) As our data show, a careful layout technique incorporating such features as guard rings, common-centroid geometry, etc., is important for good current matching.

VI. Conclusions

Theoretical formulas have been derived for random effects, both local (granular) and global (large-scale), affecting the ratios of matching capacitances and currents. For capacitors, the effects considered were edge variations and oxide variations. Both single-plate and multiple-plate capacitances were analyzed. For MOSFET current sources, the error sources considered included edge variations, surface-state and ion-implanted charge variations, oxide-variation effects, and carrier mobility variations.

Using a 3.5 μm NMOS technology, a large number of matched capacitors and MOSFET current sources were fabricated and tested using on-chip measurement circuits. The results confirmed the general conclusions drawn from the theoretical relations, and also gave the optimal dimensions and bias conditions for the process.

The importance of random effects is that they represent the ultimate limitation on the achievable accuracy of MOS analog circuits in the absence of systematic errors. Our results indicate that random errors permit a 9-bit accuracy for capacitance matching and an 8-bit accuracy for MOSFET current-source matching without excessive area requirements.

Acknowledgment

The authors are grateful to S. Law and S. Eckert of the Xerox Corporation and to Prof. K. Yao of UCLA for useful discussions, and to the Xerox Microelectronics Center for fabricating the test circuits described.

References

[1] P. R. Gray, D. A. Hodges, and R. W. Brodersen, *Analog MOS Integrated Circuits*. New York: IEEE Press, 1980.
[2] R. Gregorian, K. W. Martin, and G. C. Temes, "Switched-capacitor circuit design," *Proc. IEEE*, vol. 71, pp. 941–966, Aug. 1983.
[3] J. L. McCreary and P. R. Gray, "All-MOS charge redistribution analog-to-digital conversion techniques—Part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, Dec. 1975.
[4] H. U. Post and K. Waldschmidt, "A high-speed NMOS A/D converter with a current source array," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 295–300, June 1980.
[5] S. Kelly and D. Ulmer, "A single-chip CMOS PCM codec," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 54–59, Feb. 1979.
[6] J.-B. Shyu, G. C. Temes, and K. Yao, "Random errors in MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1070–1076, Dec. 1982.
[7] J. T. Wallmark, "A statistical model for determining the minimum size in integrated circuits," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 135–142, Feb. 1979.
[8] J. L. McCreary, "Successive approximation analog-to-digital conversion in MOS integrated circuits," Ph.D. dissertation, Univ. California, Berkeley, 1975.
[9] W. N. Carr and J. P. Mize, *MOS/LSI Design and Application*. New York: McGraw-Hill, 1972, p. 46.
[10] B. S. Song and P. R. Gray, "Threshold-voltage temperature drift in ion-implanted MOS transistors," *IEEE J. Solid State Circuits*, vol. SC-17, pp. 291–298, Apr. 1982.
[11] J.-B. Shyu, "The obtainable accuracy of analog MOS integrated circuit elements," Ph.D. dissertation, Univ. California, Los Angeles, 1984.
[12] P. R. Gray and R. G. Meyer, "MOS operational amplifier design—A tutorial overview," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 969–982, Dec. 1982.
[13] A. S. Grove, *Physics and Technology of Semiconductor Devices*. New York: Wiley, 1967.
[14] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*. New York: Wiley, 1977.
[15] J. L. McCreary, "Matching properties, and voltage and temperature dependence of MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 608–616, Dec. 1981.

Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design

KADABA R. LAKSHMIKUMAR, MEMBER, IEEE, ROBERT A. HADAWAY, AND MILES A. COPELAND, SENIOR MEMBER, IEEE

Abstract—This paper is concerned with the design of precision MOS analog circuits. Section II of the paper discusses the characterization and modeling of mismatch in MOS transistors. A characterization methodology is presented that accurately predicts the mismatch in drain current over a wide operating range using a minimum set of measured data. The physical causes of mismatch are discussed in detail for both p- and n-channel devices. Statistical methods are used to develop analytical models that relate the mismatch to the device dimensions. It is shown that these models are valid for small-geometry devices also. Extensive experimental data from a 3-μm CMOS process are used to verify these models.

Section III of the paper demonstrates the application of the transistor matching studies to the design of a high-performance digital-to-analog converter (DAC). A circuit design methodology is presented that highlights the close interaction between the circuit yield and the matching accuracy of devices. It has been possible to achieve a circuit yield of greater than 97 percent as a result of the knowledge generated regarding the matching behavior of transistors and due to the systematic design approach.

I. Introduction

THE DESIGN of precision analog circuits requires a thorough understanding of the matching behavior of components available in any given technology. In MOS technology, capacitors are being widely used for designing precision analog circuits such as data converters [1] and switched-capacitor filters [2], [3] because of their excellent matching characteristics [4]. The matching behavior of MOS capacitors has been discussed in detail [5]–[7]. However, all precision analog circuits cannot be designed using capacitors alone. For applications such as high-speed data conversion, capacitive techniques tend to be too slow. Further a digital VLSI process may not offer linear capacitors. These factors motivated us to study the matching behavior of MOS transistors.

Manuscript received April 21, 1986; revised July 14, 1986. The Carleton University part of this work was supported by the Canadian Commonwealth Scholarship and Fellowship Committee and the Natural Sciences and Engineering Research Council of Canada.

K. R. Lakshmikumar was with the Department of Electronics, Carleton University, Ottawa, Ont. K1S 5B6, Canada. He is now with AT&T Bell Laboratories, Murray Hill, NJ 07974.

R. A. Hadaway is with Northern Telecom Electronics Ltd., Ottawa, Ont., Canada.

M. A. Copeland is with the Department of Electronics, Carleton University, Ottawa, Ont. K1S 5B6, Canada.

IEEE Log Number 8610606.

Section II of this paper discusses the characterization and modeling of mismatch in MOS transistors. The interest in such a study is evidenced by recent publications [7], [8]. In [8] experimental results of matching of MOS current mirrors are discussed without any reference to the physical causes of mismatch. The work reported in [7] attempts to break down the causes of mismatch but the experimental results are limited to large-area n-channel devices only. The work reported here is aimed at providing a more comprehensive understanding of the causes of mismatch in both p- and n-channel devices of large and small geometry. As the circuit designer has freedom to choose only the device dimensions, analytical models have been developed that relate the electrical mismatch to the dimensions. Extensive data to verify these models are obtained from a 5-V, 3 μm, p-well CMOS process that is in use at Northern Telecom Electronics Limited, Ottawa, Canada.

Section III of the paper demonstrates the application of the knowledge of matching behavior for the design of a high-speed digital-to-analog converter (DAC). Of late, the area of high-speed data converters in MOS technology is gaining importance (for example, [23] and [24]). However, these designs do not indicate the circuit yield obtainable for a given resolution, or the possibility of extension of these techniques to higher resolution converters. Therefore a circuit design methodology is presented here that relates the achievable linearity and yield to the matching accuracy of the components. A high-performance DAC with a circuit yield of greater than 97 percent has been realized without using any post-process trimming and yet occupying a small chip area using this design methodology [9].

II. Transistor Matching Studies

In general, there are two variations to consider in an integrated circuit process. *Global variation* accounts for the total variation in the value of a component over a wafer or a batch. *Local variation* or *mismatch* reflects the variation in a component value with reference to an adjacent component on the same chip. As the design of precision analog circuits is based on component ratios rather than their

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-21, no. 6, pp. 1057–1066, Dec. 1986.

absolute values, we have concentrated our study on the mismatch behavior.

The characterization of mismatch in MOS transistors is more complex than that in the case of capacitors. The drain current matching not only depends on the device dimensions but also on the operating point. In Section II-A, a characterization methodology is developed that accurately determines the mismatch in drain current over a wide operating region, using a minimum set of measurement data. The physical causes of mismatch are discussed in Section II-B, and analytical expressions to relate the mismatch to device dimensions are developed. We call these quantitative relationships *mismatch models*.

A. Characterization Methodology

Our aim is to predict the mismatch in the drain current over a wide range of operating conditions using a minimum set of measured data, and simultaneously to throw light on the detailed causes of mismatch. This problem can be best approached by measuring the mismatch in various parameters of a suitable circuit model [10]. The model chosen should be such that it gives an adequate description of the electrical behavior of the device, and at the same time should have readily measurable parameters that are amenable to statistical description. As an elaborate circuit model may greatly exceed the accuracy of measurable data or may hamper the extraction of statistically significant model parameters, we chose the simple square-law model. The current–voltage relationship in the triode region is given by

$$I = K(V_{GS} - V_T - V_{DS}/2)V_{DS} \tag{1}$$

where I is the drain current, K is the conductance constant, V_T is the threshold voltage, and V_{DS} is the drain-to-source voltage. The statistically significant parameters of this model are V_T and K. The mismatch in V_T accounts for the variations in the different charge quantities, and in the gate oxide capacitance per unit area. The variations in the dimensions, channel mobility, and gate oxide capacitance per unit area are measured as the mismatch in K. As both V_T and K are dependent on the gate oxide capacitance per unit area, we need to measure the correlation between the mismatches in V_T and K also.

The square-law model (1) is not an accurate description of the current–voltage relationship. It should be noted that we are only looking for local variations and are not so much concerned about the estimation of the absolute value of the parameters. Therefore any small model error would cancel out to a first order while estimating the mismatch, and hence the square-law model should suffice for our application. Several assumptions are made while deriving this one-dimensional model. As some of these are not strictly applicable, a further discussion to justify the use of this model is in order.

The gradual channel approximation and the assumption that the substrate is uniformly doped do not necessarily hold for small-geometry devices. An accurate analysis calls for a two-dimensional solution of Poisson's equation. However, in order to develop analytical expressions for the mismatch behavior, we use a one-dimensional circuit model and apply appropriate corrections to account for the effects of dimensional dependence on threshold voltage and nonuniform doping of the substrate. Also we have assumed a simplified picture of the oxide–silicon interface, i.e., that oxide fixed charge and interface trap charges are considered to be smeared-out uniform charge sheets. In fact the oxide fixed charge and charged interface traps are localized, and not sheets [11]. The accurate calculation of the surface potential would then be a two-dimensional problem. We are simplifying the problem by estimating the aggregate of the localized nonuniformity over the entire area of the channel by using the simplified one-dimensional circuit model. With these approximations, we develop analytical expressions for the mismatch behavior that compare remarkably well with experimental data.

Generally, MOS transistors will be operating in the saturation region in analog circuits. Therefore we should relate the measured mismatches in V_T and K to the saturation region, where the drain current is given by

$$I = \frac{K}{2}(V_{GS} - V_T)^2. \tag{2}$$

Then the variance in the drain current may be written as

$$\frac{\sigma_I^2}{\bar{I}^2} = \frac{\sigma_K^2}{\bar{K}^2} + 4\frac{\sigma_{VT}^2}{(V_{GS} - \bar{V}_T)^2} - 4r\frac{\sigma_{VT}}{V_{GS} - \bar{V}_T} \cdot \frac{\sigma_K}{\bar{K}} \tag{3}$$

following the derivation in [12] concerning the variance of a function of two random variables. Here r is the correlation coefficient between the mismatches in V_T and K, $\bar{I}$ is the expected value of the random variable I, σ_I is the standard deviation of I, and so on. Thus the mismatch in drain current at any operating point may be estimated if σ_K, σ_{VT}, and r are known.

Experimentally, V_T and K are determined by measuring the drain current versus gate voltage for a small value of V_{DS}. The maximum slope of the I versus V_{GS} curve provides the value of K. V_T is obtained from the intercept of the maximum slope on to the V_{GS} axis. If ΔK_i is the difference in the value of K for the ith matched pair of devices, the standard deviation of K is given by

$$\sigma_K = \left[\frac{1}{N-1}\left\{\sum_{i=1}^{N}(\Delta K_i)^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\Delta K_i\right)^2\right\}\right]^{1/2} \tag{4}$$

where N is the number of matched pairs measured on each wafer. The second term in (4) is close to zero as the matched pairs are laid out in such a way as to minimize systematic mismatch. σ_{VT} is also computed in a similar way.

B. Factors Causing Mismatch

1. Threshold Voltage Mismatch: The threshold voltage of a transistor may be expressed as

$$V_T = \phi_{MS} + 2\phi_B + \frac{Q_B}{C} - \frac{Q_f}{C} + \frac{qD_I}{C} \tag{5}$$

where ϕ_{MS} is the gate–semiconductor work function difference, ϕ_B is the Fermi potential in the bulk, Q_B is the depletion charge density, Q_f is the fixed oxide charge density, D_I is the threshold adjust implant dose, and C is the gate oxide capacitance per unit area. The last term in (5) accounts for the threshold adjust implant where the implanted ions are assumed to have a delta function profile at the silicon–silicon dioxide interface [14]. The standard deviation of V_T may be determined if we can find the standard deviations of the various terms on the right-hand side of (5). The Fermi potential ϕ_B has a logarithmic dependence on the substrate doping, and ϕ_{MS} has a similar dependence on the doping in the substrate and in the polysilicon gate. Hence these terms may be regarded as constants not contributing to any mismatch.

Next we consider oxide fixed charge which is reported to have a Poisson distribution [11, p. 242]. Then its variance is given by

$$\sigma_{Qf}^2 = \frac{q\bar{Q}_f}{\bar{L}\bar{W}} \tag{6}$$

where L is the effective length and W is the effective width of the channel.

