A Clock-Based Analog Memory Element for Integrated Circuits

by

Micah G. O'Halloran

B.S., Electrical and Computer Engineering (2000)

University of Florida

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Electrical Engineering

at the

Massachusetts Institute of Technology

September 2002

Signature of Author.....
Department of Electrical Engineering and Computer Science
July 31, 2002

Certified by.....
Rahul Sarpeshkar
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by.....
Arthur C. Smith
Chairman, Department Committee on Graduate Students
Department of Electrical Engineering and Computer Science

A Clock-Based Analog Memory Element for Integrated Circuits

by

Micah G. O'Halloran

Submitted to the Department of Electrical Engineering and Computer Science
On August 9, 2002 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in
Electrical Engineering

ABSTRACT

A clock-based signal restoration scheme was developed as a method of implementing an analog memory cell for integrated circuits (ICs). The technique is similar to single-slope analog-to-digital (A/D) conversion, however, since the goal of the memory is to store an analog, rather than digital voltage, the digital memory circuits and the digital-to-analog (D/A) conversion circuits are not needed. An analog memory cell implementing this clock-based signal restoration algorithm was designed and fabricated using the MOSIS AMI 1.5μm CMOS process.

Experimental results show that the memory element possesses 8-bits of resolution from input to output, with a 2.8V input range when powered from a 3.3V rail. The experiments also show that the memory element is capable of storing an analog value with 8-bits of precision for over 2.5 hours. Storage at 8-bits of precision for indefinite lengths of time should be possible with the addition of more stable on-chip biasing circuitry and the use of better power-supply-rejection techniques in the ramp generator construction. The fabricated memory cell size on the AMI 1.5μm process is 300μm × 250μm (375λ × 310λ).

Thesis Supervisor: Rahul Sarpeshkar
Title: Assistant Professor of Electrical Engineering

Table of Figures

Table of Contents

1 Introduction to Analog Storage

While short-term analog storage is omnipresent in today's analog integrated circuits (ICs), the use of medium-term analog storage has been very limited. In fact, the sole application in which medium-term analog storage has found any popularity at all is in storing weight information in hardware neural-network implementations. The lack of use of this technique outside of the neural-network community is not due to a shortage of on-chip storage needs, but is due to the susceptibility of analog storage techniques to disturbances – both stochastic and deterministic. Digital storage is the obvious solution to this noise immunity issue. However, this method requires an analog-to-digital (A/D) converter to form a multi-bit representation of the analog value we wish to store, and a digital-to-analog (D/A) converter to reconstruct a quantized version of the analog input from the multiple stored bits. These circuits tend to be both area and power hungry, and a storage solution that does not require these blocks may be advantageous. The purpose of this research is to develop a compact, high-resolution, low-power analog memory cell utilizing area and power efficient strategies for coping with on-chip disturbances that can be used for medium-term analog storage. The cell will be used to create self-tuning analog circuits that are able to monitor and improve their own performance.

1.1 Approaches to Analog Storage

All successful integrated analog memories have taken one of two approaches to avoid corruption by outside disturbances. The first approach is to reduce the disturbance to such a value that its effect is negligible on the stored value for the desired length of storage time (application dependent). This is the strategy used in floating gate storage techniques, and technically, the same one used in simple sampled-capacitor techniques. The alternate approach is to quantize the analog information to a finite precision, and store the quantized value rather than the true analog value. An output close to, but usually not exactly the same as, the original input can be re-created from the quantized representation. The latter technique offers a distinct advantage over the former in that a finite amount of noise can be tolerated without loss of information. This is the approach used in analog-to-digital-to-analog (A/D/A) converters, which store the quantization information across many channels, and is the same approach used in a class of quantized-analog storage circuits that store the quantization information on a single channel.

1.2 Past Implementations of Analog Storage on ICs

Let's now look at how the two approaches of the previous section are applied in various analog memory topologies in the literature. Throughout these discussions, we will assume that the voltage we wish to store, V_{in}, is a DC signal. This assumption is valid for the applications in which we would use these circuits.

1.2.1 Simple Capacitive Storage

The simplest form of analog storage is capacitive storage, as shown in Figure 1.1. In this circuit, a metal-oxide-semiconductor field effect transistor (MOSFET) is used as a

switch to sample the value of V_{in} onto a sampling capacitor. This type of storage is used widely in switched-capacitor circuit design for short-term storage of signals and offset information.

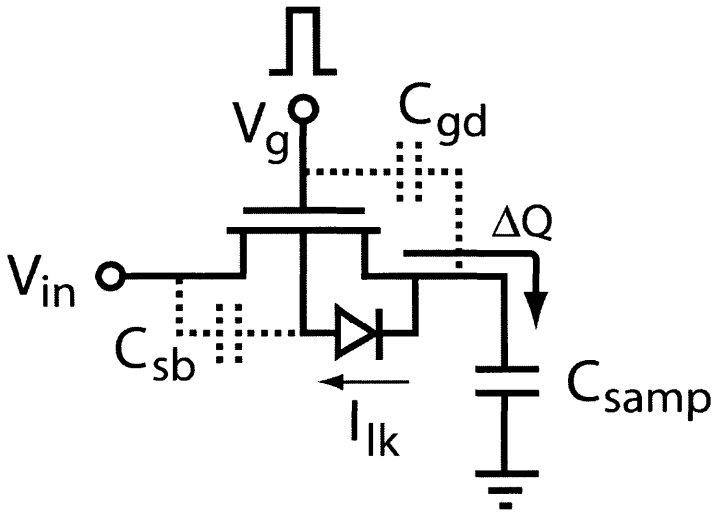


Figure 1.1-- Capacitive storage is the simplest form of analog storage. In this technique, a MOS transistor is used as a switching element to connect and disconnect the signal source V_{in} from the capacitor C_{samp}. The value of V_{in} at the time the transistor turns off is held on C_{samp}.

There are several sources of error in this circuit that cause the voltage on C_{samp} to differ from V_{in} after the switch opens. Since these errors appear in nearly all storage circuits, let's examine their behavior carefully for this simple circuit.

The first error is due to the charge injection, ΔQ, from the collapsing MOS conduction channel. It is known that the total MOS channel charge, Q_{tot}, in above-threshold operation is described by

$$Q_{tot} = WLC_{ox}(V_{gs} - V_t) \tag{1.1}$$

where W is the transistor width, L is the transistor length, C_{ox} is the gate capacitance per unit area, V_{gs} is the gate-to-source voltage of the transistor, and V_t is the transistor's

threshold voltage. Since V_{gs} is a function of V_{in}, the total channel charge is a nonlinear function of V_{in}. ΔQ is the portion of Q_{tot} that exits the transistor through its drain during turnoff. It is typically assumed that ½ of the total channel charge exits through the drain, as long as V_g switches quickly [1]. However, it seems more likely that the separation proportions are a strong function of the impedances seen looking into the source and the drain from the channel. Since C_{sb} and C_{db}, the source-to-bulk and drain-to-bulk capacitances, vary nonlinearly as V_{in} changes, the separation proportions should also vary nonlinearly with V_{in}.

A known technique that helps to reduce the magnitude of ΔQ to essentially zero is to decrease the slope of V_g's turnoff transition to such a point that equilibrium between the source and drain voltages is constantly maintained during turnoff, until $V_{gs} - V_t = 0V$. Using this technique, all of the channel charge can be forced out of the source. However, this technique is usually too slow in practice. There have also been attempts at developing a precise analytical solution to the value of ΔQ in the literature for fast turnoff. However, proper circuit design can minimize the effects of ΔQ without actually knowing its value.

The second source of error in this sampling circuit is due to the overlap capacitance C_{gd} coupling the gate signal V_g to the held voltage through the capacitive divider it forms with C_{samp}. The voltage change ΔV on C_{samp} can be described by

$$\Delta V = \frac{C_{gd}}{C_{gd} + C_{db} + C_{samp}} \times \Delta V_g \tag{1.2}$$

where ΔV_g is the voltage movement of V_g during turnoff, assuming fast switching transitions. The magnitude of this effect obviously decreases as C_{gd} is made smaller and

C_{samp} is made larger. This effect also decreases if the switching transition is slowed, for the same reason ΔQ's effect is decreased. Finally, ΔV is not a constant due to C_{db}'s nonlinear dependence on V_{in}.

The final major source of error in this circuit is due to I_{lk}, the leakage current through the drain-to-bulk reverse-biased diode. I_{lk} can be effectively modeled as a large resistor between C_{samp} and ground. This leakage causes the stored voltage to drift with time, and is the major reason that the simple capacitive storage technique is useful for only short-term storage.

1.2.2 Low-Leakage Capacitive Storage with constant ΔQ injection

In traditional sample-and-hold circuits, the nonlinear dependence of ΔQ is difficult to remove, but a constant ΔQ can be easily counteracted. The circuit of Figure 1.2 ensures a constant charge injection ΔQ regardless of the value of V_{in} [2]. In this circuit, V_{in} is applied at the negative terminal of the left op-amp. When the switching transistor is closed, the feedback loop requires that the output of the right op-amp equal V_{in}. In order to accomplish this, charge is integrated on C_{samp} until this requirement has been met. After the loop has settled, no current flows through the switching transistor, and its source and drain are both grounded due to the virtual ground created by the right op-amp, regardless of the value of V_{in}. Therefore, as long as the loop is allowed to fully settle, the amount of charge injected into the virtual ground, ΔQ, will be independent of V_{in}. The capacitive coupling term is still present in this topology, but it is also constant now, completely independent of V_{in}. Also, I_{lk} is minimized in this circuit. The switching transistor's bulk is connected to ground, and during hold mode so is the negative terminal of the right op-amp. I_{lk} is zero in this ideal case, and the voltage on C_{samp} would not

change with time. In reality, the op-amp offset voltages cause the leakage to be nonzero, but it is still reduced in this topology.

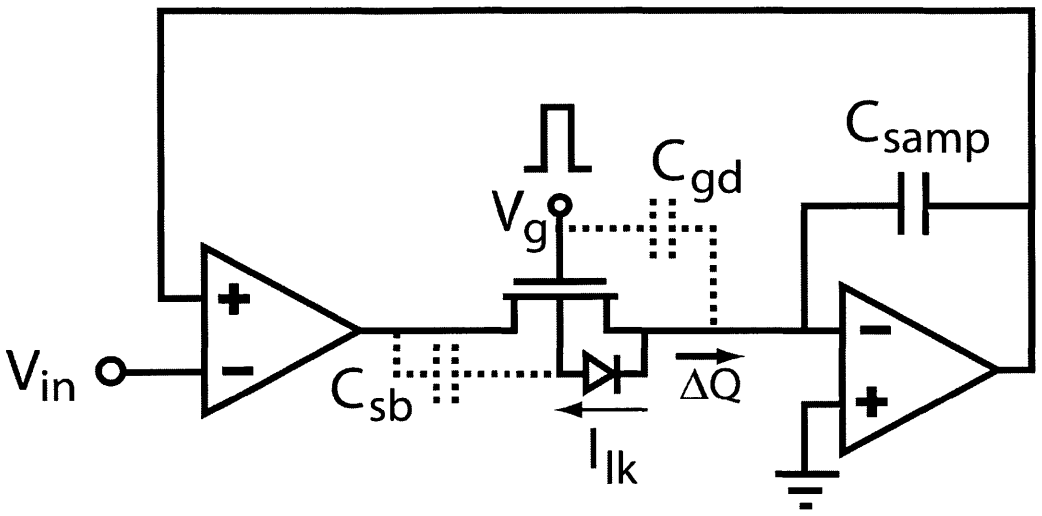


Figure 1.2– A possible way of implementing low-leakage capacitive storage with constant charge injection. In this circuit, the input voltage V_{in} drives the negative terminal of the left op-amp. When the transistor is switched on, the output voltage of the second op-amp must integrate current on C_{samp} until the voltage on C_{samp} is equal to V_{in}. Notice that after the circuit settles, both the source and drain of the switching transistor are at ground, because of the virtual ground established by the second op-amp at this node. The charge injection ΔQ is still nonzero in this circuit, however it is now a constant value independent of the value of V_{in}.

The nonlinearities of the simple capacitive storage circuit have been removed by this new topology at the expense of the speed of operation – the loop must be kept stable during sampling, and thus bandwidth must be sacrificed.

1.2.3 Floating Gate Storage

Floating-gate storage offers several advantages over simple capacitive storage. First, the leakage currents in floating gate cells are many orders of magnitude smaller than the leakage currents through reverse-biased P/N junctions, thus the storage time is increased by orders of magnitude for the same level of precision. Second, floating-gate

cells are nonvolatile – they don't lose their information when the power is turned off. Finally, floating gate cells can be just as compact as capacitive storage cells.

A typical floating-gate transistor is shown in Figure 1.3. The transistor is different from a conventional transistor in two ways. The first difference is that there are

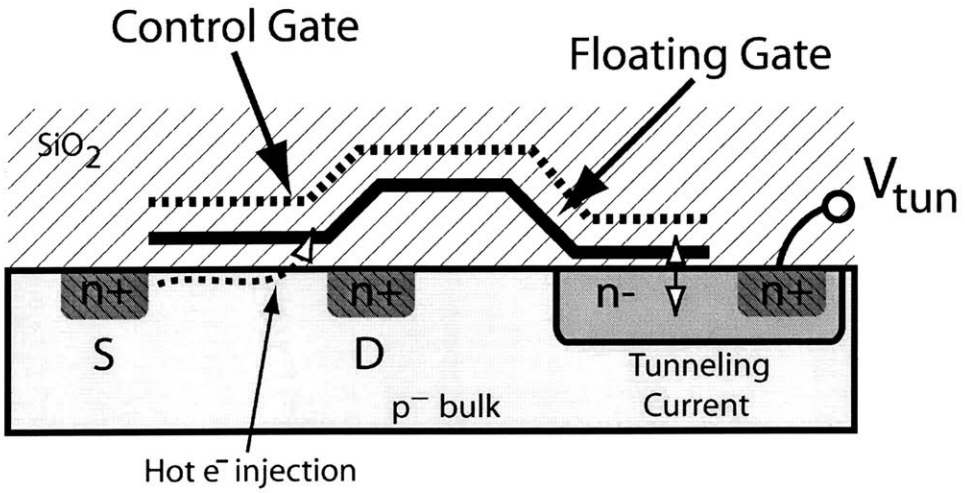


Figure 1.3-- Overview of floating-gate transistor operation. Two different mechanisms are typically used to transfer charge onto and off of a floating gate. Hot-electron injection occurs when electrons are accelerated to such a high speed at the drain of a transistor that they can pass through the gate oxide energy barrier and enter the floating gate. Tunneling occurs when a large positive voltage difference is created between the V_{tun} terminal and the control gate, causing electrons to tunnel from the floating gate through the thin gate oxide to the n^- region. The disadvantage of these techniques is that they require large voltages to establish reasonable current levels, gate oxide trapping slowly degrades the efficiency of current injection as the cell is exercised, and the current levels are ill-defined, requiring special tuning methods to achieve high-resolution floating gate voltages.

two gates instead of one. The first is the floating gate, which is completely isolated inside a layer of SiO_2, and makes no connections to any other layers; the second is the control gate, which is used in a manner similar to a normal transistor gate. The second difference between this floating gate transistor and a conventional transistor is the tunneling structure, which is typically used to tunnel electrons from the floating gate

(occasionally it is also used to tunnel electrons onto the floating gate). The underlying principle of this transistor structure that is exploited in its use as a memory element is that the floating gate controls the transistor exactly like a normal gate does. Unlike a normal gate, however, a DC voltage can be established on the floating gate and left for long periods of time without leakage corrupting it significantly – hence its memory property.

By incorporating this transistor into a conventional negative feedback amplifier structure, as shown in Figure 1.4, a memory cell capable of driving loads is created [3]. The cell consists of a PMOS current source and a floating-gate NMOS which together

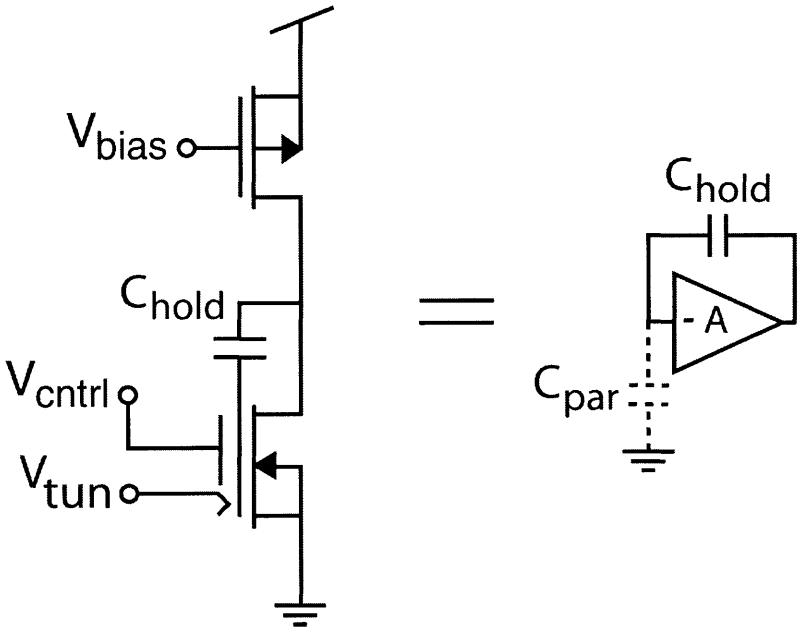


Figure 1.4-- A floating-gate NMOS transistor and a PMOS current source are used to create a common-source amplifier, with C_{hold} in the feedback path. A simple representation of the circuit is shown on the right. Through the impedance divider formed by C_{hold} and C_{par}, the input voltage is pegged. Any charge added or removed from the input floating node is immediately counteracted by a movement of the output of the amp. Thus, this structure acts as an analog storage cell capable of driving loads.

form a common-source amplifier, along with the feedback capacitor C_{hold}, which connects the output of the amplifier with its input. A block representation of what has

been created is shown in the right half of the figure. The feedback capacitor C_{hold} and the parasitic capacitance C_{par} form an impedance divider in the feedback path of a negative gain amplifier. The circuit operates like an op-amp with resistive feedback around it, except that with the capacitive feedback there is no DC path to the input terminal of the amplifier. Suppose the capacitor initially has 0V across it. When the amplifier is powered on, the input and output of the amplifier will settle to the voltage necessary such that the NMOS carries the same current as the PMOS transistor, due to the negative feedback. Now, using the floating gate structure, electrons can be added to or removed from the input node of the amplifier. As long as the gain of the amplifier is large, the DC operating point of the amplifier input will not move during this process. Instead, the output of the amplifier will move in the proper direction such that, through the capacitive divider, it forces the input voltage to remain pegged. Thus the output of the amplifier can be set to any voltage within V_{DSAT} (= V_{gs}-V_t for an above-threshold MOS) of the rails, and that voltage will be held virtually leakage-free.

Despite the numerous advantages of floating gates as analog storage elements, they possess several key disadvantages [4]. First, the injection and tunneling currents used to add and remove charge from the floating gate are ill-defined as a function of the control variables. The only way of achieving accurate analog voltages at the output of the amplifier is through an iterative read/write process, which requires support circuitry and is typically very slow (on the order of one second for high accuracies). Also, in order to achieve tunneling or injection, voltages much higher than the standard rail voltages must be applied to the tunneling region and control gate. These voltages must be either supplied from off chip or created on chip, and special MOS transistors capable of

switching large voltages must be included. Finally, the oxide through which tunneling is achieved degrades over time, causing higher leakage currents and lower intentional currents the more it is stressed.

1.2.4 Analog-to-Digital-to-Analog Storage

The most robust method of storing an analog variable for long time spans is through analog-to-digital-to-analog (A/D/A) conversion. In this scheme, shown in Figure 1.5, an analog signal is converted to a digital representation using an analog-to-digital (A/D) converter, the resulting digital value is stored in flip-flops, and a digital-to-analog (D/A) converter is used to create a quantized analog version of the original analog input. A recent example of this form of converter is presented in [5]. This method has all of the advantages of digital storage, including high noise immunity and storage density. The

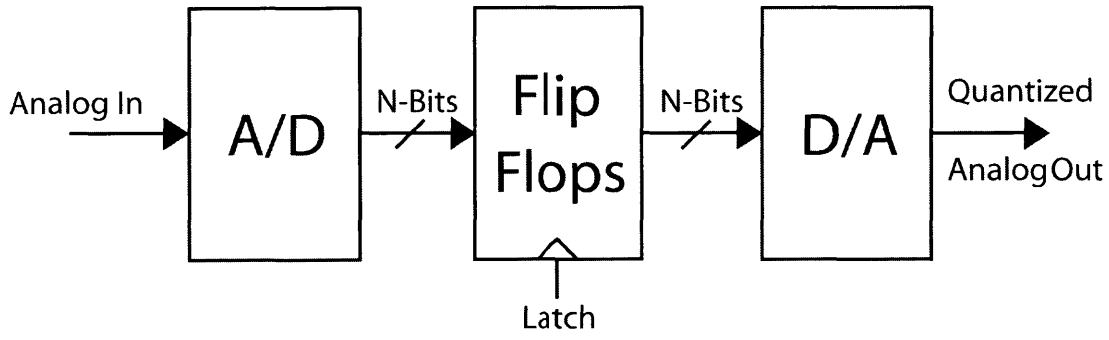


Figure 1.5– This figure shows a top-level view of an analog-to-digital-to-analog (A/D/A) converter. The converter takes in an analog signal, creates a digital representation of it using an analog-to-digital (A/D) converter, stores the digital representation in digital flip-flops, and re-creates a quantized version of the original signal using a digital-to-analog (D/A) converter.

size and power consumption of the A/D and D/A converters, however, can be large. In situations where the analog information is needed only for medium time scales, it may be more power-efficient to store the analog variable using some other technique. Of course,

on some long time scale, digital storage will always be the cheapest and the most likely to preserve the information in the face of noise.

1.2.5 Quantized-Analog Storage

Quantized-analog storage schemes were developed as a form of synaptic weight storage by researchers in the field of neural networks [6], [7], [8]. These circuits accept an analog input voltage, quantize and store the analog voltage on a capacitor, and periodically re-quantize the capacitor voltage to combat leakage. This process is conceptually a form of A/D/A conversion, however the flip-flops have been replaced by a single capacitor and the D/A converter is not necessary. Because it lacks both of these components, quantized-analog storage can potentially offer an area and power advantage over a full A/D/A scheme in situations where the required resolution on the analog variable is not too high.

Several implementations of the quantized-analog scheme have been proposed in the literature. One of the earliest versions of this idea, shown in Figure 1.6 was presented by Hochet [6]. Part (a) of the figure shows the basic circuitry that is required to implement Hochet's analog memory, and part (b) shows typical waveforms that are present during the circuit's operation. Clk is the basic system clock and Run is a signal which, in this case, remains high for eight clock periods. The periodic restoration process begins with an unknown analog voltage to be quantized on C_{store} and with C_{ramp} initialized

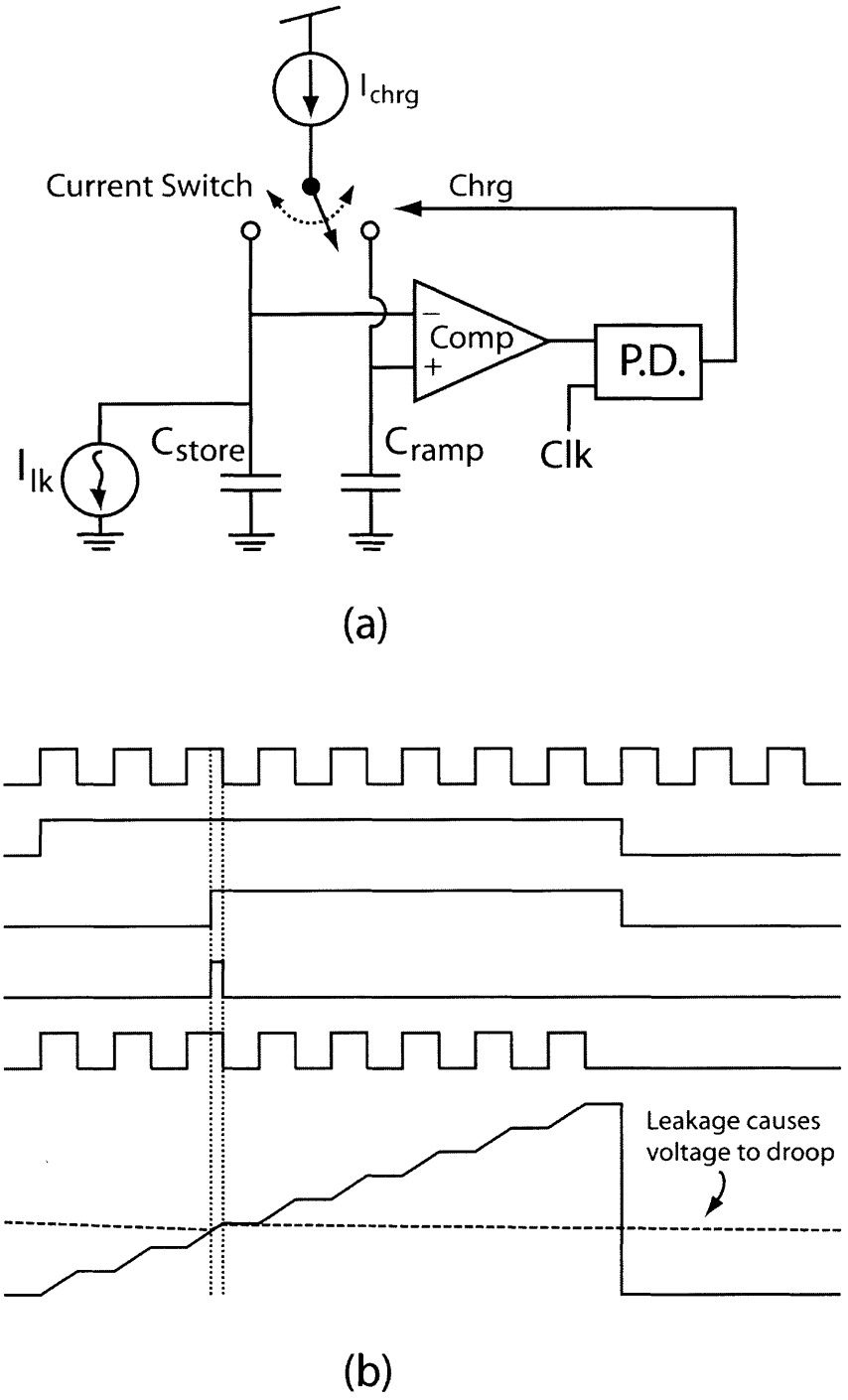


Figure 1.6– (a) The analog circuitry required for implementing Hochet's quantized-analog memory. (b) Typical waveforms from the circuit's operation.

to zero volts. The quantization occurs during the time frame when Run is high. During this time, whenever Clk is high, I_{chrg} is connected to C_{ramp}, and the voltage on C_{ramp} ramps

to a new voltage level which is one quantized voltage step higher than it was previously. The magnitude of this voltage step is defined by

$$\Delta V = \frac{I_{chrg}}{C_{ramp}} \times \frac{T}{2} \tag{1.3}$$

where T is the period of Clk. In this example, ΔV should be set such that the voltage on C_{ramp} covers the full possible range of voltages that could be present on C_{store} in eight clock cycles. The comparator compares the voltage on C_{ramp} with that on C_{store} throughout this process, and the comparator output, Comp, switches high when the voltage on C_{ramp} is larger than that on C_{store}. The time difference between the rising edge of Comp and the falling edge of Clk represents the voltage difference between the voltage on C_{store} and the next higher quantized voltage level of C_{ramp}. This time difference is represented as a pulse on the signal Chrg, and is produced by the phase detector, (P.D.) circuit. While Chrg is high, I_{chrg} is allowed to charge C_{store}. If C_{store} and C_{ramp} are of the same value, the voltage change on C_{store} should be of the exact amount necessary to bring its value back to the nearest upper quantized level. If this restoration is performed often enough, and the leakage is always off of the capacitor, leakage can be counteracted and no information is lost.

Hochet built a discrete version of this circuit, and was able to demonstrate 5-bit accurate analog storage. Several algorithmic choices severely limited his ability to achieve better performance. First, he requires the capacitors to be well-matched, which is hard to ensure for discrete circuits, and is not trivial in IC circuits without sacrificing area. Second, comparator delay directly affects the width of the pulse on Chrg, and thus affects how much C_{store} is charged. Hochet suggests that the comparator delay should be tuned to be a multiple of the clock period, T, which would cancel its effect on the

algorithm. This is not a trivial task, and a scheme in which the outcome is not sensitive to comparator delay is preferable. Finally, we would like to be able to counteract leakage regardless of whether it is onto or off of the storage capacitor.

An algorithm which solves the first two problems, but does not solve the bi-directional leakage issue, was proposed by Vittoz et al. [7]. The authors recognized that it was possible to generate one global staircase signal, rather than a local one in each cell. Each local cell monitors the staircase bus and waits for its local comparator to trigger, as before. However, instead of using the imprecise phase information to update the storage capacitor, simply copy the quantized voltage from the bus during the flat portion of the staircase. This technique allows us to eliminate our dependence on the comparator delay because after each ramping portion we have half of a clock cycle to decide we have exceeded our stored voltage and to copy the new quantized level onto our storage capacitor. Thus, as long as the comparator delay plus the voltage copying time is shorter than T/2, the comparator delay adds no error. In addition, this technique does not depend on matching between any components for its accuracy. The authors did not fabricate this circuit, and so no claim of its accuracy was made. It is reasonable to expect that it will have more resolution than the previous circuit due to the algorithmic improvements.

A third algorithm which solves all of the above problems was introduced by Cauwenberghs [8]. In his implementation, shown in Figure 1.7 each local cell consists of a negative-feedback amplifier memory cell like the one introduced in Figure 1.4, an incrementing and a decrementing current source used to inject small charge packets onto the memory cell, and bus interface circuitry. Each cell periodically sends its analog information to a central A/D/A converter, which then instructs the cell to inject or extract

a small amount of charge to or from the storage capacitor in order to bring its value closer to the nearest quantization level. Notice that the stored value is not forced to a particular

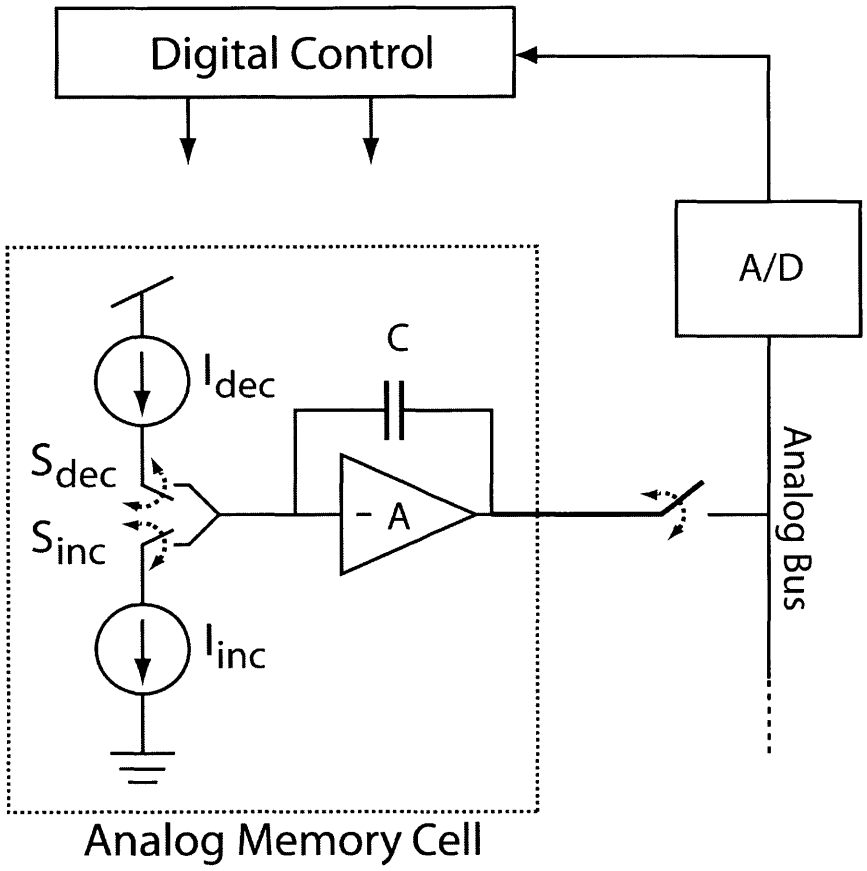


Figure 1.7– Cauwenbergh's partial-refresh oversampling analog memory cell.

quantized level in each cycle, as it was in the previous two circuits, but only pushed in the proper direction. This strategy guarantees that occasional errors in the quantization process do not result in a complete loss of information, as it would in the previous circuits. However, because only small updates are made, they must occur more often, which means it is necessary to spend more power to implement this scheme than the previous schemes. The IC implementation of this circuit achieved 8-bits of storage resolution, and was able to hold voltages to 8-bit accuracy for over 24 hours.

1.3 Chapter Summary

In this chapter, we introduced the theory behind the most popular forms of IC analog storage, including: simple capacitor storage and low-leakage capacitive storage, floating gate storage, A/D/A converters, and quantized-analog memories. In the next chapter, we will draw upon our knowledge of these circuits to aid us in developing an algorithm that will meet the design goals of this project.

2 System-Level Algorithm & Hardware Design

The designs presented in Chapter 1 span the majority of known methods of storing a representation of an analog variable on an integrated circuit. There are other possibilities available in more exotic and experimental fabrication processes [9], but we will ignore them here for reasons that will be presented shortly. This chapter establishes the design goals for the memory cell, and explains the rationale behind the system-level algorithmic and hardware choices that were made.

2.1 Design Objectives

The first, and probably most stringent, restriction on the memory cell is that it should not require components that are not available in a vanilla CMOS process. The explosion in popularity of digital IC fabrication has made CMOS the least expensive of all IC processes. As a result, there is an economic push for both digital and analog designs to be implemented in CMOS. A second requirement is that the cell should occupy as little die area as possible. It is difficult to establish numbers a priori, however, area should be considered in all design decisions. All of the implementations discussed in Chapter 1 were designed to consume minimum area. The final requirements are that the cell should consume at most a few microwatts of power from a 3.3V rail, it should

possess at least eight bits of precision, it should be capable of storing a new memory value in a few milliseconds or less, and it should be able to store a memory value for at least one hour. For easy referencing, these requirements have been compiled in Table 2.1.

