Design for Reliability of Low-voltage, Switched-capacitor Circuits

by

Andrew Masami Abo

B.S. (California Institute of Technology) 1992

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy
in
Engineering—Electrical Engineering and
Computer Sciences
in the
GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA, Berkeley

Committee in charge:
Professor Paul R. Gray, Chair
Professor Bernhard E. Boser
Professor Ilan Adler

Spring 1999

The dissertation of Andrew Masami Abo is approved

Chair Date

Date

Date

University of California, Berkeley

Spring 1999

Design for Reliability of Low-voltage, Switched-capacitor Circuits

Andrew Masami Abo

Abstract

Design for Reliability of Low-voltage, Switched-capacitor Circuits

by

Andrew Masami Abo

Doctor of Philosophy in Engineering
University of California, Berkeley
Professor Paul R. Gray, Chair

Analog, switched-capacitor circuits play a critical role in mixed-signal, analog-to-digital interfaces. They implement a large class of functions, such as sampling, filtering, and digitization. Furthermore, their implementation makes them suitable for integration with complex, digital-signal-processing blocks in a compatible, low-cost technology–particularly CMOS. Even as an increasingly larger amount of signal processing is done in the digital domain, this critical, analog-to-digital interface is fundamentally necessary. Examples of some integrated applications include camcorders, wireless LAN transceivers, digital set-top boxes, and others.

Advances in CMOS technology, however, are driving the operating voltage of integrated circuits increasingly lower. As device dimensions shrink, the applied voltages will need to be proportionately scaled in order to guarantee long-term reliability and manage power density.

The reliability constraints of the technology dictate that the analog circuitry operate at the same low voltage as the digital circuitry. Furthermore, in achieving low-voltage operation, the reliability constraints of the technology must not be violated.

This work examines the voltage limitations of CMOS technology and how analog circuits can maximize the utility of MOS devices without degrading relia-

bility. An emphasis is placed on providing circuit solutions that do not require process enhancements. The specific research contributions of this work include (1) identifying the MOS device reliability issues that are relevant to switched-capacitor circuits, (2) introduction of a new bootstrap technique for operating MOS transmission gates on a low voltage supply without significantly degrading device lifetime, (3) development of low-voltage opamp design techniques. With these design techniques building blocks necessary for switched-capacitor circuits can be implemented, enabling the creation of sampling, filtering, and data conversion circuits on low-voltage supplies. As a demonstration, the design and characterization of an experimental 1.5V, 10-bit, 14.3MS/s, CMOS pipeline analog-to-digital converter is presented.

Paul R. Gray, Chair

Contents

Acknowledgments **vii**

Chapter 1 Introduction **1**

Chapter 2 Switched-Capacitor Building Blocks **5**

2.1 Sample-and-hold (S/H) 5

2.1.1 Top-plate S/H 5

2.1.2 Bottom-plate S/H 8

2.2 Gain stage 11

2.3 Integrator 14

2.4 Comparator 14

Chapter 3 Switched-Capacitor Applications **19**

3.1 Filters 19

3.2 Sigma-delta analog-to-digital converters 27

3.3 Pipeline Analog-to-digital converters 28

3.4 Capacitor digital-to-analog converters 30

Chapter 4 CMOS Technology Scaling **33**

4.1 CMOS Scaling 34

4.2 Voltage scaling for low power 35

4.3 Voltage scaling for reliability 38

4.3.1 Gate oxide breakdown 38

4.3.2 Hot-electron effects 41

4.4 Fundamental scaling limits 44

4.5 Analog circuit integration 45

Chapter 5 Low-voltage Circuit Design 49

5.1 Low-voltage, switched-capacitor design issues 49
5.2 Reliable, high-swing MOS switch 53
5.2.1 Operation 54
5.2.2 Design guidelines 56
5.2.3 Layout considerations 59
5.3 Opamp 62
5.3.1 Application 63
5.3.2 Topology 64
5.3.3 Biasing 64
5.3.4 Linear Settling time 67
5.3.5 Slew rate 74
5.3.6 Noise 76
5.3.7 DC gain 77
5.3.8 Common-mode feedback 77
5.4 Comparator 78
5.4.1 Offset 80
5.4.2 Meta-stability 83
5.4.3 Pre-amplifier bandwidth 84