The depletion charge per unit area Q_B is also a random variable dependent on the distribution of the dopant atoms. This is an important difference between the treatment given here and the one reported in [7], where Q_B is treated as a constant. In fact, it is reported in [11, p. 237] that the dopant ions are nonuniformly distributed in MOS devices. No theoretical treatment of fluctuations in dopant ion density is available. However, we shall show that the physical conditions in the substrate favor a Poisson distribution [13]. The number of atoms per unit volume in silicon is $5\times10^{22}\ \text{cm}^{-3}$. Only a very small fraction of these sites are occupied by the dopant atoms. The number of dopant atoms in nonoverlapping volumes is independent. Further, for domains of sufficiently small volume, the probability of finding exactly one dopant atom in a domain is proportional to the volume, and the probability of finding more than one atom is negligible. Hence the dopant ions may be considered to have Poisson distribution. Then the variance in Q_B may be shown to be [15]

$$\frac{\sigma_{QB}^2}{\bar{Q}_B^2} = \frac{1}{4\bar{L}\bar{W}\bar{W}_d\bar{N}_A} \tag{7}$$

where W_d is the depletion layer width and N_A is the substrate doping.

Similarly, assuming the implanted ions to follow a Poisson distribution, the variance of D_I is given by

$$\sigma_{DI}^2 = \frac{\bar{D}_I}{\bar{L}\bar{W}}. \tag{8}$$

The variance in C may be determined by estimating the variances in oxide thickness and permittivity [7]. It can be shown that [15]

$$\frac{\sigma_C^2}{\bar{C}^2} = \frac{1}{\bar{L}\bar{W}} A_{ox} \tag{9}$$

where A_{ox} is a parameter to be determined from measurements.

The random variables Q_f, Q_B, D_I, and C are all independent. Hence the variance in V_T may be written using (5) as

$$\sigma_{VT}^2 = \frac{1}{\bar{C}^2}\left(\sigma_{QB}^2 + \sigma_{QF}^2 + q^2\sigma_{DI}^2\right) + \frac{\sigma_C^2}{\bar{C}^2}\left(\frac{\bar{Q}_B^2}{\bar{C}^2} + \frac{\bar{Q}_f^2}{\bar{C}^2} + \frac{q^2\bar{D}_I^2}{\bar{C}^2}\right). \tag{10}$$

Substituting (6)–(9) into (10) we have

$$\sigma_{VT}^2 = \frac{1}{\bar{L}\bar{W}\bar{C}^2}\left[q\left(\bar{Q}_B + \bar{Q}_f + q\bar{D}_I\right) + A_{ox}\left(\bar{Q}_B^2 + \bar{Q}_f^2 + q^2\bar{D}_I^2\right)\right]. \tag{11}$$

Now let us examine the importance of the various terms on the right-hand side of (11), for both p- and n-channel devices. Consider n-channel devices first. In our process, the threshold adjust implant is carried out for p-channel transistors only. Therefore $qD_I = 0$ for n-channel devices. The depletion charge density per unit area is

$$Q_B = 7.7\times10^{-8}\ \text{C/cm}^2. \tag{12}$$

In a well-controlled process the number of fixed oxide charges can be reduced to about $2\times10^{10}/\text{cm}^2$, and hence

$$Q_f = 3.2\times10^{-9}\ \text{C/cm}^2. \tag{13}$$

Comparing (12) and (13) we may infer that the contribution of the variability of the fixed oxide charges to threshold voltage mismatch (11) may be neglected.

The measured relative standard deviation of the threshold voltage ($\sigma_{VT}/\bar{V}_T$) is plotted against the reciprocal of the square root of the effective channel area in Fig. 1 for n-channel devices with six different W/L (drawn) values. $\sigma_{VT}/\bar{V}_T$ is chosen as the ordinate so as to express the variation as a percentage, independent of the operating point. However, if one is interested in current mismatch only, then $\sigma_{VT}/(V_{GS} - \bar{V}_T)$ should be plotted so that it may be directly used in (3).

For collecting statistics, 128 device pairs of each size were measured on every wafer. The vertical error bars reflect the spread in measured values over four wafers. Although the error bars appear large in this figure, in reality they represent relatively small deviations. For ex-

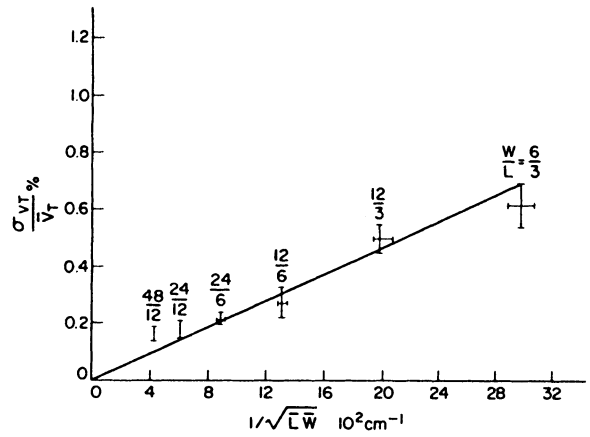


Fig. 1. Threshold voltage mismatch versus dimensions for n-channel devices.

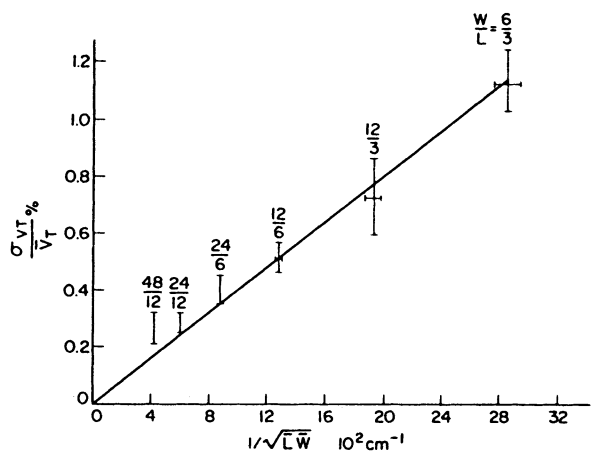


Fig. 2. Threshold voltage mismatch versus dimensions for p-channel devices.

ample, devices with $W/L = 6\ \mu m/3\ \mu m$ have a spread in the standard deviation of matching of threshold voltage of 3.5–4.9 mV. This spread is partly due to the nominal process variations from wafer to wafer, and partly due to the dependence of the matching on the electrical dimensions of the device. The effective channel length and width of devices were measured electrically at different places on a wafer and also on different wafers. The spread in the values is indicated by the horizontal error bars. The measured data fit well with the theoretical straight line relationship given by (11), which may be approximated for n-channel devices as

$$\sigma_{VT}/\bar{V}_T = \frac{1}{\sqrt{L\bar{W}}}\left(2.5875\times10^{-12} + 1.2421\, A_{ox}\right)^{1/2}/\bar{V}_T. \tag{14}$$

Comparing the slope of the line in Fig. 1 with that in (14), it is found that

$$A_{ox} = 6.4631\times10^{-14}\ \text{cm}^2. \tag{15}$$

Then from (9), $\sigma_C/\bar{C} = 0.02$ percent for a 24×6-μm^2 gate. This low value agrees with the extremely uniform nature of the gate oxide thickness observed in other measurements [16].

Now we consider p-channel devices. As the threshold adjust implant is a very shallow one, a considerable portion of the implanted ions is retained in the gate oxide. Although this results in charged states, they are readily annealed during subsequent processing [17]. However, the presence of these impurity atoms in the oxide may cause a degradation in the capacitance matching of p-channel devices as compared to n-channel ones. For our process $Q_B = 4.810\times10^{-8}$ C/cm^2 and $qD_I = 8.0\times10^{-8}$ C/cm^2. Hence the contribution of Q_f to threshold voltage mismatch may be ignored and (11) may be written as

$$\sigma_{VT}/\bar{V}_T = \frac{1}{\sqrt{L\bar{W}}}\left(4.2945\times10^{-12} + 1.8463\, A_{ox}\right)^{1/2}/\bar{V}_T. \tag{16}$$

The numerical coefficients in (16) are larger than the corresponding ones in (14) indicating a larger mismatch in p-channel devices, owing to the additional threshold adjust implant. This may be physically interpreted as a larger variation in the surface concentration due to the differential doping occurring at the surface.

The mismatch in threshold voltage of p-channel devices is plotted in Fig. 2. The data fit very well into the theoretical straight line relationship, and it is found that

$$A_{ox} = 3.0369\times10^{-12}\ \text{cm}^2. \tag{17}$$

Comparing (15) and (17) we may infer that the gate oxide capacitance matching is poorer for p-channel devices than that for n-channel ones.

We will now summarize the threshold voltage mismatch behavior. These findings are particular to this work.

a) The standard deviation of mismatch is inversely proportional to the square root of the effective channel area.

b) In a well-controlled process the nonuniform distribution of the fixed oxide charges has negligible effect on threshold voltage mismatch.

c) The nonuniform distribution of the dopant atoms in the bulk is a major contributor to the threshold voltage mismatch. The assumption that these atoms follow a Poisson distribution has resulted in excellent agreement with measurements.

d) Devices which use a compensating threshold adjust implant have a higher mismatch in threshold voltage due to the differential doping occurring at the surface. This is the major reason for the significantly larger mismatch noticed in p-channel devices as compared to n-channel transistors.

e) The gate oxide capacitance is quite uniform and hence has little influence on the threshold voltage mismatch. However, between n- and p-channel devices, the gate oxide capacitance of the latter has slightly poorer matching characteristic. This could be due to the nonuniform distribution of the threshold adjust implant atoms in the gate oxide.

2. Conductance Constant Mismatch: The conductance constant is given by

$$K = \mu C\, W/L \tag{18}$$

where μ is the channel mobility. We can express the variance in K in terms of the variances in μ, C, W, and L. Let us first consider the length of the device.

Electrically, the channel length is the average distance between the source and the drain diffusions. Any raggedness in the definition of the polysilicon may not be exactly reproduced in the source and drain diffusion edges. Further, we are not so much concerned about this raggedness; rather we are interested in the difference in the electrical length from one device to the next. Such a mismatch in length may not be due to the nonuniformity of the edge alone. In the absence of a complete knowledge of the causes of variation in length, we will simply indicate the variance of the length by σ_L^2 and make no attempt to derive any expression for it. In [7] the nonuniformity of the edges is the only cause considered for the mismatch and an expression for σ_L^2 is derived which is inversely proportional to the width of the device. This would mean that the mismatch in length would tend to zero for very wide devices. We have observed results that contradict this. For example, the conductance constant matching of devices with $W = 100$ μm and $L = 2$ μm is not all that different from devices with $W = 200$ μm and $L = 2$ μm. In fact we have noticed that σ_L is more or less independent of the device width.

The width of the device may be treated similarly and thus we let the variance in W be σ_W^2. The definitions of the length and width occur during different stages of processing and under different conditions. Hence L and W may be treated as independent random variables.

To determine the variance in mobility, a knowledge of the factors that affect it is required. It is reported in [18] that at room temperature and moderate gate bias the electron mobility is mainly governed by scattering due to interface charge centers and phonons. An empirical relationship for μ is [18]

$$\mu = \frac{\mu_0(N_A)}{1+\alpha(N_A)N_f} \tag{19}$$

where $\mu_0(N_A)$ and $\alpha(N_A)$ are empirical constants with very little dependence on the dopant concentration. Thus the mismatch in μ may be approximated to be entirely due to the nonuniformity of Q_f. As the fixed oxide charges have a Poisson distribution, we may write

$$\sigma_\mu^2 = \frac{\mu_0^2\alpha^2}{(1+\alpha\bar{Q}_f)^4}\cdot\frac{\bar{N}_f}{\bar{L}\bar{W}}. \tag{20}$$

The discussion given above for the mobility mismatch is for electrons only. We are not aware of any model that relates the mobility of holes to the doping concentration in the bulk and the fixed oxide charge density. The situation in the case of p-channel devices is further complicated by the threshold adjust implant. This could cause some damage in the substrate which may not be completely annealed, resulting in a poorer mobility matching than in the case of n-channel devices. In spite of these uncertainties, it is still reasonable to assume that the mobility mismatch has a similar dependence on channel dimensions as given by (20). Then

$$\frac{\sigma_\mu}{\bar{\mu}} = \left(\frac{A_\mu}{\bar{L}\bar{W}}\right)^{1/2} \tag{21}$$

where $\sqrt{A_\mu} = 4.95\times10^{-7}$ cm for n-channel devices and is not known for p-channel transistors.

The factors on the right-hand side of (18) are all independent. Thus

$$\frac{\sigma_K^2}{\bar{K}^2} = \frac{\sigma_L^2}{\bar{L}^2} + \frac{\sigma_W^2}{\bar{W}^2} + \frac{\sigma_\mu^2}{\bar{\mu}^2} + \frac{\sigma_C^2}{\bar{C}^2}. \tag{22}$$

From (9) and (21)

$$\frac{\sigma_K^2}{\bar{K}^2} = \frac{1}{\bar{L}\bar{W}}(A_\mu + A_{ox}) + \frac{\sigma_L^2}{\bar{L}^2} + \frac{\sigma_W^2}{\bar{W}^2}. \tag{23}$$

After substituting the values of A_μ and A_{ox} for n-channel devices, (23) may be solved for σ_L and σ_W using the measured values of σ_K of different sized devices. It is found that σ_L and σ_W are approximately the same and in the range of 0.01–0.03 μm. To provide a feel for the relative importance of the factors causing mismatch in K, we may substitute $\sigma_L = \sigma_W = 0.02$ μm in (23). Then

$$\frac{\sigma_K^2}{\bar{K}^2} = (2.46\times10^{-13} + 0.646\times10^{-13})\cdot\frac{1}{\bar{L}\bar{W}} + 4\times10^{-12}\left(\frac{1}{\bar{L}^2}+\frac{1}{\bar{W}^2}\right) \tag{24}$$

where the effective dimensions $\bar{L}$ and $\bar{W}$ have the units of centimeters. $\sigma_K/\bar{K}$ is plotted against $(1/\bar{L}^2+1/\bar{W}^2)^{1/2}$ in Fig. 3. The plotted relationship is not linear as shown by (24), with the curvature increasing for smaller geometry devices due to the increasing contribution of the $1/\bar{L}\bar{W}$ term. A similar plot for p-channel devices is shown in Fig. 4. The p-channel devices have a larger mismatch in conductance constant. One reason for this is the poorer gate oxide capacitance matching as has already been pointed out in connection with threshold voltage mismatch. Another factor could be a larger mobility variation.