Desired Memory Cell Properties

1. The design should be vanilla CMOS compatible.
2. Each cell should occupy as little die area as possible.
3. Each cell should consume at most a few microwatts.
4. The power supply voltage should be 3.3V.
5. The precision of the cell must be at least eight bits.
6. The cell should store new values in < a few milliseconds.
7. The memory should be able to hold state for one hour.

Table 2.1– This table outlines the design objectives for the analog storage cell.

2.2 Algorithm and Hardware Design

Looking back at the different strategies of Chapter 1, we can now cull our options. Let's take a brief look at how well each can achieve the design objectives.

2.2.1 Simple Capacitive Storage

Capacitive storage is compatible with vanilla CMOS, would consume very little power, and could potentially hold more than eight bits of information. However, the capacitor would have to be too large to store eight bits of information for one hour. Assuming a (best case) 10fA leakage current, and a full-scale range of 3.3V, we would need a capacitor whose value is

$$C = I_{lk} \times \frac{\Delta t}{\Delta V} = 10fA \times \frac{3600\sec}{\left(3.3V/2^8\right)} = 174.5pF \tag{2.1}$$

to store eight bits of precision for one hour. This is far too large for IC applications, and this technique alone will not meet our needs.

2.2.2 Low-Leakage Capacitive Storage

Like simple capacitive storage, low-leakage capacitive storage meets all of our needs except that it also requires a large capacitor in order to store eight bits for an hour. In the best possible case we may reduce the leakage current from 10fA to 1fA, meaning the cell will still require a 17.45pF storage capacitor. This capacitor is more reasonable in size, but is still much too large for use in our memory cell.

2.2.3 Floating Gate Storage

Floating gate storage is capable of storing more than 14-bits of precision for more than an hour using only nanowatts of power, and the cell size for floating gate storage can be as small as 70μm x 70μm [3]. However, the process layers needed to construct floating gates are not available in all CMOS technologies, and the write times for accurate analog storage are usually much more than a millisecond because of the iterative process needed for writing. Also, the cost of the iterative storage circuitry must be accounted for in the total system cost. Finally, probably the biggest disadvantage of floating gate circuits is the need for a voltage much higher than the inherent voltage rail of the circuit for adjusting the charge on the floating gate.

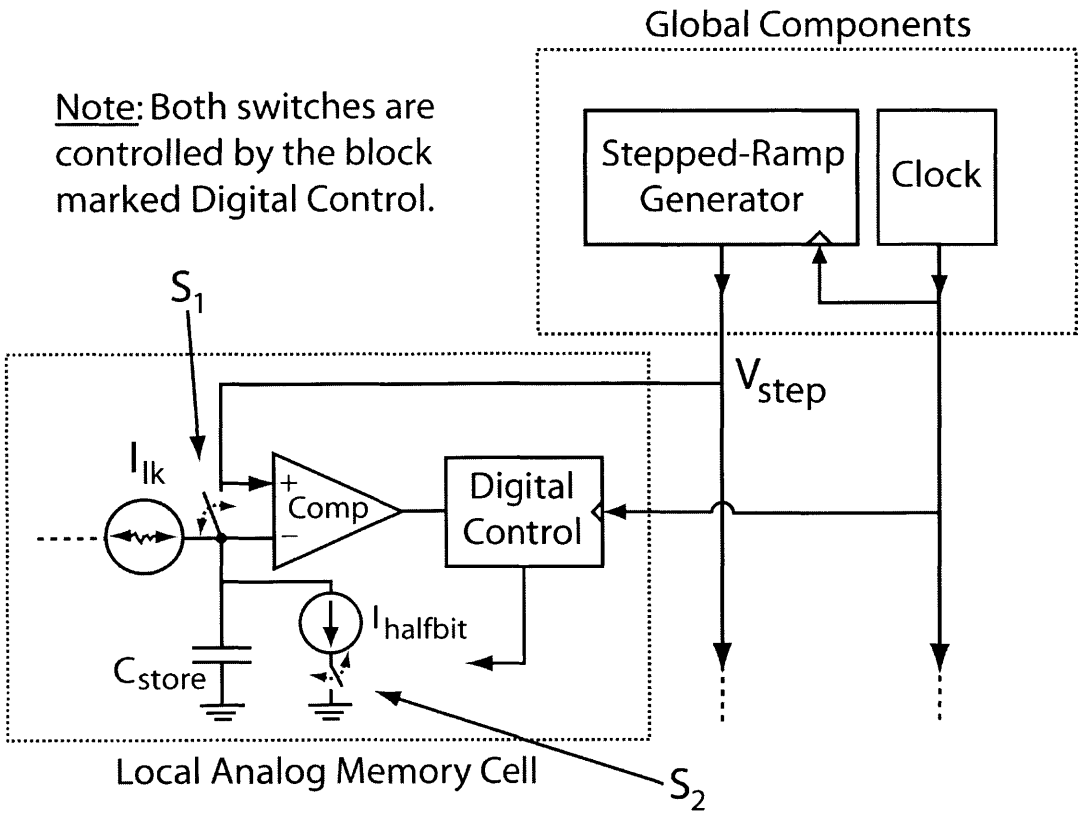
2.2.4 Analog-to-Digital-to-Analog Storage

This technique is CMOS compatible, and can meet the precision, write time, and hold time specifications. However, the components needed for its implementation consume a large amount of area and power. If a large array of analog memory cells share the cost of the A/D and D/A, the technique is feasible, otherwise it is not.

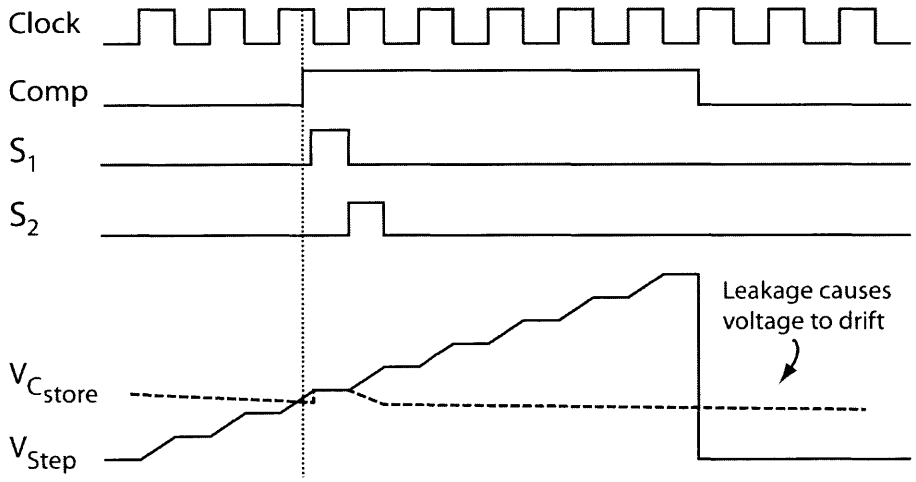
2.2.5 Quantized-Analog Storage

This technique is CMOS compatible, it offers the potential of small cell size and microwatt level operation, and it should be able to achieve at least 8 bits of precision. The write time can theoretically be much less than a millisecond, and if designed properly, indefinite hold times should be possible. The author actually independently developed a quantized-analog algorithm to solve this design problem long before a literature search revealed that this class of analog storage circuits had already been proposed. However, despite the fact that the technique has been known for at least a decade, it has received little attention in the literature, so not much is known about its advantages or limitations. This project should help expose some of the limitations of these circuits, as well as hopefully prove that they possess some advantages over other analog storage methods.

The independently developed algorithm is very similar to that of Vittoz et al. [7], except that a method of rejecting leakage in both directions was devised. A pictorial representation of the algorithm is shown in Figure 2.1. Part (a) of the figure shows a system-level view of the hardware needed for this scheme. The stepped-ramp generator



(a)



(b)

Figure 2.1– (a) System-level view of the circuitry needed to implement the chosen algorithm. (b) Typical waveforms present during the operation of the circuit in (a).

and the clock provide global signals to all of the memory cells, and their overhead cost is shared by these cells. The system is designed such that all local cells can operate in parallel based on the information they receive from the global circuits. The local cells consist of a storage capacitor, comparator, current source, and digital control block. Part (b) of the figure shows typical waveforms that would be present during normal circuit operation.

A storage cycle begins with an unknown voltage that is to be quantized stored on C_{store}. The switches S_1 and S_2 are both open, V_{step} is initialized to zero volts. At this point a conversion begins. The stepped-ramp generator produces the waveform V_{step} shown in part (b) by ramping up one quantization level in voltage during every high portion of the global clock, and holding its output steady during every low portion of the global clock. The ramp generator operates in this manner until it has reached the highest possible voltage it can output, and then it resets for a new cycle. All of this occurs regardless of what happens in the local cells. During this ramping process, the memory cells' asynchronous comparators compare V_{step} with their local version of V_{Cstore}, and output the result to their respective Digital Control block. Once a cell has determined that V_{step} has exceeded V_{Cstore}, the switch S_1 closes during the next low cycle of the clock. Thus, the quantized value provided by V_{step} is stored on C_{store}. Finally, during the following high cycle of the clock, switch S_2 enables the current source $I_{halfbit}$, which subtracts ½ of a bit's worth of charge from C_{store}. This final step allows leakage either onto or off of C_{store} between restorations to be counteracted, as will be shown below. A magnified view of the critical transition points on V_{Cstore} and V_{ramp} are shown in Figure 2.2.

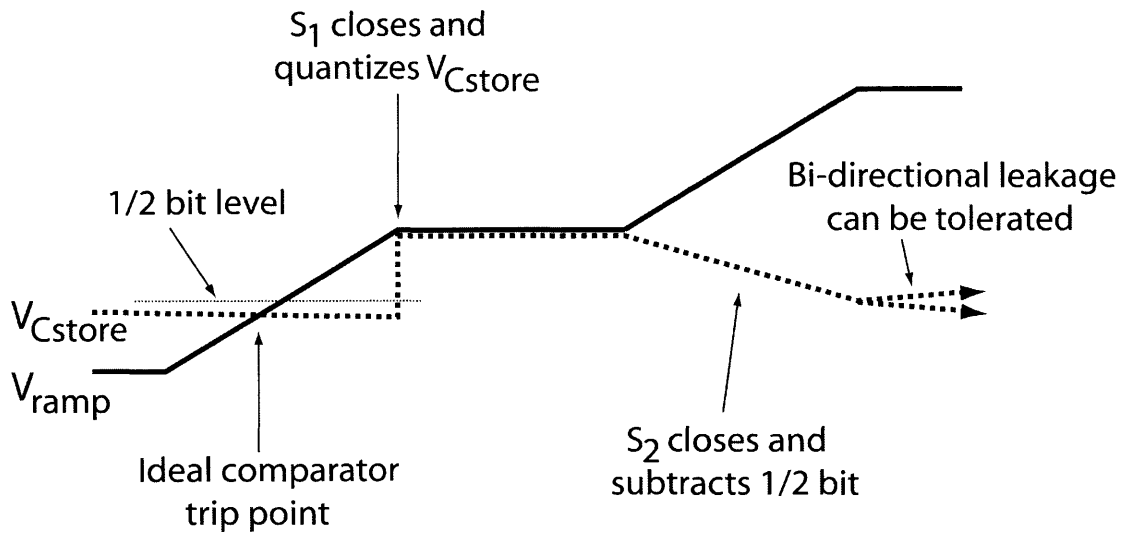


Figure 2.2– Magnified view of V_{ramp} and V_{Cstore} signal behavior during the critical portion of the restoration cycle.

As was discussed above, limited amounts of leakage in either direction can be counteracted with this analog storage technique. However, this is true only if certain conditions are met. Figure 2.3 illustrates how much leakage can be tolerated between the N^{th} and the $N+1^{th}$ quantization cycle when the leakage is off of the capacitor. As long as

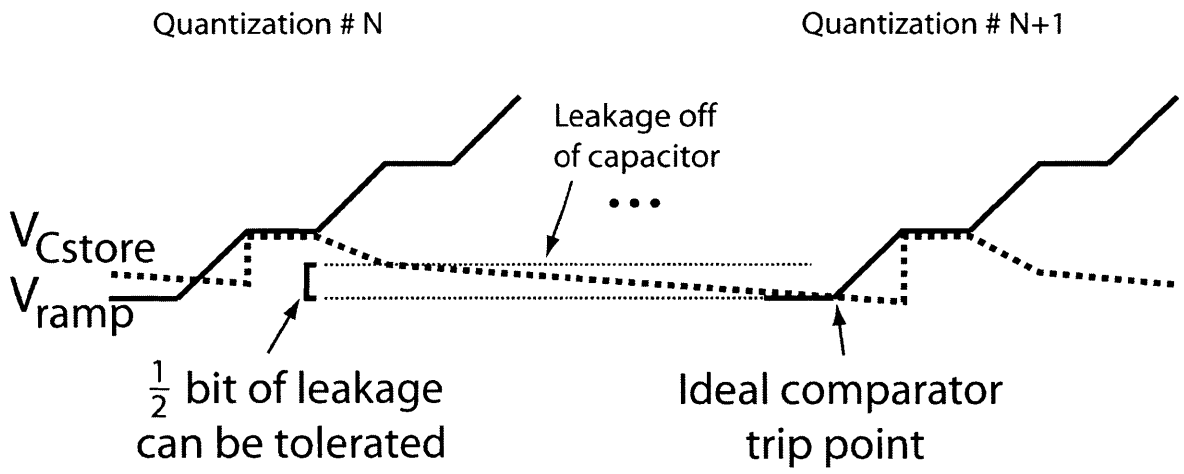


Figure 2.3 – A full ½ bit of charge leakage off of the capacitor can be counteracted with the quantized analog storage technique.

the leakage does not cause the stored voltage to move below the next lower quantized voltage value during the time of a full conversion cycle, the stored value can be re-quantized correctly. Thus, a full ½ bit of voltage leakage can be counteracted by the memory element. Figure 2.4 illustrates the complementary situation where charge is leaking onto the storage capacitor between the N^{th} and $N+1^{th}$ quantization. Again, a full ½ bit of voltage leakage can be counteracted by the next quantization cycle. Thus, the condition that must be met in order to assure that leakage does not corrupt the stored

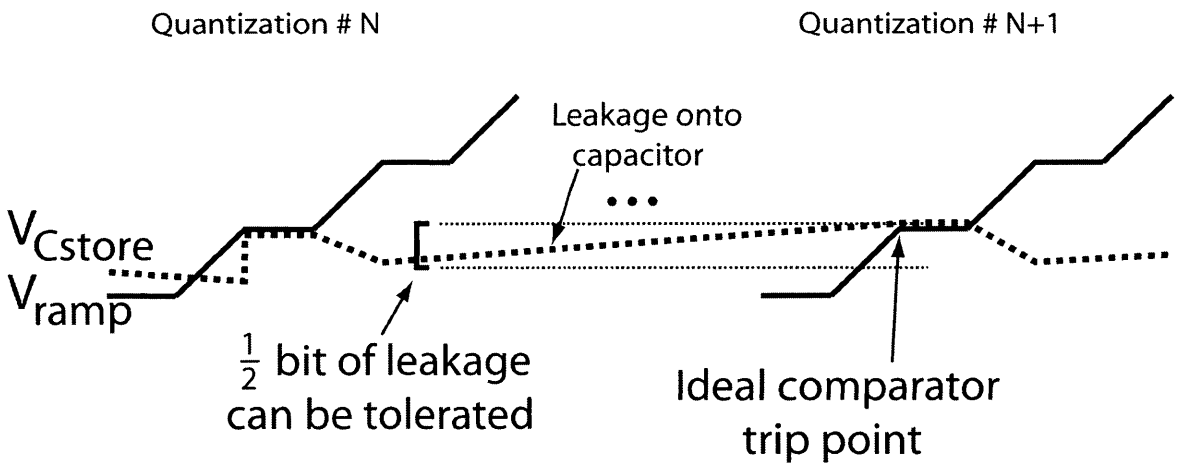


Figure 2.4 – A full ½ bit of charge leakage onto the capacitor can be counteracted with the quantized analog storage technique.

analog voltage is that the leakage must be less than ½ of a bit's worth of voltage over a full conversion cycle's worth of time, but the direction of the leakage does not matter. This condition assures that the stored analog voltage will never drift more than ½ bit from its quantized level.

The above arguments proved that ½ of a bit of leakage can be counteracted with the quantized-analog technique on a value that was assumed to have been quantized the step before. However, there is also a boundary condition that must be analyzed in this

circuit. The cyclical storage pattern is broken when a new analog voltage is stored in the cell, as shown in Figure 2.5. In this figure, the value of V_{Cstore} is seen dropping from its previous value (somewhere above the current view) to its new value. In this worst-case example, the value of V_{ramp} has just surpassed the new stored voltage value in the previous half-cycle. Thus, no quantization of the stored voltage will occur until a full conversion cycle later. Since the new voltage value is just larger than the ½ bit level of the ramp signal, it will drift (up in this case) by ½ of a bit to just above the next quantized voltage level before the first time it is quantized. It is then erroneously quantized up, rather than down, leading to a full bit of quantization error. The value then remains in this bin from this point on, as would be expected from Figure 2.4. Thus, in the worst-case, the quantized-analog memory element may introduce a full bit of initial quantization error.

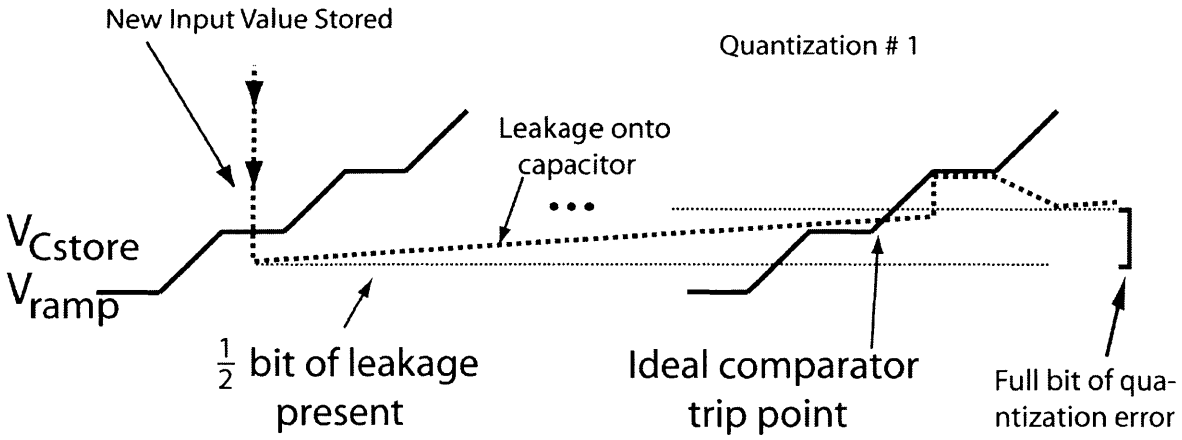


Figure 2.5 – This figure illustrates how a full bit of quantization error can occur when a new value is stored in the memory cell.

This algorithm should be much more robust to process variation and circuit non-idealities than Hochet's implementation introduced in Section 1.2.5. The global ramp

generator ensures that a single analog voltage stored in any local cell will be quantized to the same voltage level (assuming the comparators have negligible input-offset voltages). The comparator's delay does not affect circuit operation as long as this delay, plus the time it takes to copy the ramp voltage to C_{store}, is less than ½ of the clock period. Finally, because of the ½ bit subtraction, we no longer have to guarantee that the leakage is unidirectional. In comparison with Cauwenbergh's implementation from Section 1.2.5, this cell is designed such that all of the information flow is from global circuits to local ones. Since the local circuits do not need to be able to drive a large analog bus with at least an N-bit precise voltage, where N is the accuracy of the storage cell, they do not need to consume as much power in this implementation. The global ramp generator is then the only circuit which must be capable of driving an N-bit accurate analog voltage across the chip, and so the power is spent in only one circuit block.

2.3 Chapter Summary

In this chapter, a list of design goals was developed for the analog memory cell. Some are merely qualitative, e.g. use as little area as possible, while others give more specific limits on power consumption, acquisition time, and hold time. The goals are listed in Table 2.1. After the design objectives were established, each of the analog memory techniques introduced in Chapter 1 was compared with the design goals, and a quantized-analog algorithm was chosen for the memory cell's implementation. Chapter 3 presents in detail the circuits used to implement this algorithm, SPICE simulations of these circuits, and SPICE simulations of the entire quantized-analog memory system.

3 Circuit Design & Simulations

The first two chapters have laid the foundation on which the remaining chapters are based. We examined existing approaches to analog storage and past implementations of analog storage circuits. Lessons learned from these implementations will aid in the design of the analog memory cell in this project. An algorithm and underlying circuit topology which should be able to meet the various design objectives that are required of the memory cell have also been developed. This chapter will first present the models used to describe MOS transistors in above-threshold operation. Next, it will introduce the basic circuit building blocks used in the memory cell, highlighting the most important features of each, and will compare hand-calculated parameters for the cells with SPICE simulations. After the constituent circuits have been examined, they will be used as "black boxes" to construct the overall topology, and SPICE simulations will be used to verify the overall system operation.

3.1 Large-Signal and Small-Signal Device Models

Before presenting the circuits used in this design, the large-signal and small-signal above-threshold MOS models are reviewed. The parameters used in the rest of this chapter and the next are all defined in this section.

3.1.1 Large-Signal Above-Threshold MOS Model

Figure 3.1 shows an NMOS transistor and defines its commonly-used forms of voltage and current variables. Figure 3.2 shows a family of operating curves depicting I_D as a function of V_{DS} for different values of V_{GS}, with $V_{SB} = 0V$. These transfer curves were simulated in SPICE using parameters extracted from the MOSIS 1.5μm AMI CMOS process that will be used to fabricate the final memory cell. The figure shows two regions of transistor operation based on the external voltages: the triode region and the saturation region. These regions are separated by a dotted line plotting the points where $V_{DS} = V_{GS}-V_t$. This particular value of drain-to-source voltage is known as V_{Dsat} because it is the voltage at which a transistor operating at a constant V_{GS} moves from the saturation to the triode region. The equation defining transistor operation in the triode region is

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS}-V_t)V_{DS}-\frac{V_{DS}^{\ 2}}{2\kappa}\right] \tag{2.2}$$

where μ_n is the mobility for electrons, C_{ox} is the gate capacitance per unit area, W is the transistor gate width, L is the transistor gate length, κ is a parameter representing the ratio of control over the channel potential of the gate versus the bulk, V_t is the transistor's threshold voltage, and the remaining voltages are as they were defined in Figure 3.1. In the triode region, the transistor behaves like a resistor as long as the $V_{DS}^2/2\kappa$ term is much smaller the $(V_{GS} - V_t)V_{DS}$ term. This is true in the area near the origin, where the I_D versus V_{DS} curves look locally like straight lines – the same characteristic as a resistor. When this condition holds, the effective resistance of the MOSFET is given by the

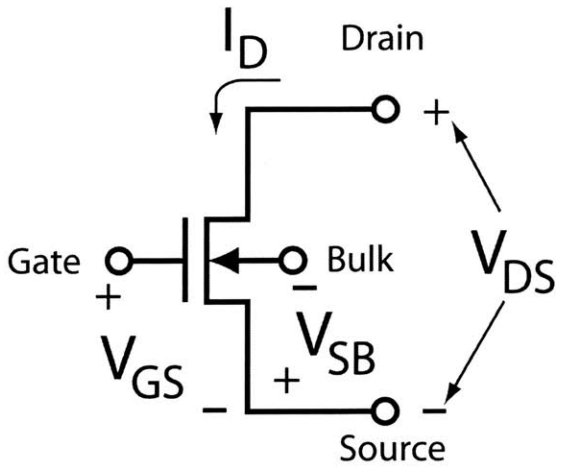


Figure 3.1 – NMOS transistor symbol labeled with the voltage and current variables typically used in describing its operation.

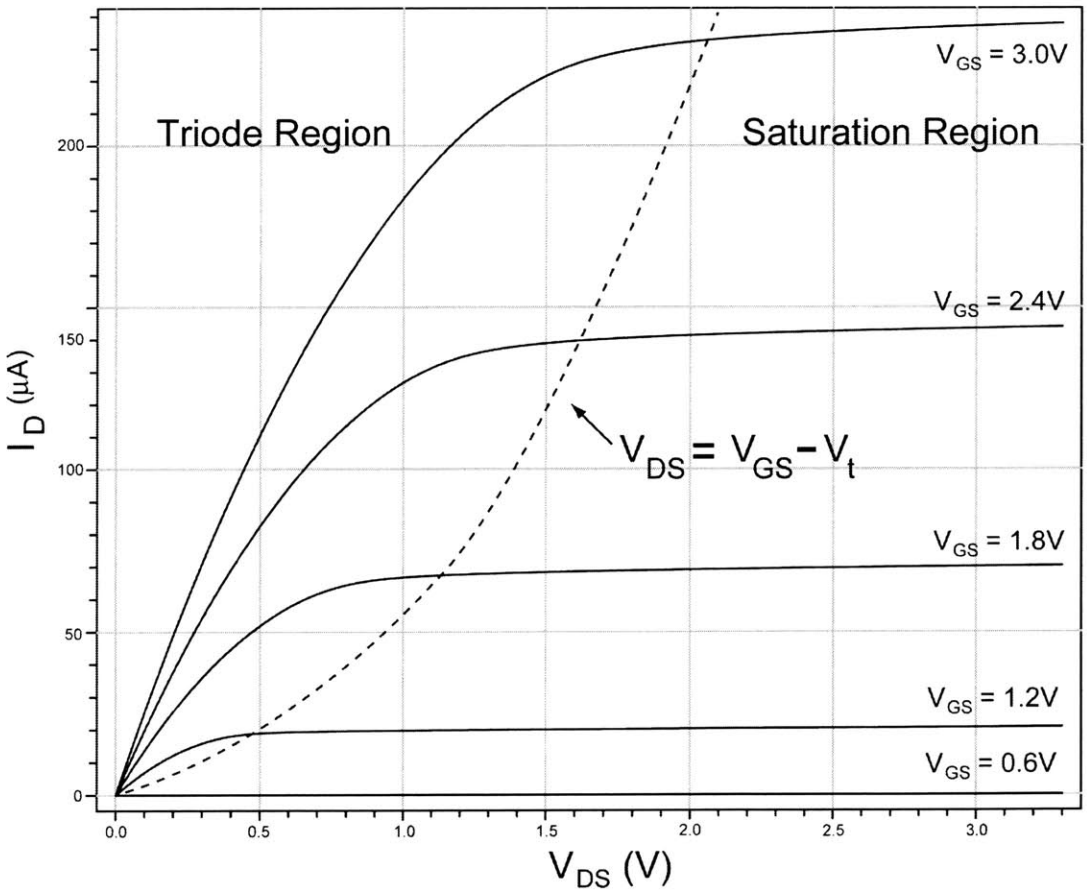


Figure 3.2 – I_D versus V_{DS} for different values of V_{GS}. This data was simulated in SPICE based on models from the MOSIS 1.5μm CMOS process that will be used for fabricating the memory cell.

equation

$$R_{eff} = \frac{V_{DS}}{I_D} \approx \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_t)}. \tag{2.3}$$

This region of operation is typically encountered in situations where the MOSFET is used as a switch.

The second region of operation for a MOSFET is the saturation region. Notice that in this portion of the curves, the value of V_{DS} has only a small effect on the current flowing through the transistor -- effectively, the transistor looks like a current source with fairly high output impedance. This is the region of the MOS transistor that is used for amplification. The equation defining operation in the saturated region is

$$I_D = \frac{\mu_n C_{ox} \kappa}{2}\left(\frac{W}{L}\right)(V_{GS} - V_t)^2 \left[1 + \lambda\left(V_{DS} - V_{Dsat}\right)\right] \tag{2.4}$$

where all of the variables are the same as before, and λ is a variable describing the change in drain current as a function of a change in the drain-to-source voltage. The parameter λ is inversely proportional to transistor length, and is bias-point dependent.

An effect that was not taken into account in the previous two equations is the body effect. When the bulk is biased at a different voltage than the source, it causes the transistor exhibit the same characteristics as above, except with a different threshold voltage. The new threshold voltage is defined as

$$V_t = V_{to} + \gamma\left(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \tag{2.5}$$

where V_t is the overall threshold voltage, V_{to} is the threshold voltage when $V_{SB} = 0V$, γ is the body-effect constant, and $2\phi_F$ is the surface inversion potential.

As for PMOS transistors, all of the above equations are valid if a negative sign is placed in front of every voltage variable, μ_n is replaced by μ_p, and I_D is defined as the source-to-drain current of the transistor.

3.1.2 Small-Signal Above-Threshold NMOS Model

As was mentioned in the previous section, a triode transistor looks very similar to a resistor for drain-to-source voltages close to zero. Therefore, its small-signal model is just a resistor from drain-to-source, with the resistance given by Equation 3.2. In addition to the channel resistance, there is capacitance from the gate to both the source and drain, through the conduction channel. The values of these capacitors are usually estimated as

$$C_{gs} = C_{gd} = \frac{WLC_{ox}}{2}. \tag{2.6}$$

There are also source-to-bulk and drain-to-bulk capacitances which will be very similar to the values in the saturated region of operation, so we will defer their defining equations until then. A simple triode region small-signal model with the parameters described above is shown in Figure 3.3.

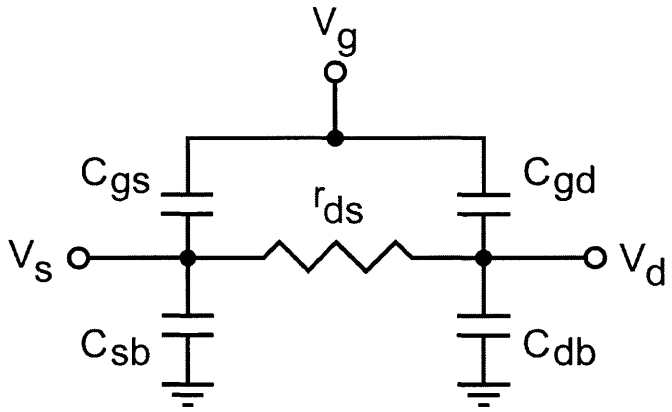


Figure 3.3 – Small signal model of a MOSFET transistor operating in the triode region.

The small-signal model for saturation region operation is by far more interesting. This model is shown in Figure 3.4. There are now two voltage-dependent current sources and an output resistance in the model. The primary transconductance, g_m, is defined as

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu_n C_{ox} \kappa \frac{W}{L}(V_{GS} - V_t) = \sqrt{2\mu_n C_{ox} \kappa \frac{W}{L} I_D} \,. \tag{2.7}$$

The g_m term models the incremental change in channel current due to an incremental change in gate-to-source voltage. The second transconductance, g_{mb}, models an incremental change in drain current due to an incremental change in source-to-bulk voltage. This term is described by

$$g_{mb} = \frac{\partial I_D}{\partial V_{SB}} = \frac{\partial I_D}{\partial V_t}\frac{\partial V_t}{\partial V_{SB}} = \frac{\gamma g_m}{2\sqrt{V_{SB} + |2\phi_F|}} \,. \tag{2.8}$$

The r_{ds} term models the finite slope of the I_D vs. V_{DS} curves we noticed in Figure 3.2, and is defined as

$$\frac{1}{r_{ds}} = \frac{\partial I_D}{\partial V_{DS}} = \lambda\left(\frac{\mu_n C_{ox} \kappa}{2}\right)\left(\frac{W}{L}\right)(V_{GS} - V_t)^2 = \lambda I_{DSAT} \cong \lambda I_D \tag{2.9}$$

where I_{DSAT} is the current flowing as the transistor just leaves the triode region and enters

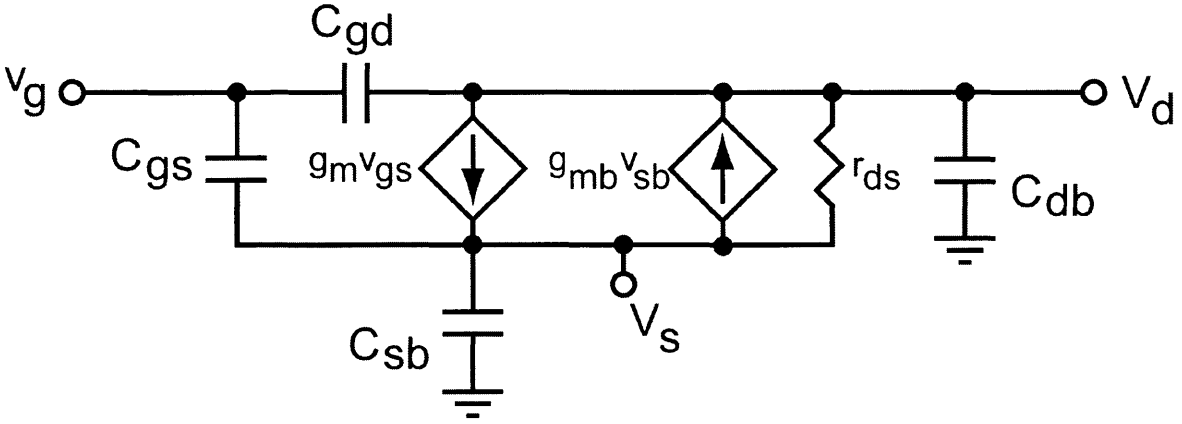


Figure 3.4 – Small signal model of a MOSFET transistor operating in the saturation region.

the saturation region, and it is assumed that λ is small and thus I_{DSAT} is roughly the same as I_D. The gate-to-source capacitance is defined as

$$C_{gs} = WC_{ox}(\frac{2}{3}L + L_{ov}) \tag{2.10}$$

where L_{ov} is the effective length of overlap between the gate and the source region. The source-to-bulk capacitance is due to the reverse-biased p-n junctions that the source forms with the bulk, and is defined by

$$C_{sb} = \frac{C_{jo}}{\sqrt{1+\frac{V_{SB}}{\Phi_o}}}(A_s + A_{ch}) + \frac{C_{j-swo}}{\sqrt{1+\frac{V_{SB}}{\Phi_o}}}P_s \tag{2.11}$$

where C_{jo} is the zero-voltage depletion capacitance per unit area, $C_{j\text{-}swo}$ is the zero-voltage sidewall capacitance per unit area, A_s and A_{ch} are the areas that the source and channel form with the bulk, respectively, P_s is the length of the perimeter of the source excluding the side facing the channel, and Φ_o is the built-in voltage of the p-n junction. The only change to this equation for triode operation is that only half of the channel area is included then. Similarly, C_{db} is given by

$$C_{db} = \frac{C_{jo}}{\sqrt{1+\frac{V_{DB}}{\Phi_o}}}A_d + \frac{C_{j-swo}}{\sqrt{1+\frac{V_{DB}}{\Phi_o}}}P_d \tag{2.12}$$

where A_d is the area that the drain junction forms with the bulk, and P_d is the length of the perimeter of the drain junction excluding the side facing the channel. Again, only half of the channel area would be included in the triode equation. The gate-to-drain capacitance is small, but very important, especially in circuits with gain. It is given by

$$C_{gd} = C_{ox}WL_{ov}. \tag{2.13}$$

All of these models have been derived for an NMOS transistor. However, the PMOS behaves as the "mirror image" of the NMOS transistor, and its operation should be clear from its use in circuits in the next section.