Chapter 6 Pipeline ADC Architecture 85

6.1 Pipeline ADC architecture 85
6.2 1.5-bit/stage architecture 86
6.3 1.5 bit/stage implementation 89
6.4 Pipeline stage accuracy requirements 90
6.4.1 Capacitor matching 90
6.4.2 Capacitor linearity 91
6.4.3 Opamp DC gain 92
6.4.4 Opamp settling 93
6.4.5 Thermal noise 93

6.4.6 Error tolerances 93
6.4.7 Design example 96

Chapter 7 Prototype Implementation 99
7.1 Technology 99
7.2 Layout 99
7.3 Master bias 102
7.4 Clock generator 103
7.5 Capacitor trimming 105
7.6 Gain stage 108
7.7 Sub-ADC/DAC 109

Chapter 8 Experimental Results 111
8.1 Test setup 111
8.2 Dynamic linearity and noise performance 113
8.3 Static linearity 114
8.4 Summary 116

Chapter 9 Conclusion 119

Bibliography 121

Index 128

Acknowledgments

It has been a real privilege to be a graduate student in the EECS department at the University of California, Berkeley. My experience here has been full of opportunities to learn from a faculty and student body with a wide and deep expertise in engineering. I would like to directly thank those people in the department who were particularly instrumental in contributing to my experience at Berkeley.

I would like to thank my advisor, Professor Paul Gray, for his invaluable guidance throughout my long journey as a graduate student. His many years of circuit design experience have allowed him to focus in on the critical and interesting issues of any problem. It has been a unique privilege to be in his group, and undoubtedly the credibility and prestige of his name will continue to open doors for me in the future.

I would like to thank the members of my thesis committee for their valuable feedback. I would like to thank Professor Boser for his insightful comments and discussions on opamp design. I would like to thank Professor Ilan Adler for also reading my thesis. Professor Chenming Hu was a very valuable member of my qualifying exam committee. I would like to thank him for adding the device technology perspective, which directly contributed to the chapters on CMOS technology.

One of the real assets of the EECS department is the working environment of 550 Cory Hall. Graduate students really spend more hours that they would like to in that room, but the people there make it fun at least *some* of the time. It has been invaluable to have fellow students who are immediately accessible and who freely give there help you have a technical question. I probably learned as much in 550 as I did in the classroom. I would like to thank all the members of Professor Gray's group past and present. When I was a rookie, the veterans such as Dave Cline, Cormac Conroy, Robert Neff, Ken Nishimura, and Greg Uehara really helped me when I was truly clueless. I would like to thank the pipeline

ADC veterans Thomas Cho, Dave Cline, and Cormac Conroy from whose work my own work stems. The current (and recently graduated) group has been a real pleasure to work with. I would like to thank my immediate cube-mates Sekhar Narayanaswami, Keith Onodera, and Jeff Ou for their company and numerous donuts. It been a pleasure working with the more recent students, and I wish them a speedy graduation (if they haven't already) Danelle Au, George Chien, Kelvin Khoo, Li Lin, Luns Tee, and Martin Tsai. I would like to thank Arnold Feldman for our useful discussions on opamps and switched-capacitor circuits. Thanks to Jeff Weldon and Carol Barrett who have been good friends. Thanks to Dennis Yee, Tom Burd, and Chin Doan for all the golf outings to Alameda. I'd like to also thank Arthur Abnous, Anna Ison, Dave Lidsky, and Lisa Guerra who have been good friends in 550.

Living at 920 Keeler has definitely been a lot of fun, and an unforgettable part of my Berkeley experience. Thanks to the original gang of Arya Behzad, Srenik Mehta, and Tony Stratakos. I'll never forget Arya's Target bikes, Srenik living in the kitchen, and Tony's demonstration of flammable gases. The last few years have been equally fun. Thanks to Chris Rudell and Sekhar Narayanaswami for being great friends. Who's going to order the Nino's? Now, I hand the torch to Joe Seeger to carry on the tradition.