We will now summarize the mismatch behavior in the conductance constant. These results are particular to this work.

a) The mismatch in K due to edge variations is proportional to $(1/\bar{L}^2+1/\bar{W}^2)^{1/2}$. The standard deviation of mismatch in length and width is in the range 0.01–0.03 μm. For n-channel devices, this is the dominant source of mismatch in K.

b) The larger gate oxide capacitance variation observed in p-channel devices in connection with V_T mismatch agrees with the larger mismatch in K.

c) For n-channel devices, the variation in mobility has little effect on the mismatch in K. The corresponding quantity for p-channel transistors, however, could be larger

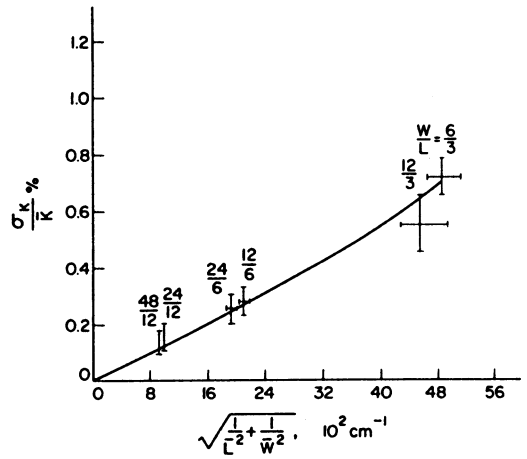


Fig. 3. Conductance constant mismatch versus dimension for n-channel devices.

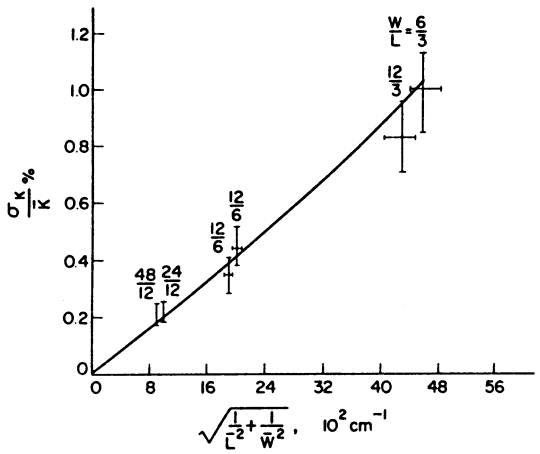


Fig. 4. Conductance constant mismatch versus dimension for p-channel devices.

due to any damage in the substrate caused by the threshold adjust implant.

3. Correlation Between Mismatches in V_T and K: A common contributing factor to the mismatches in V_T and K is the variation in the gate oxide capacitance. Hence we can expect a dependence between the mismatches in V_T and K. A theoretical expression for the correlation coefficient is derived in [15]. Also the value has been experimentally measured. The agreement is excellent for n-channel devices and fair for p-channel ones. However, both the theoretical and experimental values are close to zero indicating that the mismatches in V_T and K are almost independent.

4. Mismatch in Drain Current: The drain current mismatch in the saturation region is given by (3). As the correlation coefficient is nearly equal to zero, we have

$$\frac{\sigma_I^2}{\bar{I}^2}=\frac{\sigma_K^2}{\bar{K}^2}+4\frac{\sigma_{VT}^2}{(V_{GS}-\bar{V}_T)^2}. \tag{25}$$

At low values of gate-to-source voltage the dominant factor causing the mismatch in drain current is the threshold voltage variation. For bias levels approaching the mid-rail,

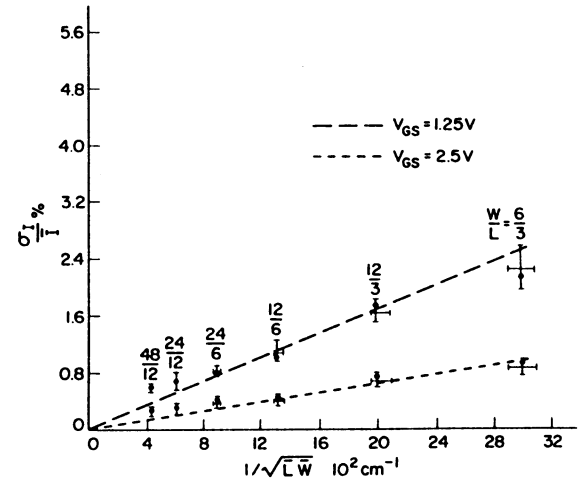


Fig. 5. Drain current mismatch versus dimension for n-channel devices. The dots are the estimated values using (25).

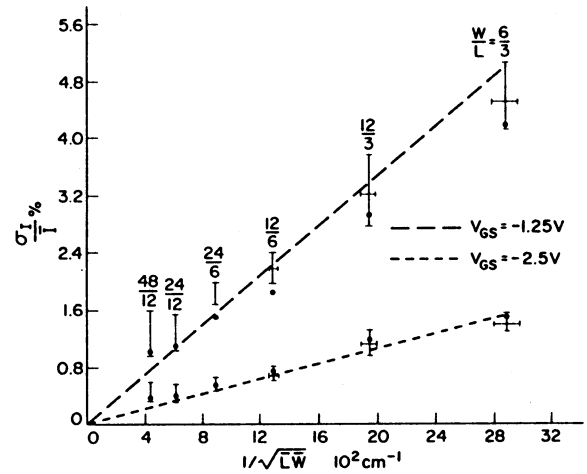


Fig. 6. Drain current mismatch versus dimension for p-channel devices. The dots are the estimated values using (25).

the conductance constant and the threshold voltage mismatches have almost equal contributions to the drain current mismatch. From (25) σ_I may be estimated from the measured values of σ_K and σ_{VT}. Also we have actually measured σ_I at different gate biases to compare with the estimated values. Fig. 5 is a plot of $\sigma_I/\bar{I}$ versus $1/\sqrt{\bar{L}\bar{W}}$ for n-channel devices for two gate voltages. The estimated values obtained from (25) using the measured average values of σ_{VT} and σ_K are indicated by the dots. A similar result is shown for the p-channel devices in Fig. 6. The excellent agreement between the measured values and the estimated ones for both p- and n-channel devices validates the characterization methodology and also verifies the mismatch models.

5. Range of Applicability: It is important to consider the dimensional range over which the mismatch models we have developed in the preceding sections are accurate. In our analysis we have assumed that the dimensional variations are accounted for entirely by the mismatch in conductance constant and have no influence on the threshold voltage mismatch. As the threshold voltage of small-geom-

etry devices is a function of channel length and width, it is necessary to estimate the mismatch in threshold voltage brought about by the dimensional variation to validate the above assumption and hence the mismatch models. To this end, the shift in threshold voltage brought about due to short-channel and narrow-width effects was estimated for a device with effective dimensions 2×2 μm^2 fabricated in our process, using the expressions in [14]. It was found that the mismatch component of the threshold voltage brought about by the dimensional variations is only 10 percent of the total threshold voltage mismatch, in the worst case [15]. We also verified this fact for even smaller device geometries using the process parameters given in [19]. Hence we may attribute the dimensional variations entirely to the mismatch in K and not to V_T. Thus we may conclude that the characterization methodology and the mismatch models are valid for small-geometry devices also. However, as new processes emerge permitting smaller geometry devices, further experimental work is needed to characterize the mismatch.

6. Effect of Temperature: As the threshold voltage and conductance constant vary with temperature, it is interesting to know their matching behavior as a function of temperature. In the case of the threshold voltage, as expressed by (5), the only terms that are dependent on temperature are ϕ_{MS} and ϕ_B. We have seen that the contribution of these terms to the threshold voltage mismatch is negligible. Therefore we may expect the matching behavior of threshold voltage to be almost independent of temperature.

The only factor through which the conductance constant matching can be affected is the temperature dependence of mobility. For n-channel devices we have seen that the mismatch in conductance constant is largely due to photolithographic edge variations, and mobility variations have the least effect. Thus temperature variations should have very little effect on the conductance constant matching of n-channel devices. Since the mismatch in drain current is due to mismatches in threshold voltage and conductance constant, we can expect the current mismatch in n-channel devices to be almost unaffected over a wide temperature range. Limited experimental results seem to agree with this prediction. As far as the p-channel devices are concerned, since the mobility behavior of holes is not clearly understood, no theoretical explanation of the temperature effect is possible.

III. Design Methodology

To demonstrate the usefulness of the study of the matching behavior of transistors and the related models, we took up the task of designing an 8-bit current-steering CMOS DAC. Circuit details of the DAC have already been presented [9]. Here we only indicate the design methodology. In Section III-A, a brief description of the DAC configuration is presented. Section III-B discusses statistical error analysis. The close interaction between the DAC configuration, the matching accuracy of devices, and the circuit yield is brought out here. Finally, yield results are presented in Section III-C.

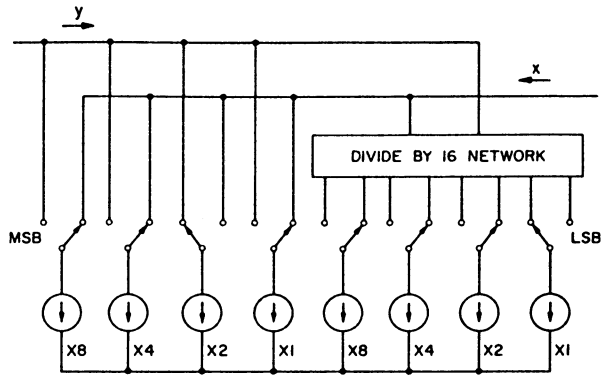


Fig. 7. Schematic of a multiple current-source DAC.

A. DAC Configuration

A multiple current-source approach was chosen to realize the binary-weighted currents. This was done primarily to overcome the problems of nonlinear relationship between the drain current and aspect ratio of small-geometry MOS devices [14], [15], [20], and the voltage coefficient of resistance of R-$2R$ networks. The configuration is shown in Fig. 7 and is similar to that reported in [21]. The least significant bit (LSB) has one *unit* current source, the next significant bit has two unit current sources connected in parallel, and so on. The exponential growth in the number of unit current sources is overcome by having an interstage 16:1 resistive current divider.

B. Statistical Error Analysis

In general, the errors generated by a DAC consist of linearity, offset, and gain errors. Usually, a DAC may be calibrated for zero gain and offset errors. Linearity error, however, occurs due to the random mismatch in the conversion elements. Hence, the circuit yield is a function of the matching accuracy of the unit current sources.

Integral nonlinearity of a DAC is generally defined as the difference between the actual output to the desired output normalized to the full-scale output of the DAC. This enables the nonlinearity to be expressed in terms of fractions of LSB or as a percentage.

Let x be the output of the DAC for a given input word and y be the analog complement of the output. The full-scale output of the converter is the sum $x+y$. To express the error as a fraction of the full scale, first we normalize the output to the full scale as follows:

$$z(x,y)=\frac{x}{x+y} \tag{26}$$

where z is the normalized output, and x and y are dependent on the input digital word and the DAC configuration.

For the 8-bit interstage divider DAC shown in Fig. 7

$$x = \left[8D_1 + 4D_2 + 2D_3 + D_4 + \frac{1}{16}(8D_5 + 4D_6 + 2D_7 + D_8)\right] I_{\text{unit}} \quad (27)$$

and

$$y = \left[8(1-D_1) + 4(1-D_2) + 2(1-D_3) + (1-D_4) + \frac{1}{16}\{8(1-D_5) + 4(1-D_6) + 2(1-D_7) + (1-D_8)\}\right] I_{\text{unit}} \quad (28)$$

where

$$D_i = 0 \text{ or } 1$$
$$i = 1, 2, \cdots, 8$$

D_1 is the most significant bit (MSB), D_8 is the LSB, and I_{unit} is the unit current source with a mean value $\bar{I}$ and standard deviation of matching σ. As the unit current sources are random and uncorrelated, we may treat x and y to be independent random variables with standard deviations σ_x and σ_y, respectively. Now we may determine the standard deviation of z in terms of σ_x and σ_y [12]

$$\sigma_z^2 = \frac{\bar{y}^2\sigma_x^2 + \bar{x}^2\sigma_y^2}{(\bar{x}+\bar{y})^4} \quad (29)$$

where σ_x and σ_y are evaluated as follows:

$$\bar{x} = \left[8D_1 + 4D_2 + 2D_3 + D_4 + \frac{1}{16}(8D_5 + 4D_6 + 2D_7 + D_8)\right] \bar{I}$$

and

$$\sigma_x^2 = \left[8D_1 + 4D_2 + 2D_3 + D_4 + \frac{1}{16}(8D_5 + 4D_6 + 2D_7 + D_8)\right] \sigma^2 = \frac{\bar{x}}{\bar{I}} \cdot \sigma^2. \quad (30)$$

Similarly

$$\sigma_y^2 = \frac{\bar{y}}{\bar{I}} \cdot \sigma^2. \quad (31)$$

Substituting (30) and (31) into (29)

$$\sigma_z^2 = \frac{1}{\bar{I}} \cdot \frac{\bar{x}\bar{y}}{(\bar{x}+\bar{y})^3} \cdot \sigma^2. \quad (32)$$

The expected value of z may be shown to be [12]

$$\bar{z} = \frac{\bar{x}}{\bar{x}+\bar{y}}. \quad (33)$$

Using (33) in (32), we have

$$\sigma_z^2 = \frac{\bar{z}(1-\bar{z})}{\bar{I}(\bar{x}+\bar{y})} \cdot \sigma^2. \quad (34)$$

The above result may also be derived by determining the joint probability density function of z as shown in [22]. Equation (34) expresses the variances of the D/A output for different digital words. The function $\bar{z}(1-\bar{z})$ will have a maximum value when $\bar{z} = 1/2$, i.e., when the output is halfway through the full scale, and falls off towards zero for minimum and maximum input word combinations. This observation suggests that the MSB current could be the most critical and should have the highest accuracy. Error contributions of the bits taper off towards the LSB, and hence the relative error contributions of all the bit current sources need not be the same. Such an error distribution is indeed a natural consequence in the multiple current-source approach where the relative accuracy of the bits improves towards the MSB. This may be shown as follows. If the unit current source has a mean value $\bar{I}$ and standard deviation of matching σ, connecting n such sources in parallel would produce an equivalent current source with mean value $n\bar{I}$ and standard deviation $\sqrt{n}\,\sigma$, as the current sources are uncorrelated. Thus there is an improvement in accuracy by a factor $\sqrt{n}$.