The extracted values for all of the parameters mentioned above are listed in Table 3.1. Note that these are rough values since they typically depend on the bias point of the circuit, but they should give a ballpark estimate of what to expect from a circuit using hand calculations.

Partial List of MOSIS 1.5μm AMI Process SPICE Parameters

Parameter	Units	NMOS Value	PMOS Value
V_{to}	V	0.544	-0.796
$\mu_{n,p}$	$cm^2/V \cdot s$	648.05	258.67
C_{ox}	F/m^2	1.135875×10^{-3}	1.135875×10^{-3}
γ	-	1.2497	0.3952
$\lvert 2\phi_F \rvert$	V	0.80850	0.68920
$C_{ox}L_{ov}$	F/m	1.57×10^{-10}	2.03×10^{-10}
C_{jo}	F/m^2	2.812164×10^{-4}	2.876295×10^{-4}
$C_{j\text{-}swo}$	F/m	1.488744×10^{-10}	2.013717×10^{-10}
Φ_o	V	0.97858	0.7500708
κ	-	0.7	0.7
λ *	V^{-1}	0.031	0.100

** This λ is valid for a minimum length transistor of 1.6μm. The value of λ should scale as 1/L.*

Table 3.1 – This table lists the SPICE parameters that are typically used in hand calculations for the MOSIS 1.5μm AMI CMOS process.

3.2 Analog Circuit Building Blocks Used In the Memory Cell

This section introduces the basic analog circuit blocks that were used to construct the memory cell, and compares hand-calculated performance parameters for these blocks against SPICE simulations.

3.2.1 PMOS-Input Common-Source Amplifier

The PMOS-input common-source amplifier is shown in Figure 3.5. It consists of a single PMOS transistor, which acts as an amplifier, and a single NMOS transistor, which acts as a current source and active load for the PMOS. From a large-signal perspective, the amplifier has a very limited range of input voltages. V_{in}'s DC value must be exactly the voltage necessary to ensure that the PMOS transistor carries the same DC current as the NMOS transistor supplies. Otherwise, the output of the amplifier will saturate to one of the rails and will not amplify at all. Its use in this system will always be coupled with feedback techniques in order to solve this problem. A second point that should be made is that the output voltage can swing within $V_{DSAT,n}$ of ground and $V_{DSAT,p}$ of V_{DD} (remember, $V_{DSAT} = V_{GS}-V_t$). This is known as rail-to-rail operation, and will help maximize the dynamic range of the overall system. For the remainder of this section, ignore DC biasing and assume V_{in} has the proper DC value to position the amplifier in its high-gain region.

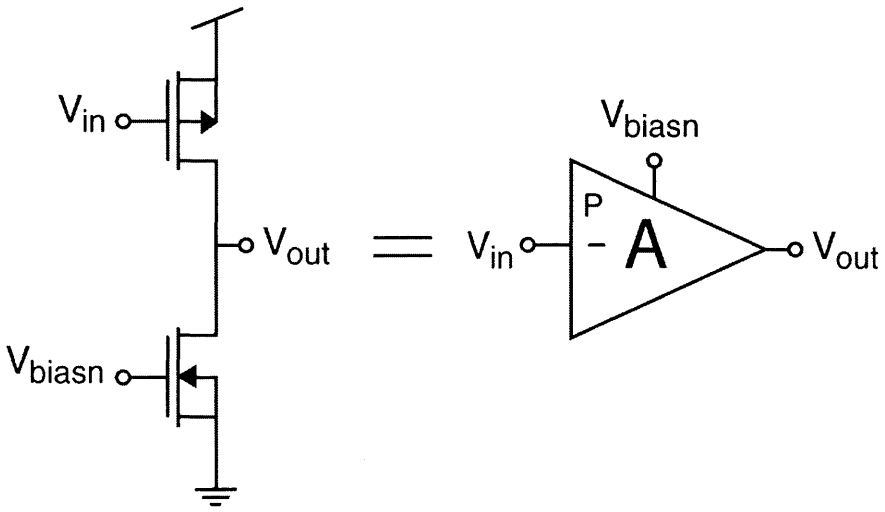


Figure 3.5 – PMOS input common-source amplifier circuit (left) and the symbol used to represent it (right).

For our application, we would like the circuit to achieve a DC gain of 40dB with a -3dB bandwidth of 1×10^6 rad/s, all while the amplifier is loaded with ≈70fF output capacitance (which will be introduced by other circuits loading its output when the system is built). A simplified small-signal model of this amplifier with the load capacitor included is shown in Figure 3.6. There are several small-signal components that have no effect on the circuit and therefore have been removed. First, neither transistor has a g_{mb} generator or a source-to-bulk capacitance. This is because both transistors have their

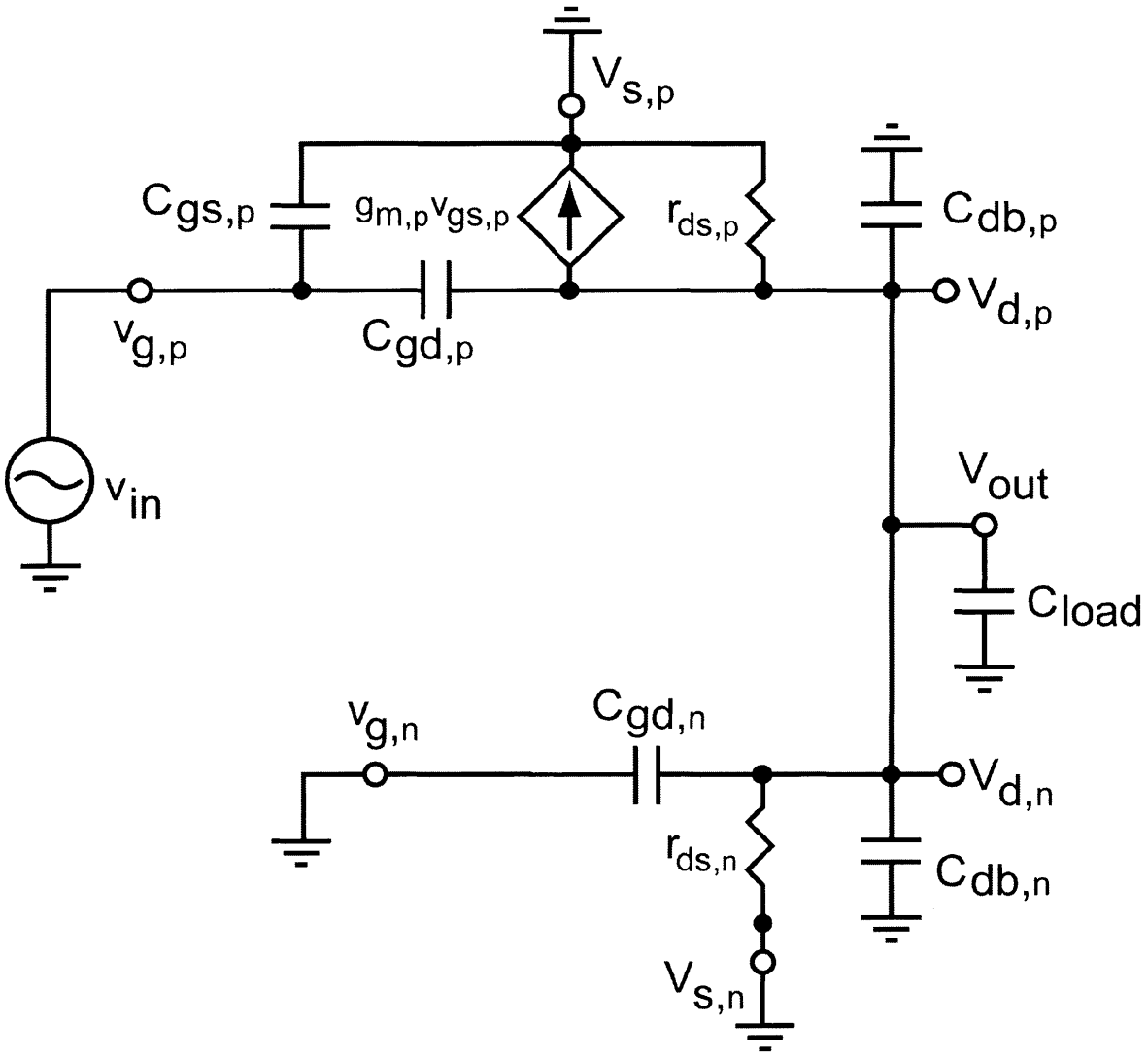


Figure 3.6 – Simplified small-signal model of PMOS-input common-source amplifier.

bulk and source terminals connected together. Second, since the NMOS is biased at a constant gate voltage, its gate is at small signal ground. Its source is also at small-signal ground, so neither the gate-to-source capacitance nor the g_m generator of the NMOS has an effect on circuit operation. The transfer function for this circuit is

$$\frac{v_{out}}{v_{in}} = -g_{m,p}\left[\frac{r_{ds,n}\left\|r_{ds,p}\right.}{1+s(C_{gd,n}+C_{db,n}+C_{db,p}+C_{load})(r_{ds,n}\left\|r_{ds,p}\right.)}\right], \tag{2.14}$$

and the resulting equations for the DC gain and -3dB bandwidth in rad/s are

$$A_{DC} = -g_{m,p}\left(r_{ds,n}\left\|r_{ds,p}\right.\right) \tag{2.15}$$

$$\omega_{-3dB} = \left[\frac{1}{(C_{gd,n}+C_{db,n}+C_{db,p}+C_{load})(r_{ds,n}\left\|r_{ds,p}\right.)}\right]. \tag{2.16}$$

All of the terms in Eqns. 3.14 and 3.15 depend on each other through the basic transistor parameters. It is usually easiest to simply choose an initial value for a few of the variables, and then solve for the others to see if the results for the other parameters are reasonable. For example, begin by choosing W = 8μm and L = 4μm for both the PMOS and NMOS transistors. These are small transistors by analog standards, and therefore the required power to meet the specifications should also be small. Next, using Eqns. 3.11 and 3.12, calculate the capacitive term in the denominator of Eqn. 3.15 for an output voltage half-way between the rail voltages. Assume the transistor drains measure 4.8μm × 8μm (standard layout rules require 4.8μm length, minimum). The gate-to-drain capacitance given by Eqn. 3.12 is

$$C_{gd,n} = \left(1.57\times10^{-10}\,F/m\right)\left(8\times10^{-6}\,m\right) = 1.26\,fF. \tag{2.17}$$

The drain-to bulk capacitances are

$$C_{db,n} = \frac{\left(2.81\times10^{-4}\,F/m^2\right)}{\sqrt{1+\frac{1.65\text{V}}{0.979\text{V}}}}\left(4.8\times8\times10^{-12}m^2\right)+\frac{\left(1.49\times10^{-10}\,F/m\right)}{\sqrt{1+\frac{1.65V}{0.979V}}}(17.6\times10^{-6}m)$$

$$=6.58fF+1.6fF=8.18fF \tag{2.18}$$

and

$$C_{db,p} = \frac{\left(2.88\times10^{-4}\,F/m^2\right)}{\sqrt{1+\frac{1.65\text{V}}{0.75\text{V}}}}\left(4.8\times8\times10^{-12}m^2\right)+\frac{\left(2.01\times10^{-10}\,F/m\right)}{\sqrt{1+\frac{1.65V}{0.75V}}}(17.6\times10^{-6}m)$$

$$=6.18fF+1.98fF=8.16fF. \tag{2.19}$$

Thus, using Eqn. 3.15, in order to achieve a -3dB bandwidth of at least 1×10^6 rad/s the net output resistance of the amplifier should be

$$r_{ds,n}\left\|r_{ds,p}\right. \le \frac{1}{(1.26fF+8.18fF+8.16fF+70fF)(10^6\,rad/s)} = 11.4\times10^6\Omega \tag{2.20}$$

and in turn, using Eqn. 3.14, $g_{m,p}$ should be at least

$$g_{m,p} \ge \frac{100\,V/V}{11.4\times10^6\Omega} = 8.8\,\mu A/V. \tag{2.21}$$

Using Eqn. 3.6, we see that Eqn. 3.20 requires an I_D of at least

$$I_D \ge \frac{g_{m,p}{}^2}{2\mu_p C_{ox}\kappa\frac{W}{L}} = \frac{\left(8.8\,\mu A/V\right)^2}{(2)\left(0.02587\,m^2/V\cdot s\right)\left(1.14\times10^{-3}\,F/m^2\right)(0.7)\left(\frac{8\mu m}{4\mu m}\right)} = 0.937\mu A. \tag{2.22}$$

Since the output resistance of the amplifier is composed of the parallel combination of the output impedances of the PMOS and NMOS, we need each of their r_{ds}'s to be roughly 2 × 11.4MΩ = 22.8MΩ. Using Eqn. 3.8 and $r_{ds} \approx 22.8\text{M}\Omega$, we get

$$I_D \geq \frac{1}{\frac{4\mu m}{1.6\mu m} \times 0.031 \times 22.8 \times 10^6 \Omega} = 0.57\mu A. \tag{2.23}$$

An I_D of 1μA will meet the constraints of Eqns. 3.21 and 3.22, and is a reasonable current value. A SPICE simulation of the transfer function of this circuit with the NMOS biased to supply 1μA is shown in Figure 3.7. The DC gain according to the simulation is 39dB, with a -3dB bandwidth of 167kHz, or 1.05×10^6 rad/s. These values are in very good agreement with the hand calculations, and thus our model of the circuit is bolstered.

Another parameter of importance for this circuit is the input-referred noise power

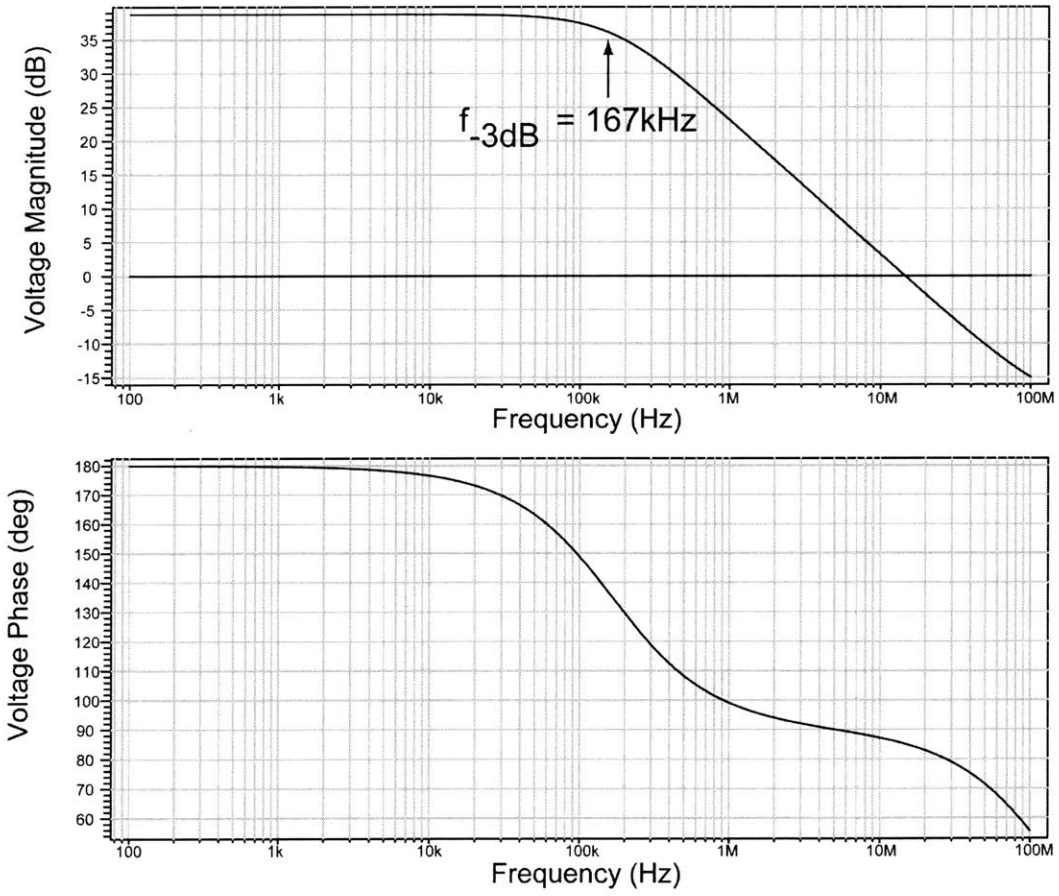


Figure 3.7 – Bode plot of the transfer function of the common-source amplifier shown in Figure 3.5, with the NMOS transistor biased to supply 1μA.

per unit bandwidth. The channel resistance white noise model for a saturated MOS transistor is

$$I_d^2(f) = 4kT\left(\frac{2}{3}\right)g_m \tag{2.24}$$

where k is Boltzmann's constant (1.38×10^{-23} JK^{-1}) and T is the temperature in Kelvins. The noise model for this circuit is shown in Figure 3.8. It is assumed that these two noise sources are uncorrelated, and thus their current power adds at the output of the amplifier.

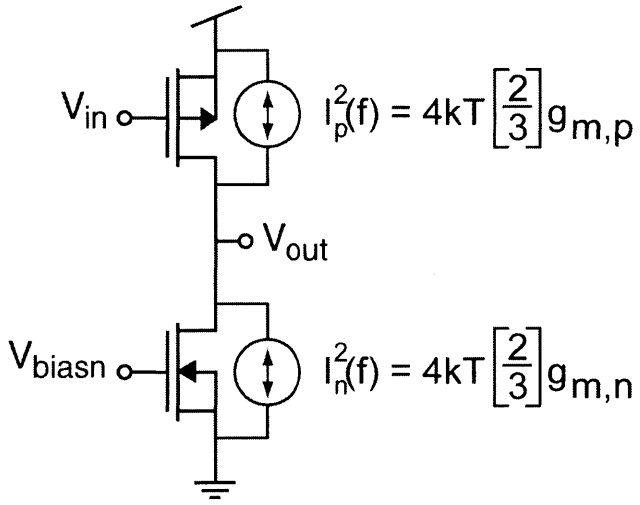


Figure 3.8 – Noise model for the common-source amplifier.

This current noise power is referred back to the input of the amplifier by dividing by $g_{m,p}^2$. Thus, the overall input-referred voltage white noise should be

$$V_{in}^2(f) = \frac{I_p^2(f) + I_n^2(f)}{g_{m,p}^2} = \frac{4kT\left(\frac{2}{3}\right)\left(g_{m,p} + g_{m,n}\right)}{g_{m,p}^2}. \tag{2.25}$$

Evaluating Eqn. 3.24 at the operating point of the amplifier, we find the input white noise

$$V_{in}^2(f) = \frac{4\left(1.38\times10^{-23}\ \mathrm{JK}^{-1}\right)(300K)\left(\frac{2}{3}\right)\left(9.09\,\mu A/V + 14.38\,\mu A/V\right)}{\left(9.09\,\mu A/V\right)^2} = 3.13\times10^{-15}\,V^2/Hz. \quad (2.26)$$

The noise is usually expressed in its root form

$$V_{in}(f) = 56\,nV/\sqrt{Hz}. \quad (2.27)$$

Simulating the input-referred noise spectral density using SPICE, we obtain

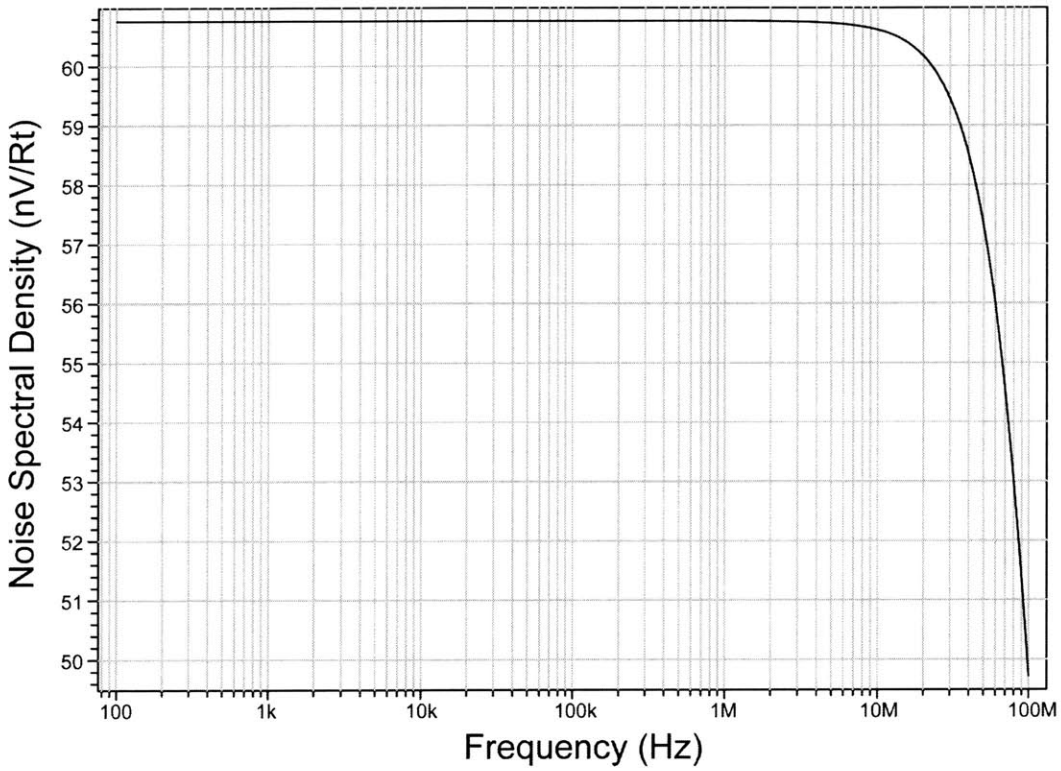


Figure 3.9 – SPICE simulated input-referred noise spectral density for the PMOS-input common-source amplifier biased at I_D = 1μA.

which agrees with the hand-calculated results to within 10%. A second important noise source in MOS circuits is 1/f noise. However, this circuit will be used in an auto-zeroing amplifier configuration, and therefore most of the 1/f noise will be eliminated. This topic is addressed more fully in Appendix A.

3.2.2 NMOS-Input Common-Source Amplifier

The second analog circuit used in the memory cell is the complement of the amplifier introduced in the previous section. The amplifier topology is identical to that of the PMOS-input amplifier, except the PMOS is now the bias transistor and the NMOS is the input transistor. This new circuit is shown in Figure 3.10. Again, for the moment, ignore the issue of biasing this amplifier in its high-gain region.

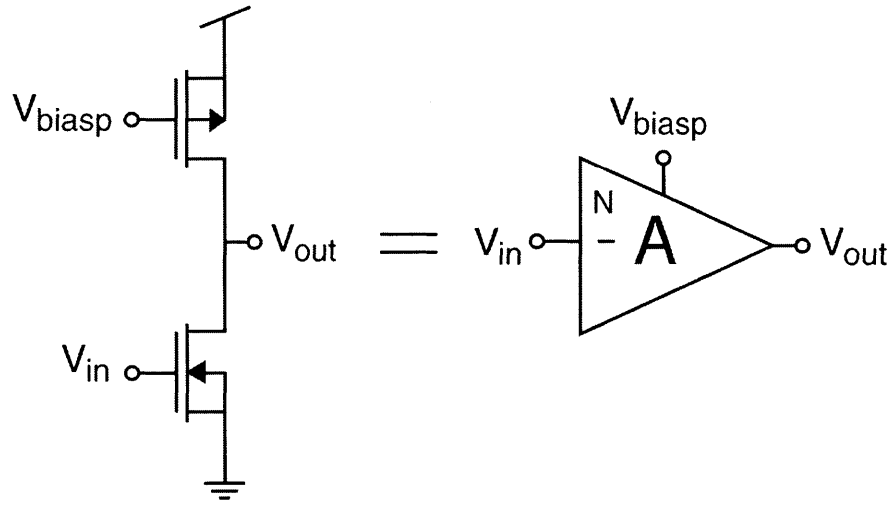


Figure 3.10 -- NMOS input common-source amplifier circuit (left) and the symbol used to represent it (right).

What should be expected to change with this new circuit versus the previous one if all sizing is kept constant? First, since the mobility of the NMOS electrons is greater than the PMOS holes, $g_{m,n} > g_{m,p}$, and the gain should increase. Also as a result of this fact, the input-referred noise should be lower than for the previous amplifier. There are also a few parasitic capacitors which switch roles in this new topology, but their effects will be small. Hand analysis predicts that for the same bias of 1μA, the new gain should be

$$A_{new} = A_{old}\frac{g_{m,n}}{g_{m,p}} = \left(100\frac{V}{V}\right)\left(\frac{14.38\,\mu A/V}{9.09\,\mu A/V}\right) = 158\,V/V = 44dB. \qquad (2.28)$$

The simulation of this amplifier plus the 70fF load in Figure 3.11 shows that the new DC gain and -3dB breakpoint are 44.5dB and f_{-3dB} = 147kHz, or ω_{-3dB} = 924 krad/s. These are within 10% of predicted values – definitely close enough for this design.

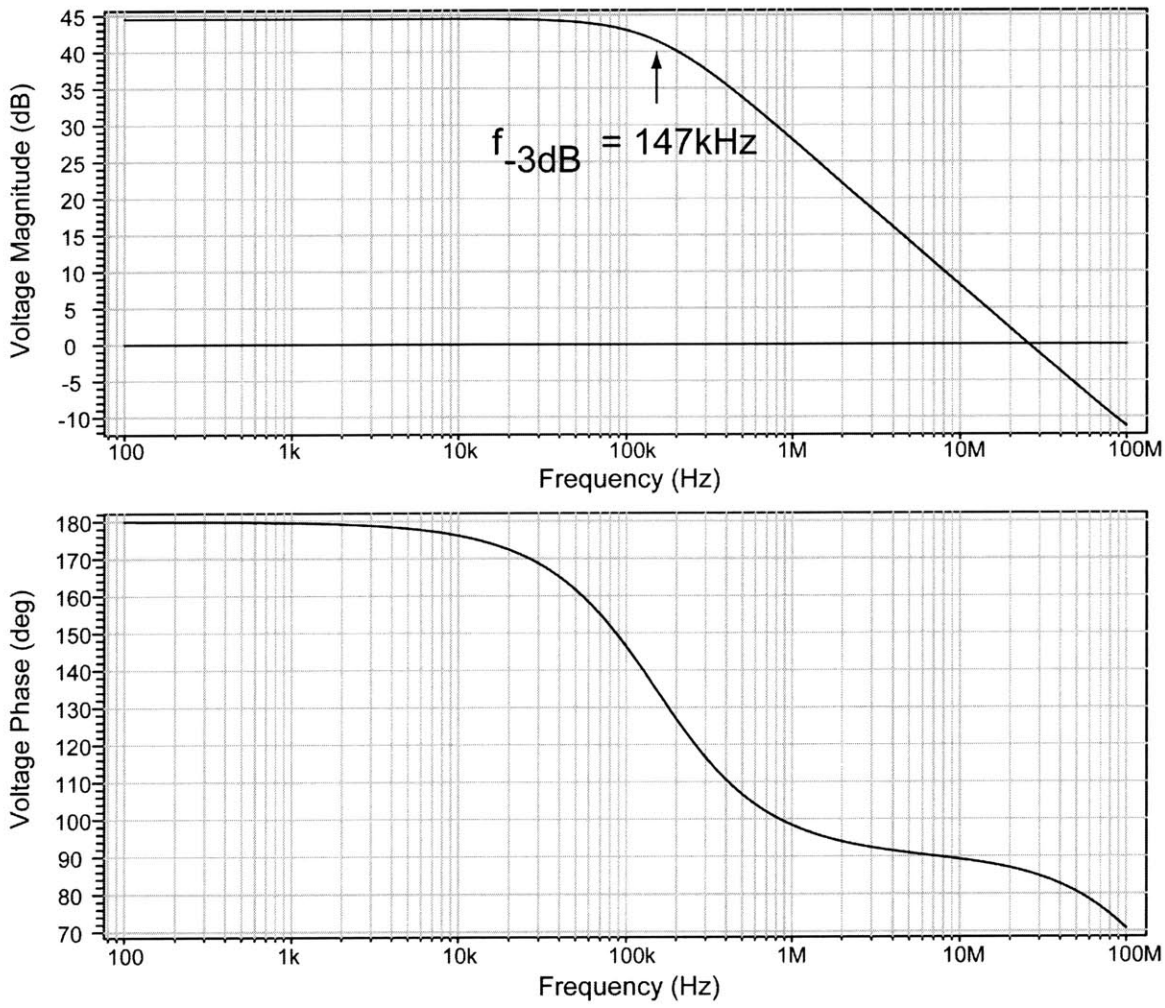


Figure 3.11 -- Bode plot of the transfer function of the common-source amplifier shown in Figure 3.10, with the PMOS transistor biased to supply 1μA.

Hand analysis predicts that the noise of the new topology should be

$$V_{in}^2(f) = V_{in,old}^2(f)\frac{g_{m,p}^2}{g_{m,n}^2} = \left(3.13\times10^{-15}\,V^2/Hz\right)\left(\frac{9.09\,\mu A/V}{14.38\,\mu A/V}\right)^2 = 1.25\times10^{-15}\,V^2/Hz, \quad (2.29)$$

or in volts per root hertz form

$$V_{in}(f) = 35.4\,nV/\sqrt{Hz}. \quad (2.30)$$

SPICE simulation gives the input-referred noise result shown in Figure 3.9, which agrees with the hand-calculated value to within 5%.

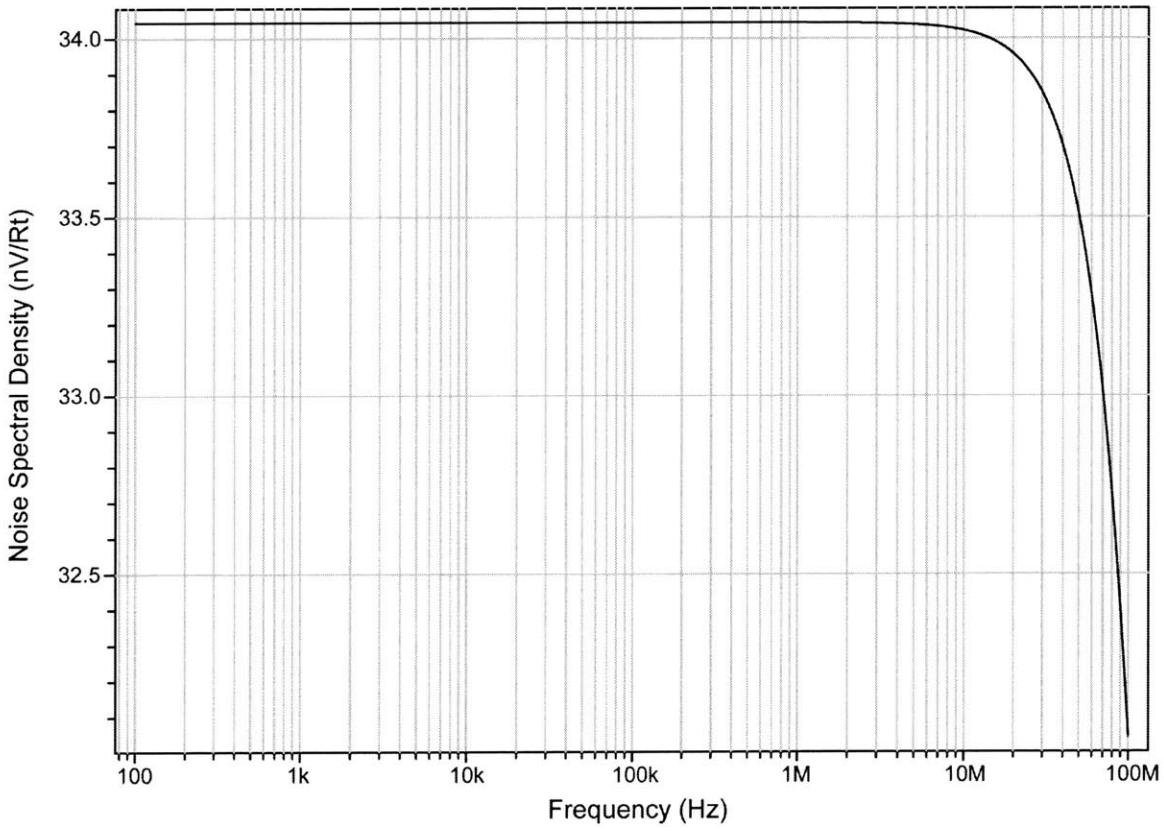


Figure 3.12 -- SPICE simulated input-referred noise spectral density for the NMOS-input common-source amplifier biased at I_D = 1μA.