I would like to thank my family, for being very supportive and patiently listening to me trying to explain my research. I would like to thank my parents, Mary and Joe Abo, for believing me every time I said I was graduating "next year." I also would like to thank my sister, Julie, and my brother-in-law, Aaron Dominguez, for always being there and providing me with vacations that were such important and fun breaks from my work.

Lastly, I would like to thank the Semiconductor Research Corporation for their financial support through the Graduate Fellowship Program. I would also like to thank the student liaison, Ginny Poe, for helping with all the administration issues that go along with having a fellowship.

Chapter 1

Introduction

SWITCHED-CAPACITOR circuits fill a critical role in analog/digital interfaces–particularly highly integrated applications. In these applications, a complex, digital-signal-processing core is often interfaced to real-world inputs and outputs. Such applications include voiceband modems, disk drive read channels, set-top cable television receivers, baseband processing in wireless transceivers, and others.

For these high-volume, dedicated applications, cost is often the most important factor. Increasing levels of mixed-signal integration have been instrumental in lowering the fabrication, packaging, and testing costs of these products.

CMOS has proven to be the most cost-effective technology for achieving high levels of integration. For analog circuits this technology typically does not have the same raw performance as bipolar or BiCMOS, but for complex, mixed-signal applications CMOS offers a distinct cost advantage, as evidenced by the wide commercial acceptance of CMOS for analog signal processing. In particular, switched-capacitor circuits exploit the charge storing abilities of CMOS to achieve precision signal processing. Thus, high-performance filters and data converters can be implemented in CMOS. Although an increasing amount of signal processing is performed in the digital domain, the analog-digital interface will remain a fundamentally necessary element.

There are numerous examples of mixed-signal applications that employ switched-capacitor circuits to perform front-end analog pre-processing and digitization. Figure 1.1 shows a voiceband codec used in telecommunications networks. This example is a typical mixed-signal application comprised of several switched-capacitor (SC) interface circuits. Switched-capacitor circuits are used to implement low-pass (LPF) high-pass (HPF) filters, analog-to-digital convert-

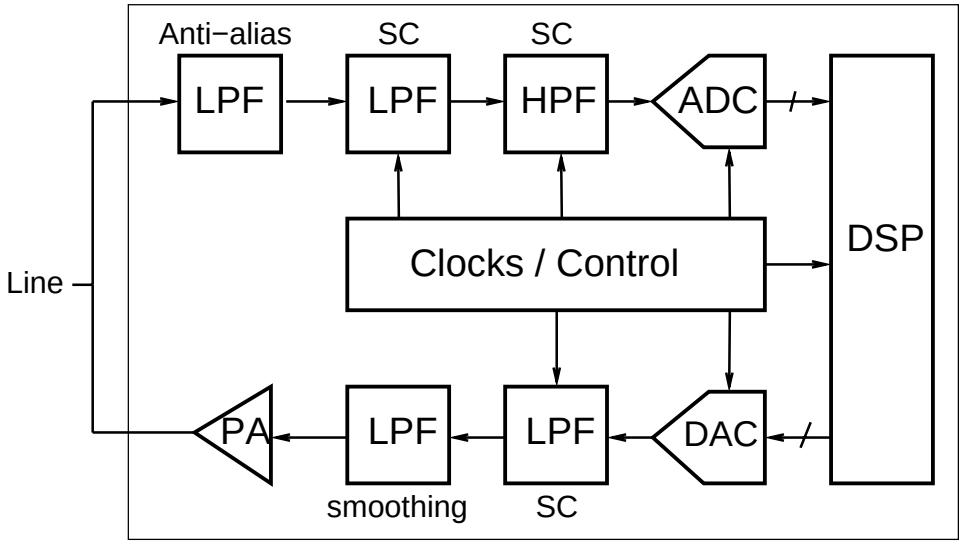


Figure 1.1 Integrated PCM voiceband codec

ers (ADC) and digital-to-analog converters (DAC). Figure 1.2 shows an integrated baseband processing unit suitable for a cellular phone or wireless LAN application. Disk drive read channels have also evolved to high levels of integration as shown in figure 1.3.