The analysis so far has shown that maximum error occurs halfway through the full scale and the error contributions of the individual bits taper off towards the LSB. Now we proceed further to relate the circuit yield to the standard deviation of the unit current sources. Here we define circuit yield as the percentage of functional devices that have integral nonlinearity less than 1/2 LSB. In other words, we are eliminating catastrophic device failures due to defects, etc. With this definition, a theoretical estimate of the circuit yield of the DAC is obtained by multiplying the probabilities that each of the 256 outputs of the DAC have less than 1/2 LSB error. For normal distribution with variances given by (34), the yield is

$$G = \prod_{i=2}^{255} \frac{1}{\sqrt{2\pi}\,\sigma_z} \int_{\bar{z}-1/512}^{\bar{z}+1/512} \exp\left\{-\frac{(z-\bar{z})^2}{2\sigma_z^2}\right\} \cdot dz = \prod_{i=2}^{255} \operatorname{erf}(Q/\sqrt{2}) \quad (35)$$

where 1/512 is the normalized 1/2 LSB value and

$$Q = \frac{1}{512\left[\dfrac{\bar{z}(1-\bar{z})}{15+15/16}\right]^{1/2}} \frac{\sigma}{\bar{I}}. \quad (36)$$

The method used to derive (35) is quite general and may be easily extended to converters of different resolutions and accuracies. The circuit yield as given by (35) is plotted as a function of $\sigma/\bar{I}$ in Fig. 8.

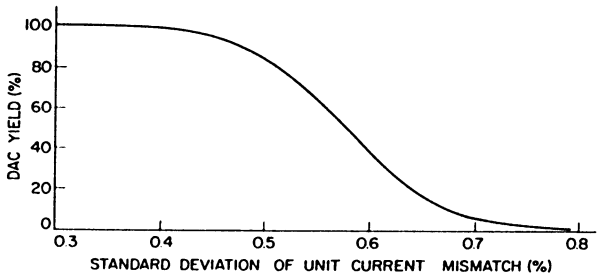


Fig. 8. DAC yield versus current-source mismatch.

TABLE I
DAC YIELD

Wafer Number	Devices Tested	Functional Devices	Good Devices	Circuit Yield %
1	32*	32	31	97
2	55*	55	55	100
3	64	60	60	100
4	64	52	52	100
5	64	60	60	100

*Broken Wafers. Hence the number of devices tested are less than 64.

Because of the error function nature of the relation between yield and current-source matching, there is an almost flat region close to the 100-percent yield level, followed by a very steep region and finally the yield asymptotically approaches zero. To avoid the possibility of any process variations from batch to batch affecting the yield adversely, the design should be such as to avoid the region where the yield is very sensitive to the matching accuracy. On the other hand, a very conservative matching tolerance is also not desirable because the improvement in yield is marginal with improvement in matching accuracy beyond a certain point. Therefore we chose 95-percent yield level as an optimum value that would not appreciably shift due to process variations from batch to batch. From the theoretical relationship shown in Fig. 8, the standard deviation of matching of the current sources should be about 0.45 percent to achieve this yield level. It may be noted that for an 8-bit DAC, an integral nonlinearity of $\pm 1/2$ LSB is equivalent to 0.2 percent of full scale. Thus, with this configuration, it is possible to provide good integral linearity associated with a high circuit yield without requiring an equivalent degree of component matching. Further it is shown in [15] that the divide-by-16 network is highly tolerant to component mismatch and hence is not a potential source of yield loss.

The analysis given above provides a systematic design approach for data converters. A similar approach may be used to design any precision analog circuit in general.

C. Results

Based on the understanding of the matching behavior of MOS devices and the systematic design methodology, an 8-bit high-speed DAC has been designed. The electrical performance results are reported in [9]. The unit current source used in the DAC is made up of a 24-μm-wide and 6-μm-long n-channel transistor in combination with a 4.7-kΩ source degradation resistor. This configuration has a better matching accuracy than a transistor alone for the same current value, owing to the better matching of resistors and the local negative feedback offered by the resistor [9]. The combination has a standard deviation of current matching of 0.45 percent when biased at a current of 128 μA. This should result in a circuit yield of approximately 95 percent. It may be noted that the same degree of matching may be obtained without using the source degradation resistor by choosing larger area devices, and/or operating the devices with a larger gate-to-source voltage. Knowledge of the mismatches in V_T and K in conjunction with (25) may be used to obtain curves such as those in Figs. 5 and 6, for any process and operating condition. This information when used with (35) or Fig. 8 completes the design cycle.

Circuit yield statistics of the DAC are presented in Table I. Column 2 indicates the number of devices that are tested on each wafer. The next column shows the number of devices that are functional. In other words, we are eliminating catastrophic failures here. The fourth column shows the number of devices with integral nonlinearity less than $\pm 1/2$ LSB. Finally, the circuit yield is shown in the last column. In most cases the circuit yield is 100 percent, demonstrating the accuracy of the device characterization and the circuit design methodology. We have been able to achieve this high level of yield using relatively small devices and without using any trimming because of the knowledge we generated regarding the matching behavior of the devices as a function of dimensions, and the systematic design methodology we have followed.

IV. Conclusion

The design of precision analog circuits presents challenges in the areas of device matching characterization and circuit design. Novel methodologies relevant to both aspects of the design are presented in this paper. Section II is devoted to the study of transistor matching behavior. The overall objective has been not only to provide a clear understanding of the random mismatch, but also to develop a comprehensive design approach for precision analog circuits. The parameters a circuit designer will have freedom to choose are the dimensions of the devices. Therefore analytical models have been developed that relate the mismatch to device dimensions.

Section III of the paper discusses the application of the matching characterization in precision analog design. Design methodology for a high-performance DAC is illustrated. This is presently important because of the need for high-speed data converters in MOS technology. The close interaction between device matching and circuit yield is discussed. Experimental results of circuit yield are also presented.

ACKNOWLEDGMENT

The authors are indebted to the management of Northern Telecom Electronics Ltd., for extending their facilities to carry out this research. They would like to thank Dr. M. Simard-Normandin for her help with device characterization. Thanks are also due to M. King for many suggestions and discussions. The assistance of N. Prasad concerning statistical analysis is gratefully acknowledged.

REFERENCES

[1] J. L. McCreary and P. R. Gary, "All MOS charge redistribution analog-to-digital conversion techniques—Part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, Dec. 1975.

[2] J. T. Caves, M. A. Copeland, C. F. Rahim, and S. D. Rosenbaum, "Sampled analog filtering using switched capacitors as resistor equivalents," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 592–599, Dec. 1977.

[3] B. J. Hosticka, R. W. Brodersen, and P. R. Gray, "MOS sampled data recursive filters using switched capacitor integrators," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 600–608, Dec. 1977.

[4] D. A. Hodges, P. R. Gray, and R. W. Brodersen "Potential of MOS technologies for analog integrated circuits," *IEEE J. Solid-State Circuits*, vol. SC-13, pp. 285–294, June 1978.

[5] J. L. McCreary, "Matching properties and voltage and temperature dependence of MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 608–616, Dec. 1981.

[6] J. B. Shyu, G. C. Temes, and K. Yao, "Random errors in MOS capacitors," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1070–1076, Dec. 1982.

[7] J. B. Shyu, G. C. Temes, and F. Krummenacher, "Random error effects in matched MOS capacitors and current sources," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 948–955, Dec. 1984.

[8] M. Akyia and S. Nakashima, "High-precision MOS current mirror," *Proc. Inst. Elec. Eng.*, vol. 131, pt. I, pp. 170–175, Oct. 1984.

[9] K. R. Lakshmi Kumar, R. A. Hadaway, M. A. Copeland, and M. I. H. King, "A high-speed 8-bit current steering CMOS DAC," in *Proc. IEEE 1985 Custom Integrated Circuits Conf.*, pp. 156–159.

[10] J. Logan, "Characterization and modelling for statistical design," *Bell Syst. Tech. J.*, vol. 50, pp. 1105–1147, Apr. 1971.

[11] E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*. New York: Wiley, 1982.

[12] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. Tokyo: McGraw-Hill, Kogakusha, 1965.

[13] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. I. New York, Wiley, 1957, p. 146.

[14] S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. New York: Wiley, 1981.

[15] K. R. Lakshmi Kumar, "Characterization and modelling of mismatch in MOS devices and application to precision analog design," Ph.D. dissertation, Carleton Univ., Ottawa, Ont., Canada, 1985.

[16] M. Simard-Normandin, private communication, 1985.

[17] S. K. Ghandhi, *VLSI Fabrication Principles Silicon and Gallium Arsenide*. New York: Wiley, 1983, p. 353.

[18] S. C. Sun and J. D. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 562–573, Aug. 1980.

[19] J. R. Brews, W. Fitchner, E. H. Nicollian, and S. M. Sze, "Generalized guide for MOSFET miniaturization," *IEEE Electron Device Lett.*, vol. ED.-1, pp. 2–4, Jan. 1980.

[20] P. Yang and P. K. Chatterjee, "SPICE modelling for small geometry MOSFET circuits," *IEEE Trans. Computer-Aided Des.* vol. CAD-1, pp. 169–182, Oct. 1982.

[21] P. H. Saul *et al.*, "An 8-bit, 5-ns monolithic D/A converter subsystem," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 1033–1039, Dec. 1980.

[22] S. Kuboki *et al.*, "Nonlinearity analysis of resistor string A/D converters," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 383–390, June 1982.

[23] P. H. Saul *et al.*, "An 8b CMOS video DAC," in *ISSCC Dig. Tech. Papers*, 1985, pp. 32–33.

[24] T. Miki *et al.*, "An 80 MHz 8b CMOS D/A converter," in *ISSCC Dig. Tech. Papers*, Feb. 1986, pp. 132–133.

Measurement and Modeling of Charge Feedthrough in n-Channel MOS Analog Switches

WILLIAM B. WILSON, STUDENT MEMBER, IEEE, HISHAM Z. MASSOUD, MEMBER, IEEE, ERIC J. SWANSON, MEMBER, IEEE, RHETT T. GEORGE, JR., MEMBER, IEEE, AND RICHARD B. FAIR, SENIOR MEMBER, IEEE

Abstract — **Charge feedthrough in analog MOS switches has been measured. The dependence of the feedthrough voltage on the input and tub voltages, device dimensions, and load capacitances was characterized. Most importantly, it was observed that the feedthrough voltage decreases linearly with the input voltage. The significance of this observation when considering harmonic distortion in sample-and-hold circuits is discussed. A first-order computer simulation based on the quasi-static small-signal MOSFET capacitances shows good agreement with experimental results.**

I. Introduction

CMOS CIRCUITS have historically been used in digital logic and memory applications as a result of their high packing density and low power consumption. Recently, analog CMOS has combined the benefits of both analog and digital circuits on the same die. This combination offers many advantages in signal processing and analog-to-digital conversion. A key component in interfacing digital and analog circuitry is the analog switch, or transmission gate, where a signal is fed from the input side of the device to a load connected at the output. The on–off switching capability is provided by the gate voltage governing the presence of an inversion channel.

MOSFET's are not ideal switches. Consider a typical sample-and-hold circuit shown in Fig. 1. Ideally, when the switch is turned off, the output voltage on the load capacitance should remain at its value at the time of switching. However, the MOSFET actually couples some of the charge in the inversion layer onto the load capacitor. Thus the final capacitor voltage consists of two components, one directly related to the input voltage and an error component arising from the charge stored in the switch. This error component is the feedthrough voltage. In this paper, this feedthrough voltage is measured in MOS switches implemented in a 3.5-μm CMOS technology. In addition, first-order simulations based on quasi-static small-signal models of MOSFET capacitances have been made to predict the magnitude of the feedthrough voltage. Experimental results and computer simulations are presented and the conclusions drawn lead to a first-order understanding of the phenomenon of charge feedthrough.

Manuscript received March 4, 1985; revised July 8, 1985. W. B. Wilson was supported by a Microelectronics Center of North Carolina Fellowship.

W. B. Wilson, H. Z. Massoud, R. T. George, Jr., and R. B. Fair are with the Department of Electrical Engineering, Duke University, Durham, NC 27706.

E. J. Swanson was with AT&T Bell Laboratories, Reading, PA 19604. He is now with Crystal Semiconductor, Austin, TX 78744.