3.2.3 PMOS-Input Operational Transconductance Amplifier (P-OTA)

A third analog component in this system is the PMOS-input operational transconductance amplifier (P-OTA), and is shown in Figure 3.13. This amplifier is

similar to the previous two amplifiers in its function, but differs in the fact that it has two input terminals and one output terminal. Let's take a moment to consider the general behavior of this circuit. M_1 acts as a current source. M_2 and M_3 form a differential pair and determine how the bias current from M_1 is divided between the two legs of the amplifier. M_4 sinks whatever current M_2 provides and forces M_5 to sink the same current from V_{out}, thus forming a current mirror. If V_+ and V_- are biased at the same DC voltage, then the differential pair will force the same current down both legs. If a small differential voltage is placed between V_+ and V_-, then slightly different currents will flow

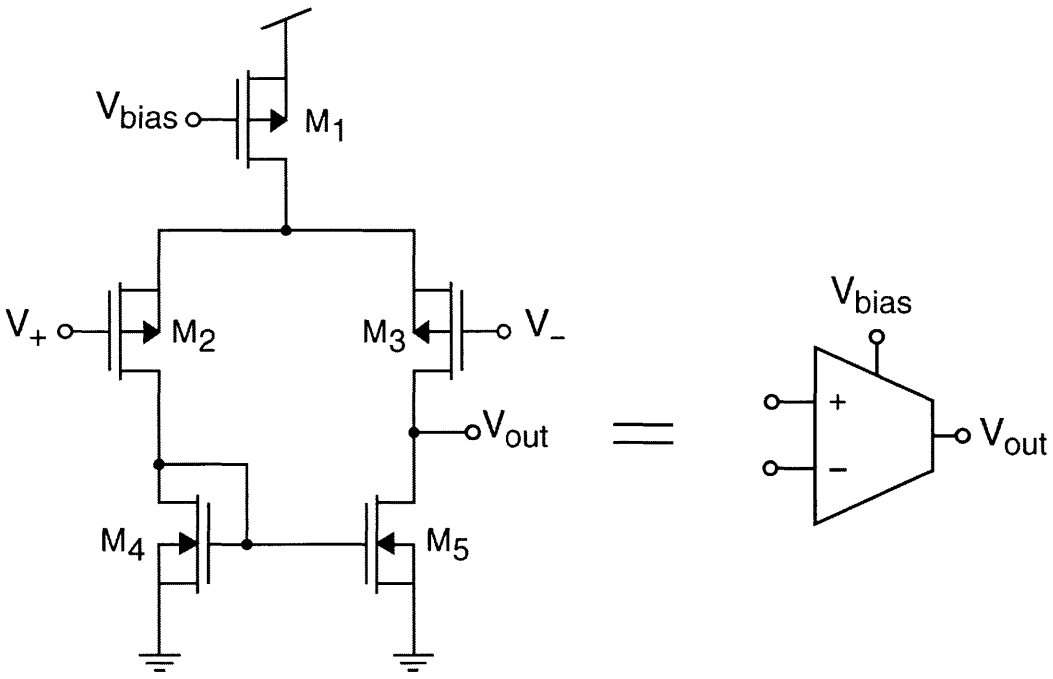


Figure 3.13 – PMOS-input operational transconductance amplifier (P-OTA) circuit (left) and the symbol used to represent it (right).

in each leg. These two leg currents are subtracted from each other at the output of the amplifier because of the current mirror, and are multiplied by the very large impedance at the output node to produce V_{out}.

3.2.3.1 Large-Signal Operation

The basic picture of OTA operation described above depends on all transistors remaining saturated. Because of this requirement, the common-mode range of the input voltage is limited. The common-mode of the input voltage is defined as

$$V_{cm} = \frac{V_+ - V_-}{2}. \tag{2.31}$$

In this circuit, M_1 enters the triode region of operation if V_{GS2} or V_{GS3} makes the following equation true

$$V_{cm} + \left|V_{GS2,3}\right| + \left|V_{DSAT1}\right| \geq V_{DD}. \tag{2.32}$$

Thus, this forms an upper-bound on the common-mode range of the circuit. On the lower end of operation, V_{cm} can equal ground and even go below ground before it causes any transistors to leave saturation, so the common-mode range includes ground. However, notice that the lower the common-mode voltage resides, the less the output of the amplifier can swing.

3.2.3.2 Small-Signal Operation

As was mentioned above, if all of the transistors remain saturated, then the OTA amplifies differential small-signal input voltages, producing a single-ended large-signal output voltage. A complete small-signal model hand-analysis of this circuit is extremely tedious. Instead, let's develop intuition about its small-signal behavior from Figure 3.14 and our knowledge of the small-signal behavior of the common-source amplifiers. If a differential voltage, Δv, is placed across the differential pair's gates, it will divide equally

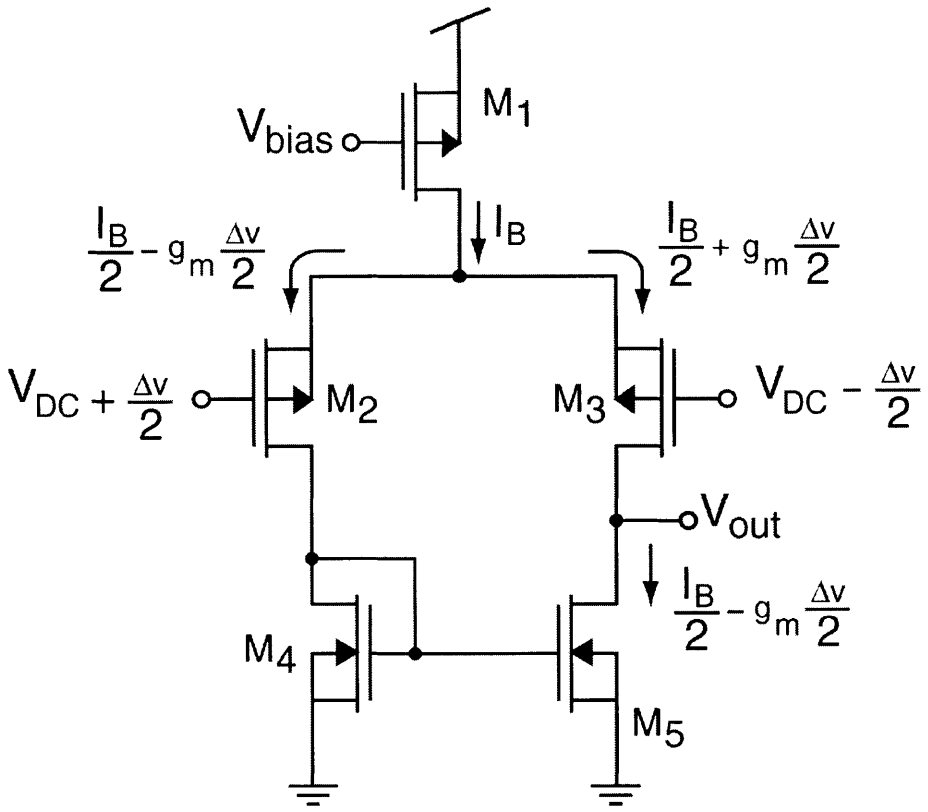


Figure 3.14 – Small-signal behavior of OTA currents given a small-signal voltage input.

between the two transistors' v_{gs} voltages, giving a change of $\Delta v/2$ across each, but in opposite directions. Each of these transistors has the same g_m, if they are sized the same, and so a current change of $g_m(\Delta v/2)$ develops in each, again in opposite directions. The increased current through M_3 arrives directly at the output, while a copy of the decreased current through M_2 arrives at the output due to the current mirror. Thus, the net current change at the output will be twice the value of the current change in one of the differential pair transistors, or

$$\Delta i_{out} = 2 \times g_m \times \frac{\Delta v}{2} = g_m \Delta v. \tag{2.33}$$

This current change will be multiplied by the impedance seen at the node V_{out} to give the basic transfer function of the amplifier, much like it did in the common-source amplifier presented earlier

$$\frac{v_{out}}{v_{in}} \approx \frac{g_{m2,3}\left(r_{out,3} \| r_{out,5}\right)}{1+s\left(C_{db,3}+C_{db,5}+C_{load}\right)\left(r_{out,3} \| r_{out,5}\right)}. \tag{2.34}$$

There are also secondary poles in this amplifier which can affect its operation. The first pole is due to the diode-connected (M_4) transistor's input impedance and a significant parasitic capacitance on this node. This time constant is roughly defined by

$$\tau_1 \approx \frac{1}{g_{m,4}}\left(C_{gs,4}+C_{gs,5}+C_{db,4}+C_{db,2}\right). \tag{2.35}$$

The other parasitic pole is due to the common-source node of the differential pair. This pole is defined by transistors M_2 and M_3's input impedance, and the parasitic capacitances at this node. The time constant is roughly

$$\tau_2 \approx \frac{1}{2g_{m2,3}}\left(C_{gs,2}+C_{gs,3}+C_{bsub,2}+C_{bsub,3}+C_{db,1}\right), \tag{2.36}$$

where $C_{bsub,2}$ and $C_{bsub,3}$ are due to the fact that the bulk terminals of M_2 and M_3 are tied to this common-source node, and these N-well bulks have a non-negligible capacitance to the substrate. An important point to make here is that SPICE *does not model* this capacitance as part of the PMOS model, and it must be included manually in the circuit netlist if its effects are to be included.

3.2.3.3 OTA Specification Development

Before the OTA design can be completed, a short description of its intended use is needed so that appropriate design specifications can be set. In one part of the memory

circuit we will need to sample a sensitive voltage onto a capacitor, and hold the voltage with low leakage. Figure 1.2 presented a method of accomplishing this task which required two op-amps. An alternate method is shown in Figure 3.15. In this circuit, a single PMOS transistor is used as the sampling switch for the capacitor C_{samp}. In order to reduce I_{lk} to the minimum possible value after the PMOS switch has opened, the well of the PMOS is bootstrapped, using a voltage buffer, to the same voltage as is being held on C_{samp}. If the voltage buffer is offset-free, I_{lk} will theoretically be zero. However, in all real circuits there is offset, and the leakage will be nonzero, but minimized.

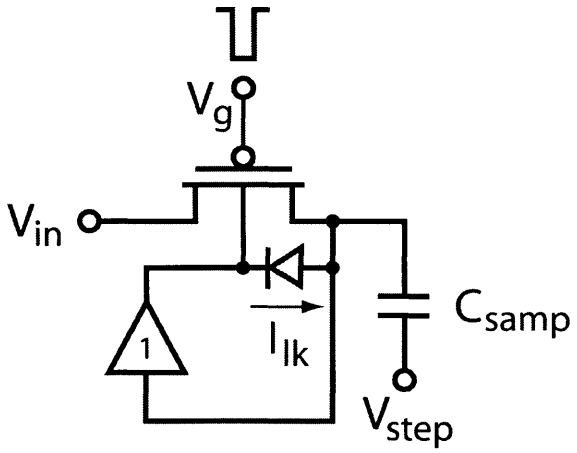
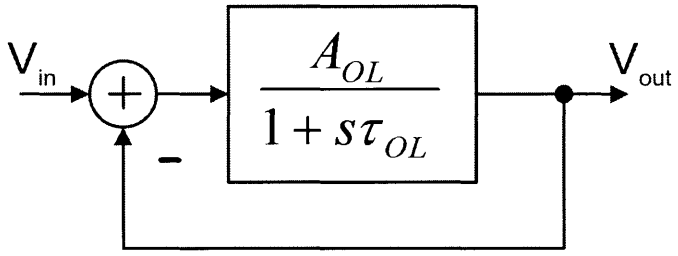


Figure 3.15 – PMOS transistor with bootstrapped well to reduce parasitic junction leakage while the transistor is off.

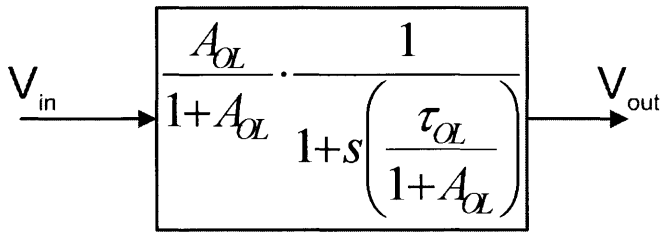
Not only should the buffer be capable of driving a DC voltage onto the well, but it must also meet certain transient requirements. The voltage at the top node of C_{samp} will move with a stepped-ramp characteristic (see Figure 2.1) between 0V and 2.2V as V_{step} is driven by a stepped-ramp generator. The buffer should be able to drive the well with this stepped-ramp signal with an error of less than 10mV at any instant in time. If an intentional error of +10mV is then included in the buffer, the diode will not become

forward biased during these transients. The buffer should be able to meet the 10mV specification for ramps with slopes as high as 10^5 V/s.

An OTA will be used in a unity negative feedback configuration as the buffer. The OTA in Figure 3.13 can be placed in unity-negative feedback by connecting V- to V_{out}. For the DC requirement, it is wise to include an intentional offset in the OTA so that the output is always at a slightly higher voltage than the input. This will ensure that the random input-offset voltage of the OTA doesn't accidentally forward-bias the leakage diode. For the transient requirement, we need to model a unity negative feedback OTA's



(a)



(b)

Figure 3.16 – Block diagram of an OTA in unity negative feedback (a), and a reduced block diagram of the same system (b).

response to an input ramp. Assume that the OTA's open-loop response is first order, as was derived in Eqn. 3.33, with an open-loop time constant τ_{OL}. If the OTA is now placed

in unity negative feedback, the block diagram in Figure 3.16 (a) models its behavior. This block diagram can be reduced using Black's formula to the block diagram shown in Figure 3.16 (b). We see that the magnitude of the gain through this system is close to one, as long as A_{OL} is large, and that the dominant pole's time constant has been reduced by a factor of $(1 + A_{OL})$.

The response of a single-pole transfer function to a ramp input is described in Figure 3.17. The first term in the output is what would be expected out of a "perfect" gain stage. The second term is an error term arising from the fact that the gain stage possesses a single-pole dynamic. Notice that the error is zero at zero time, and grows

$$V_{in} \cdot t \cdot u(t) \longrightarrow \boxed{\frac{A}{1 + \tau s}} \longrightarrow \underbrace{A \cdot V_{in} \cdot t \cdot u(t)}_{\text{Ideal Term}} \underbrace{- A \cdot \tau \cdot V_{in}\left(1 - e^{-\frac{t}{\tau}}\right)}_{\text{Error Term}}$$

Figure 3.17 – Response of a single-pole transfer function to a ramp input. The first term in the output would be the response of an amplifier with no dynamics. The second term in the output settles to a constant error term in steady state.

with time as the exponential term dies away. After several time constants, the exponential term is zero, and we are left with a constant error between the input and output ramp with a value of $A \cdot \tau \cdot V_{in}$. In order for this error to remain less than 10mV for a ramp with a slope of 10^5 V/s, we find the following restriction on τ (assume A = 1)

$$\tau \cdot V_{in} \leq 10mV \quad \rightarrow \quad \tau \leq \frac{10mV}{10^5 \, V/s} \quad \rightarrow \quad \tau \leq 100ns. \tag{2.37}$$

Referring again to Figure 3.16, we see that this corresponds to an open-loop -3dB bandwidth for the OTA of

$$\frac{\tau_{OL}}{1+A_{OL}} \leq 100ns \quad \rightarrow \quad \omega_{-3dB} = \frac{1}{\tau_{OL}} = \frac{1}{(100ns)(1+A_{OL})}. \tag{2.38}$$

We would like the OTA to achieve an open-loop gain of 40dB, and thus, from Eqn. 3.37, an ω_{-3dB} of

$$\omega_{-3dB} = \frac{1}{\tau_{OL}} = \frac{1}{(100ns)\left(1+100^{V}/_{V}\right)} \approx 10^{5} rad/s, \tag{2.39}$$

or a -3dB frequency of greater than 16kHz.

3.2.3.4 OTA Design

In the previous section, the intended use of the OTA was described. Several requirements on the DC and transient operation of the OTA were either stated outright, or implied. Each of these requirements will be addressed separately in this section.

The DC specification requires that the OTA maintain a reverse bias on the bulk-to-drain diode of the PMOS switch at all times, but that this reverse-bias voltage should not be any larger than necessary. There are several factors that determine how much, if any, intentional DC offset should be included in the OTA design. First, as mentioned earlier, there is always some random error in the input-offset voltage due to process variation. This error can be as large as 5mV. An intentional offset should be added to the DC design to ensure that even with this random error present, the OTA will never have a net offset that leads to a forward bias on the diode.

A second reason for including an intentional DC offset was also mentioned earlier. During ramp transients, the input and output voltages of the OTA will be

different because of the error term shown in Figure 3.17. A solution to this problem is to include a fixed DC offset with a magnitude larger than this transient error term.

A third reason for including a DC offset is that the finite gain of the OTA introduces a gain error. Looking at Figure 3.16 and assuming A_{OL} is 40dB, the magnitude of the transfer function will be 100/101, which is roughly 1% lower than the ideal magnitude of 1. This non-ideality creates a gain error in the input-output transfer function of the OTA, which means that the difference between the input voltage and output voltage changes as the magnitude of the input value changes. The worst-case error occurs when the magnitude of the input is largest, which in this design will be about 3V. At this value, the gain error will cause a difference between the input and output of

$$Gain\ Error = V_{in} - V_{out} = V_{in} - \frac{100}{101}V_{in} = (3V)\left[1 - \frac{100}{101}\right] = 29.7mV. \qquad (2.40)$$

Without increasing the gain, an easy way of counteracting this error is to add 30mV to the intentional offset of the OTA.

The last two errors both cause the output voltage of the OTA to be less than the input voltage. The first error can be in either direction, so assuming the worst-case scenario, it also causes the output to be less than the input. Thus, the intentional DC offset should be positive, and should have a magnitude larger than all three of the errors combined. A DC offset of 50mV should be enough. This intentional DC offset can be added to the OTA by sizing M_3's W/L ratio larger than M_2's W/L ratio by an appropriate amount. The reason why this sizing adds an offset in the positive direction can be seen by again referring to Figure 3.13, and remembering the OTA has been placed in unity-negative feedback. The negative feedback forces the currents in each leg of the OTA to equalize. Thus, the currents in M_2 and M_3 must be equal. Since both transistors are

saturated, and since the W/L ratio of M_3 is larger than that of M_2, the gate-to-source voltage of M_3 will be smaller than that of M_2 for the same level of current. This implies that the gate of M_3 is closer to V_{DD} than the gate of M_2, and thus the output has a positive DC offset from the input.

This gain error term was overlooked in the original system design, and may actually introduce some of the experimental problems seen in this implementation. We will address this issue in the next chapter. For the rest of this section, however, assume the intentional DC offset should be 15mV.

All of the design specifications are now known except for the value of the load capacitor. The primary capacitor will be the bulk-to-substrate capacitance of the PMOS switch, whose value should be roughly 400fF, based on process data from MOSIS. Using SPICE, a reasonable set of parameters was established for the OTA's transistor sizes and biasing. The sizes used for the OTA transistors are shown in Table 3.2, and the

OTA Transistor	Width (μm)	Length (μm)
M_1	11.2	3.2
M_2	9.6	3.2
M_3	14.4	3.2
M_4	6.4	3.2
M_5	6.4	3.2

Table 3.2 – Transistor sizing for the P-OTA that was actually implemented on-chip.

bias current was set to 0.8μA. The resulting open-loop transfer function is shown in Figure 3.18. The simulations predict a -3dB frequency of 21kHz, which is equivalent to 1.3×10^5 rad/s, and a DC gain of 39.9dB. Although it is not shown here, the phase sharply decreases between 20MHz and 100MHz; this happens due to the parasitic poles that were

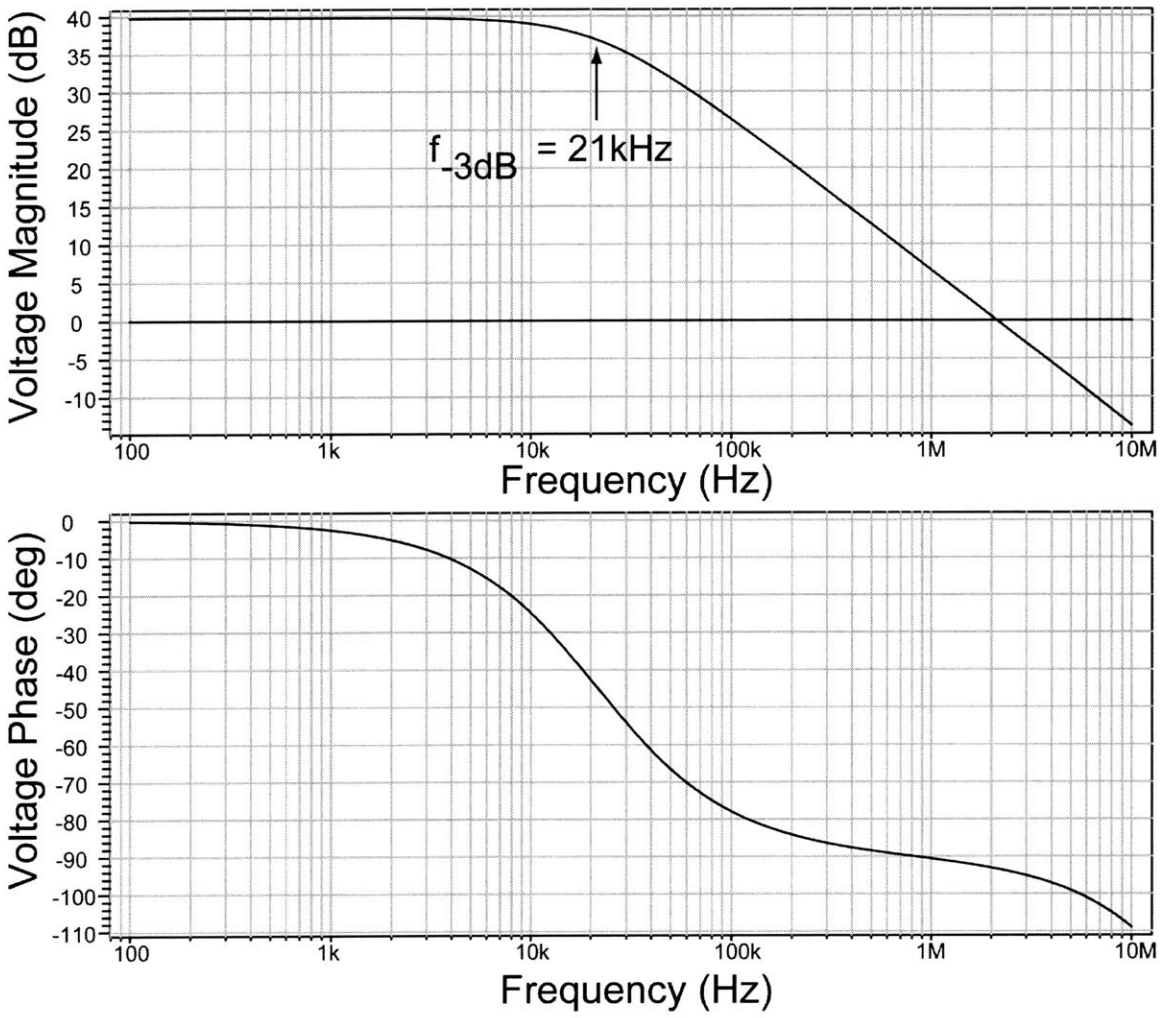


Figure 3.18 -- Bode plot of the transfer function of the P-OTA shown in Figure 3.13, biased at 1μA.

predicted in Eqns. 3.34 & 3.35. The simulations also show that the offset of the amplifier is +15.5mV, which was the desired offset, and that the common-mode spans the range from 0V to 2.2V. The input-referred noise voltage per root hertz of the open-loop OTA is shown in Figure 3.19.

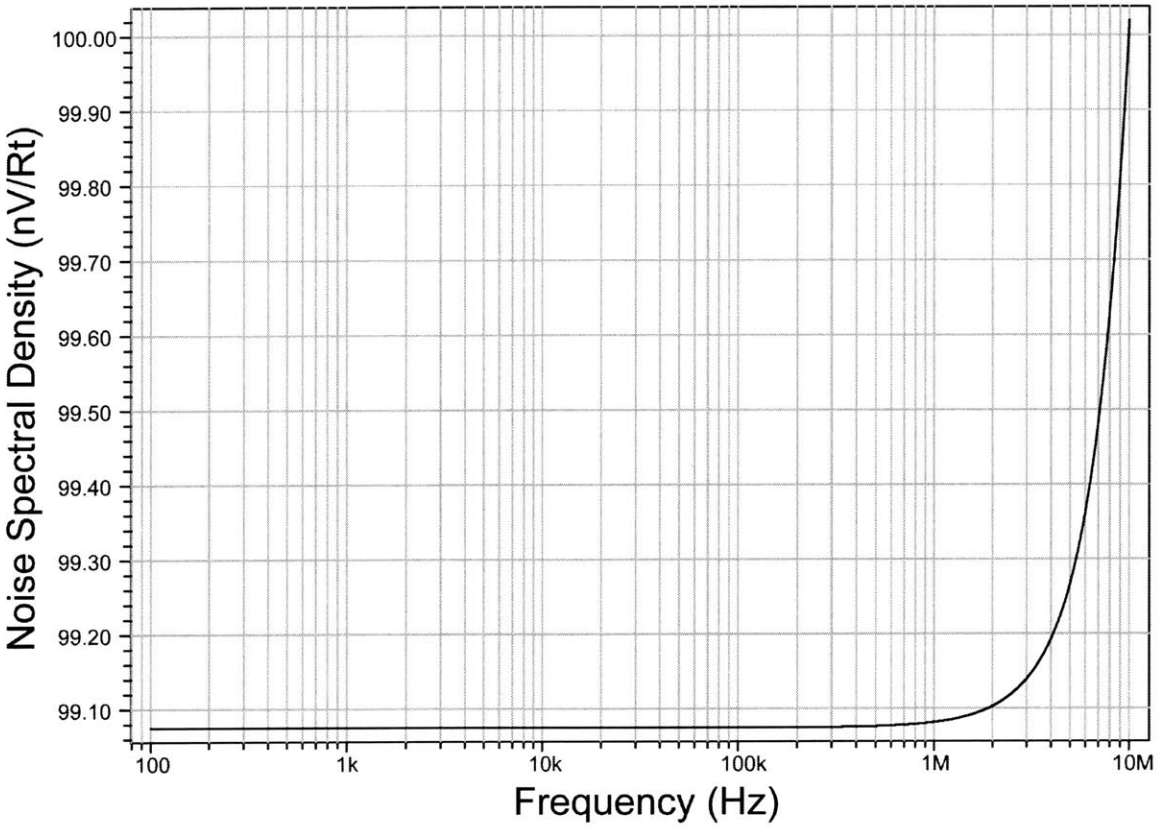


Figure 3.19 -- SPICE simulated input-referred noise spectral density for the PMOS-input OTA biased at $I_D = 1\mu A$.

3.2.4 Transmission Gate

The final analog building block used to construct the memory cell is the transmission gate. The circuit structure and symbol for this device are shown in Figure 3.20. The transmission gate is made of an NMOS and PMOS transistor whose sources and drains are tied together to form an input and an output node. The gate of the NMOS is connected to the digital signal V_{on}, while the gate of the PMOS is connected to the digital signal $/V_{on}$. When $V_{on} = V_{DD}$, both the NMOS and PMOS conduct, and when $V_{on} = 0V$, both the NMOS and PMOS are off. Thus, the transmission gate is a voltage-controlled switch. The reason that an NMOS and PMOS are placed in parallel to create the switch is so that the switch is able to pass rail-to-rail voltages. NMOS transistors can

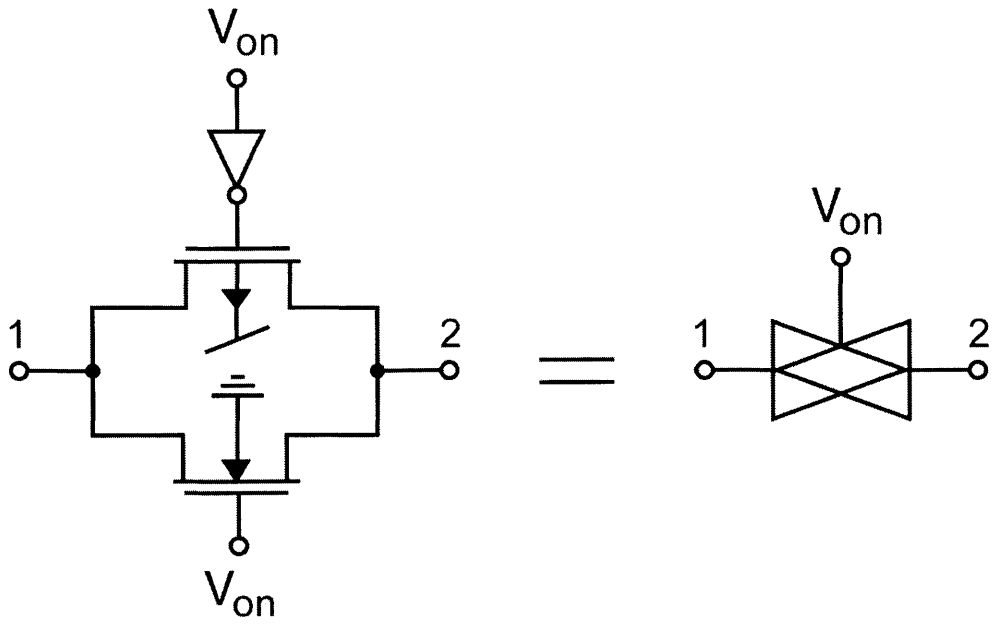


Figure 3.20 – Transmission gate circuit (left) and the symbol used to represent it (right).

pass voltages from 0V to V_{DD}-$|V_{t,n}|$ effectively, and PMOS transistors can pass voltages from $|V_{t,p}|$ to V_{DD} effectively. Thus, in parallel, they can pass voltages from 0V to V_{DD}. The parameter of primary importance in this circuit is the "ON" resistance, R_{ON}. This parameter is a function of the voltages that are applied to the terminals of the switch (nodes 1 and 2), and is lowest when the applied voltage is near the rails and highest when the applied voltage is between the rails. Both the NMOS and PMOS are minimum-size transistors with W=2.4μm and L = 1.6μm for all of the transmission gates used in this design. A SPICE simulation of the value of R_{ON} as a function of the applied terminal voltage is shown in Figure 3.21.

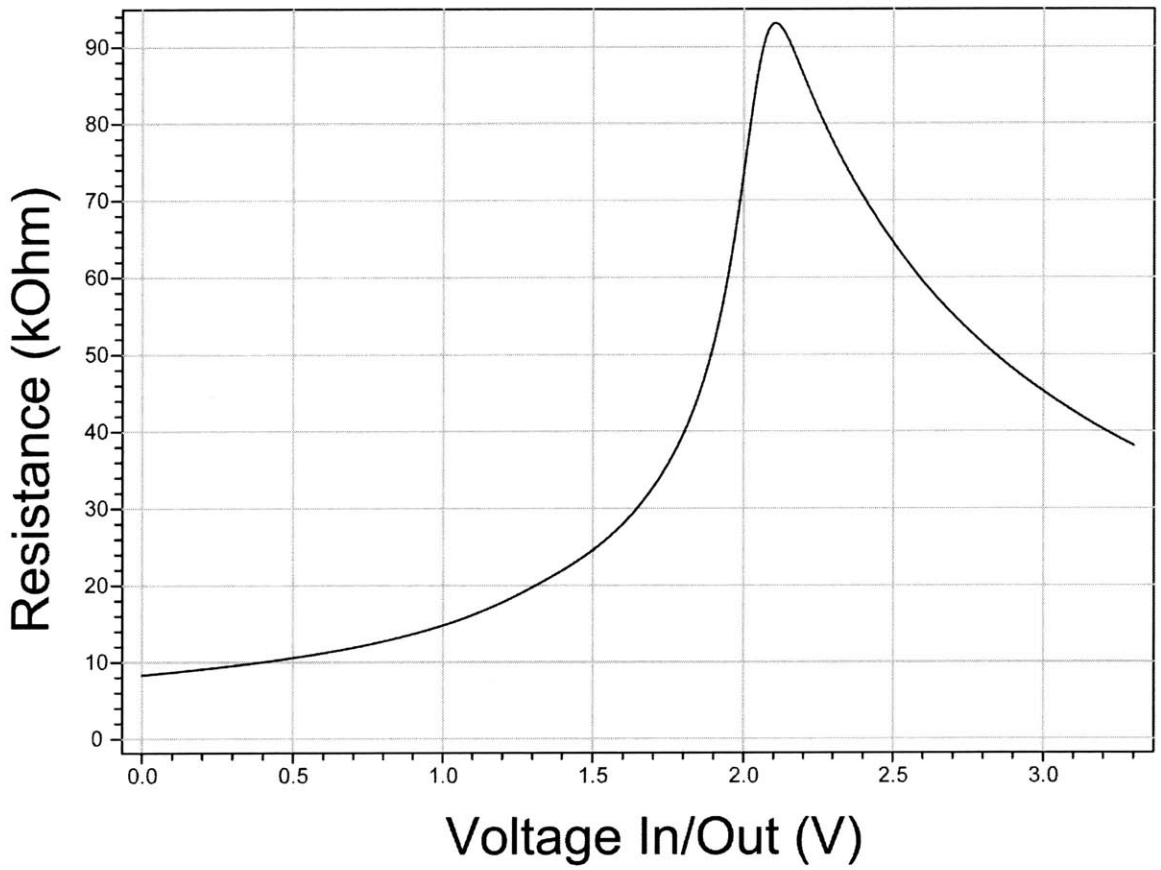


Figure 3.21 – The "ON" resistance of a minimum-size transmission gate as a function of the voltage level it is passing.

3.3 The Complete Memory Cell System

All of the major analog components that will be used to build the analog memory system were introduced in the previous section. This section will outline how these blocks are pieced together to form the local memory cell and the global ramp generator.

3.3.1 Local Memory Cell Design

The primary functions of the local memory cell were introduced in Figure 2.1. To review, the required functions were:

1. Sample-and-hold an input analog voltage with low leakage.

2. Monitor the global stepped-ramp bus to determine when the held voltage has been surpassed by the ramping signal.
3. Sample-and-hold the nearest quantized level from the stepped-ramp bus as soon as the local value has been surpassed.
4. Repeat this procedure for as long as the voltage needs to be stored.

To accomplish these tasks, the local cell must contain circuitry capable of performing a high-precision, low-offset comparison, and circuitry that implements a very precise and low-offset sample-and-hold. The total combined input-referred offset of the comparator and the sample-and-hold, along with leakage currents, directly limits our maximum storage precision, and should be made as small as possible. To understand why the input offset of the comparator must be small, consider a sample set of waveforms from a storage cell with a non-negligible DC offset, as shown in Figure 3.22. Notice that the true capacitor voltage $V_{Cstore,actual}$ looks like the DC offset voltage $V_{Cstore,offset}$ to the comparator due to its input-offset. Thus, the comparator switches one cycle too early.

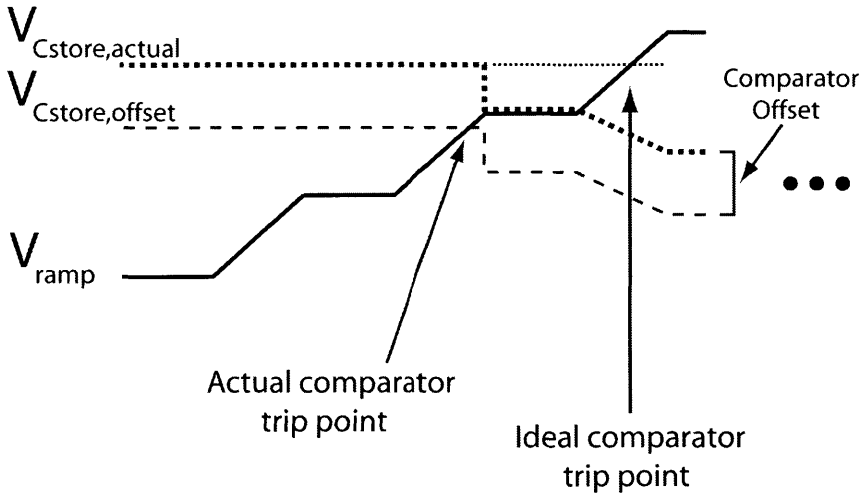


Figure 3.22 – Sample waveforms from a quantization step in which the input-offset of the comparator causes a loss of information.