The continued miniaturization of CMOS devices presents new benefits as well as obstacles to the implementation of switched-capacitor circuits. Similar to digital circuits, analog switched-capacitor circuits benefit from inherently faster tran-

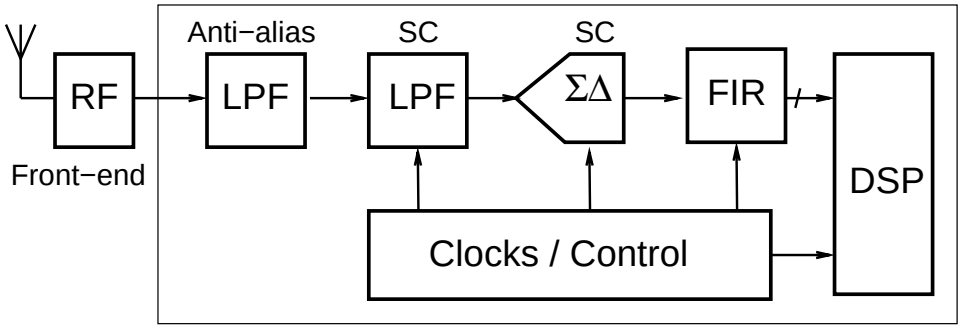


Figure 1.2 Baseband signal processing in wireless datacom receiver

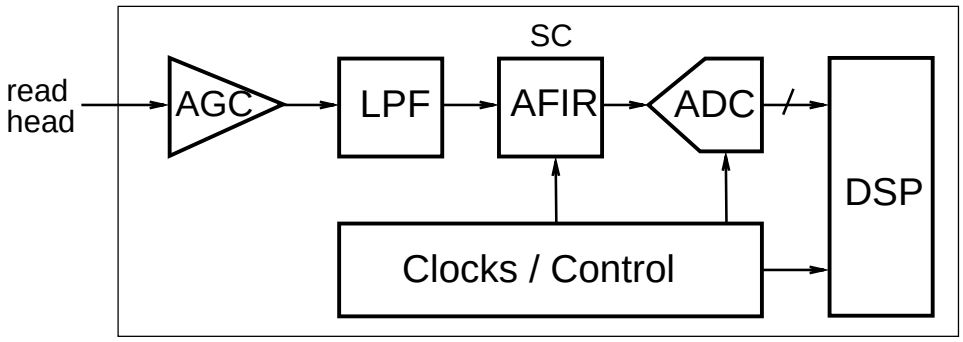


Figure 1.3 Disk drive read channel

sistors at smaller geometries. These scaled devices, however, place new reliability constraints on allowable operative voltage. As a result, future voltage supplies will be greatly reduced from current levels. This reduction greatly complicates analog, switched-capacitor circuit design.

This research has focused on examining the reliability issues that demand low-voltage operation, and the implementation of reliable, low-voltage, switched-capacitor circuits in CMOS. An emphasis was placed on circuit solutions that do not rely on additional enhancements to the technology, such as multiple threshold voltages, or thick oxides capable of supporting large voltages. Such circuit solutions tend to be more flexible and cost effective.

Chapter 2

Switched-Capacitor Building Blocks

SWITCHED-CAPACITOR circuits are pervasive in highly integrated, mixed-signal applications. This chapter describes the basic building blocks that comprise these circuits. These blocks are the sample-and-hold (S/H), gain stage, integrator, comparator. From these elements more complex circuits can be built such as filters, analog-to-digital converters (ADC), and digital-to-analog converters (DAC). All sampled-data circuits, such as these, require a pre-conditioning, continuous-time, anti-alias filter to avoid aliasing distortion. A more detailed discussion of this continuous-time block can be found in [31]. After the theory of operation of each block is described, a brief discussion of practical non-idealities follows. This chapter is not intended as a rigorous and detailed analysis; it is a brief overview. A more rigorous analysis can be found in the references.

2.1 Sample-and-hold (S/H)

The sample-and-hold is the most basic and ubiquitous switched-capacitor building block. Before a signal is processed by a discrete-time system, such as an ADC, it must be sampled and stored. This often greatly relaxes the bandwidth requirements of following circuitry which now can work with a DC voltage. Because the S/H is often the first block in the signal processing chain, the accuracy and speed of entire application cannot exceed that of the S/H.