II. Measurement System

When an n-channel MOSFET is turned on, an inversion channel of electrons is formed underneath the gate. When the device is turned off, these electrons are dispersed, flowing either into the substrate of the device, or to the loads at the source and drain. The effect produced by electrons flowing into the substrate is called charge pumping [1]. The flow of electrons to the source and drain nodes is called charge or clock feedthrough. It was first discussed by Stafford *et al.* in 1974 [2]. This study is concerned with measuring charge feedthrough in an n-channel MOSFET in a sample-and-hold circuit under a variety of bias and loading configurations. The circuit used is shown in Fig. 2, where it can be seen that both the source (connected to nC) and the drain (connected to $2C$) nodes are floating. This is in contrast with recent studies of charge feedthrough where only one node was floating [3], [4]. The value of C is 1.15 pF and the index n in the load capacitance nC can be varied from $n=1$ to $n=96$. For $n=96$, the source load capacitance simulates a voltage source. This permits the measurement of the feedthrough voltage in a situation where only one node is allowed to float, and makes possible comparisons with other studies [3], [4]. The feedthrough voltages at the source and drain are measured as V_{OUT1} and V_{OUT2} after source followers M_3 and M_4, respectively. The distinction between source and drain becomes irrelevant once steady-state conditions are attained and the labels V_{OUT1} and V_{OUT2} are used to identify one side of the device from the other. The source followers supply the power to drive the inherent capacitances of the package and the cable connections. Thus these capacitances do not need to be considered. Bias circuitry (not shown in Fig. 2) maintains the transistors M_1 and M_2 of the source followers biased in saturation with a 25-μA current. The gate length

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1206–1213, Dec. 1985.

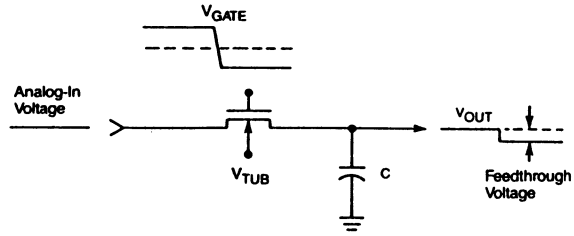


Fig. 1. Charge-feedthrough definition in a sample-and-hold circuit.

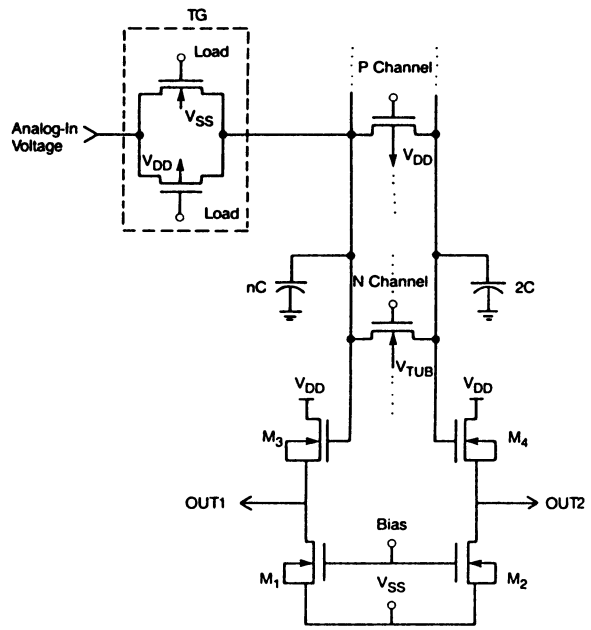


Fig. 2. Basic circuit used in measuring charge feedthrough.

in both the n- and p-channel test devices varies between 4 and 20 μm.

III. Measurement Procedure

In this investigation, only the charge feedthrough on n-channel devices was characterized. The parallel p-channel device was used to initialize the system ensuring that the prefeedthrough voltage on each load capacitor is identical. This is necessary when n-channel devices are body-effect limited.

The basic measurement procedure is started by applying a gate voltage of +5 V to the n-channel test transistor, which turns the switch on. Then the two load capacitors ($2C$ and nC) are charged to their initial value called the analog-in voltage V_{AI}. At this point, the gate voltage is quickly driven to −5 V turning the switch off. As this occurs, the stored charge within the device is dumped onto the now-floating load capacitors shifting the voltage across each of them. The difference between the measured voltages at each node (V_{OUT1} and V_{OUT2}) before and after the gate transition yields the feedthrough voltage at the corresponding node. This is shown graphically in Fig. 1.

The following variables affect charge feedthrough and will be studied.

1) The initial voltage on each of the load capacitors is called the analog-in voltage V_{AI}. It is placed on the two load capacitors by closing the CMOS transmission gate, labeled *TG* in Fig. 2, and by turning on both the n-channel and p-channel test devices. The magnitude of V_{AI} affects the amount of charge stored in the n-channel device in two ways. First, it directly affects the amount of charge on the gate through its effect on the gate-to-source voltage V_{gs}. Secondly, it affects the amount of charge stored in the bulk through the body effect. It should be noted that this latter effect is nonlinear.

2) The tub voltage V_{TUB} also has an important effect on the amount of charge stored in the device. However, unlike V_{AI}, V_{TUB} only contributes via the nonlinear body effect. Thus the change in the feedthrough voltage with V_{TUB} is less dramatic than with V_{AI}.

3) The gate length L of the device has an obvious effect. By increasing the area of the gate, the total amount of charge stored in the device is increased, as is the feedthrough voltage.

4) Finally, the value of the load capacitance is expected to affect the feedthrough voltage.

Several other parameters, which also have a pronounced effect on the amount of charge feedthrough, were held constant in this study. The fall time of the gate signal, which directly affects the amount of charge pumping [1], was held constant at 10 ns. Changing the fall time of the gate signal has been shown both theoretically and experimentally to affect the measured feedthrough voltage [3]–[5]. Such effects are currently being investigated. Also, the gate voltage was always pulsed from +5 to −5 V. The device width was the same (20 μm) for all devices, with the overlap capacitance estimated to be 11.5 fF, based on an assumed overlap distance of 1 μm. Accurate measurements of the overlap distances at the source and drain are presently being made. The p-well was doped with an acceptor concentration of 1.1×10^{16} cm^{-3} resulting in a zero-bias threshold voltage (V_{T0}) of 0.7 V.

A large matrix of feedthrough voltage measurements was obtained where each data point was the average of ten measurements. The accuracy of the measurements was 5 mV, and the standard deviation between the ten measurements was consistently less than one least significant bit (5 mV).

Due to the layout of the circuit, there are several additional parasitic capacitances in parallel with the source and drain load capacitances. These capacitances have different values and, as a consequence, for the symmetrical load case of $n = 2$, the feedthrough voltages measured at the source and drain are not identical.

IV. Experimental Results

In this section, the measured dependence of the feedthrough voltage at the source and drain is presented as a function of the analog-in voltage, the tub voltage, the device dimensions, and the value of the load capacitances.

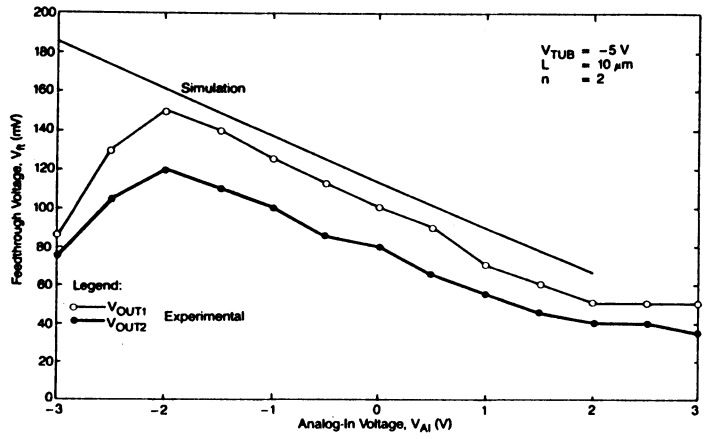


Fig. 3. Dependence of the feedthrough voltage V_{ft} on the analog-in voltage V_{AI}. Circles (experiment), solid lines (simulation).

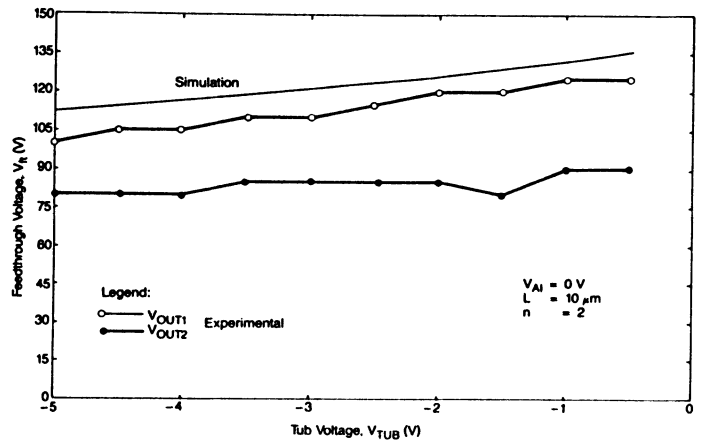


Fig. 5. Dependence of the feedthrough voltage V_{ft} on the tub voltage V_{TUB}. Circles (experiment), solid lines (simulation).

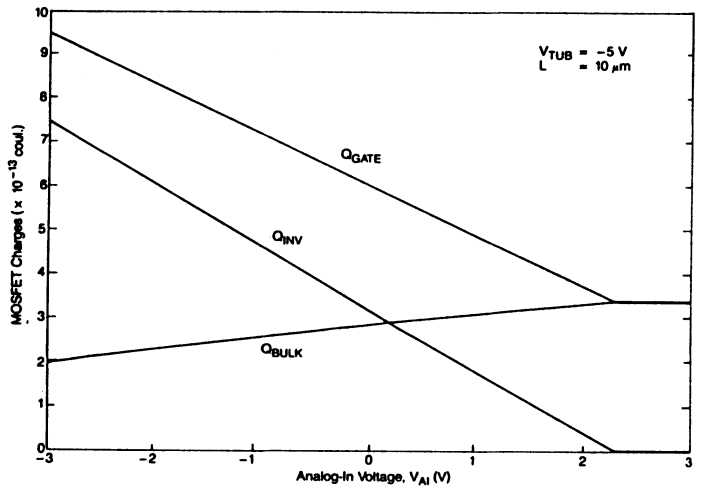


Fig. 4. Dependence of the gate charge Q_{GATE}, the inversion charge Q_{INV}, and the bulk charge Q_{BULK} on the analog-in voltage V_{AI}.

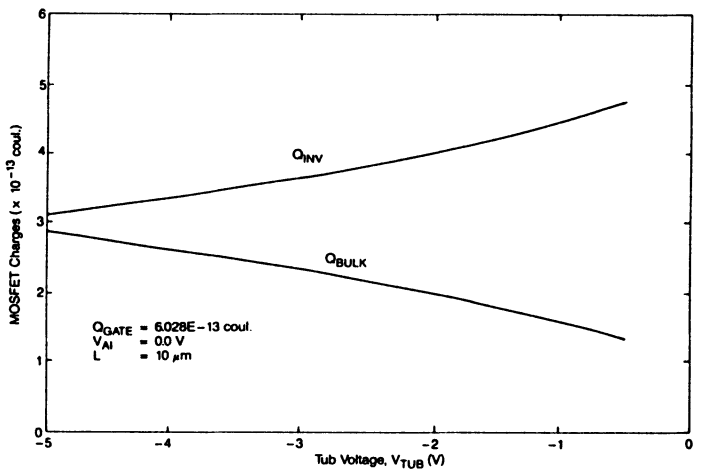


Fig. 6. Dependence of the inversion charge Q_{INV}, and the bulk charge Q_{BULK} on the tub voltage V_{TUB}.

The results of the simulations, which will be described later, are also shown in Figs. 3, 5, 7, and 8.

A. Effect of the Analog-In Voltage

Before discussing the measurement results, it is important to mention that the gain of the source-follower circuit is nearly unity (0.974) over the entire range of values of the input voltage V_{AI}. A typical plot of the dependence of the feedthrough voltage on V_{AI} is shown in Fig. 3. The important observation here is the linearity of the feedthrough voltage from -2 to $+2$ V. In this range, the nonlinear body effect, dependent on the bulk-to-source voltage, does not significantly affect the feedthrough voltage. This suggests that the amount of inversion channel charge initially stored in the device is the predominant source of the feedthrough voltage. This is particularly important as it implies that the inversion charge varies linearly with the channel potential. This dependence is demonstrated in Fig. 4 where the calculated inversion charge is plotted as a function of the analog-in voltage V_{AI}. This linear dependence of Q_{INV} on V_{AI} confirms the inversion channel charge as the predominant source of feedthrough charge.

There are two portions of the plots that deviate from linearity. The first is for low values of V_{AI}. It is suspected that this dropoff is a function of the fall time of the gate signal, and that the curve would be extended if the fall time was increased. The second deviation from linearity is for large values of V_{AI}. In this case, there is no inversion charge stored in the device in its initial state. The difference in feedthrough voltage at V_{OUT1} and V_{OUT2}, in this case, can be attributed to different overlap capacitances at the source and drain. The cause of the deviation is shown clearly in Fig. 4 where stored charge is plotted against the initial channel voltage V_{AI}. This graph represents the initial amount of stored charge in the intrinsic device only. The equations describing the stored charge are listed in the Appendix. The gate voltage is +5 V, and as can be seen, the inversion charge goes to zero as V_{AI} reaches 2.3 V. Because of the dependence of the feedthrough voltage on the inversion channel charge, the remaining discussion of the dependence of V_{ft} on V_{AI} will be limited to $V_{AI} \leq 2$ V where the inversion charge goes to zero.

B. Effect of the Tub Voltage

Fig. 5 shows the feedthrough voltage dependence on the tub voltage V_{TUB}. Here, the effect of the backgate bias is demonstrated. Again, it is remarkable to note that the curve is reasonably linear given that a nonlinear effect is involved. The cause of this linearity can be shown by examining a plot of the initial conditions as shown in Fig. 6 It is evident here that the body effect is nonlinear. However, the nonlinearity is quite small and becomes even less significant in the feedthrough process.

It is also interesting to note that increasing the backgate bias decreases the feedthrough voltage. As V_{TUB} is increased, the stored depletion charge increases and the inversion charge decreases. This is significant because V_{TUB} can be fixed at a large negative value, minimizing V_{ft}. It is also desirable to have a large negative V_{TUB} because it increases the allowed range of input voltages to the switch. Consequently, all remaining discussion and data will be for a switch with $V_{TUB} = -5$ V, minimizing the feedthrough voltage and maximizing the allowable input range.