This causes the actual capacitor voltage to be quantized in the wrong direction. Since this process repeats continuously in order to store the analog voltage, successive quantization steps will continue to reduce the stored voltage by one bit until it is zero. A similar situation would occur if the offset were in the positive direction, with the stored voltage settling at V_{DD}. Similarly, if the sample-and-hold circuit possesses an offset that is too large, the same railing process will occur. Thus, unlike a normal A/D converter in which a constant offset voltage can be tolerated, because of the cyclical operation of this circuit, offsets must be made very small in comparison to the desired storage precision.

The auto-zeroing sample-and-hold comparator circuit shown in Figure 3.23 (a) was designed to accomplish all of the required tasks. Part (b) of the figure shows the timing of the digital signals that control the circuit. Let's examine how the cell operates. The sampling cycle begins when $V_{off,1}$ switches low, turning on the PMOS transistors M_1 and M_2, and turning off T_1. At the same time that $V_{off,1}$ switches from high to low, $V_{off,2}$ switches from low to high turning T_2 and M_3 on and T_3 off. A simplified schematic of the circuit during this state of operation is shown in Figure 3.24. In this state, both the P-input and N-input common-source amplifiers are connected in unity-negative feedback. Referring to the P-input common- source amplifier in Figure 3.5, we see that connecting the amplifier in unity negative feedback temporarily diode connects the PMOS transistor. Thus, the input and output settle to the gate voltage which ensures that the PMOS carries the same current as the NMOS provides. Applying unity negative feedback to the N-input common-source amplifier yields complementary results. Note that these gate

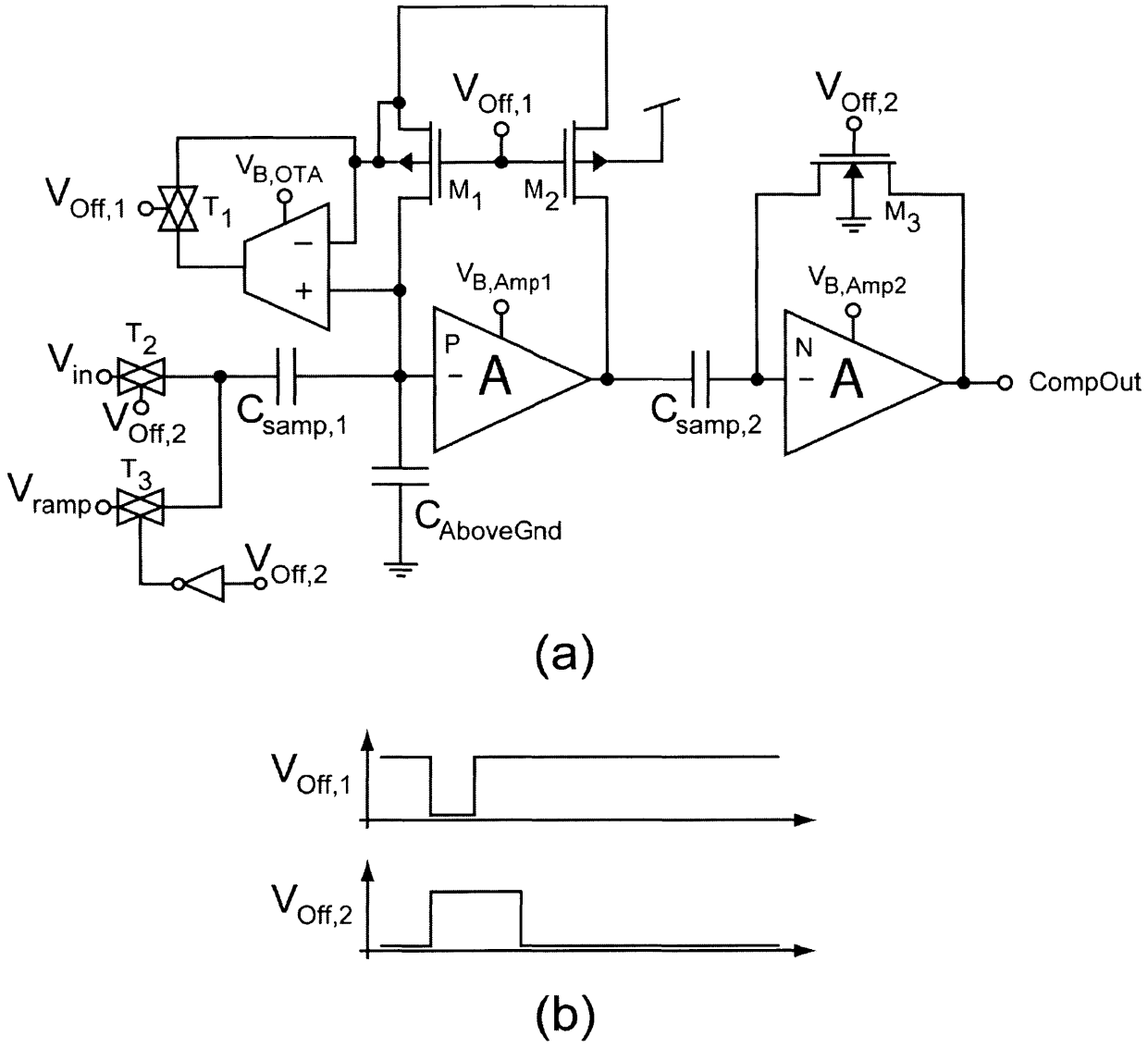


Figure 3.23 – (a) The local memory cell's auto-zeroing sample-and-hold comparator. (b) The waveforms that drive the auto-zeroing sample-and-hold comparator.

voltages are exactly the voltages which will bias both of the amplifiers in their high-gain regions during open-loop operation. Thus, in this state, the left plate of $C_{samp,1}$ is charged to the voltage V_{in} and the right plate of $C_{samp,1}$ is charged to the voltage which biases the P-input amplifier in its high-gain region. Also, the left plate of $C_{samp,2}$ is charged to the output voltage of the P-input amplifier, and the right plate of $C_{samp,2}$ is charged to the voltage which biases the N-input amplifier in its high-gain region.

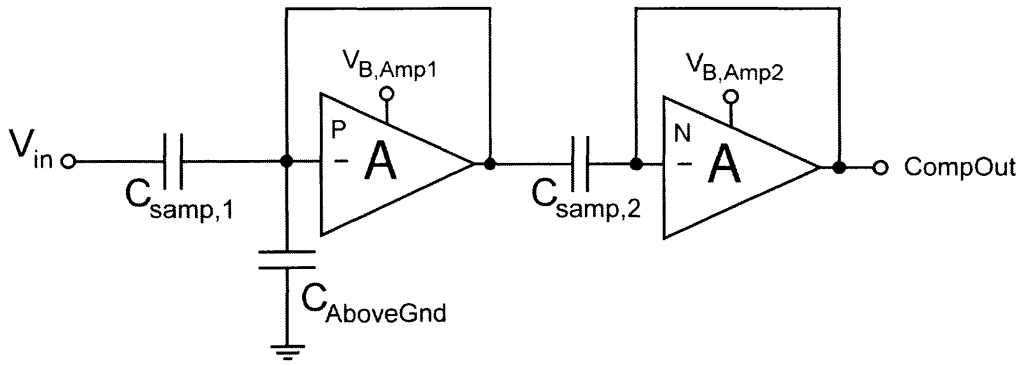


Figure 3.24 – Simplified schematic showing connectivity present in the local memory cell during the first stage of sampling.

The second half of the sampling cycle begins when $V_{off,1}$ switches from low to high. At this transition point, T_1 turns on, and M_1 and M_2 turn off. The other transmission gates and transistors remain as they were during the previous state. A simplified schematic of the circuit during this state of operation is shown in Figure 3.25. The first important change that occurs during this transition is that the feedback around the P-input amplifier is broken due to M_1 and M_2 turning off. Also, the OTA now bootstraps the well terminal of M_1 to reduce leakage, as was described in Section 3.2.3.3. An important error term is introduced in this step also: charge is injected onto the right

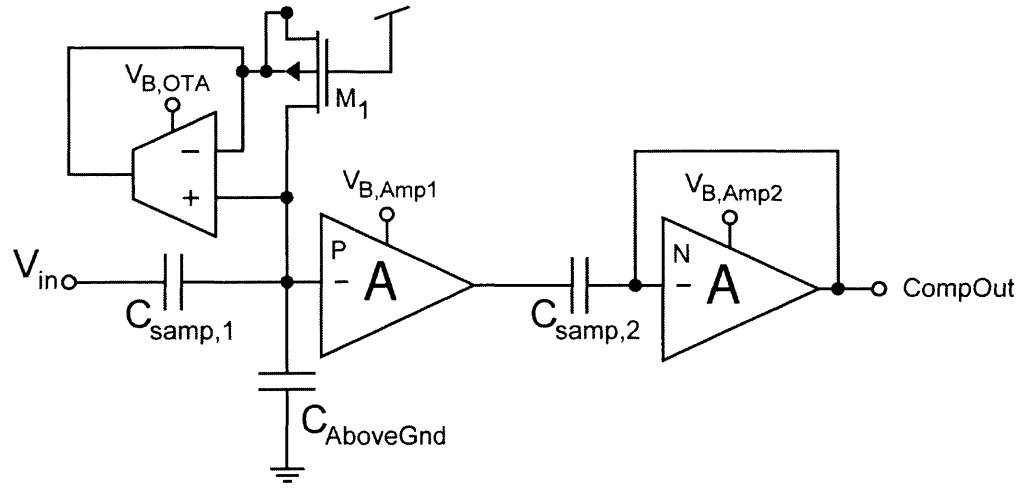


Figure 3.25 – Simplified schematic showing connectivity in the local memory cell during the second stage of sampling.

plate of $C_{samp,1}$ due to M_1 and M_2 turning off, causing a voltage change Δv on this node. The P-input amplifier will amplify Δv by its open loop gain, and was designed with only 40dB of gain to ensure that amplifying this charge injection will not cause its output to rail.

Suppose that $C_{samp,2}$ and the N-input amplifier were not present in this circuit, and we simply provided more gain in the first stage to create the high-gain comparator. After the above steps, if the input voltage on the left plate of $C_{samp,1}$ is swept from 0V to V_{DD}, the output of the P-input amplifier would transition from V_{DD} to 0V when the input ramp reaches V_{in} - Δv rather than at V_{in}, which is the desired comparison point. Thus $C_{samp,1}$ must be made large to keep Δv small enough for the desired precision of the memory. Rather than use a single-stage high-gain comparator, a second moderate-gain amplification stage was added. After the output of the P-input amplifier settles, the N-input amplifier is still sampling its own bias point and the output of the P-input amplifier. Thus, the error due to charge injection from M_1 and M_2 has been output referred, and sampled onto $C_{samp,2}$, and thus does not affect the final comparator transition point. Notice that the charge injection from M_3 turning off is injected onto the right plate of $C_{samp,2}$, and will affect the comparator transition point. However, the voltage change from this charge injection must be input-referred to determine its effect on the overall comparator transition point, and is thus divided by the gain of the P-input stage, and the capacitive attenuator at the input. The trip point of the comparator is described by

$$V_{trip} = V_{in} - \frac{\Delta v_{M3}}{\left(A_{amp,P} \cdot \frac{C_{samp,1}}{C_{samp,1} + C_{AboveGnd}} \right)} \tag{2.41}$$

where Δv_{M3} is the voltage change on $C_{samp,2}$'s right plate due to charge injection from M_3, $A_{amp,P}$ is the gain of the P-input amplifier, and the capacitor ratio is the attenuation factor of the input capacitive divider. Even with a charge injection from M_3 of 20mV (much higher than will actually occur), the input-referred offset due to this charge injection is less than 0.5mV.

The third and final state of operation for this circuit is the comparison state, which occurs while $V_{off,1}$ is high and $V_{off,2}$ is low. In this state of operation, T_1 and T_3 are on, while T_2 and M_1-M_3 are off. A simplified schematic of the local memory cell during its third state of operation is shown in Figure 3.26. In this state, V_{ramp} is now connected to the left plate of $C_{samp,1}$, and moves in a stepped-ramp movement from 0V to V_{DD}. At a certain point in time, the stepped-ramp voltage will surpass the value of V_{in} that was used to originally tune the circuit to the proper bias point. CompOut will then transition from low to high with a gain of over 74dB (less than 80dB due to the capacitive attenuator).

$C_{AboveGnd}$ is included in this circuit for the sole purpose of limiting the voltage

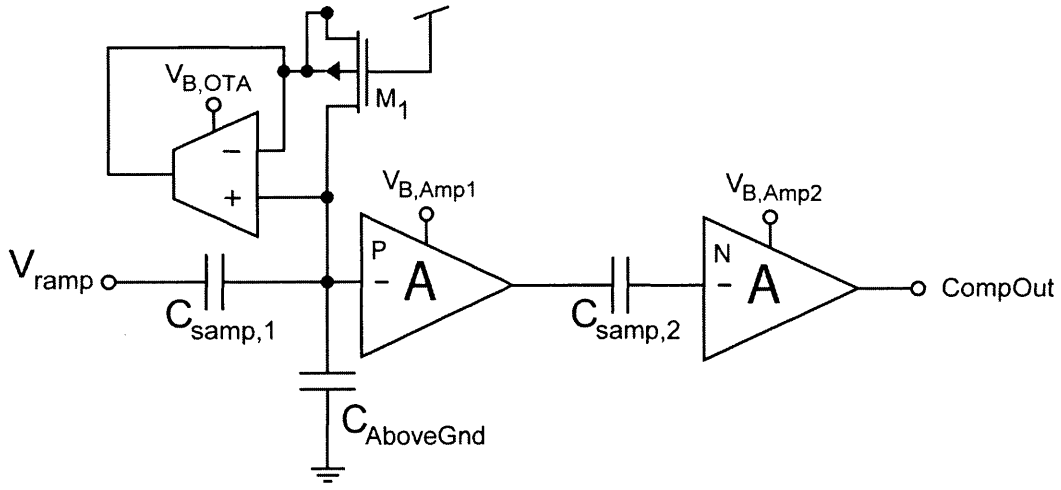


Figure 3.26 -- Simplified schematic showing connectivity in the local memory cell during the third stage of sampling.

swing at the input node to the P-input amplifier. To see why it is necessary, assume V_{in} = 3.0V, and that the high-gain bias point of the P-input amplifier is 2.2V in the sampling phase. Thus, $C_{samp,1}$ will store 3.0V on its left plate and 2.2V on its right plate. When the cell transitions to the comparison state, the left plate of $C_{samp,1}$ is pulled to 0V, and, without $C_{AboveGnd}$, the right plate would be pulled to -0.8V. The OTA is not able to drive voltages below ground onto the well of M_1, and thus, the circuit will not operate properly. Therefore, $C_{AboveGnd}$ should be sized so that the input to the P-input amplifier never falls below the common-mode range of the OTA follower.

Table 3.3 lists the sizes of all of the components used to construct the local memory cell of Figure 3.23. The memory cell was simulated in SPICE using these parameters to ensure that the system operates correctly. The digital signals used to control the system during simulation are shown in Figure 3.27. The voltage waveform present on the right plate of $C_{samp,1}$ is shown in Figure 3.28. A tiny glitch is visible on this waveform at the rising edge of $V_{off,1}$, due to the charge injection into this node from M_1 and M_2 turning off. This portion of the curve is magnified in Figure 3.29, where we can see that the magnitude of the charge injection is roughly 2.5mV. In order to arrive at this value, the capacitors $C_{samp,1}$ and $C_{AboveGnd}$ had to be chosen large enough. Looking at

Device	Parameters
T_1-T_3	All transistors: W=2.4μm, L=1.6μm
Inverter	All transistors: W=2.4μm, L=1.6μm
OTA	See Table 3.2
P & N-Input Amplifiers	All transistors: W=8μm, L=4μm
$C_{samp,1}$	400fF
$C_{AboveGnd}$	200fF
$C_{Samp,2}$	250fF
M_1-M_3	All transistors: W=2.4μm, L=1.6μm

Table 3.3– List of component sizes and values for the fabricated memory cell.

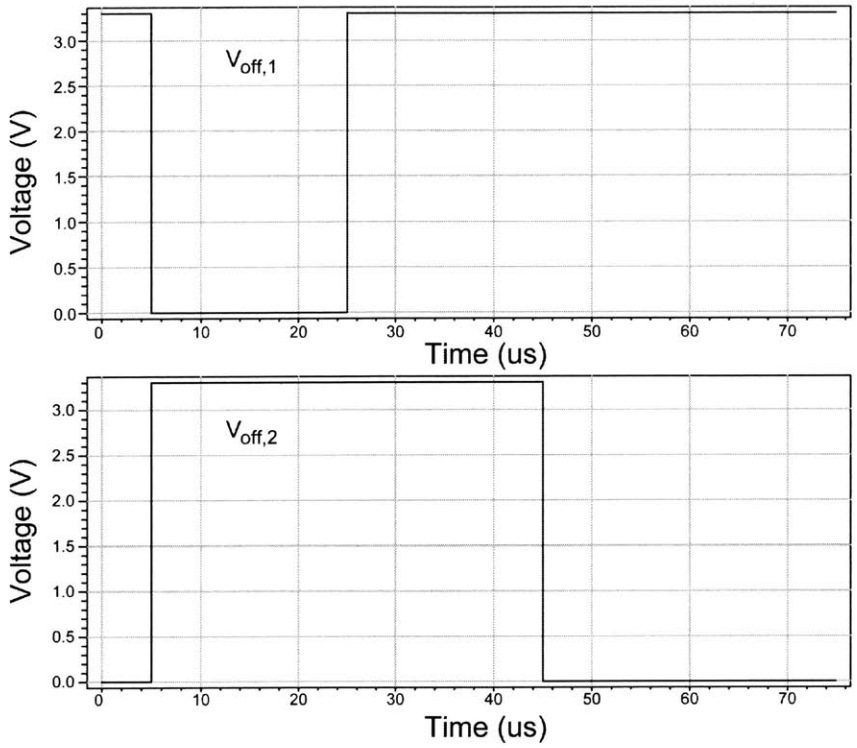


Figure 3.27 – The digital control signals that were used to test the memory cell's operation.

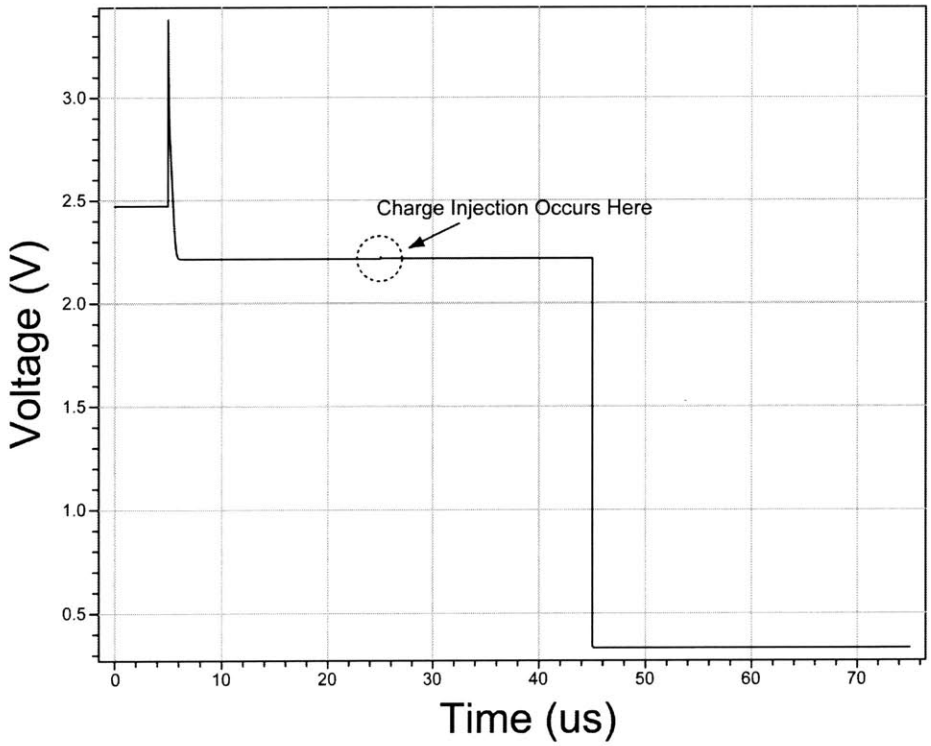


Figure 3.28 – Voltage on the right plate of $C_{samp,1}$ during the sampling cycle. The circled area is where M_1 and M_2 turn off, and the charge injection onto this node occurs.

the simulated voltage waveform at the output of the P-input amplifier, shown in Figure 3.30, we see that the tiny voltage change due to charge injection at the input is amplified by a factor of 100 at the output. However, the amplifier is still not close to railing. Notice that the output voltage takes much longer to settle in this portion of the curve than in other portions. This is because during this phase of operation, the P-input amplifier is operating open-loop, and thus the -3dB time constant determines its settling time, while during other portions of the curve, the amplifier is in unity negative feedback, and the time constant of settling is the reciprocal of the unity-gain frequency of the amplifier. Looking at the time constant of this settling, we notice an inconsistency between what was presented earlier, and the amplifier behavior now. The settling time, which should

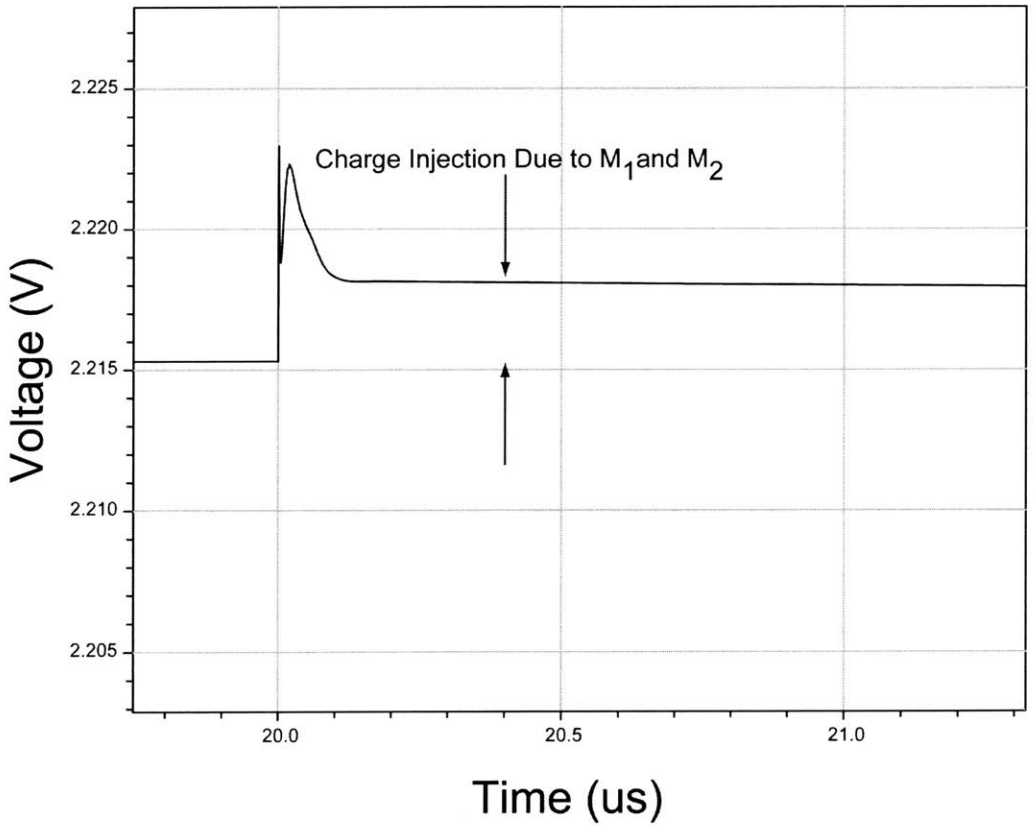


Figure 3.29 – Charge injection due to M_1 and M_2 turning off. We see the magnitude of the injection is about 2.5mV.

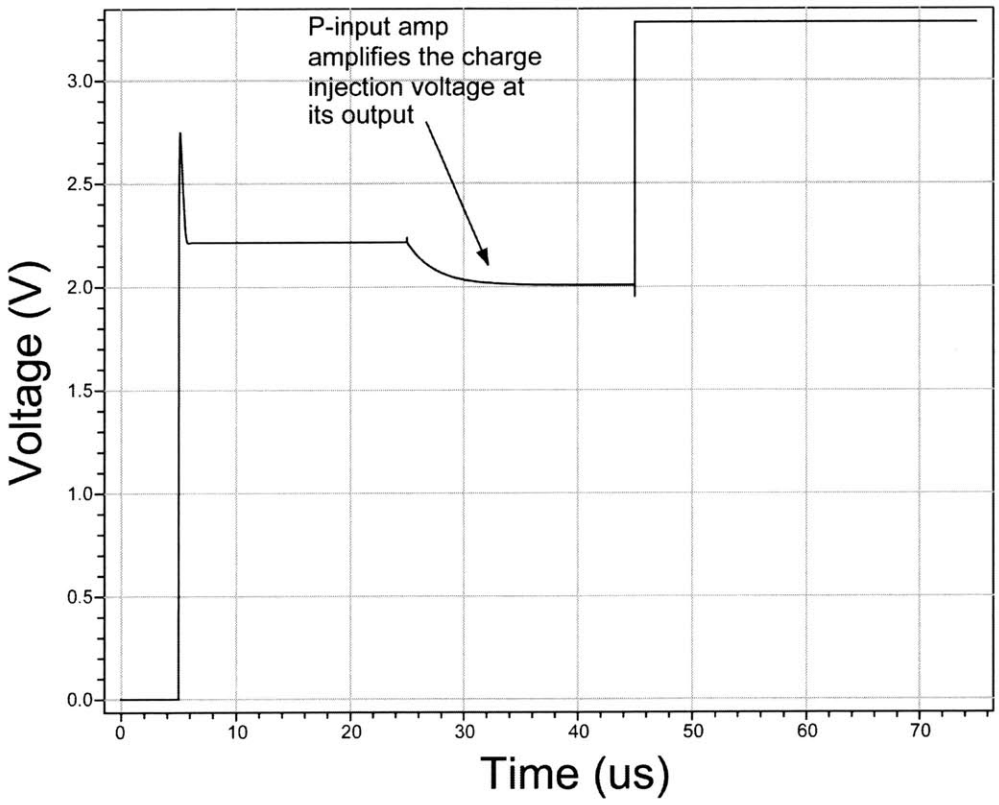


Figure 3.30 – The output voltage from the P-input amplifier. The curved section of the waveform is the portion in which the charge-injection induced voltage at the input is amplified at the output.

roughly equal 5τ, looks to be around $15\mu s$. This implies $\tau = 3\mu s$, rather than the original $1\mu s$. The answer to this dilemma is that the capacitance loading the P-input amplifier is larger when the N-input amplifier is in unity negative feedback than when it is open loop. In closed loop, the input to the N-input amplifier looks like a virtual ground, and thus the full value of $C_{samp,2}$ loads the P-input amplifier's output. Since this load capacitance is 250fF, while the amplifier design used a 70fF load capacitor, an increase in settling time by a factor of three makes sense.

The precision of the comparator when the sampling phase is over is shown in Figure 3.31. The top plot in this figure shows the input voltage as a function of time. The input voltage present during the sampling phase of operation was 3.0000V. After

sampling, the input is ramped from 2.9995V to 3.0005V -- this area of the plot is marked by the vertical dashed lines. The bottom plot shows the output of the comparator due to this change in input voltage. The 1mV change at the input is amplified to a nearly rail-to-rail output from the comparator. This proves that the auto-zeroing sample-and-hold comparator can sample an input voltage, compare it with a second input voltage, and do so with input-referred offset of <1mV. Since the input to the circuit is capacitively coupled, the allowed voltage input range is rail-to-rail, and the precision is over 11 bits.

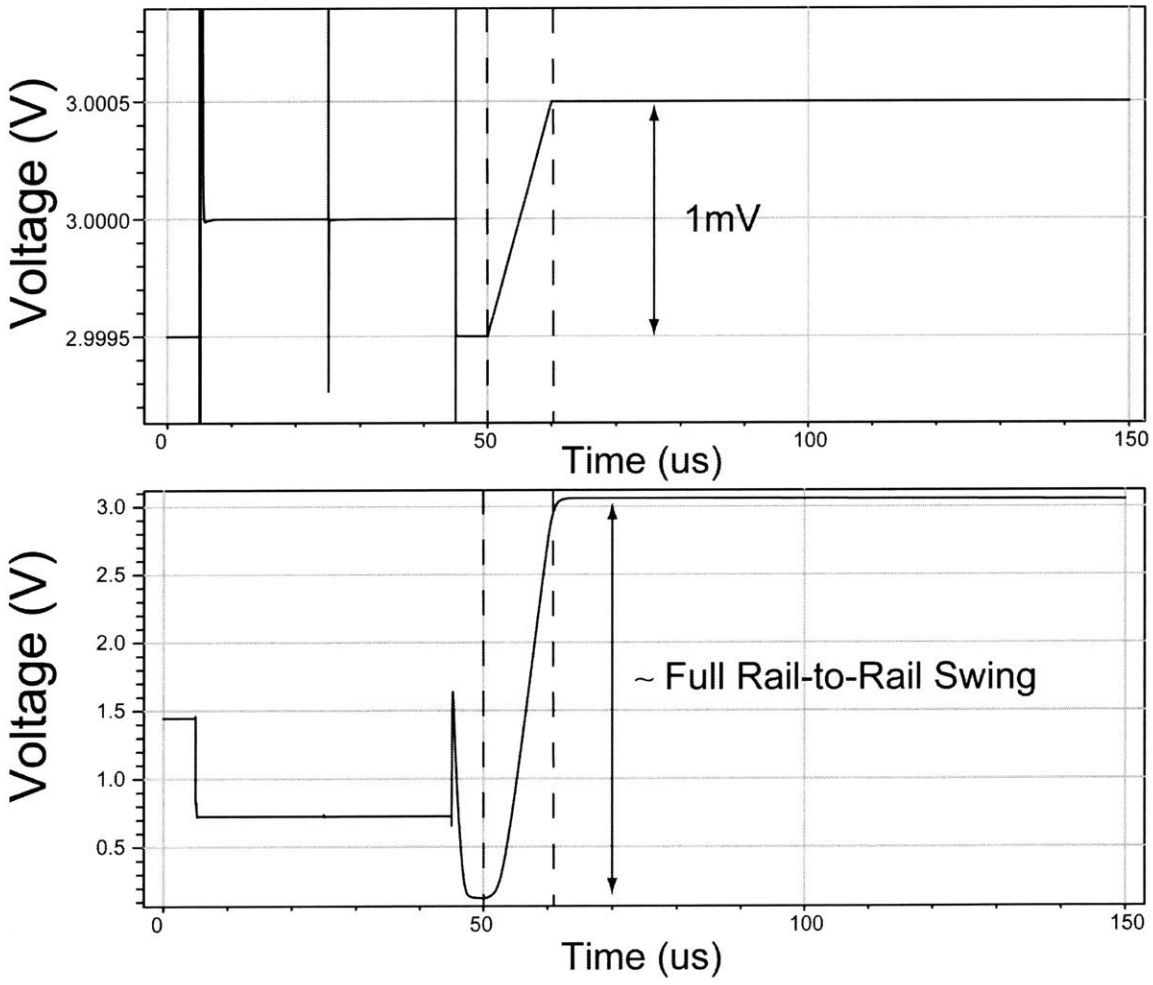


Figure 3.31 – The top plot shows the input voltage to the comparator, while the bottom plot shows the output voltage. Between the vertical dashed lines, the input is varied around the tuned input voltage of 3V by 1mV while the output swings nearly rail-to-rail.

The final parameter that should be determined is the total input-referred noise of the comparator after the sampled-bias points have been established. The first assumption that will be made is that auto-zeroing has eliminated all of the low-frequency 1/f noise. This is not entirely valid, but is a good place to begin the analysis. For a discussion of auto-zeroing and its effects on noise, refer to Appendix A. Thus, we are left with only white noise in this amplifier. A second assumption which also seems reasonable is that the noise of the N-input amplifier can be ignored. This is true because the noise of this amplifier is divided by the gain of the P-input amplifier when it is input-referred. Thus, the total input-referred noise of the amplifier should be (assume 1st order behavior)

$$V_{inoise,tot} = \sqrt{V_{in}^2 \cdot f_o \cdot \frac{\pi}{2}} = \sqrt{\left(60 nV/\sqrt{Hz}\right)^2 (160kHz)\left(\frac{\pi}{2}\right)} = 0.03mV_{rms}. \qquad (2.42)$$

The plot shown in Figure 3.32 shows the SPICE noise simulation results for the comparator circuit. The total output noise is shown to be 180mV. The overall gain of the amplifier, including the capacitive divider at the input is 5.5×10^3 V/V, which gives a total input-referred noise of

$$V_{inoise,tot} = \frac{180mV}{5.5\times10^3 V/V} = .0327mV_{rms}. \qquad (2.43)$$

This result is in excellent agreement with theory, so our two assumptions are most likely correct. The value of input-referred noise that this circuit exhibits is much lower than is necessary for this application, and has basically no impact on the overall resolution of our circuit.