2.1.1 Top-plate S/H

In CMOS technology, the simplest S/H consists of a MOS switch and a capacitor as shown in figure 2.1. When V_g is high the NMOS transistor acts like a linear resistor, allowing the output V_o to track the input signal V_i. When V_g transitions

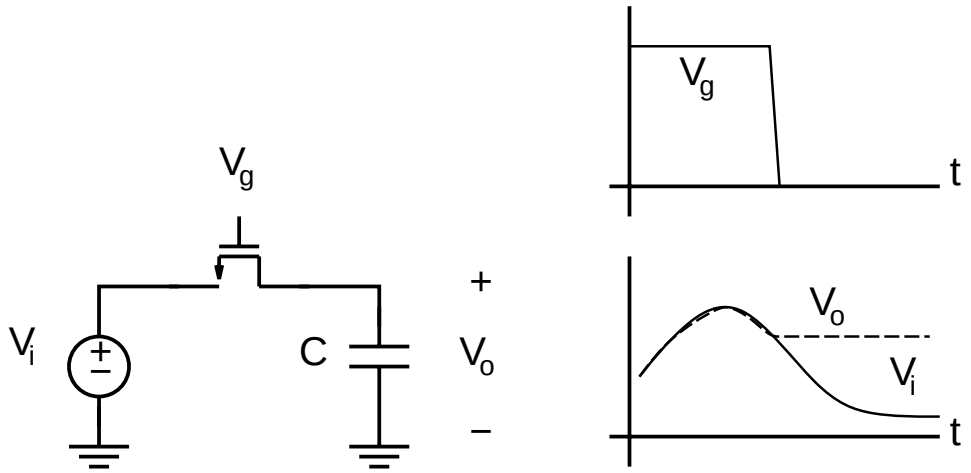


Figure 2.1 MOS sample-and-hold circuit

low, the transistor cuts off isolating the input from the output, and the signal is held on the capacitor at V_o.

There are several practical limitations to this circuit. Because the RC network has finite bandwidth, the output cannot instantaneously track the input when the switch is enabled. Therefore, a short acquisition period must be allocated for this (exponentially decaying) step response. After the S/H has acquired the signal, there will be a tracking error due to the non-zero phase lag and attenuation of the sampling network. The latter linear, low-pass filtering does not introduce distortion and is usually benign for most applications. The on-conductance, however, of the transistor is signal dependent:

$$g_{ds} = \mu C_{ox} \frac{W}{L}(V_g - V_i - V_t) \tag{2.1}$$

Thus the transfer function from input to output can become significantly nonlinear if $(V_g - V_i - V_t)$ is not sufficiently large. A detailed analysis of these dynamic errors can be found in [49].

When the switch turns off, clock feed-through and charge injection introduce error in the output. When the gate signal V_g transitions from high to low, this step AC couples to the output V_o via parasitic capacitances, such as C_{gs} and C_{gd}. Be-

cause the output is a high impedance node, there is no way to restore the DC level. This coupling is called clock feed-through. This error is usually not a performance limitation because it is signal-independent and therefore only introduces an offset and not distortion. To first order this error can be eliminated using a differential configuration. Charge injection, however, is a signal-dependent error. When switch is turned off quickly, the charge in the channel of the transistor is forced into the drain and source, resulting in an error voltage. The charge in the channel is approximately given by equation 2.2. Because q is signal dependent, it represents a gain error in the S/H output. There have been several efforts to accurately characterize this error [49, 69, 76, 77].

$$q = WLC_{ox}(V_g - V_i - V_t) \tag{2.2}$$

This circuit is also sensitive to parasitic capacitance. Any parasitic capacitance at the output change the amount of signal charge sampled, which is often the critical quantity in switched-capacitor circuits. Bottom-plate sampling can greatly reduce these errors.