C. Effect of the Channel Length

The width of the experimental devices has been fixed at 20 μm, while the length ranged from 4 to 20 μm. In Fig. 7, the systematic alternation of measured values of V_{ft} at V_{OUT1} and V_{OUT2} is a direct result of the different overlap capacitances, as discussed earlier, at the diffusions used as source and drain. Due to the special layout of this circuit, these diffusions were alternately used as source or drain with increasing channel length. The feedthrough voltage appears to have a linear dependence on device dimensions as expected. Theoretically, this can be attributed to the linear dependence of the inversion channel charge on the gate area. From this, it follows that to minimize the feedthrough voltage, a minimum-sized switch should be used. If such devices are not available, then feedthrough cancellation schemes should be used [6].

D. Effect of the Load Capacitance

The feedthrough voltage is plotted against the inverse of the load capacitance in Fig. 8. It can be seen that the amount of charge fed out of the switch is independent of the value of the load capacitance. This is demonstrated in two ways. First, the feedthrough voltage is constant on the drain side, where the load capacitance is fixed ($2C$). Thus, changing the load capacitance on one side (nC) of the device does not appear to affect the amount of charge feedthrough on the other. The second observation comes from the feedthrough voltage at the source, where the load capacitance nC is varying. If V_{ft} is plotted versus the inverse load capacitance then the amount of feedthrough charge is independent of the load capacitance if the curve is linear. Thus the feedthrough voltage is minimized by increasing the load capacitance which would, however, increase the data capture time in a sample-and-hold circuit.

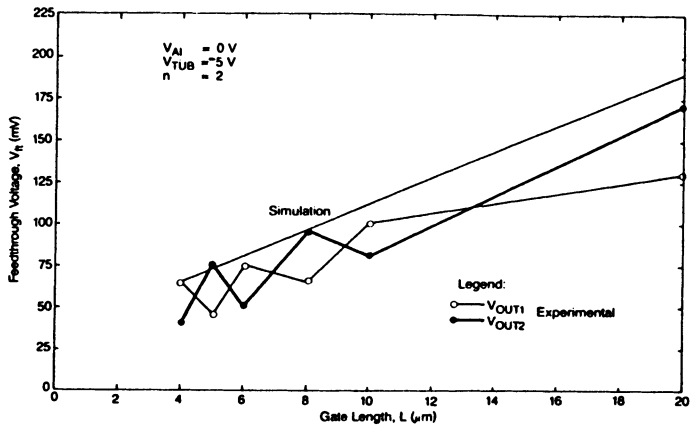


Fig. 7. Dependence of the feedthrough voltage V_{ft} on the gate length L. Circles (experiment), solid lines (simulation).

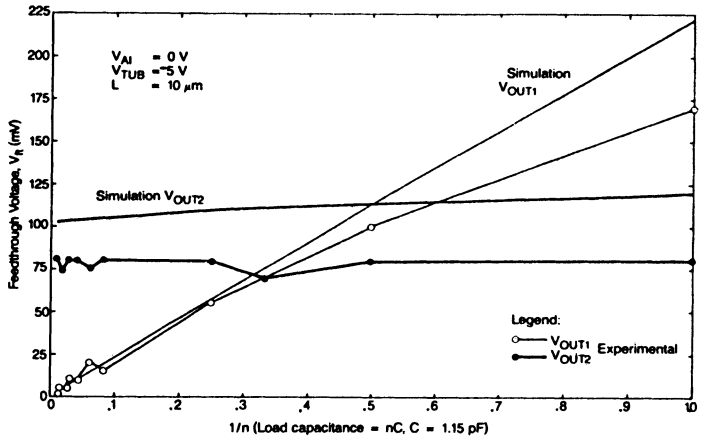


Fig. 8. Dependence of the feedthrough voltage V_{ft} on $1/n$, the inverse of the load capacitance index number. Circles (experiment), solid lines (simulation).

Consequently, a compromise must be reached based on requirements on the circuit performance.

V. Modeling Charge Feedthrough

A. Model Description

It is important to be able to predict the amount of charge feedthrough in MOSFET circuits, especially when simulating the behavior of sample-and-hold circuits. The following is a discussion of a simple program which simulates charge feedthrough in a transmission gate where both the source and drain are floating when the switch is turned off.

The simulation is based on the MOSFET quasi-static small-signal capacitance model developed by Liu and Nagel [7]. The simulation is divided into four sections corresponding to four regions of the model. Each region is simulated in turn yielding a contribution to the feedthrough voltage for that region. Finally, the voltages from each of the four regions are summed together yielding the total feedthrough voltage. In spite of the fact that the original

model has been considerably simplified, the simulations show qualitative agreement with the measured results.

The first assumption in the simulation is that charge pumping is ignored. This assumption manifests itself in two ways. First, it is not necessary to include the gate-to-bulk capacitance, since this element cannot contribute charge to the load capacitors at the source and drain. Second, an ideal 50–50 split of the inversion channel charge is assumed between the source and the drain. This assumption will cause the simulation to predict values slightly larger than will actually occur.

The second assumption in the simulation is that the drain-to-source voltage is zero when the device is biased in the linear region. In principle, this is valid because the initial voltages at the source and drain are equal, producing a uniform inversion channel. Also, from the experimental results it is observed that the load on one side of the device does not affect the other side. This implies that the device is never biased in the saturation region. As the gate voltage is decreased, the device will move from a uniformly distributed inversion layer directly into depletion. It is then possible to ignore an entire region of device operation, thus shortening the program considerably. This assumption also requires modification of the capacitance equations in the linear region. The equations are shown in Table I, where both the original and the modified expressions for the capacitances are shown for each region of the model.

The equations for the feedthrough charge increments resulting in each region are obtained in closed form. This was done by analytically integrating the capacitance equations. Each integral was then evaluated at the initial and final voltages to obtain the charge increment. This integration made it necessary to accept a discontinuity in the bulk depletion capacitance. This should not cause, however, any significant errors.

TABLE I(a)
MODEL EQUATIONS

Original Equations [7]	This Work
Inversion Region (Region 1)	
$C_{GS}=C_{OV}+\frac{2}{3}C_{OX}WL\frac{V_{SAT}[3V_{SAT}-2V_{DS}]}{[2V_{SAT}-V_{DS}]^2}$	$C_{GS}=C_{OV}+\frac{1}{2}C_{OX}WL$
$C_{GD}=C_{OV}+\frac{2}{3}C_{OX}WL\frac{[V_{SAT}-V_{DS}][3V_{SAT}-V_{DS}]}{[2V_{SAT}-V_{DS}]^2}$	$C_{GD}=C_{OV}+\frac{1}{2}C_{OX}WL$
$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}\left[1+\frac{2}{3}\frac{C_{BG}}{C_{JS}}WL\frac{V_{SAT}[3V_{SAT}-2V_{DS}]}{[2V_{SAT}-V_{DS}]^2}\right]$	$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}\left[1+\frac{1}{2}\frac{C_{BG}}{C_{JS}}WL\right]$
$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}\left[1+\frac{2}{3}\frac{C_{BG}}{C_{JD}}WL\frac{[V_{SAT}-V_{DS}][3V_{SAT}-V_{DS}]}{[2V_{SAT}-V_{DS}]^2}\right]$	$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}\left[1+\frac{1}{2}\frac{C_{BG}}{C_{JD}}WL\right]$

TABLE I(b)
MODEL EQUATIONS

Original Equations [7]	This Work
Depletion-Inversion Region (Region 2)	
$C_{GS}=C_{OV}+\frac{2}{3}C_{OX}WL\left[1-4[V_{GS}-V_{TH}]^2\right]$	$C_{GS}=C_{OV}+\frac{1}{2}C_{OX}WL\left[1-4[V_{GS}-V_{TH}]^2\right]$
$C_{GD}=C_{OV}$	$C_{GD}=C_{OV}+\frac{1}{2}C_{OX}WL\left[1-4[V_{GS}-V_{TH}]^2\right]$
$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}\left[1+\frac{2}{3}\frac{C_{BG}}{C_{JS}}WL\left[1-4[V_{GS}-V_{TH}]^2\right]\right]$	$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}$
$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$	$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$

B. Simulation Results

In Fig. 9, a plot of the feedthrough voltage predicted by the simulation for each region is shown. It should be noted that the backgate bias is -5 V, to minimize the feedthrough voltage. This constrains the device from ever actually entering the accumulation region, because the gate voltage reaches -5 V while the device is biased in the depletion–accumulation region. The linearity of the feedthrough voltage versus V_{AI} for each individual region confirms the dominance of the gate-to-source and gate-to-drain capacitances on the feedthrough process. Also, the feedthrough voltage from the inversion region dominates over the feedthrough voltage from other regions, which are not a strong function of V_{gs}. From this observation, one can reduce the problem of minimizing the feedthrough voltage to minimizing the stored inversion charge, or the use of feedthrough cancellation schemes [6].

Comparisons between the simulation and measurements are shown in Figs. 3, 5, 7, and 8, where the plots of the measured and calculated values of the feedthrough voltage are in qualitative agreement. The simulation overestimates the value of the feedthrough voltage. This is possibly due to ignoring the effects of charge pumping. The errors involved range from 15 to 40 percent in some cases. It is also interesting to note that the simulation results in a nearly linear dependence of the feedthrough voltage on each of the parameters studied. This further supports the conclusion that the feedthrough process is dominated by the inversion channel charge, as discussed earlier.

C. Discussion

Despite the discrepancy between measured results and simulations, charge-feedthrough modeling in this investiga-

TABLE I(c)
MODEL EQUATIONS

Original Equations [7]		This Work	
Accumulation-Depletion Region (Region 3)			
$C_{GS}=C_{OV}$	$C_{GD}=C_{OV}$	$C_{GS}=C_{OV}$	$C_{GD}=C_{OV}$
$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}$		$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}$	
$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$		$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$	
Accumulation Region (Region 4)			
$C_{GS}=C_{OV}$	$C_{GD}=C_{OV}$	$C_{GS}=C_{OV}$	$C_{GD}=C_{OV}$
$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}$		$C_{BS}=\frac{C_{JS}}{\left[1-\frac{V_{BS}}{\phi_B}\right]^{m_B}}$	
$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$		$C_{BD}=\frac{C_{JD}}{\left[1-\frac{V_{BD}}{\phi_B}\right]^{m_B}}$	

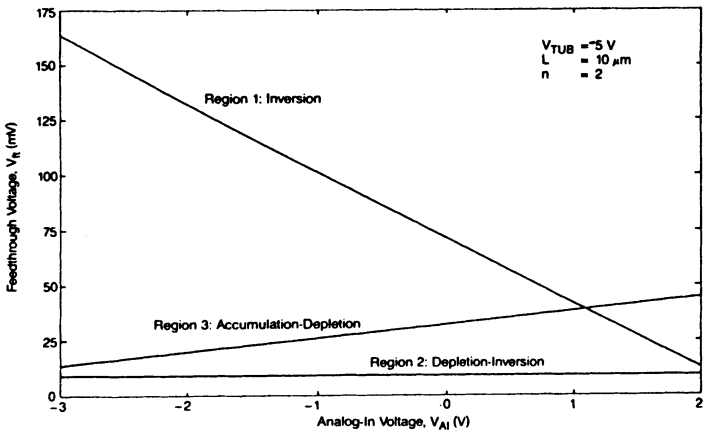


Fig. 9. Dependence of the regional feedthrough voltage V_{ft} on the analog-in voltage V_{AI}. Circles (experiment), solid lines (simulated).

tion offers the following enhancements over previous studies [3]–[5].

1) It allows for both the source and drain nodes to be floating after the switch is turned off. This allows measurement of all transient charges flowing from the source and drain, where previous efforts cannot account for the charge exiting at the node held at a fixed potential. The special case of $V_{TUB} = 0$ has been treated by Vittoz [5].

2) It is based on the quasi-static small-signal MOSFET capacitances and within the limits of this study offers good agreement with the measured results. This indicates the capability of quasi-static small-signal models to accurately simulate transient effects in a MOSFET under various doping and bias conditions.

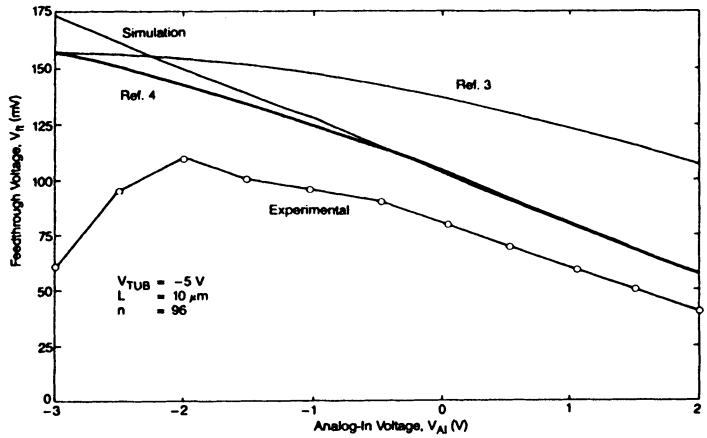


Fig. 10. Comparison of the simulation and experimental results obtained in this study with results from [3] and [4].

3) It results in the linear dependence of V_{ft} on V_{AI}, which was observed experimentally but not predicted by MacQuigg [3] (see Fig. 10).

4) It includes the influence of V_{TUB} on V_{ft}, which was measured experimentally. This was not accounted for in the lumped model of Sheu and Hu [4]. It should be remembered that although simulation results in this work and those obtained using the lumped model in [4] agree in Fig. 10, these two calculations of V_{ft} were made for different values of V_{TUB}.

5) It accounts for the junction capacitances associated with the source and drain diffusions.