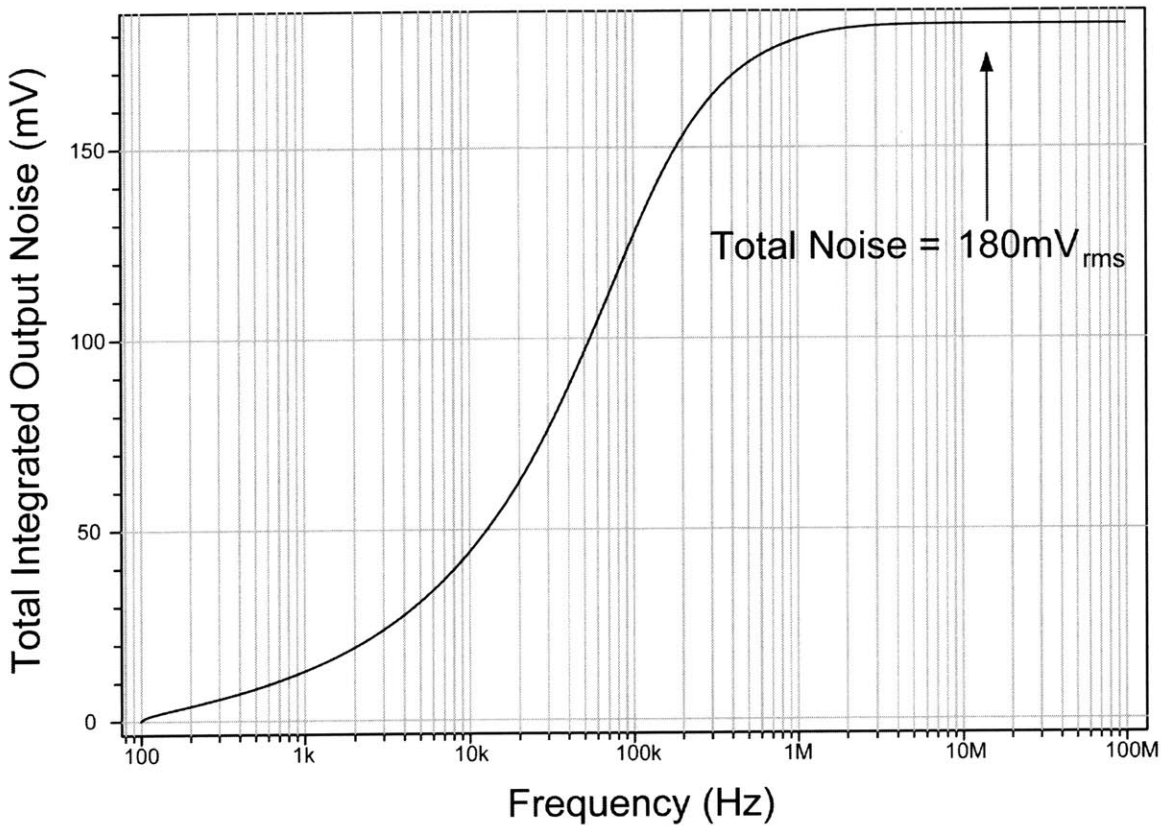


Figure 3.32 – The integrated total output white noise of the comparator in its high gain region, with all amplifiers biased at 1μA. The resulting input-referred noise is $33\mu V_{rms}$.

3.3.2 Global Stepped-Ramp Generator Design

The stepped-ramp generator forms the other half of the analog memory system. The purpose of this generator was outlined earlier in Figure 2.1. A simple scheme for producing the ramp might look something like the circuit shown in Figure 3.33. Operation of this circuit begins with the ramping capacitor, C_{ramp}, being reset by the NMOS when V_{reset} strobes high. Next, S_{inc} periodically connects the current source I_{ramp} to the capacitor. During the time when S_{inc} is closed, the voltage on C_{ramp} ramps up with a slope of I_{ramp}/C_{ramp}. During the time when S_{inc} is open, the capacitor voltage is held constant. Finally, the unity negative feedback OTA buffers the voltage out to the local

memory cells. A disadvantage of this circuit is that the P-input OTA's common-mode input range does not include a significant portion of the available rail voltage, as was discussed in Section 3.2.3.1. Thus, the range over which the capacitor voltage can be accurately reproduced at V_{out} is equally limited. Typically, when a circuit possesses common-mode limitations, feedback can be used to correct the problem by pegging the common-mode limited node at a fixed voltage. This circuit can be remedied with this technique. The resulting stepped-ramp generator that was designed for this project is shown in Figure 3.34. The ramping cycle begins when V_{reset} switches from low-to-high, and thus T_1 and T_2 are turned on, while T_3 is turned off. In this state, the left plate of

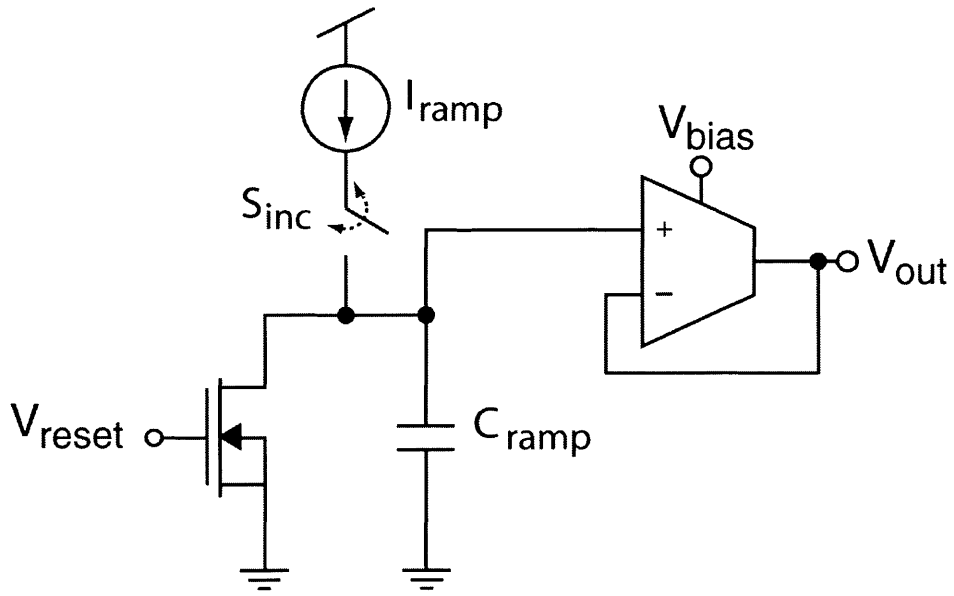


Figure 3.33 – Example of the components required to create the stepped-ramp generator.

C_{ramp} is set to 0V, while the right plate of this capacitor is set to a voltage which biases the N-input amplifier in its high-gain region. Next, V_{reset} transitions from high-to-low, turning T_1 and T_2 off, and turning T_3 on. Thus, the capacitor is wrapped in feedback

around the N-input amplifier. This is the configuration of the components during stepped-ramp signal generation. A simplified schematic of this configuration is shown in Figure 3.35.

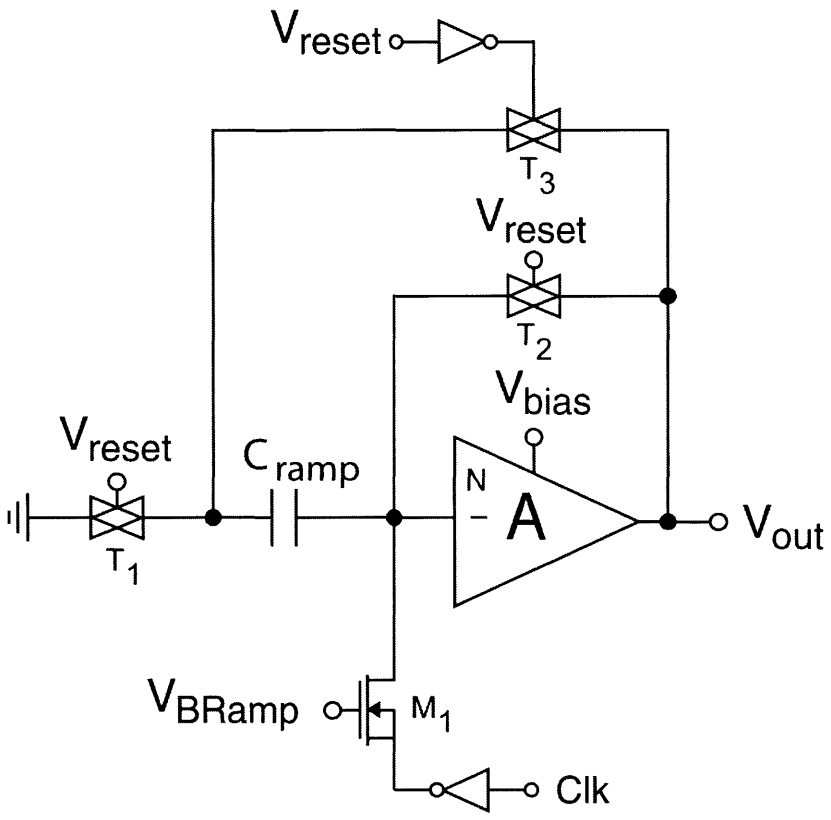


Figure 3.34 – Diagram showing the components that make up the stepped-ramp generator which was implemented on-chip.

Transistor M_1 is a switched-current source that provides a fixed current when Clk is high, and provides no current when Clk is low. The pulsing clock signal provides the control necessary to create a stepped-ramp signal. The current sourced by M_1 is extracted from the virtual ground node set up at the input to the amplifier. Additionally, M_1 injects a small amount of channel charge into the virtual ground node every cycle, but since the amount of charge injected should remain constant, it is considered to be part of the net current extraction per cycle. The feedback present due to C_{ramp} forces the input voltage to

the amplifier to remain basically constant, which is why this node is a virtual ground. Thus, if current is extracted from the virtual ground node, it must be compensated for by an increase in the voltage at the output of the amplifier. This voltage movement is

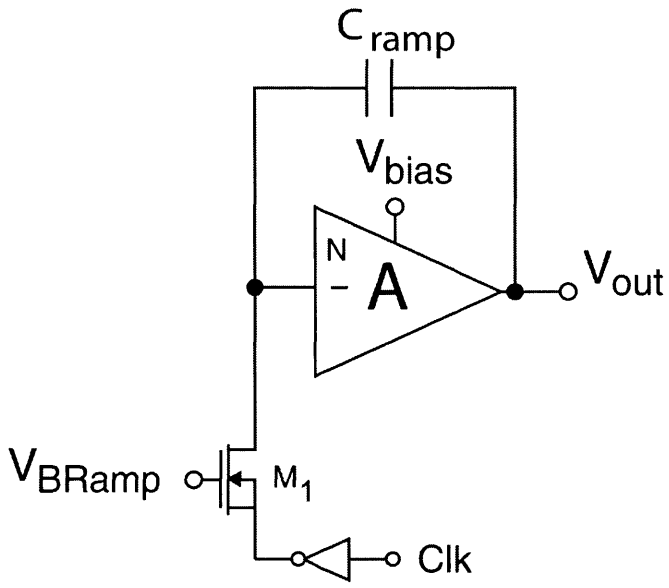


Figure 3.35 – Simplified schematic of stepped-ramp generator during stepping-ramp portion of operation.

capacitively coupled back to the amplifier's input to oppose the initial change in voltage.

Before discussing the component values and performance of the stepped-ramp generator, a final component needs to be added to this circuit. The implementation of the analog memory circuit in this project was not exactly the same as the proposed algorithm from Chapter 2. Unlike the algorithm presented in Chapter 2, the ½ bit voltage removal on the test chip was implemented by stopping the ramping waveform, pulling down ½ bit, and then sampling the ramp value with the comparator circuit. This was an oversight in the original design, as it was not yet discovered that the stepped-ramp generator could be made into a global component. Switching this function to the local cell is trivial, and will be implemented in future designs. The ½ bit subtraction was accomplished by

connecting the top plate of a 50fF capacitor to the virtual ground node of the amplifier, and the bottom plate to two transmission gates. One transmission gate can connect the bottom plate to ground, the other can connect it to a voltage slightly above ground. Initially, it is connected to ground. When the proper value on the stepped-ramp waveform is detected, the capacitor plate is switched to the slightly more positive voltage. This injects a small positive charge onto the virtual ground node, and causes the output of the amplifier to decrease slightly. This decrease can be tuned to ½ of a bit by tuning the value of the positive voltage.

Device	Parameters
T_1-T_3	All transistors: W=2.4μm, L=1.6μm
Inverter	All transistors: W=2.4μm, L=1.6μm
M_1	W=16μm, L=8μm
N-Input Amplifier	All transistors: W=16μm, L=8μm
C_{ramp}	400fF
$C_{1/2bit}$	50fF

Table 3.4 -- List of component sizes and values for the fabricated stepped-ramp generator.

Table 3.4 lists the component sizes and values that were used in the fabricated stepped-ramp generator of Figure 3.34. Figure 3.36 shows the lower portion of the stepped-ramp waveform, highlighting the transition from triode to saturated operation of the ramp amplifier's NMOS transistor. Figure 3.37 shows the upper portion of the stepped- ramp waveform, illustrating the transition from saturated to triode operation of the amplifier's PMOS transistor. Together, these two simulations predict a maximum operating voltage range of more than 2.8V on a 3.3V scale. This important parameter will be measured on-chip in the next chapter.

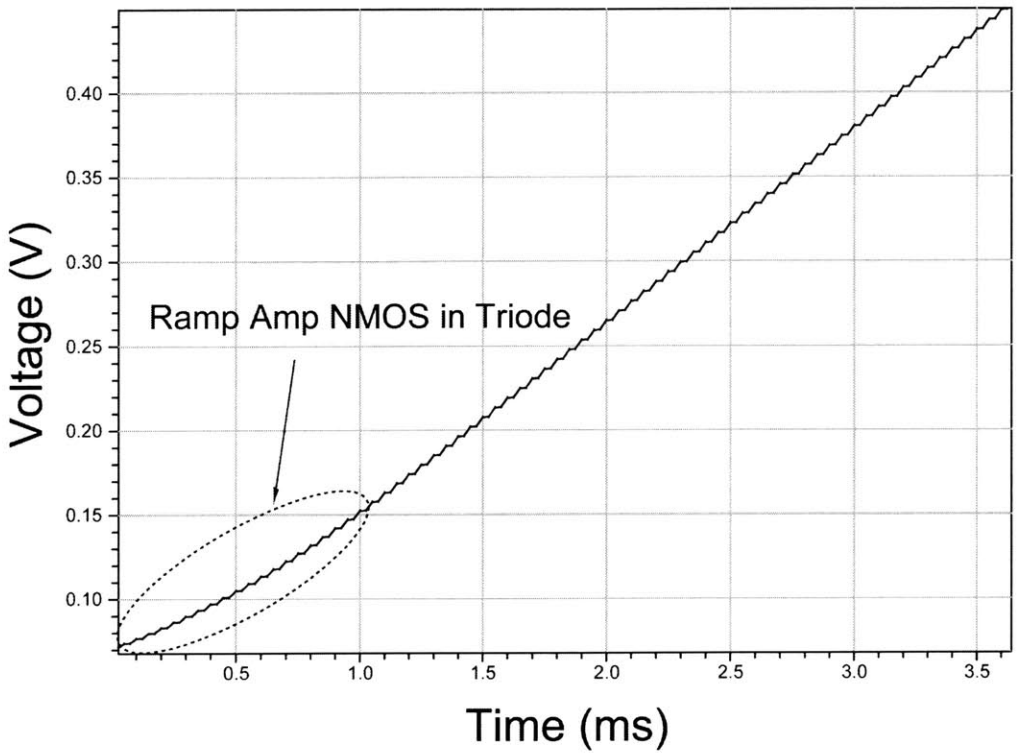


Figure 3.36 – Simulation of lower portion of the stepped-ramp waveform depicting the transition of the NMOS transistor from triode to saturation region.

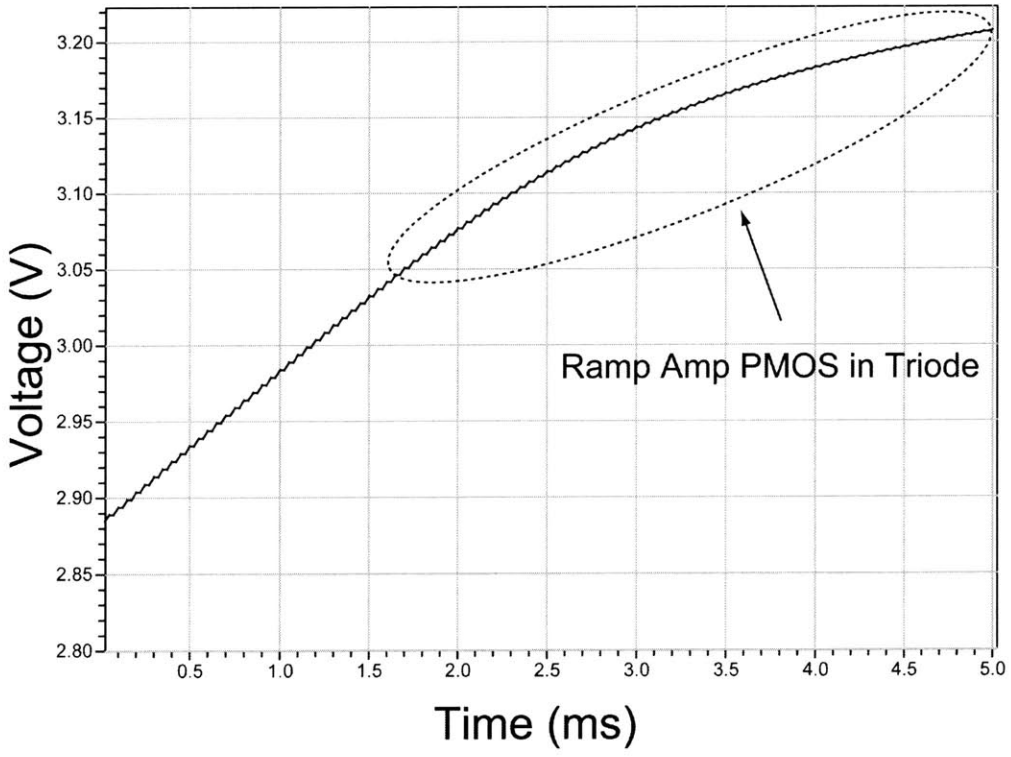


Figure 3.37 – Depiction of the top section of the same stepped-ramp waveform, showing the PMOS transistor entering the triode region of operation.

Another important factor that will impact the stepped-ramp generator's operation is the stability of the current source transistor that provides the ramping current. This current must be very stable in order to ensure that the step size remains uniform throughout the ramping cycle. One source of error in this current is due to the fact that over the course of the ramping cycle, the virtual ground node of the amplifier will move approximately

$$\Delta V_{vg} = \frac{2.8V}{100^{V}/_{V}} = 28mV \tag{2.44}$$

because of the finite gain of the amplifier. This voltage variation will directly affect the current source's supplied current by modulating the V_{DS} voltage of the current source transistor. The transistor M_1 in Figure 3.34, functions as the current source transistor in the ramp generating circuit. We can express the change in current as a function of the supplied current in the following equation

$$\Delta I_D = \frac{\Delta V_{DS}}{r_{out}} = \frac{\Delta V_{DS}}{\frac{1}{\lambda I_D}} \rightarrow \frac{\Delta I_D}{I_D} = \lambda \cdot \Delta V_{DS} \tag{2.45}$$

Substituting in $\lambda = 0.031V^{-1}$ from Table 3.1, and scaling this value appropriately for L = 8μm (the length of M_1), we find that independent of the value of the ramping current, we can expect the ratio of ΔI_D to I_D to be

$$\frac{\Delta I_D}{I_D} = \left(0.031V^{-1}\right)\left(\frac{1.6\mu m}{8\mu m}\right)\left(28\times10^{-3}V\right) = 1.736\times10^{-4}. \tag{2.46}$$

This is a change of only 0.01736% over the full range of operation, which is approximately 12.5 bits of linearity.

It is clear that the above error is much less than our required noise floor, and can be safely ignored. An error that may not be ignored, however, is the random white and 1/f noise. There are two primary sources of noise that must be considered. The first is the amplifier noise itself, and the second is the noise from transistor M_1. Referring back to Figure 3.35 as a guide, we can arrive at the block diagram of this circuit's internal noise behavior, not including M_1, shown in Figure 3.38. V_{in}^2 is the input-referred noise of the amplifier itself, the forward block is the amplifier model, and the feedback block models the capacitive divider between C_{ramp} and any parasitic capacitance, C_{par}, at the input node. The overall transfer function for this circuit can be rewritten as shown in Figure 3.39. This block diagram reveals the fact that any parasitic capacitance at the input node causes the total output-referred noise to increase. Hand calculation of the total

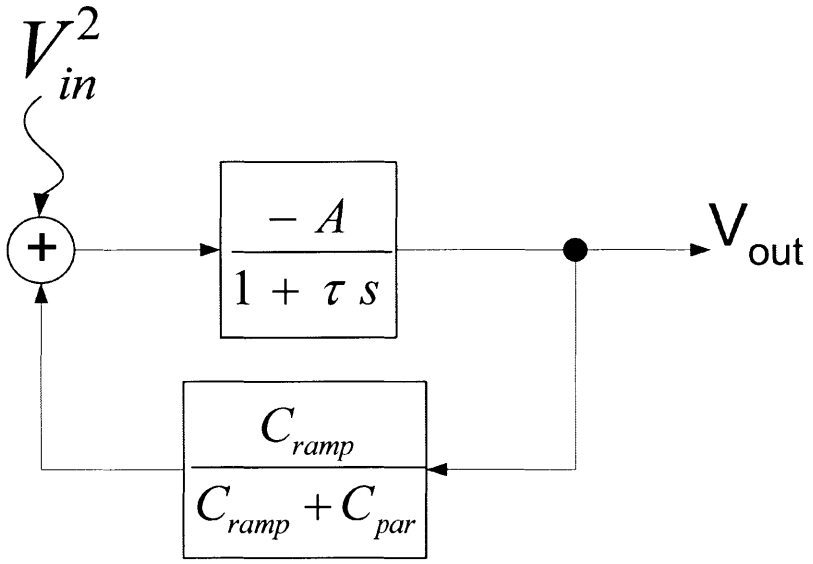


Figure 3.38 – Block diagram of stepped-ramp generator internal noise behavior.

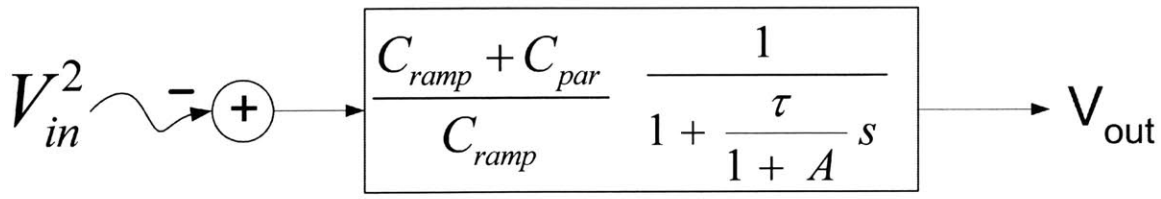


Figure 3.39 – Reduced block diagram of stepped-ramp generator internal noise behavior.

output noise due to this noise source shows

$$V_{onoise,tot} = \sqrt{V_{in}^2 \cdot f_o \cdot \frac{\pi}{2} \cdot A_{CL}^2} = \sqrt{\left(35^{nV}/\sqrt{Hz}\right)^2 (13.0MHz)\left(\frac{\pi}{2}\right)(1.2)^2} = 189\mu V_{rms}, \quad (2.47)$$

where the closed-loop bandwidth f_o and the closed loop gain A_{CL} were derived from simulation results. Noise simulation in SPICE gives the total noise plot shown in Figure 3.40, which agrees well with hand calculations.

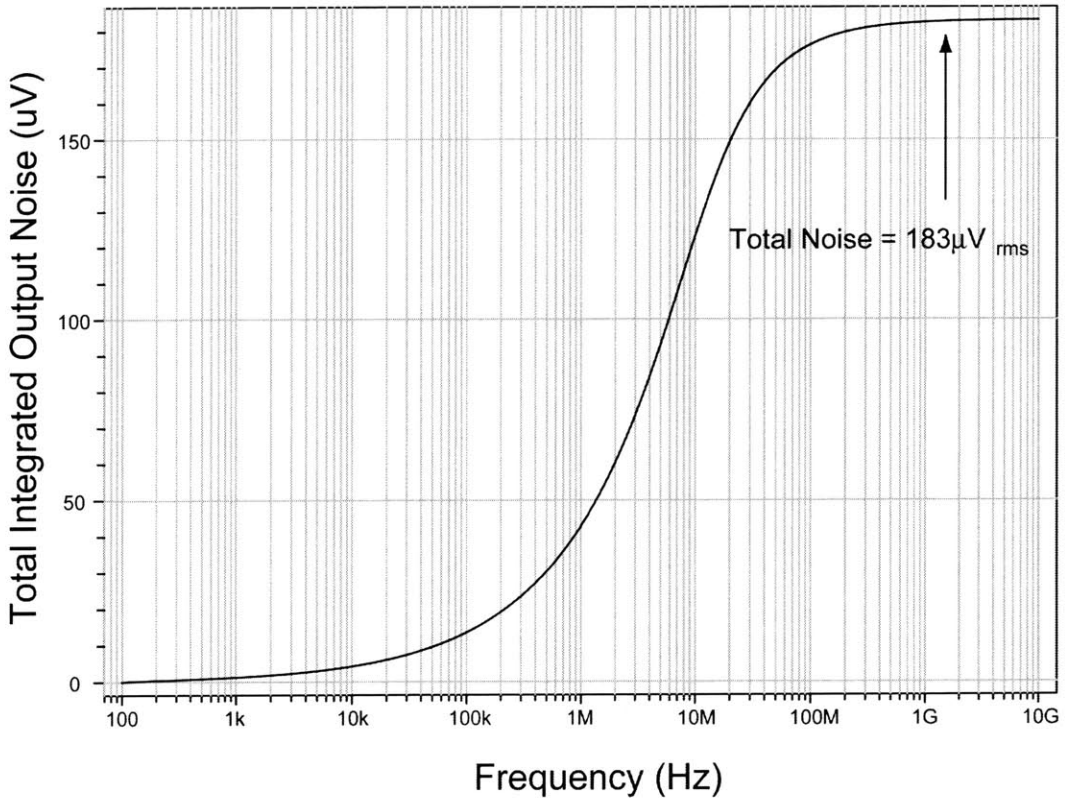


Figure 3.40 – The integrated total output white noise of the stepped-ramp generator with the generator in the configuration shown in Figure 3.35. The amplifier is biased at 1µA.

This internal amplifier noise is only one source of noise in this circuit. Another source is the transistor M_1 in Figure 3.35. A block diagram of the signal path for this noise source is shown in Figure 3.41. In this figure, the delta train, s(t), is convolved with the square pulse shown in the box below it. Together, this creates a gating waveform for the input current noise, $i_n(t)$, which turns the current noise on and off. Next, this pulsed

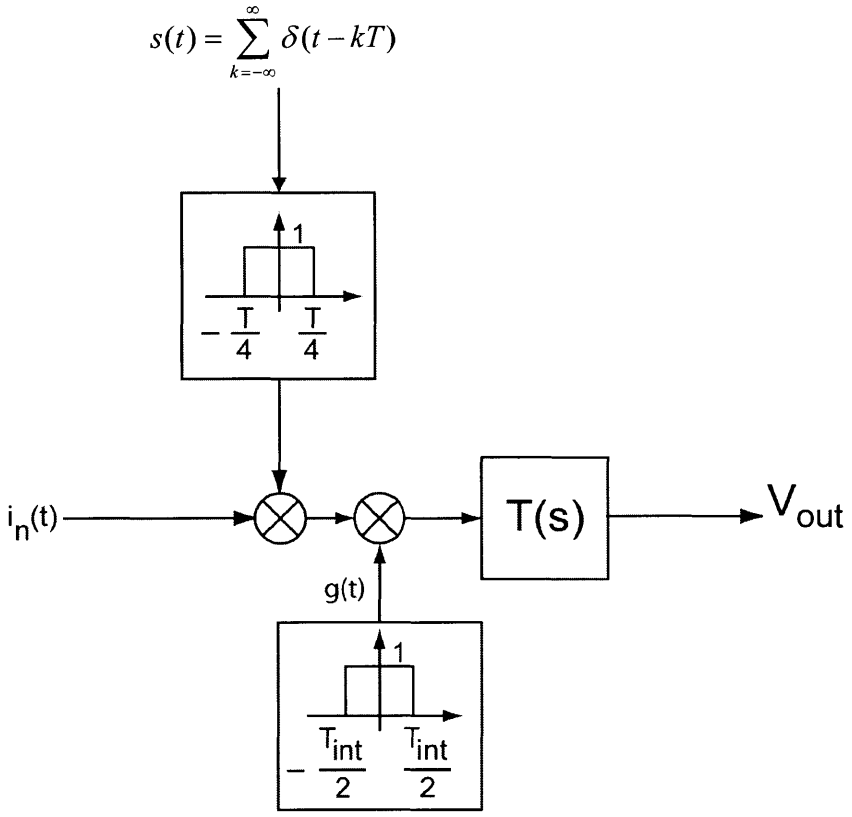


Figure 3.41 – Block diagram of current signal path for the stepped-ramp generator while in the configuration shown in Figure 3.35.

current waveform is gated on for a finite time, T_{int}, by the signal g(t). Finally, the finite pulsed current waveform is sent through the integrating amplifier's transfer function T(s). It can be shown that the Fourier series representation of s(t) is

$$s(t) = \sum_{k=-\infty}^{\infty} \frac{1}{T} e^{jk\omega_o t} \tag{2.48}$$

and that the Fourier transform of the resulting Fourier series representation of the delta train can be written as

$$S(f) = \sum_{k=-\infty}^{\infty} \frac{2\pi}{T} \cdot \delta(\omega - k\omega_o). \tag{2.49}$$

It is also known that the frequency response of the pulse blocks are described by

$$P(f) = \frac{2\sin\left(\omega \frac{T}{4}\right)}{\omega} \quad \& \quad G(f) = \frac{2\sin\left(\omega \frac{T_{\text{int}}}{2}\right)}{\omega} \tag{2.50}$$

The frequency responses of S(f) and P(f) are shown separately in Figure 3.42.

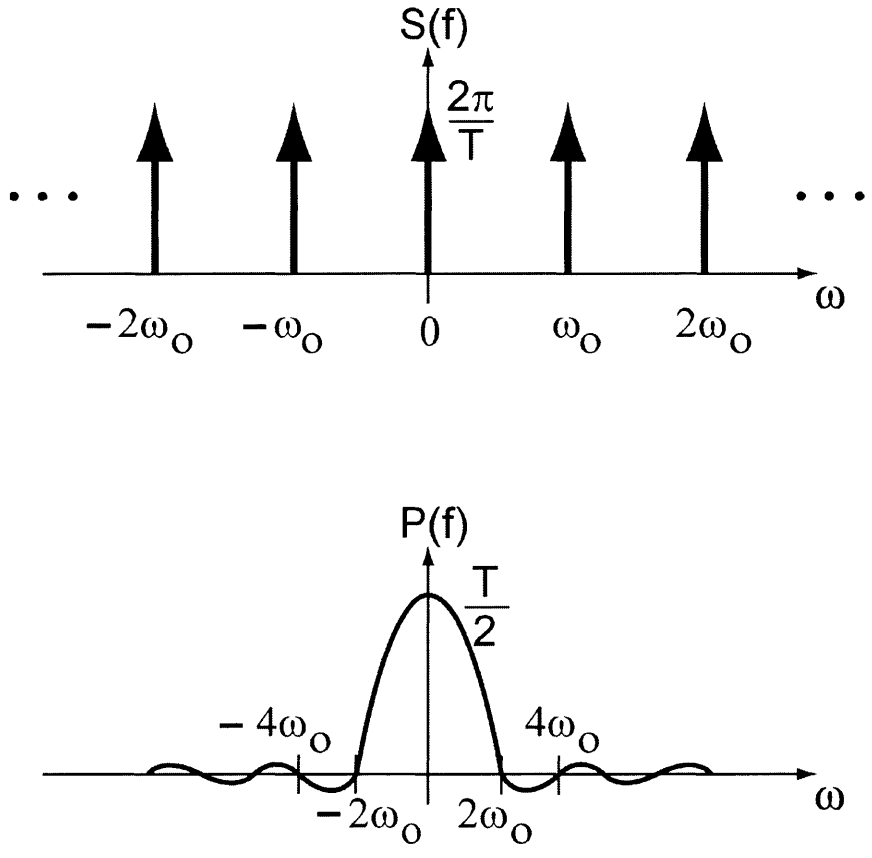


Figure 3.42 – Frequency response of the delta train, S (f), and of the pulse signal, P(f).

The overall frequency response of the sampling leg of the system, N(f), is the product of the frequency responses of S(f) and P(f), which gives the signal shown in Figure 3.43. This signal then convolves in the frequency domain with the spectrum of the noise source, $I_n(f)$, and the integration window, G(f), and then passes through T(s).

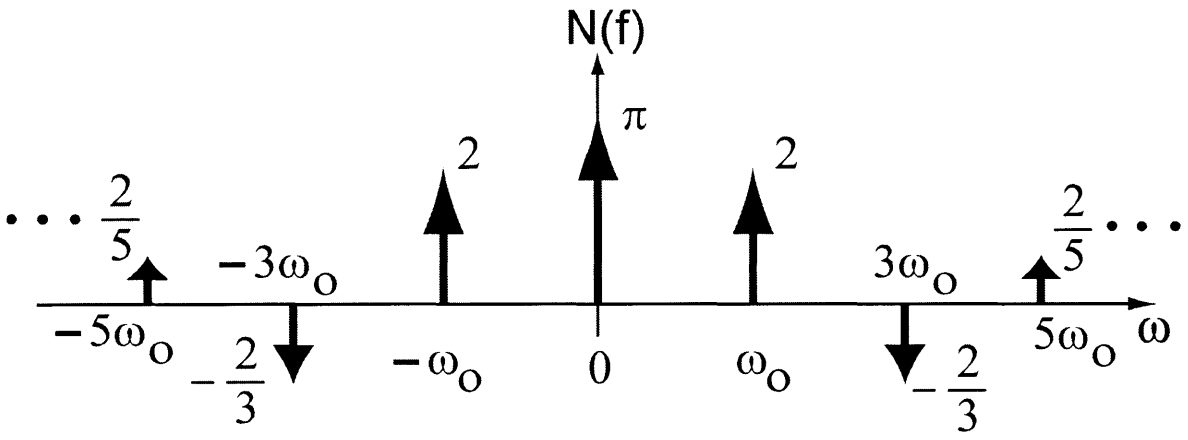


Figure 3.43 – Net frequency response of S(f) and P(f) multiplied in the frequency domain.