The channel resistance of the switch contributes thermal noise to the output. This random error sets an upper bound on the signal-to-noise ratio (SNR). If the wide-band thermal noise is integrated over all frequencies, the resulting variance in the output voltage is only dependent on the sampling capacitance C [15].

$$\overline{v_o^2} = \frac{kT}{C} \tag{2.3}$$

Jitter in the sampling clock or aperture error also introduces a random error component to the output. If the sampling edge has an variance in time of σ_t^2, the worst case voltage variance while sampling a sine wave $V_i = \hat{V}\sin(\omega t)$ is [45] is

$$\overline{v_o^2} \leq (\hat{V}\omega)\sigma_t^2. \tag{2.4}$$

2.1.2 Bottom-plate S/H

A technique called bottom-plate sampling to first order eliminates some of the errors in the top-plate S/H circuit. Figure 2.2 shows the bottom-plate sampling configuration. While clocks ϕ' and ϕ are high, V_o tracks the input voltage V_i. When clock ϕ' transitions from high to low, switch M2 turns off, and the charge on node x is trapped. Because charge is conserved, the charge on capacitor C is now fixed $q = CV_i$. This is the defines sampling instant. When clock ϕ transitions from high to low, switch M1 is turned off and the output is isolated from the input.

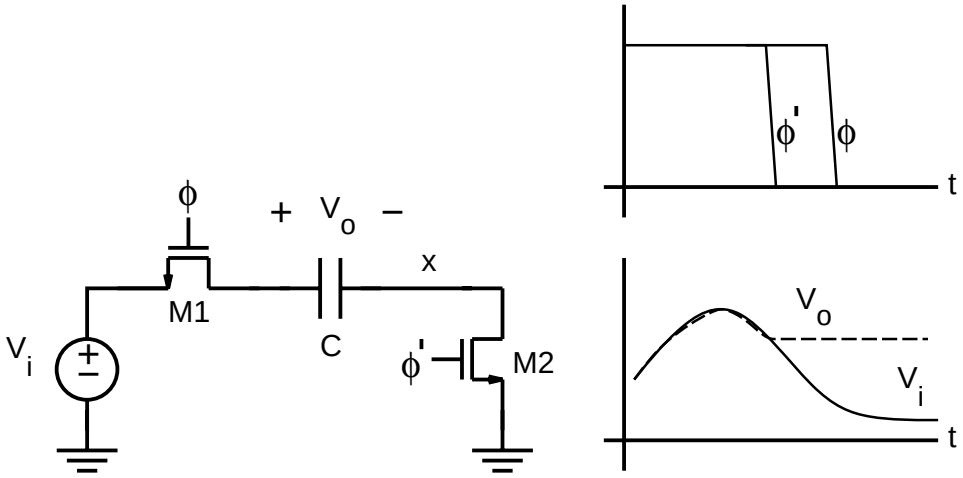


Figure 2.2 Bottom-plate sample-and-hold circuit

When M2 turns off, the voltage at node x is perturbed due to clock feed through and charge injection. In this case, however, the charge injection is signal-*independent* because drain and source see a fixed potential (ground). To first order this eliminates signal-dependent charge injection distortion. The remaining offset can be further rejected with a differential configuration. The charge injection from M1 does not alter the charge stored on capacitor C due to charge conservation.

Figure 2.3 shows a practical implementation of bottom-plate sampling in a differential configuration. This circuit uses a two-phase, non-overlapping clock. During phase ϕ_1 the input V_i is sampled differentially onto C_p and C_n as described

above. During phase ϕ_2 the opamp is put into a unity gain configuration. This drives $(V_{xp} - V_{xn}) \rightarrow 0$. Due to conservation of charge, the sampled input appears at the opamp output $V_o = V_i(\mathrm{sampled})$. Because the summing nodes of the opamp are driven to the same potential, no differential signal charge is stored on parasitic capacitance at the opamp input. Furthermore, the opamp provides a low-impedance output for driving any following signal processing blocks.

Although the opamp greatly improves the performance of the S/H circuit, it adds substantial complexity to its design. Any offset in the opamp will appear directly at the output in this configuration. For CMOS technologies this offset can be in the range of 10-50mV typically. For offset sensitive applications, there are several auto-zeroing techniques applicable to switched-capacitor circuits [24].