It should be again noted, however, that charge-feedthrough modeling in this study did not include the influence of the fall time of the gate voltage, currently under investigation, or that of charge pumping on the feedthrough voltage.

The observed linearity of the feedthrough voltage implies that the MOS switch can be used effectively in analog design. If the transfer function of a switch is considered, it will have two terms. First, the unity gain of an ideal switch will yield the input voltage at the output. Second, the linear feedthrough term is added. Thus, the transfer function will remain linear, and the analog switch should introduce little harmonic distortion into the signal passing through the switch over a limited range of input voltages. These conclusions could only be made because the source–follower circuit has a constant gain of nearly unity over the active range of values for the input voltage and all measurements are made after the transients have settled as in a typical sample-and-hold application.

VI. CONCLUSION

This paper presents measurements and first-order modeling of charge feedthrough. The feedthrough voltage in an n-channel MOSFET was measured with respect to several parameters, including analog-in voltage, tub voltage, device dimensions, and load capacitances. It was found that the feedthrough voltage is nearly linear with respect to the

analog-in voltage. Thus a typical analog switch appears to introduce practically no harmonic distortion in the output signal over a limited input range. Also, the feedthrough voltage is minimized for a large back bias, concurrently allowing for a wide input voltage swing. Minimum device dimensions are desirable to reduce the feedthrough voltage. Also, it was observed for the first time that the magnitude of the load capacitance on one side of the device has virtually no effect on the amount of charge feedthrough at the other side. In other words, the amount of charge feedthrough is independent of the load capacitance. A simulation program that provides first-order qualitative agreement with the measured data, using a quasi-static MOSFET capacitance model, was developed.

Appendix

This appendix lists the relationships of the different charge components stored in an MOS switch [8]. The charge stored on the gate Q_{GATE} is given by

$$Q_{GATE} = C_{OX}\left[V_{GATE} - V_{fb} - 2\phi_F - V_{AI}\right] \quad (1)$$

where C_{OX} is the gate oxide capacitance, V_{GATE} the gate voltage, V_{fb} the flatband voltage, ϕ_F the bulk potential difference between the Fermi and intrinsic levels, and V_{AI} the analog-in voltage.

The bulk charge Q_{BULK} is given by

$$Q_{BULK} = -\sqrt{2\epsilon_{si} q N_A [\phi_s]} \quad (2)$$

where N_A is the substrate doping, and ϕ_s the band bending in the silicon, which is given by

$$\phi_s = 2\phi_F + V_{AI} - V_{TUB}. \quad (3)$$

The inversion charge is the difference of these two charge components and is given by

$$Q_{INV} = -Q_{GATE} - Q_{BULK}. \quad (4)$$

As V_{AI} increases, the amount of inversion charge Q_{INV} decreases as the bulk charge increases, and the charge on the gate decreases. This continues until the inversion channel disappears, and the bulk charge mirrors the gate charge. Now the source and drain do not affect the surface potential and, therefore, do not influence the amount of charge stored underneath the gate of the device. Consequently, Q_{BULK} and Q_{GATE} will not change with continued increase in V_{AI}, where the gate and substrate voltages are held constant. This is shown graphically in Fig. 4, and is the cause of the nonlinearity in Fig. 3 for large values of V_{AI}. Equations (1)–(3) and (4) are used in Fig. 6 where the tub voltage is varied. Here, the effect is to shift the relative distribution of charge between the inversion channel and the bulk charge, where the total amount of charge stored in the silicon remains unchanged as the tub voltage is varied.

Acknowledgment

The authors would like to thank Dr. J. J. Paulos of North Carolina State University for many valuable discussions, and the reviewers for their comments.

References

[1] J. S. Brugler and P. G. A. Jespers, "Charge pumping in MOS devices," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 297–302, Mar. 1969.
[2] K. R. Stafford, P. R. Gray, and R. A. Blanchard, "A complete monolithic sample/hold amplifier," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 381–387, Dec. 1974.
[3] D. MacQuigg, "Residual charge on a switched capacitor," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 811–813, Dec. 1983.
[4] B. J. Sheu and C. Hu, "Switched-induced error voltage on a switched capacitor," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 519–525, Aug. 1984.
[5] E. Vittoz, "Microwatt switched capacitor circuit design," *Electrocomponent Sci. Technol.*, vol. 9, pp. 263–273, 1981.
[6] E. Suarez, P. Gray, and D. A. Hodges, "All-MOS charge redistribution analog-to-digital conversion techniques—Part II," *IEEE J. Solid-State Circuits*, vol. SC-10, p. 379, 1975.
[7] S. Liu and L. W. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 983–998, Dec. 1982.
[8] A. S. Grove, *Physics and Technology of Semiconductor Devices*. New York: Wiley, 1969.

Measurement and Analysis of Charge Injection in MOS Analog Switches

JE-HURN SHIEH, STUDENT MEMBER, IEEE, MAHESH PATIL, STUDENT MEMBER, IEEE, AND BING J. SHEU, MEMBER, IEEE

Abstract—Charge injection in MOS switches has been analyzed. The analysis has been extended to the general case including signal-source resistance and capacitance. Universal plots of percentage channel charge injected are presented. Normalized variables are used to facilitate usage of the plots. The effects of gate voltage falling rate, signal-source level, substrate doping, substrate bias, switch dimensions, as well as the source and holding capacitances are all included in the plots. A small-geometry switch, slow switching rate, and small source resistance can reduce the charge injection effect. On-chip test circuitry with a unity-gain operational amplifier, which reduces the disturbance imposed by measurement equipment to a minimum, is found to be an excellent monitor of the switch charge injection. The theoretical results agree with the experimental data.

I. Introduction

IN A monolithic sample and hold, a signal is stored on a capacitor. The accuracy of sample-and-hold circuits is disturbed by charge injected when the sampling switch turns off. The majority of sample-and-hold circuits are implemented using MOS technologies because the high input impedance of MOS devices performs excellent holding function. When the switch connecting the signal-source node and the data-storage node is turned on, the sampling function is performed. When the switch is turned off, the data stored in the storage node will be held until the next operation step occurs. However, an MOS switch is not an ideal switch. A finite amount of mobile carriers are stored in the channel when an MOS transistor conducts. When the transistor turns off, the channel charge exits through the source, the drain, and the substrate electrodes. The charge transferred to the data node during the switch turning-off period superposes an error component to the sampled voltage. In addition to the charge from the intrinsic channel, the charge associated with the feedthrough effect of the gate-to-diffusion overlap capacitance also enlarges the error voltage after the switch turns off [1]. This charge injection problem was identified in the early stage of switched-capacitor circuit development. Various compensation schemes [2],[3] have been used to reduce the switch-induced error voltage. As the design of higher precision sample-and-hold circuits progresses, the need for effective test patterns to accurately monitor the switch charge injection becomes increasingly important. In 5-V technologies, the resolution in a 10-bit analog-to-digital (A/D) converter is 4.88 mV, while the resolution in a 16-bit A/D converter is only 76 μV. The error voltage caused by the switch charge injection is usually in the millivolt range. The fully differential circuit approach [3] cancels the switch charge injection to the first order. Precise characterization and detailed analysis of the switch charge injection is of prime interest in designing high-performance integrated circuits.

There have been some attempts to model the switch charge injection. MacQuigg [4] made a qualitative observation and did SPICE simulation on a simplified case. Sheu and Hu [1] developed an analytical model corresponding to infinite source capacitance. A two-transistor source follower was used by Wilson *et al.* [5] with an attempt to improve the measurement accuracy. In this paper, analysis on the general case of switch charge injection is described in Section II. Analytical models of special cases are also presented. A better test structure for monitoring switch charge injection is proposed in Section III. Experimental results are presented in Section IV. A conclusion is given in Section V.

II. Analysis

We assume that the charge pumping phenomenon due to the capture of channel charge by the interface traps is insignificant and all the channel charge exits through the source and drain electrodes when the transistor turns off. The turn-off of an MOS switch consists of two distinct phases. During the first phase, the gate voltage is higher than the transistor threshold voltage. There is a conduction channel that extends from the source to the drain of the transistor. As the gate voltage decreases, mobile carriers exit through both the drain end and the source end and the channel conduction decreases. During the second phase, the gate voltage is below the transistor threshold voltage and the conduction channel does not exist any more. The coupling between the gate and the data-holding node is merely through the gate-to-diffusion overlap capacitance. In our analysis, attention is focused on the switch charge

Manuscript received August 18, 1986; revised October 13, 1986. This work was supported by the Defense Advanced Research Projects Agency under Contract MDA903-81-C-0335.

The authors are with the Department of Electrical Engineering and Information Sciences Institute, University of Southern California, Los Angeles, CA 90089.

IEEE Log Number 8613218.

Reprinted from *IEEE J. Solid-State Circuits*, vol. SC-22, no. 2, pp. 277–281, Apr. 1987.

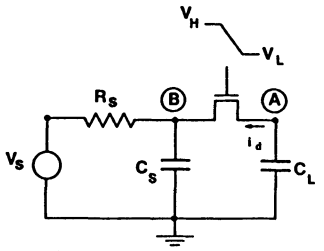


Fig. 1. Circuit for analysis of switch charge injection.

injection due to the first phase of the switch turn-off. The circuit schematic corresponding to the general case of switch charge injection is shown in Fig. 1. Capacitance C_L is the lumped capacitance at the data-holding node. Resistance R_S could be the output resistance of an operational amplifier, while capacitance C_S could be the lumped capacitance associated with the amplifier output node.

Let C_G represent the total gate capacitance, including both the channel capacitance and gate-to-source/gate-to-drain overlap capacitances:

$$C_G = WLC_0 + C_{0vs} + C_{0vd}. \tag{1}$$

By following the derivation presented in [1], Kirchkoff's current law at node A and node B requires

$$C_L \frac{dv_L}{dt} = -i_d + \frac{C_G}{2}\frac{d(V_G - v_L)}{dt} \tag{2}$$

and

$$\frac{v_S}{R_S} + C_S\frac{dv_S}{dt} = i_d + \frac{C_G}{2}\frac{d(V_G - v_S)}{dt} \tag{3}$$

where v_L and v_S are the error voltages at the data-holding node and the signal-source node, respectively. Gate voltage is assumed to decrease linearly with time from the ON value V_H:

$$V_G = V_H - Ut \tag{4}$$

where U is the falling rate. When the transistor is operated in the strong inversion region

$$i_d = \beta(V_{HT} - U)(v_L - v_S) \tag{5}$$

where

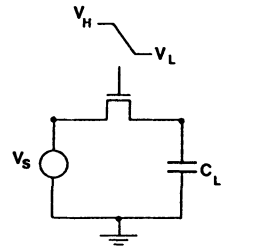
$$\beta = \mu C_0 \frac{W}{L}$$

and

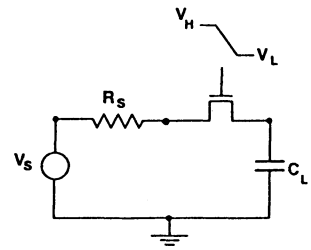
$$V_{HT} = V_H - V_S - V_{TE}. \tag{6}$$

Here V_{TE} is the transistor effective threshold voltage including the body effect. For small-geometry transistors, narrow- and short-channel effects should be considered in determining the V_{TE} value. Under the condition $|dV_G/dt| \gg |dv_L/dt|$ and $|dv_S/dt|$, (2) and (3) simplify to

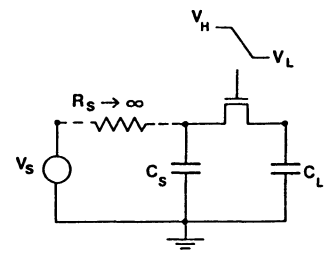
$$C_L\frac{dv_L}{dt} = -\beta(V_{HT} - Ut)(v_L - v_S) - \frac{C_G}{2}U \tag{7}$$



(a)



(b)



(c)

Fig. 2. Special cases of switch charge injection. (a) No source resistance and capacitance. (b) No source capacitance. (c) Infinitely large source resistance.

and

$$\frac{v_S}{R_S} + C_S\frac{dv_S}{dt} = \beta(V_{HT} - Ut)(v_L - v_S) + \frac{C_G}{2}U. \tag{8}$$

No closed-form solution to this set of equations can be found. Numerical integration can be employed to find the results. Analytical solutions to special cases are given below.