A block diagram of the amplifier transfer function, T(s), is shown in Figure 3.44. In this figure, $r_{o,M1}$ is the output impedance of transistor M_1, C_{vg} is the parasitic capacitance at the amplifier virtual ground node, $g_{m,Amp}$ is the transconductance of the ramp amplifier, $r_{o,Amp}$ is the ramp amplifier output impedance, C_{Amp} is the ramp amplifier output capacitance, and C_{ramp} is the feedback capacitor. This block diagram can be derived by considering the circuit in two steps. First, eliminate C_{ramp}, and derive the input-output block diagram. Next, add C_{ramp} back into the circuit and add its effects to the existing block diagram in terms of its two terminal voltages. The block diagram can be redrawn as shown in Figure 3.45. The parameters for the implemented circuit are shown in Table 3.5. The MATLAB simulated transfer function shown in Appendix C and is plotted by the solid line in Figure 3.46. This response is nearly identical to the

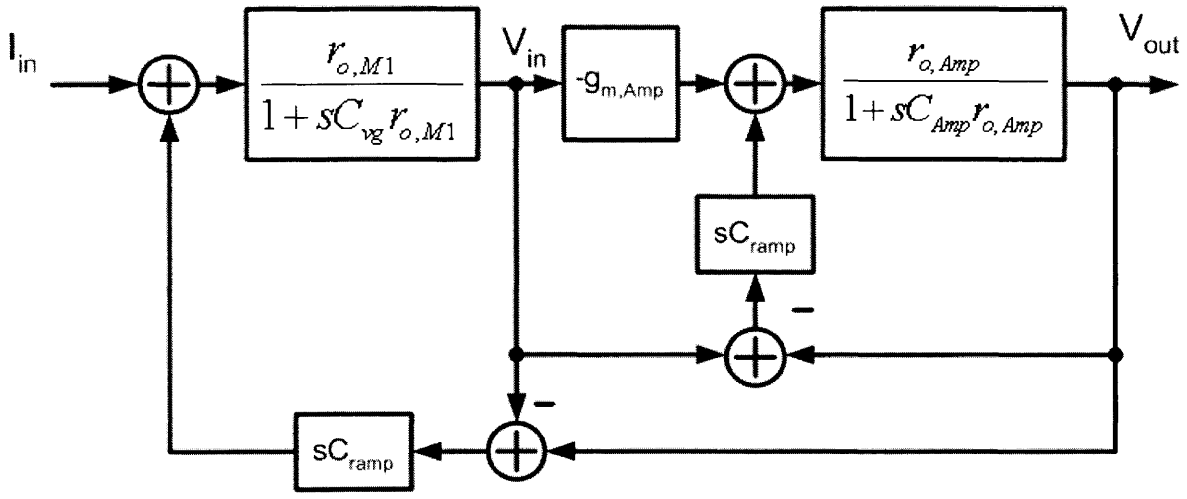


Figure 3.44 – Block diagram of capacitive-feedback ramp amplifier circuit.

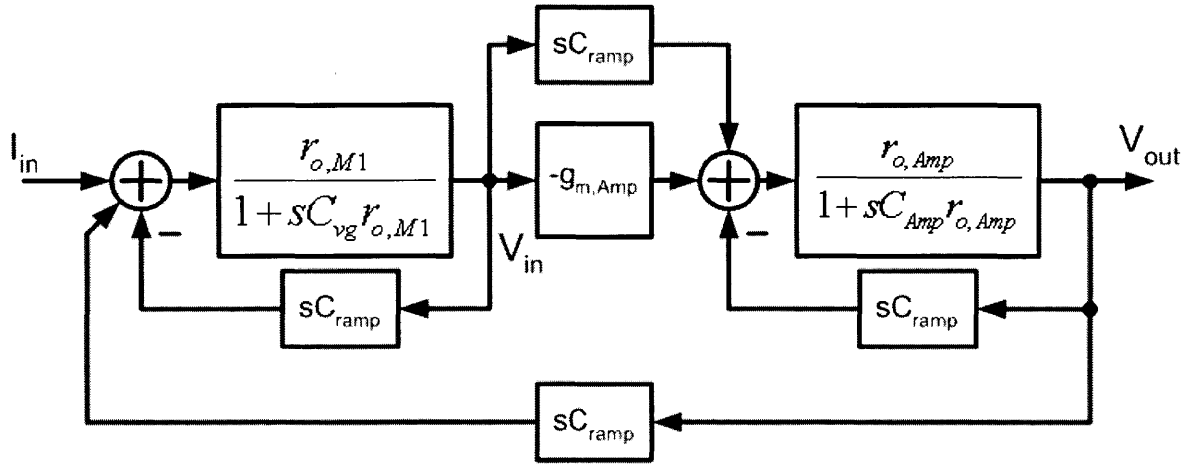


Figure 3.45 – The reduced version of the block diagram shown in Figure 3.44.

Component	Value
$r_{o,Amp}$	$6.2\times10^{6}\ \Omega$
$r_{o,M1}$	$2.9\times10^{8}\ \Omega$
$g_{m,Amp}$	8.71×10^{-6} A/V
C_{ramp}	400fF

Table 3.5 – Values of the various components for the capacitive-feedback ramp amplifier.

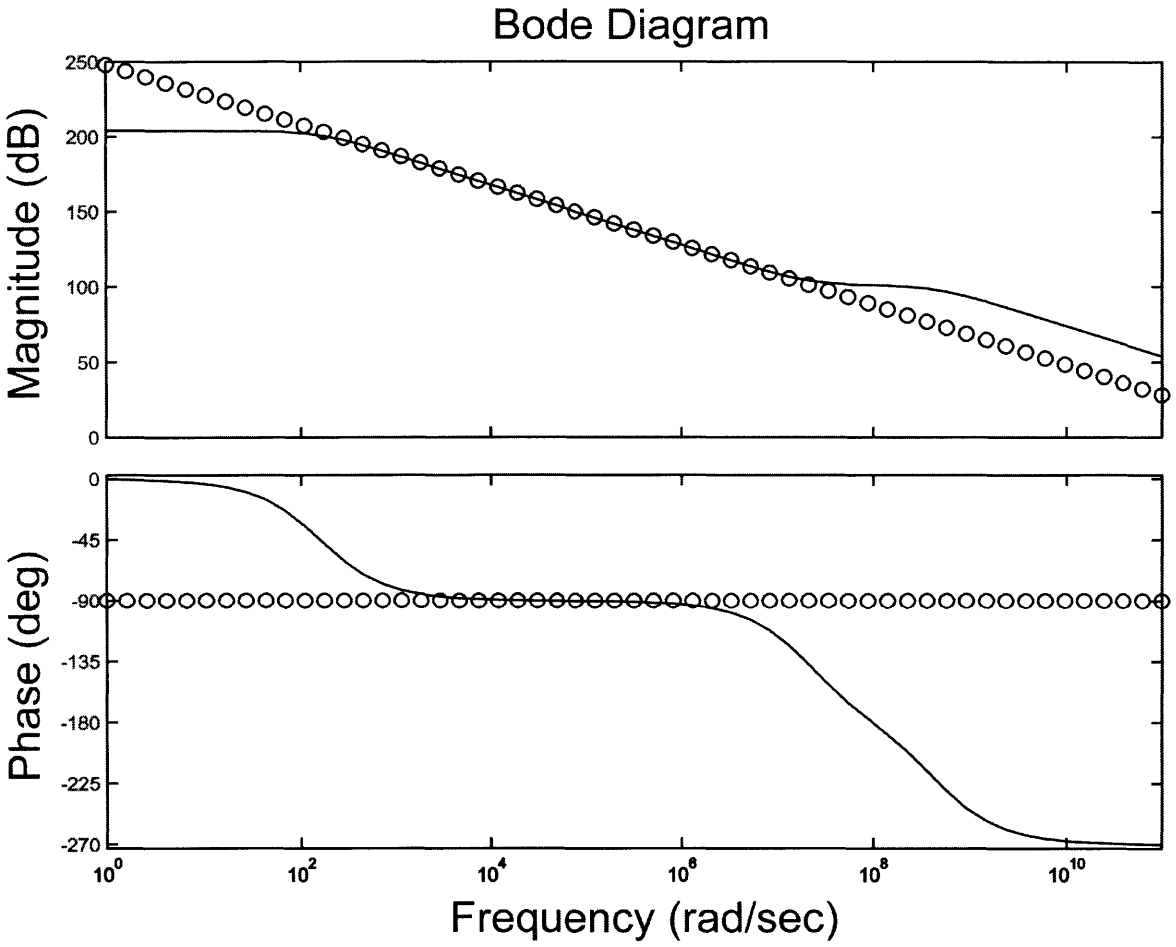


Figure 3.46 – MATLAB simulated frequency response of the capacitive-feedback ramp amplifier circuit (solid). The ideal transfer function the circuit should approximate (circles).

ideal response of $1/sC_{ramp}$ (shown by the circles in the same figure) for frequencies up to 10^7 rad/s. Thus, a good model for this topology is

$$T(s) = \frac{1}{sC_{ramp}}. \tag{2.51}$$

The noise source associated with M_1 has both a white and a 1/f components, and is described by

$$I_n^2(f) = 2qI_{bias} + \frac{K_f}{WLC_{ox}f} g_m^2 \tag{2.52}$$

because of the fact that the transistor is operating in the sub-threshold region. For this NMOS, $g_m = 2.24\times10^{-10}$ A/V, $K_f = 2\times10^{-22}$, W=16μm, and L=8μm.

Using MATLAB, the convolution of N(f), $I_n(f)$, and W(f) can be derived, and the result can then be passed through the amplifier transfer function. The result of the MATLAB script shown in Appendix C for the total noise due to transistor M_1 at the output of the amplifier is

$$V_{no} = 57\mu V_{rms}. \tag{2.53}$$

This represents the noise due to M_1 under typical operating conditions. The total noise of this ramp block will then be the combination of the inherent noise of the amplifier and the noise due to transistor M_1

$$\left(183\mu V_{rms}\right)^2 + \left(57\mu V_{rms}\right)^2 = 3.674\times10^{-8} V^2 \quad \rightarrow \quad \underline{192\mu V_{rms}}. \tag{2.54}$$

This corresponds to 0.55mV peak-to-peak noise due to these noise sources.

3.3.3 The Complete Circuit

The final step in constructing the memory cell is to simply close the loop around the components we have discussed, and add digital control circuitry. The digital circuitry is presented in Appendix B, and will not be discussed further here. The functionality of the control circuitry has been described in detail already, and sifting through the actual construction of these waveforms is uninteresting. As was mentioned in the previous section, the version of the memory cell that was implemented in this project differs slightly from the one introduced in Chapter 2. Specifically, the ½ bit subtraction circuit was included in the ramping cell rather than the local cells. This can be easily fixed in

future versions, using the same switched-capacitor technique that was used in this version, just on a local level.

4 Test Results

The implemented analog memory element was fabricated in a MOSIS 1.5μm CMOS process. The cell size is 300μm × 250μm (375λ × 310λ), and the layout for the complete chip is shown in Figure 4.1. The division of the digital control circuitry between the local

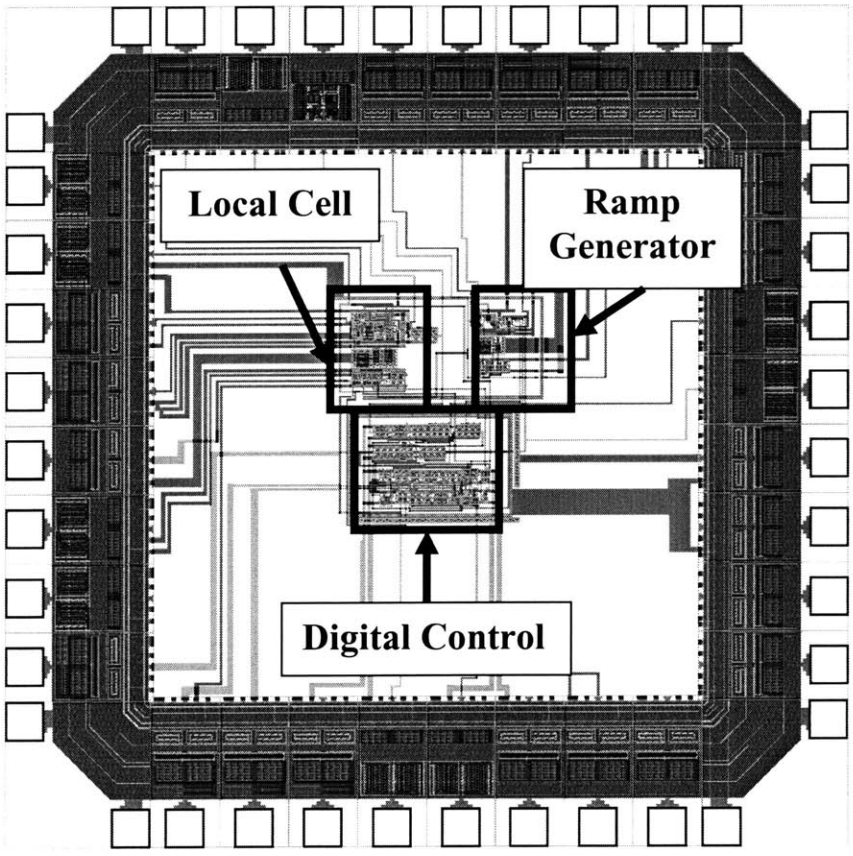


Figure 4.1 – Layout of complete analog memory cell in a MOSIS 1.5μm CMOS process.

cell and ramp generator is roughly ½ of the area for each. As was shown in the previous chapter, many of the interesting nodes in this circuit design are also very sensitive to stray capacitance, and will not operate correctly if connected to analog buffers for outside viewing. In particular, none of the nodes internal to the auto-zeroing sample-and-hold comparator are available for outside viewing. As a result, the data that can be obtained from the chip is not as rich as it could be, but still is informative.

4.1 Stepped-Ramp Generator Output Range

The first two tests that were performed were to verify that the stepped-ramp generator achieves the maximum possible linear output range. The simulated results

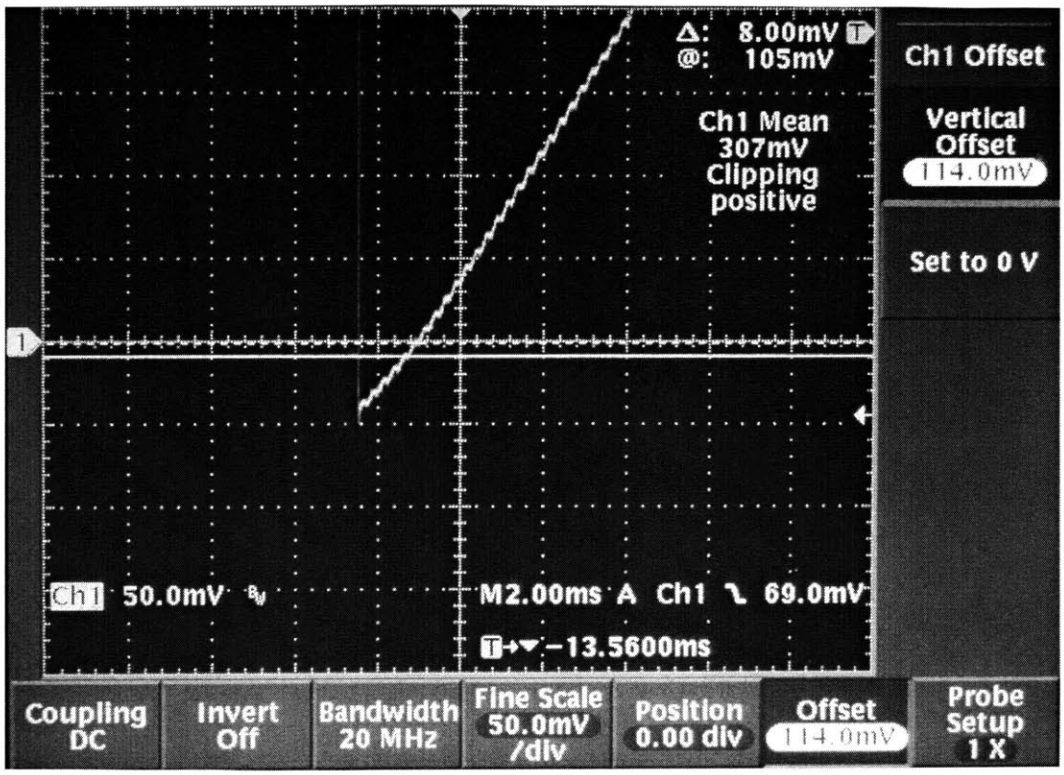


Figure 4.2 – Lower portion of the stepped-ramp generator waveform.

addressing this issue are shown in Figure 3.36 and Figure 3.37. Figure 4.2 shows test results from the chip under the same bias condition of $I_B = 1\mu A$. The solid horizontal line in the figure is located at 105mV. It can be argued exactly where the ramping waveform becomes linear, but it is clearly within 50mV of this cursor in the upwards direction. For safety, let's call the lower bound 150mV -- this is the voltage at which the NMOS enters triode operation on chip. Referring back to simulation, we see an identical waveform. The NMOS enters triode at 150mV, and we can begin to see a slight curve below this point.

The upper portion of the stepped-ramp generator output is shown in Figure 4.3.

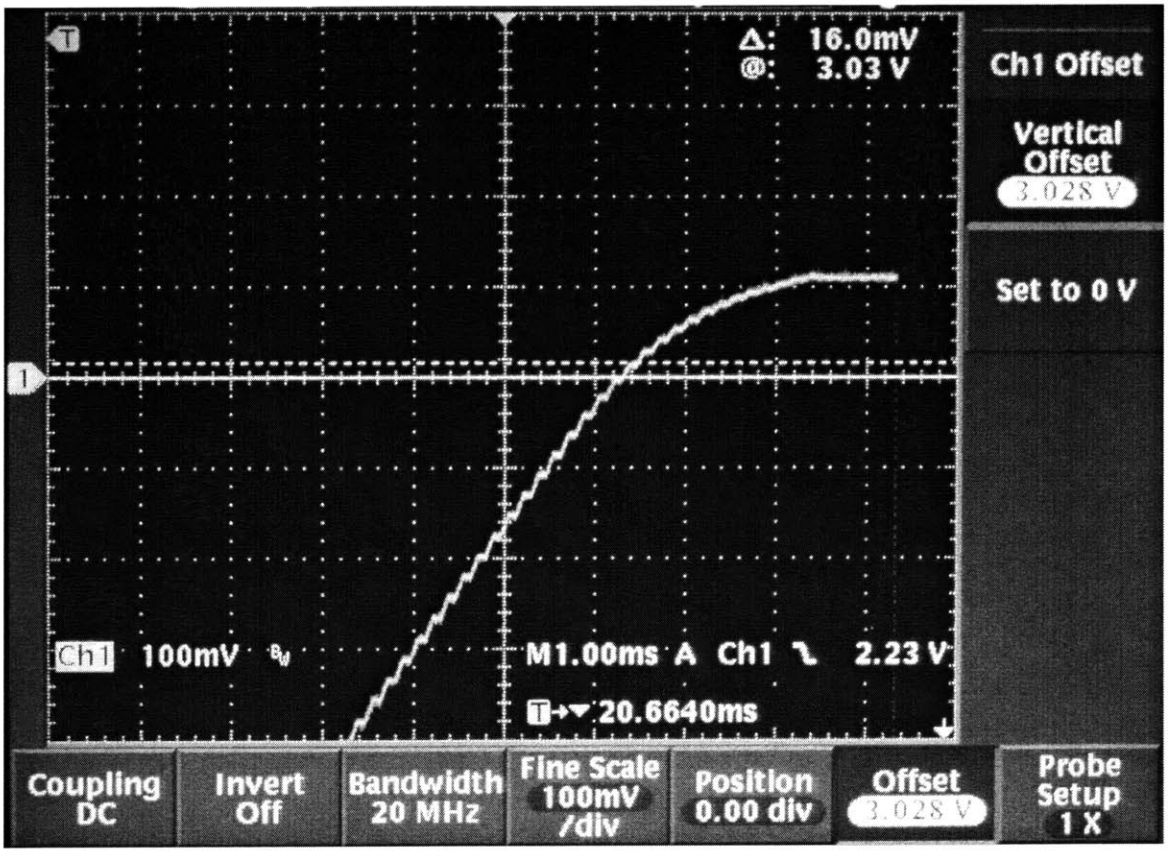


Figure 4.3 – Upper portion of the stepped-ramp generator waveform.

The solid horizontal line in this plot is coincident with the X-axis, and is located at 3.03V on the ramping waveform (offset has been added to the waveform inside the oscilloscope). After this point on the ramp, we can see definite curvature forming. Thus, it is at this point that the PMOS is entering saturation operation. Again, referring back to the simulation data, we see that it predicts nearly exactly the same waveform. Saturation in simulation occurs at somewhere near 3.05V. Thus, again the test data and actual data are in excellent agreement.

4.2 Stepped-Ramp Generator Non-Idealities

The data in the previous section matches simulation results perfectly. However, the stepped-ramp generator in general does not behave as ideally as the simulations predict. In particular, the waveform exhibits both a cycle to cycle jitter and a very low frequency drift in time. The jitter can be seen if the persistence of the oscilloscope is set to ∞. Figure 4.4 was obtained after a 5 minute hold. In this case the jitter is on the order of 7mV. More information is provided by examining different locations along the ramp. Figure 4.5 shows the jitter under the same conditions at a middle position on the ramp, while Figure 4.6 shows the jitter under the same conditions at a low position on the ramp. The jitters in these two cases were 5.2mV and 3.6mV, respectively. The variance seems to grow with the height of the ramp. There appears to be two types of jitter present in this circuit: one which adds a constant jitter everywhere along the amplifier output, and another whose effect increases the further up the ramp we travel. The fact that both types exist gives important information about the sources of the jitter.

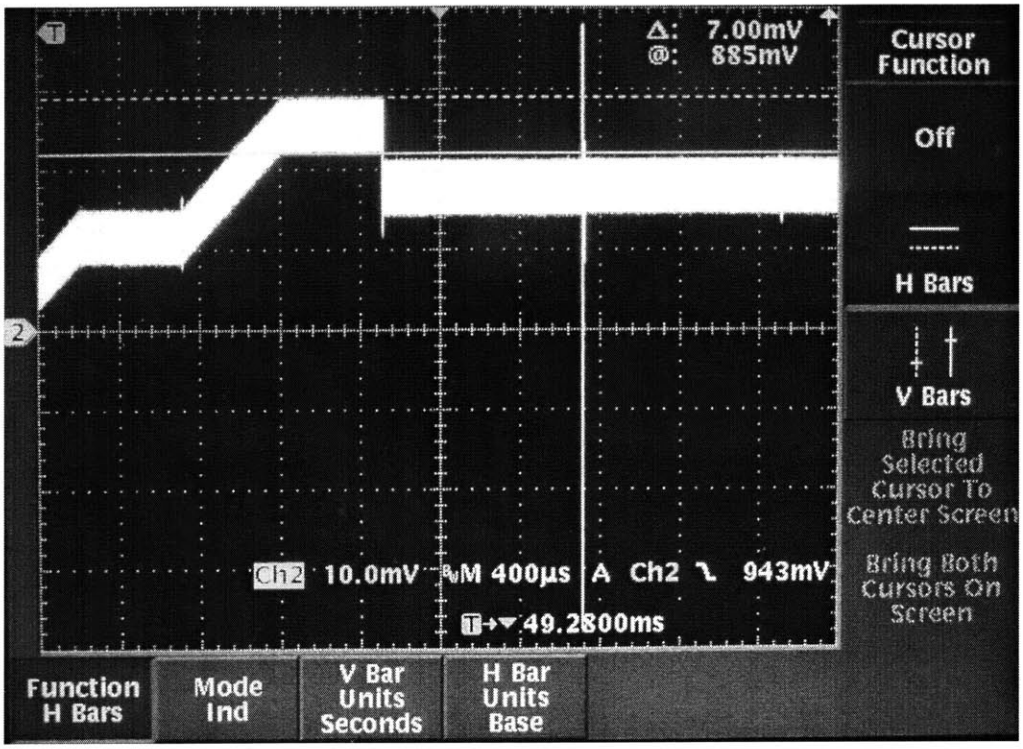


Figure 4.4 – Upper jitter with the persistence of the oscilloscope set to ∞ for 5 min.

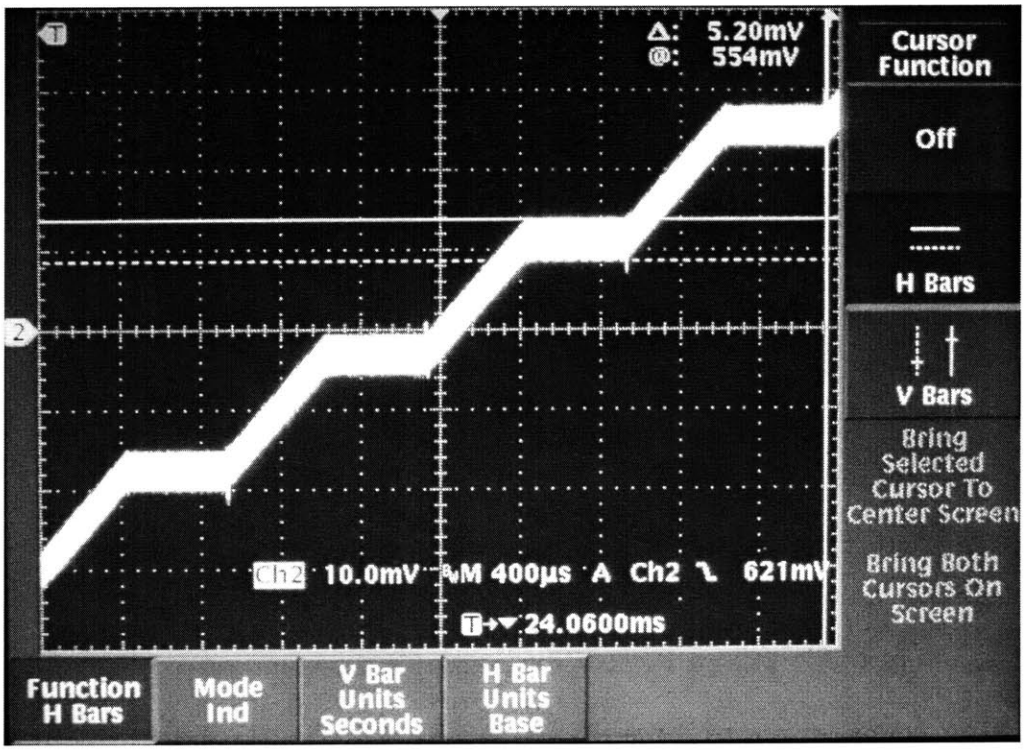


Figure 4.5 Jitter at a lower position on the ramp than in Figure 4.4, and higher than in Figure 4.6.

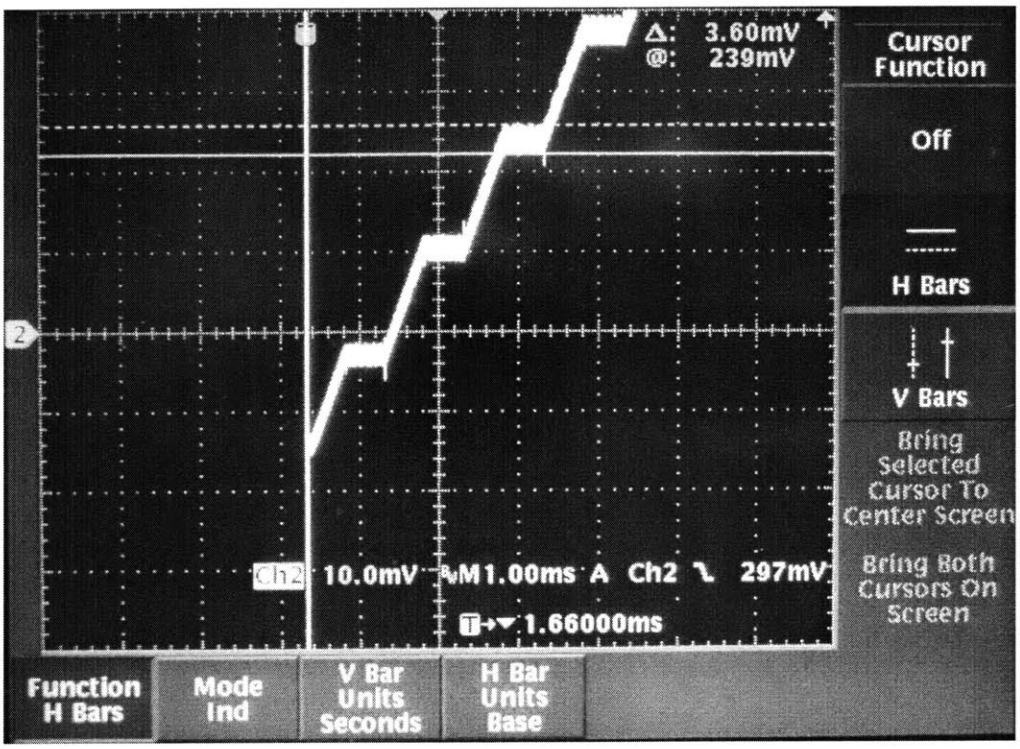


Figure 4.6 – Jitter on the ramp at a very low position. The jitter is the smallest here.

Constant jitter regardless of location along the waveform implies either inherent noise within the amplifier itself, or power-supply noise. It was shown in Eqn. 3.53 that the inherent noise of the amplifier circuitry is $192\mu V_{rms}$. This noise corresponds to 0.5mV peak-to-peak, which is not nearly enough to account for the level of noise seen. It can be shown theoretically that the power-supply-rejection-ratio (PSRR) of the implemented amplifier is very poor. Additionally, this is shown experimentally in Figure 4.7 for an intentional sinusoidal input voltage on the analog ground supplying the circuit. The noise is actually amplified from the rail to the output. This is due again to the parasitic capacitance at the input node of the amplifier, as was discussed in Section 3.3.2. We find good agreement between the predicted value of amplification, 1.2, and the value that was

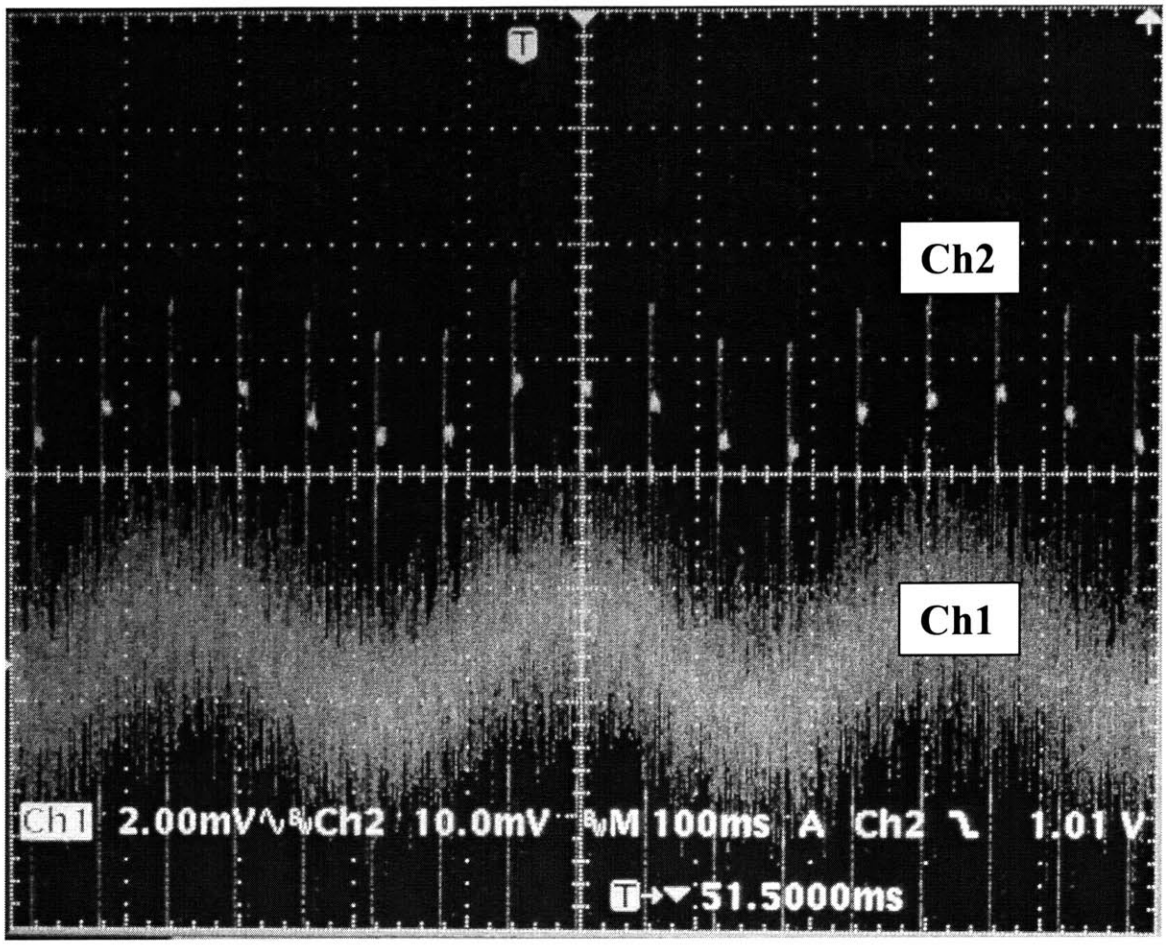


Figure 4.7 – The fuzzy sinusoidal waveform (Ch1) is the signal that was driven onto analog ground. The resulting movement in the output voltage (the brighter spots on the stick-like waveform) is a factor of 1.2 to 1.5 times larger than the input signal.

measured from the chip, 1.2 to 1.5. Most likely, a combination of poor PSRR and inherent noise of the amplifier together cause the constant portion of the jitter in the ramping waveform.

The only noise source capable of causing the ramp to diverge in time is the noise source associated with the ramp current source. This is because divergence in the ramp implies that the charging current is fluctuating in time. Thus, noise in the current source is causing this portion of the jitter. Figure 4.8 was captured to prove that the ramp

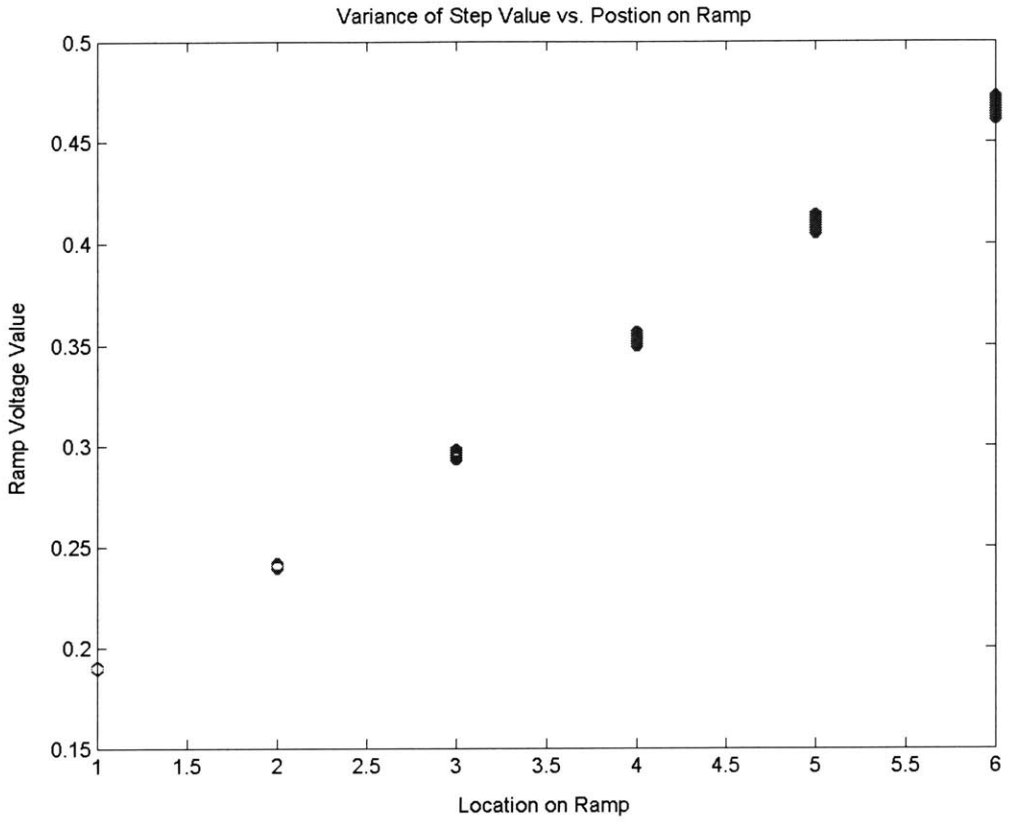


Figure 4.8 – Variance of step value versus position on the stepped-ramp waveform.

voltage does in fact diverge. It can be clearly seen that the variance in the step value is positively correlated with the step's position along the ramping waveform.

The previous analysis of Section 3.3.2 showed that the inherent noise of the transistor M_1, when referred to the output of the ramp generator, is only 0.16mV peak-to-peak. The answer to the missing noise can only come from two sources. First, cycle-to-cycle jitter can be attributed to power supply noise fluctuations. A fluctuation on the power supply of 2.5mV would be enough, after amplification to explain the noise on the ramp generator. Additionally, the divergence can be attributed to low-frequency drift in

the current supplied by M_1. This drift is known to exist (see Figure 4.11), and since all of the variance data was taken over time spans of several minutes, the low frequency drift could have affected them significantly.

4.3 Transfer Characteristic of Overall Analog Memory

The resolution to which input voltages can be captured and stored depends on where they lie in the input voltage range. This is due to the fact that the higher the ramp voltage travels, the more uncertainty there is in the steps. Thus, near the bottom of the ramp, we would expect to see higher resolution than near the top, and we do. Figure 4.9 shows a portion of the input-output transfer characteristic which exhibits 9-bits of

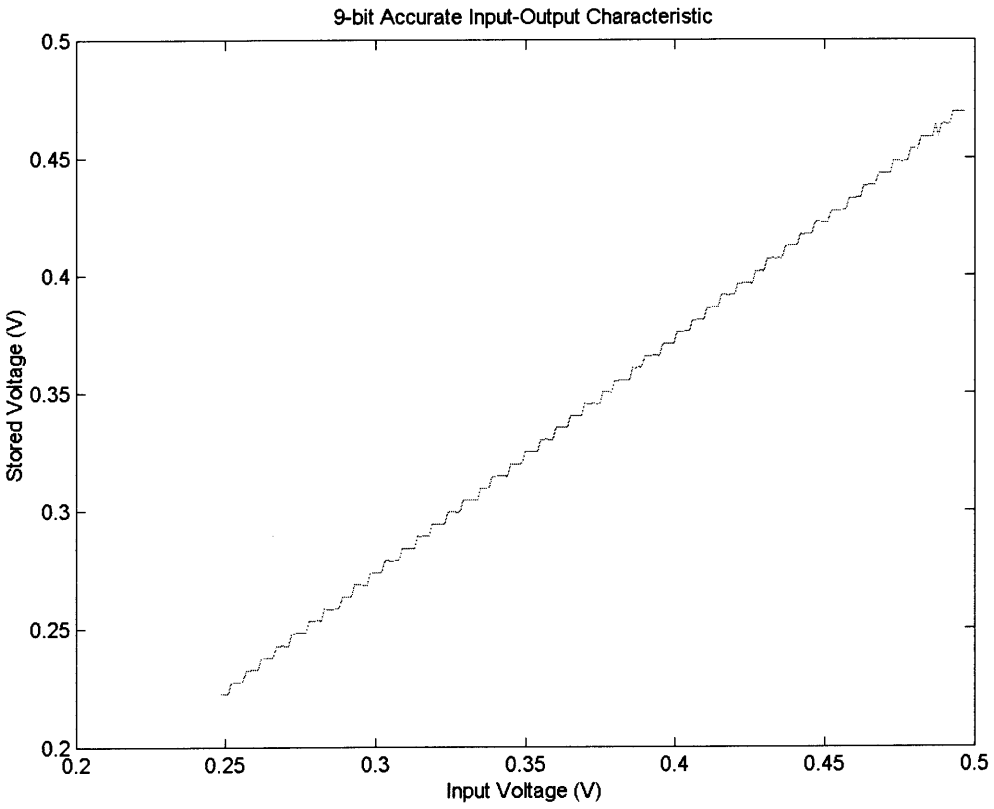


Figure 4.9 – This figure shows 9-bits of accuracy in the lower range of operation. However, the upper portion of the characteristic degrades past 0.5V.

resolution. The reader will notice that there is a significant offset error from input to output. This is the result of a mistake in the digital state machine that causes unintentional loading of the input voltage, V_{in}, during the sampling phase. Direct proof also exists that this is not comparator offset: comparator offset must be much less than the bit width at the resolution we are trying to store. Otherwise, when the signal is re-quantized during the storage phase, this error would compound and cause the stored analog voltage to rail. This does not happen, and therefore this is not offset from the storage loop or the comparator, only the input loop. This problem can be easily fixed in future implementations.

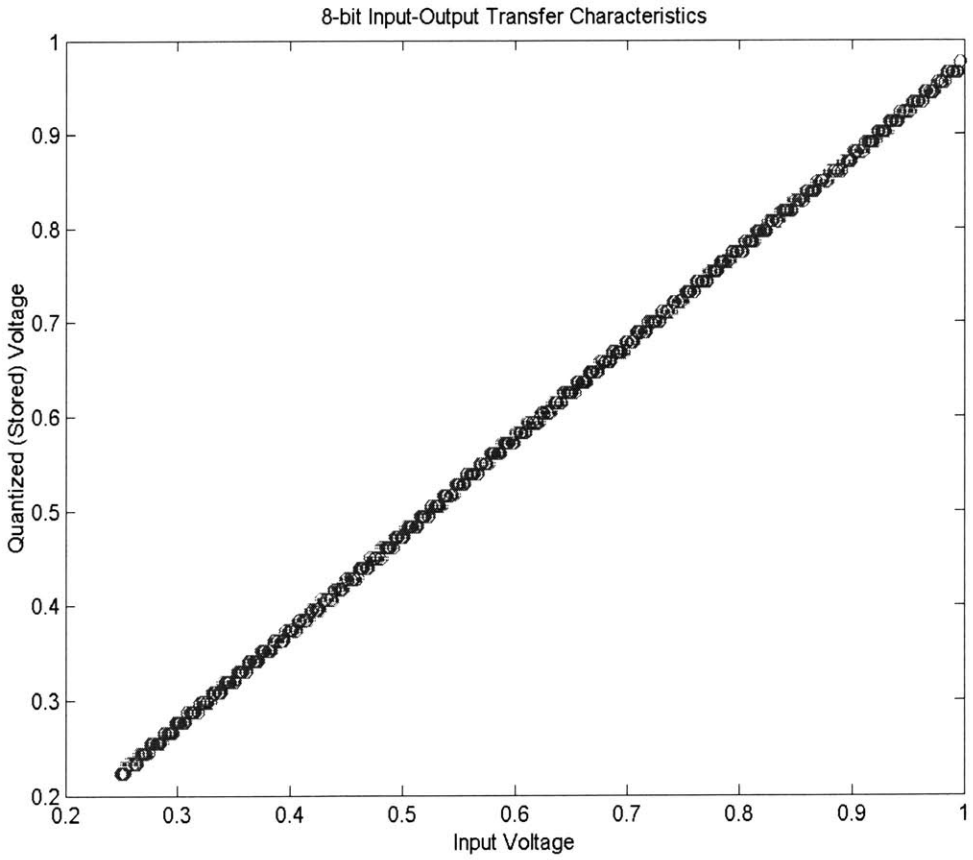


Figure 4.10 – Portion of a full 8-bit transfer characteristic.

Achieving 8 bits of resolution in the transfer characteristic and the storage characteristic simultaneously is possible. A portion of a full 8 bit transfer characteristic is shown in Figure 4.10. We can also store a value to 8 bits of resolution for almost 3 hours, as shown in Figure 4.11. The movement seen on this time scale is due to drift in the current source value with time. In order to remove this drift, a more accurate current source must be developed for the next iteration of this circuit.

The final peculiarity that was observed in the testing is that the OTA did not work properly. When the circuit is operating with small input voltages, the OTA hurt rather than helped leakage. Performance was always better without the OTA. As was alluded

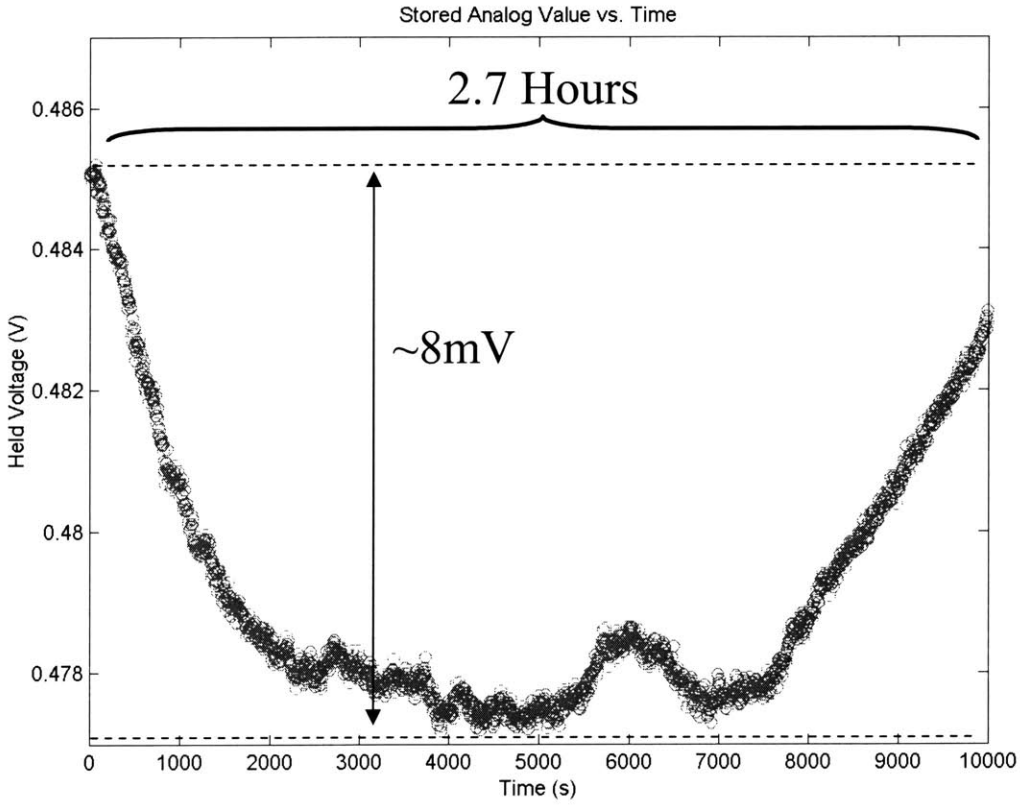


Figure 4.11 – Held voltage in analog memory vs. time. There is a drift in the ramp current source that causes voltage drift shown.

to in Section 3.2.3.4, the gain error of the OTA was originally overlooked as a potential source of offset. The implemented design contains only 15mV of intentional offset, while the analysis of Section 3.2.3.4 showed that 50mV should have been included. With this little offset included, it is possible that the PMOS is becoming forward biased at some points in its operation, and thus causing high leakage currents. However, none of the OTA signals are available for probing, so this is merely speculation.

5 Conclusions and Future Work

Referring back to Table 2.1 for a moment, we can now check to see that all of the design specifications have been met. The designed analog memory cell is completely CMOS compatible. The algorithm was chosen with small cell size in mind, and the local cells measure 300μm × 250μm (375λ × 310λ) in size. Each cell consumes 10μW, which is higher than desired, but the consumption can be lowered by eliminating the OTA from each cell, since they work better without the OTA anyway. The power supply was 3.3V. 8-bit precision was achieved in the input-output transfer characteristic. The state machine is designed such that new values can be stored in < 1ms. Finally, we have shown that the cell is capable of storing an 8-bit precise voltage on a capacitor for over 2.5 hours.

Despite the fact that the implementation has met all of the pre-defined goals of this project, there are several areas in which the design can be improved. First, the OTA buffer topology should be fabricated as a test structure so that the problems with its operation can be determined, and, if possible, the OTA should be eliminated altogether. Also, the power supply-rejection problems can be made better by designing a fully-differential version of the memory cell. With these improvements, a 10-bit version of the memory cell should be possible.

Appendix A – Auto-Zeroing and Noise

Auto-zeroing is a technique used to reduce the input-referred offset of amplifiers in switched-capacitor circuits. The way in which this technique was used in this project to create a comparator is shown in Figure A.1. In this figure, a capacitor, C_{hold}, is placed in series with a common-source amplifier, and a switch is placed from the amplifier input to its output. At the start of operation, switch S closes while V_{in} is held constant. As was explained in Section 3.3.1, this causes the amplifier input and output to settle to a voltage which balances it in its high-gain region, V_{bal}. Thus, V_{bal}-V_{in} is stored across C_{hold}. If switch S is ideal, no charge is injected into the virtual ground node of the amplifier when it opens. Now, if the input voltage V_{in} moves slightly around its original value, the movement is amplified by the gain -A of the amplifier at V_{out}. This can be performed for any input voltage value V_{in}, and thus the high-gain region of the amplifier can be centered around any input voltage value using this technique. In many switched-capacitor circuits, this technique is used with an input value of 0V, hence the term auto-zeroing.

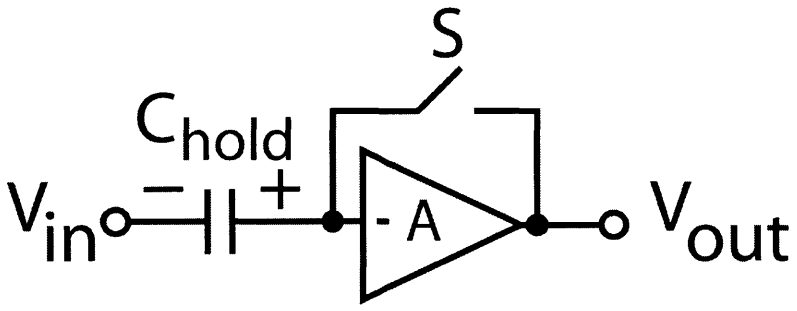


Figure A.1 – Circuit which uses auto-zeroing to establish the proper bias-point for a common-source amplifier.

Consider a real amplifier used in this configuration. As was shown in Section 3.2.1, any real amplifier has an input-referred voltage noise source which can be visualized as being in series with its input terminal. The noise source has been added to the original circuit, and the new circuit is shown in Figure A.2. The block diagram of the

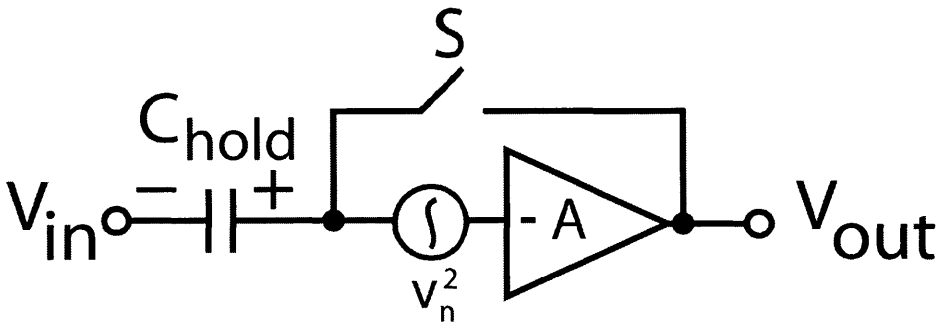


Figure A.2 – The auto-zeroing amplifier from Figure A.1 with its input-referred voltage noise source added.

transfer function of this noise source to the output of the amplifier when switch S is closed is shown in Figure A.3. We can infer from Figure 3.16 that the overall transfer function of this topology is given by

$$-\frac{A}{1+A}.$$

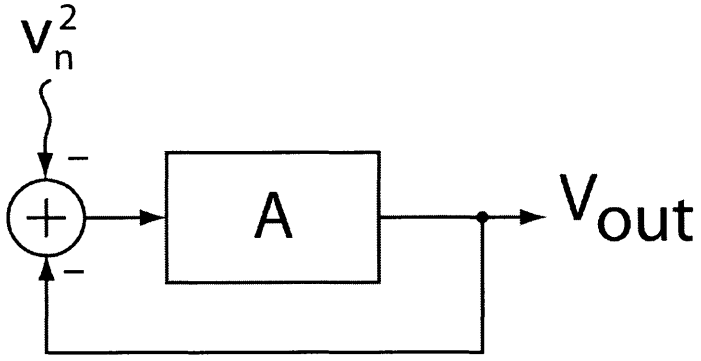


Figure A.3 – Block diagram of the transfer function of the input-referred voltage noise source to the output of the amplifier when the amplifier is placed in unity gain negative feedback.

Thus, as long as A is large in magnitude, the transfer function magnitude from input to output is very close to negative unity. Assume that the magnitude is negative unity for the remainder of this discussion. Thus, when the switch S opens, the new net voltage that has been sampled onto C_{hold} is V_{bal}-V_{in}-$V_n[nT]$. During open-loop operation, this sampled voltage is added to V_{in}, and to the current value of the noise, $V_n[(n+1)T]$. If the noise were simply a DC offset, it would not change with time and would completely cancel itself. However, the noise varies in time, and so a residual net input-referred voltage noise source still exists in the circuit, described by $V_n[(n+1)T]$- $V_n[nT]$. The goal of this appendix is to understand how the original noise is shaped in the frequency domain by auto-zeroing.

The diagram shown in Figure A.4 describes how this auto-zeroing technique's effect on noise can be viewed from a signal processing perspective. The upper path

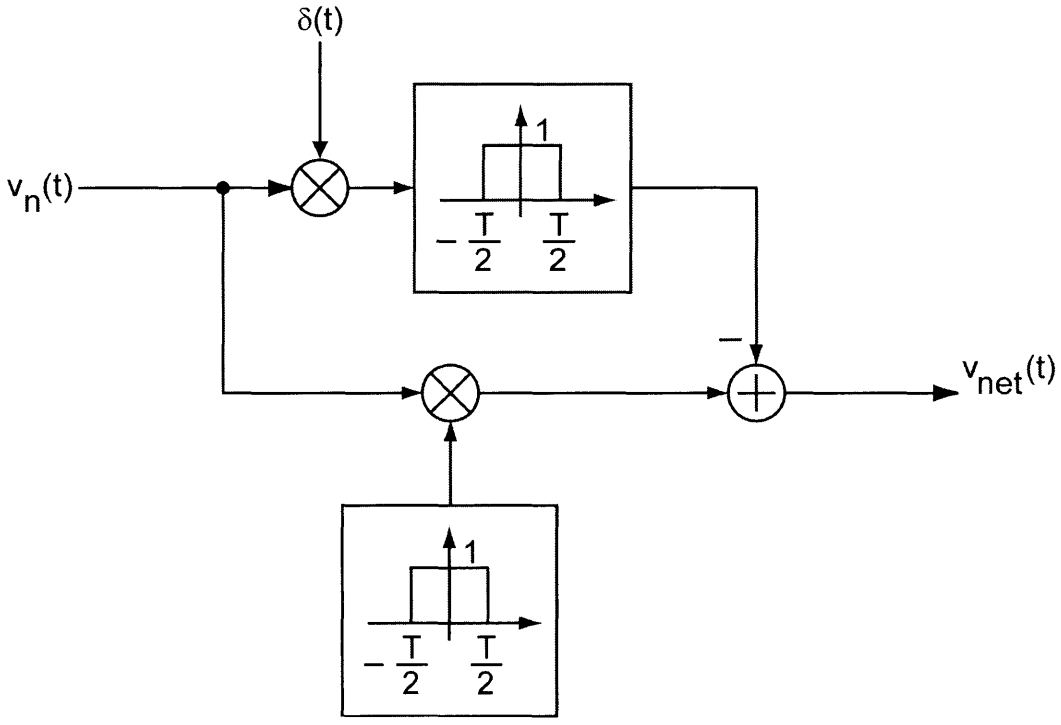


Figure A.4 – Signal processing view of the auto-zeroing technique's effect on noise.

describes a sample-and-hold of the noise signal $v_n(t)$ for a time of T. The lower path is simply a gated version of the noise source which remains on for a time T. The difference between these two paths produces the net output noise source $v_{net}(t)$, which is the effective noise of the amplifier over the time period T when it is being used as a comparator.

This diagram allows us to understand intuitively what auto-zeroing does to noise at different frequencies. At very low frequencies, the sampled value of the noise will equal the present value of the noise over the future time period T fairly well, while the higher frequency noise will oscillate and be very different than it was at the time it was sampled. Thus, the low-frequency noise component of the original signal, the 1/f noise, is killed, while the higher frequency white noise is not.

Appendix B – Digital Control Circuitry

The digital circuit block shown in Figure B.1 forms half of the chip's control circuits. *SampIn* pulsing high tells the system to sample a new input value, while *Done* pulsing high tells the system to sample the output of the ramping waveform. *Clk* is the clock signal, and *Clk/Div* is a divided version of *Clk* running at ½ its frequency. *Voff* and *Voff2* are as defined in Section 3.3.1. *SampRamp* is the same as V_{reset} from Section 3.3.2. *SampLoop* controls the sampling of the ramp output by the sample-and-hold comparator. *VSampIn* controls the sampling of the input voltage by the sample-and-hold comparator. Start was included for testing purposes, and can be ignored. The signals attached to the tags on the left are the inputs to the block, and the ones attached to tags on the right are the outputs from the block.

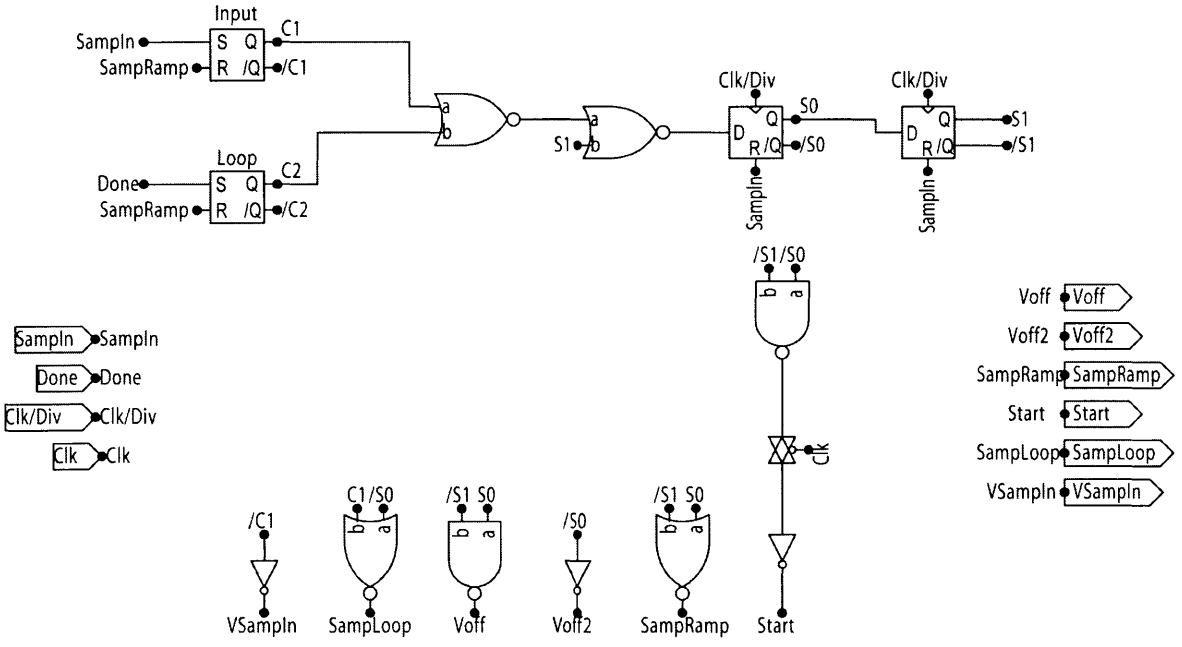


Figure B.1 – First digital circuit block used to control the analog memory.

The other half of the control circuitry is shown in Figure B.2. Trig is the output of the comparator, signaling to this circuit block that the local memory value has been surpassed by the ramp signal. Reset clears the operation of this block. Clk is the same system clock as was used in the previous digital block. Done signals to the digital block on the previous page that the ramp has surpassed the stored local value. Step serves the same function as the signal Clk shown in Figure 3.35.

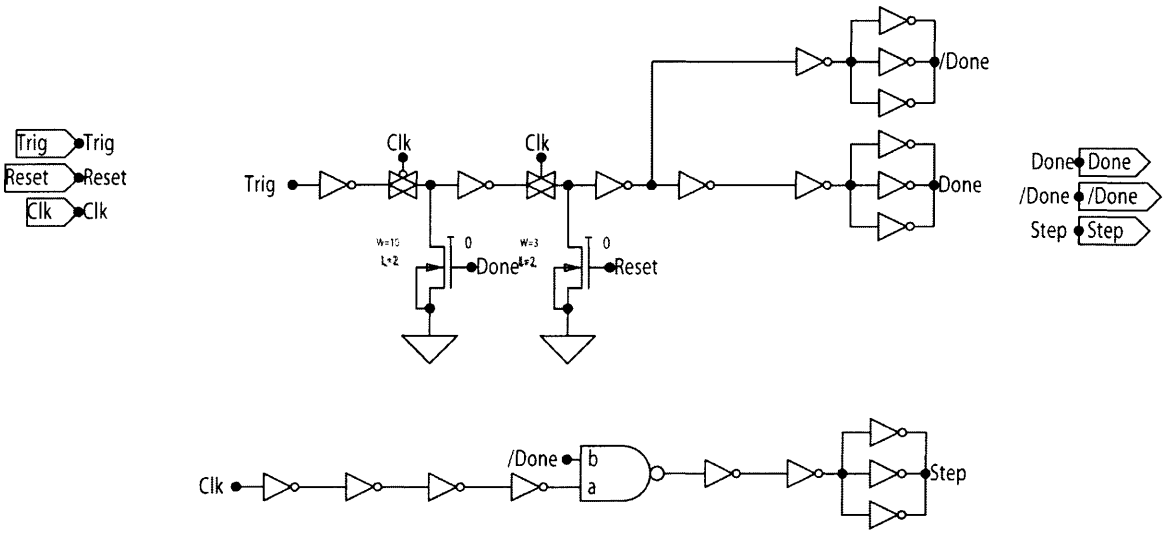


Figure B.2 – Second digital circuit block used to control the analog memory.

Appendix C – MATLAB Scripts

Script #1 – Computes the transfer function of the block diagram of Figure 3.45:

```
% This script describes and plots the capacitive-feedback ramp amplifier
% transfer function from Figure 3.44.

%Global Variables
roM1=2.9e8;                 %Output resistance of M1
roAmp=6.2e6;            %Output resistance of Amplifier
CAmp=10e-15;            %Output capacitance of Amplifier
Cvg=10e-15;                 %Output capacitance of M1
Cramp=400e-15;          %Feedback capacitor value
GmAmp=8.71e-6;              %Transconductance of Amplifier

%Create s
s=tf('s');

%Create individual transfer functions

f1=roM1/(1+s*roM1*Cvg);     %First section forward
b1=s*Cramp;                 %First section backward

ovf1=feedback(f1,b1);       %Overall TF of first section

f2=GmAmp-s*Cramp;           %Second section forward

ovf2=f2;                    %Overall TF of second section

f3=roAmp/(1+s*roAmp*CAmp);  %Third section forward
b3=s*Cramp;                 %Third section backward

ovf3=feedback(f3,b3);       %Overall TF of third section

ovftotal=ovf1*ovf2*ovf3;    %Overall forward path

ovbtotal=s*Cramp;           %Overall backward path;


%Find overall transfer function
TF=feedback(ovftotal,ovbtotal);

bode(TF,'b-',(1/ovbtotal),'ro',{10e-1,10e10});

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script #2 – Computes the total noise at the output of the capacitive-feedback amplifier in Figure 3.35 due to the current noise of transistor M_1:

```
%This script calculates the white and 1/f noise of the
%sampled-sinc frequency spectrum, and sends it through
%the integrating window and filter function of the
%ramping amplifier, as is shown in Figure 3.40.


%Global Constants
k = 1.38e-23;              %Boltzmann's constant
T = 300;                   %Operating temperature in K
q = 1.602e-19;             %Charge of an electron
gm = 2.24e-10;             %Transconductance
Ibias = 1e-11;             %Bias Current
Kf = 2e-22;                        %1/f noise parameter
Cox = 1.14e-3;             %Capacitance F/m^2
W = 16e-6;                         %Width in meters
L = 8e-6;                          %Length in meters
Cramp = 400e-15;           %Integrating capacitor value

%Global Variables
Tclk = 1e-3;               %Clock period
fclk = 1e3;                        %Clock frequency
Tint = 300e-3;             %Total integration time

%Initialize Variables
clear dtrain;
clear white_noise;
clear oneoverf;
clear total_noise;
clear sink;
clear middleo;
clear middlen;

%Create frequency vectors
fo=[-10^7:1000:10^7];
fn=[-2e7:1000:2e7];

%Find and store value of array where 0 Hz is located
for i= 1:length(fo)
        if (fo(i)==0) middleo=i;
        end;
end;

%Find and store value of array where 0 Hz is located
for i= 1:length(fn)
        if (fn(i)==0) middlen=i;
        end;
end;

%Create Transistor noise components -- give double bandwidth for convolution padding
white_noise =(2*q*Ibias);
oneoverf=(((Kf*gm^2)/(W*L*Cox))./abs(fn));
oneoverf(middlen)=0;

%Construct Total Noise
total_noise=white_noise + oneoverf;
```

```
%Create Delta Train -- S(f)
for i = 1:length(fo);
        if(mod(fo(i),fclk)==0) dtrain(i)=(2*pi/Tclk);
        else dtrain(i)=0;
        end;
end;

%Create sinc waveform for pulse block -- P(f)
sink=2*sin(2*pi*fo*Tclk/4)./(2*pi*fo);
sink(middleo)=Tclk/2;

%Create Freq Domain Representation of Delta-Sinc -- N(f)
DeltaSinc = dtrain.*sink;
DeltaSincSquared=DeltaSinc.^2;

%Convolve the DeltaSinc and the input spectrum
ConvolvedNoise=conv(total_noise,DeltaSincSquared);
TotalPulsedNoise=ConvolvedNoise(length(fo):length(fn));

%Construct Time-Limiting Block Freq Resp -- G(f)
sinkint=2*sin(2*pi*fn*Tint/2)./(2*pi*fn);
sinkint(middlen)=Tint;
sinkintSquared=sinkint.^2;

%Convolve N(f) with G(f)
ShapedNoise=conv(TotalPulsedNoise,sinkintSquared);
TimeLimitedNoise=ShapedNoise(length(fo):length(fn));

%Contruct Integrator Description
Integrator=(1/Cramp)./abs(2*pi*fo);
Integrator(middleo)=0;
IntegratorSquared=Integrator.^2;

%Shape noise with Integrator
Output=TimeLimitedNoise.*IntegratorSquared;

%Integrate the total noise power
TotalOutputNoise=(cumtrapz(fo,Output));

%Total noise power
TotalOutputNoise(length(TotalOutputNoise))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Bibliography

[1] Shieh, J., Patil., M., and Sheu, B.L., "Measurement and Analysis of Charge Injection in MOS Analog Switches," *IEEE J. of Solid-State Circuits*, Vol.22, no. 2, pp. 277-281, April 1987.

[2] Johns, D. and Martin, K., *Analog Integrated Circuit Design*. John Wiley & Sons, pp. 340-341, 1997.

[3] Diorio, C., Mahajan, S., Hasler, P., Minch, B., and Mead C., "*A High-Resolution Non-Volatile Analog Memory Cell*," IEEE International Symposium on Circuits and Systems, Vol. 3, pp. 2233-2236, 1995.

[4] Kramer, A., Hu, V., Sin, C.K., Gupta, B., Chu, R., and Ko, P.K., "EEPROM Device as a Reconfigurable Analog Element for Neural Networks," in *Tech. Dig. IEEE IEDM*, pp. 259-262, 1989.

[5] Cauwenberghs, G., "*A Micropower CMOS Algorithmic A/D/A Converter*," IEEE Trans. on Circuits and Systems I: Fund. Theory and Applications, Vol. 42, No. 11, pp. 913-919, November 1995.

[6] Hochet, B., "*Multivalued MOS Memory For Variable-Synapse Neural Networks*," Electronics Letters, Vol. 25, No. 10, pp. 669-670, May 1989.

[7] Vittoz, E., Oguey, H., Maher, M.A., Nys, O., Dijkstra, E., and Chevroulet, M., "Analog Storage of Adjustable Synaptic Weights," in *VLSI Design of Neural Networks*, Ramacher, U., and Ruckert, U., Eds. Boston, MA:Kluwer Academic Publishers, 1991.

[8] Cauwenberghs, G., "*Analog VLSI Long-Term Dynamic Storage*," IEEE International Symposium on Circuits and Systems, Vol. 3, pp. 334-337, 1996.

[9] Murray, A. F., and Buchan, L. W., "*A User's Guide to Non-Volatile, On-Chip Analogue Memory*," Electronics & Communication Engineering Journal, Vol. 10, Issue 2, pp.53-63, 1998.