Fig. 2(a) shows the case with only a voltage source at the signal-source node. Since $C_S \gg C_L$, v_S can be approximated as zero and the governing equation reduces to

$$C_L\frac{dv_L}{dt} = -\beta(V_{HT} - Ut)v_L - \frac{C_G}{2}U. \tag{9}$$

When the gate voltage reaches the threshold condition, the error voltage at the data-holding node is

$$v_L = -\sqrt{\frac{\pi UC_L}{2\beta}}\left(\frac{C_G}{2C_L}\right)\operatorname{erf}\left(\sqrt{\frac{\beta}{2UC_L}}V_{HT}\right). \tag{10}$$

Another special case is when the source capacitance is negligibly small, as is shown in Fig. 2(b). The governing

equations reduce to

$$C_L\frac{dv_L}{dt} = -\beta(V_{HT}-Ut)(v_L-v_S)-\frac{C_G}{2}U \quad (11)$$

and

$$\frac{v_S}{R_S} = \beta(V_{HT}-Ut)(v_L-v_S)+\frac{C_G}{2}U. \quad (12)$$

When the gate voltage reaches the threshold condition, (5) breaks down and the error voltage at the data-holding node is

$$v_L = -\frac{UC_G}{2C_L}\exp\left(-\frac{V_{HT}}{UC_LR_S}\right)\cdot\int_0^{V_{HT}/U}[\beta R_S(V_{HT}-U\xi)+1]^{1/C_L\beta R_S^2U}\cdot\exp\left(\frac{\xi}{C_LR_S}\right)\left(2-\frac{1}{1+\beta R_S(V_{HT}-U\xi)}\right)d\xi. \quad (13)$$

If the time constant R_SC_S is much larger than the switch turn-off time, then the channel charge will be shared between C_S and C_L, as is shown in Fig. 2(c). For the case of a symmetrical transistor and $C_S = C_L$, half of the channel charge will be deposited to each capacitor. Otherwise the following equations can be used to find out the results:

$$C_L\frac{dv_L}{dt} = -\beta(V_{HT}-Ut)(v_L-v_S)-\frac{C_G}{2}U \quad (14)$$

and

$$C_S\frac{dv_S}{dt} = \beta(V_{HT}-Ut)(v_L-v_S)+\frac{C_G}{2}U. \quad (15)$$

We now multiply (15) by the ratio C_L/C_S, and then subtract the result from (14), to obtain

$$C_L\frac{d(v_L-v_S)}{dt} = -\beta(V_{HT}-Ut)\left[1+\frac{C_L}{C_S}\right](v_L-v_S) - \frac{UC_G}{2}\left(1-\frac{C_L}{C_S}\right). \quad (16)$$

When the gate voltage reaches the threshold condition, the amount of voltage difference between the data-holding node and the signal-source node is

$$v_L - v_S = -\sqrt{\frac{\pi UC_L}{2\beta(1+C_L/C_S)}}\left(\frac{C_G(1-C_L/C_S)}{2C_L}\right)\cdot\operatorname{erf}\left(\sqrt{\frac{\beta(1+C_L/C_S)}{2UC_L}}\,V_{HT}\right). \quad (17)$$

Fig. 3 shows the calculated percentage of channel charge injected to the data-holding node when the source resistance is infinitely large. Similar plots were obtained by numerical integration after some special transformation of the problem [6]. The dimensionless quantity $V_{HT}\sqrt{\beta/UC_L}$ has been identified as the driving force of the switch charge injection effect. It has the same functional dependence as the argument of the error function in (10). A family of curves corresponding to various C_S/C_L ratios have been plotted. When the switch turns off, the channel charge exits to the signal-source node and the data-holding node under capacitive coupling and resistive conduction. In the fast switching-off conditions, the transistor conduction channel disappears very quickly. There is not enough time for the charge at the signal-source side and the charge at the data-holding side to communicate. Hence, the percentage of charge injected into the data-holding node approaches 50 percent independent of the C_S/C_L ratio. In the slow switching-off conditions, the communication between the charge at the signal-source side and the charge at the data-holding side is so strong that it tends to make the final voltages at both sides equal. This allows the majority of channel charge to go to the node with larger capacitance.

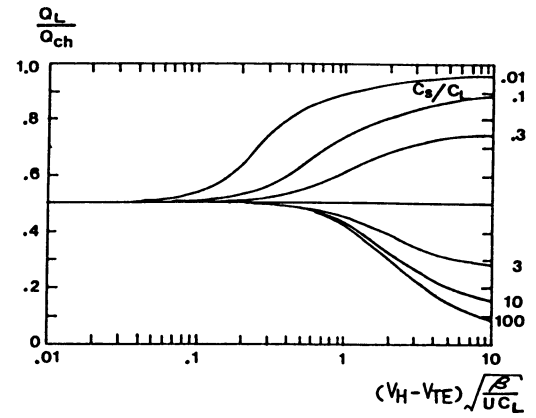


Fig. 3. Percentage of channel charge injected to the data-holding node. Source resistance is assumed to be infinitely large. A family of curves corresponding to various C_S/C_L ratios has been plotted.

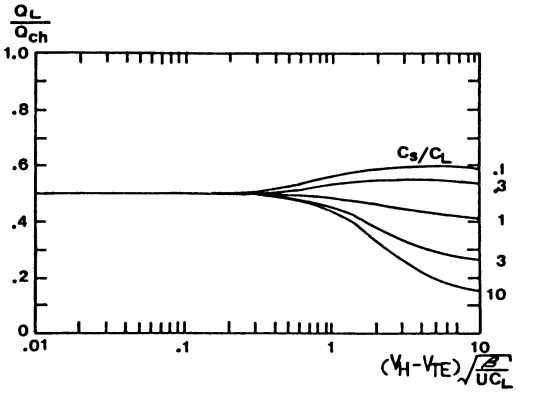


Fig. 4. Percentage of channel charge injected to the data-holding node with $V_{HT}/UR_SC_S = 1$.

Another important factor in switch charge injection is the relative magnitude of the falling rate compared with the signal time constant R_SC_S. The curves corresponding to two different V_{HT}/UR_SC_S values are shown in Figs. 4 and 5. Source resistance effectively offers a leakage path

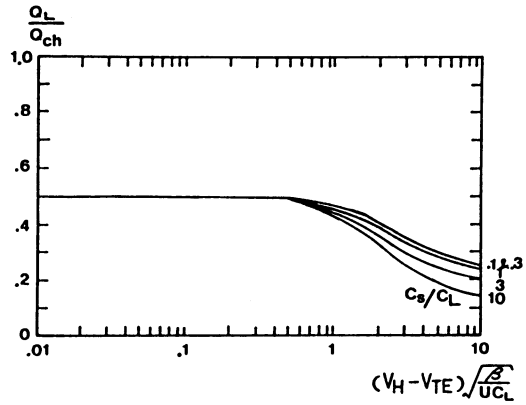


Fig. 5. Percentage of channel charge injected to the data-holding node with $V_{HT}/UR_SC_S = 5$.

for the channel charge during the switch turn-off period. Hence, a small source resistance will greatly reduce the amount of charge injected to the data-holding node.

III. Measurement

The holding capacitor is usually chosen around or above 1 pF to minimize the thermal noise voltage. Direct measurement of switch charge injection using single transistors has severe limitations. The stray capacitance of the equipment probe alters the capacitance at the interested node. When the gate voltage falling rate is high, the probe capacitance and inductance greatly perturbs measurement accuracy. On-chip circuitry can be used to circumvent the problem. It offers good buffering between the interested node and the measurement equipment. The insertion of a two-transistor source follower between the interested node and the external probe, as used by Wilson *et al.* [5], achieves the buffering function to the first order. However, a two-transistor source follower has nonlinear voltage characteristics and limited driving capability. Fig. 6 shows the two-transistor source-follower test configuration.

The unity-gain operational amplifier is found to be a better monitor of the switch charge injection. The output of the amplifier precisely tracks the input voltage. The amplifier possesses an excellent driving capability to interface with the measurement equipment. The unity-gain op-amp test configuration is shown in Fig. 7.

The circuit schematic of the operational amplifier used in the studies is shown in Fig. 8. The operational amplifier is a conventional two-stage design with a source-follower output stage [7]. It is similar to the amplifier used in the on-chip capacitance measurement of MOS transistors in some respects [8]. This circuit configuration provides good common-mode range, output swing, voltage gain, and common-mode rejection ratio. Transistors $M1-M3$ are p-channel current sources. The input stage consists of $M5-M8$. They are a p-channel differential pair with n-channel active loads and double-to-single-ended conversion. Transistors $M11$ and $M12$ are the dummy biasing string for the tracking compensation scheme and also offer dc bias to the output stage. Transistors $M4$ and $M10$ form

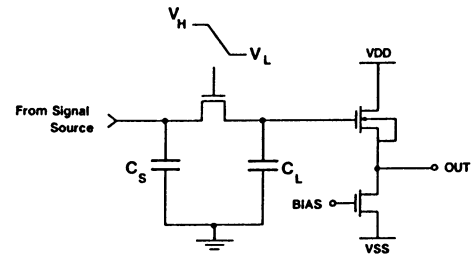


Fig. 6. A two-stage source-follower measurement approach.

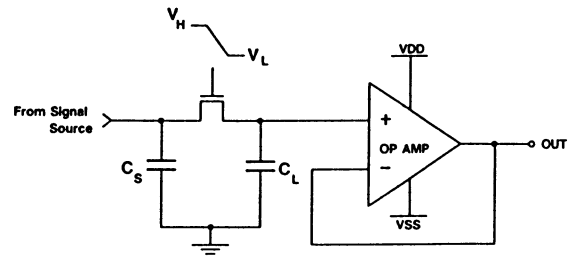


Fig. 7. A unity-gain operational-amplifier measurement approach.

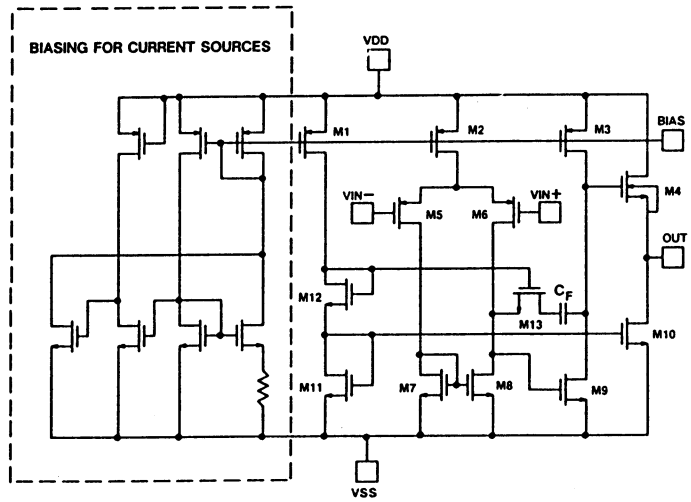


Fig. 8. Schematic of an operational amplifier suitable for charge injection monitoring.

a class-A output stage. A pole-splitting capacitor is used to compensate for the frequency response. If a fabrication process is primarily used for digital circuits and high-quality capacitors are not readily available, a thin-gate transistor can be connected to supply the necessary capacitance. Since the input-referred noise is inversely proportional to the size of the input devices, large-geometry transistors with $W/L = 99\ \mu m/6\ \mu m$ are used to keep the input-referred noise small. Notice that the substrate and source terminals of the output transistor $M4$ are connected together to eliminate the body effect. This configuration improves the amplifier output range. If an n-well process is used instead of a p-well process, then the output transistors would be changed to p-channel transistors because the transistor inside a well can have its substrate and source tied together. The bias of the current sources can be derived in two ways. A dedicated biasing circuit can be used. The other alternative is to apply an external bias to the pad BIAS. The latter approach turns out to be a good

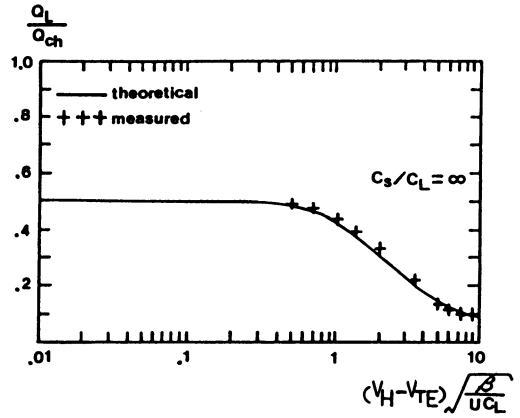


Fig. 9. Comparison of measured and theoretical charge injection results for the special case of Fig. 2(a). The gate voltage falling rate was varied in the experiments.

choice in the application because it reduces the size of the whole test pattern without sacrificing any measurement accuracy. The dotted portion in Fig. 8 denotes the optional biasing block for the current sources.

IV. Experimental Results

The transistors used in the experiments were fabricated using a 3-μm CMOS process. The transistor gate-oxide thickness is 50.0 nm, substrate doping is 10^{16} cm^{-3}, and zero-bias threshold voltage is 0.9 V. Percentage charge injection was measured against the gate voltage falling rate ranging from 1.25×10^6 to 5×10^8 V/s. Fig. 9 shows the measured data and theoretical results. Good agreement between the theoretical results and experimental data is found.

V. Conclusion

Charge injection in MOS switches has been analyzed. The analysis has been extended to the general case which includes signal-source resistance and capacitance. This extension makes the results useful for the various conditions encountered in integrated-circuit applications. Plots of the percentage charge injection corresponding to various normalized parameters are presented. The source resistance effectively offers a leakage path for the channel charge during the switch turn-off period. On-chip test circuitry with a unity-gain operational amplifier, which reduces the disturbance imposed by the measurement equipment to a minimum, is found to be an excellent monitor of switch charge injection.

Acknowledgment

The authors wish to thank Prof. P. R. Gray of the University of California, Berkeley for his suggestion of the test patterns and the anonymous reviewers for their valuable suggestions. Generous support from G. Lewicki, V. Tyree, and the MOSIS group is highly appreciated. Discussions with J. Tzeng and K.-Y. Toh were beneficial.

References

[1] B. J. Sheu and C. Hu, "Switched-induced error voltage on a switched capacitor," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 519–525, Aug. 1984.

[2] R. E. Suarez, P. R. Gray, and D. A. Hodges, "All-MOS charge redistribution analog-to-digital conversion techniques: Part II," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 379–385, Dec. 1975.

[3] R. C. Yen and P. R. Gray, "An MOS switched-capacitor instrumentation amplifier," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1008–1013, Dec. 1982.

[4] D. MacQuigg, "Residual charge on a switched capacitor," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 811–813, Dec. 1983.

[5] W. B. Wilson, H. Z. Massoud, E. J. Swanson, R. T. George, Jr., and R. B. Fair, "Measurement and modeling of charge feedthrough in n-channel MOS analog switches," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1206–1213, Dec. 1985.

[6] E. Vittoz, "Microwatt switched capacitor circuit design," *Electrocomponent Sci. and Technol.*, vol. 9, no. 4, pp. 263–273, 1982.

[7] P. R. Gray and R. G. Meyer, *Analysis and Design of Digital Integrated Circuits*, 2nd ed. New York: Wiley, 1984.

[8] J. J. Paulos and D. A. Antoniadis, "Measurement of minimum-geometry MOS transistor capacitances," *IEEE Trans. Electron Devices*, vol. ED-32, no. 2, pp. 357–363, Feb. 1985.