The finite DC gain of the opamp introduces a fixed gain error:

$$f = \frac{C}{C + C_{ip}} \tag{2.5}$$

$$V_o = \frac{a}{1 + af} \tag{2.6}$$

$$= \frac{1}{f}\frac{1}{1 + \frac{1}{af}} \tag{2.7}$$

$$\approx (1 + \frac{C}{C_{ip}})(1 - \frac{1}{af}) \tag{2.8}$$

$$\tag{2.9}$$

Where a is the DC opamp gain, f is the feedback factor, $C = C_p = C_n$, C_{ip} is the opamp input capacitance. This fixed error does not introduce distortion and is usually benign.

The finite bandwidth of the opamp limits the clock frequency of this circuit. The clock period must be sufficiently long to allow the desired level of settling accuracy at the opamp output. Typically the bias currents in the opamp can be increased to increase the opamp bandwidth at the expense of increased power consumption.

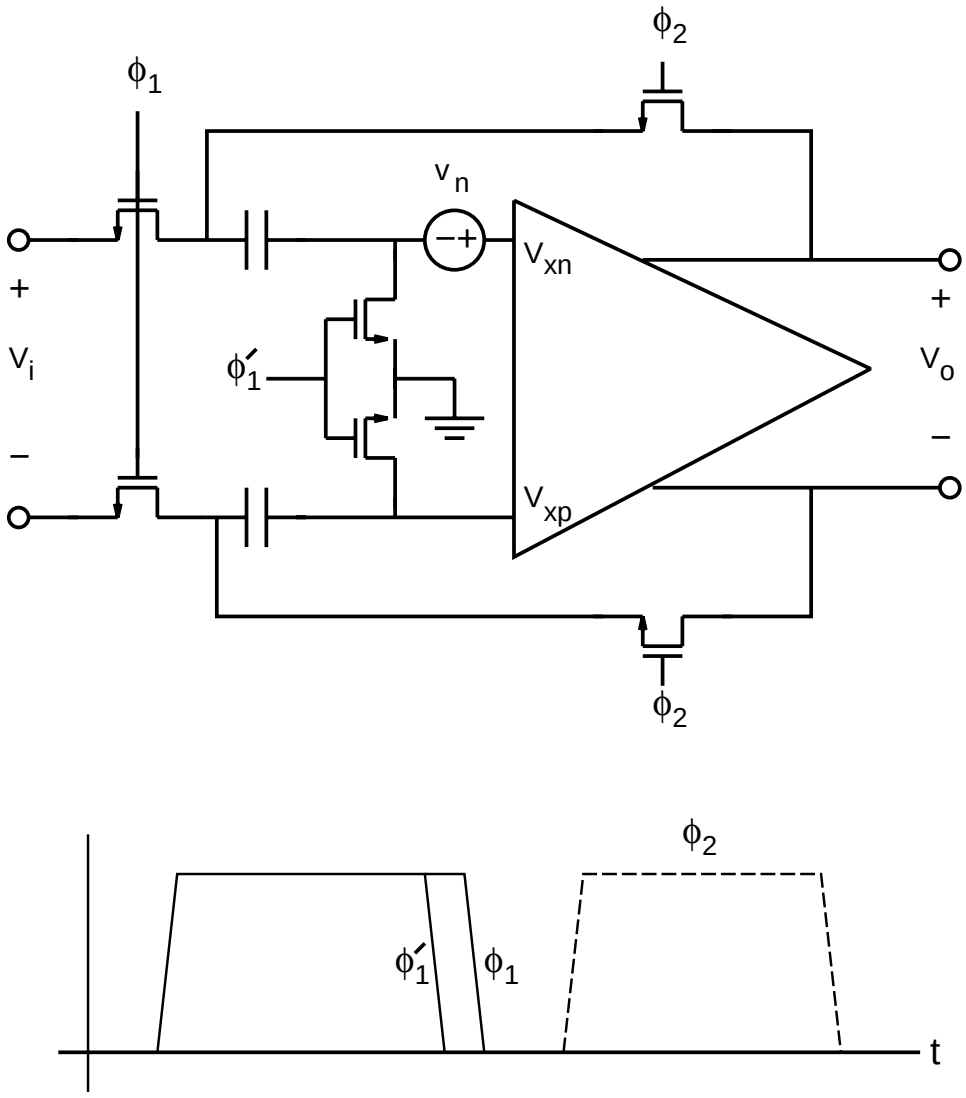


Figure 2.3 Practical fully differential sample-and-hold circuit

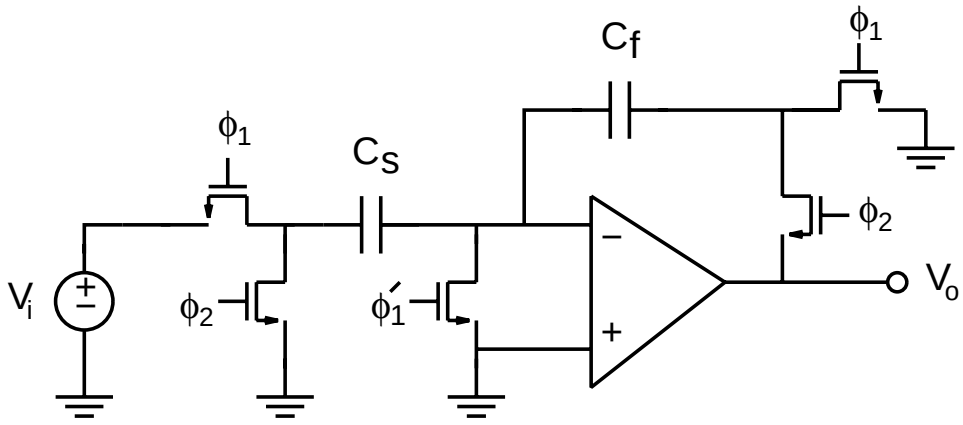


Figure 2.4 Single-ended gain stage

In addition to the fundamental kT/C sampling noise, the opamp will add thermal noise due to active elements. If the noise in the opamp can be represented as a single input-referred source v_n as shown in figure 2.3, the total output-referred noise will be:

$$\overline{v_o^2} = 2\frac{kT}{C} + \overline{v_n^2} \tag{2.10}$$

2.2 Gain stage

The sample-and-hold circuit shown in figure 2.3 can be modified to provide both gain and sample-and-hold functions. This operation is common in pipeline analog-to-digital converters (section 3.3) and filters (section 3.1). Figure 2.4 shows a gain stage that samples the input, applies gain, and holds the output value. A single-ended version is shown for simplicity, but the following analysis applies to a differential version which is most commonly used in practice.

To better understand the operation of this circuit, figures 2.2a and 2.2b show the states of the switches during phase 1 and phase 2 respectively. During phase 1 (figure 2.2a), the input V_i is sampled across C_s. The opamp is not used during this phase, and this time can be used to perform auxiliary tasks such as resetting common-mode feedback (section 5.3.8). The charge q is:

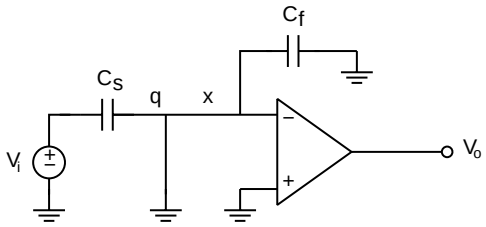


Figure 2.2a: Phase 1

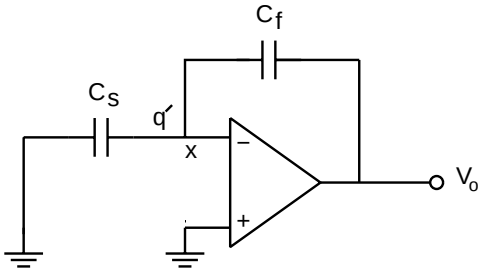


Figure 2.2b: Phase 2

$$q = C_s(0 - V_i) = -C_s V_i \tag{2.11}$$

Notice there is no charge stored on C_f since both sides are grounded. Bottom-plate sampling is employed, and the sampling instant is defined by ϕ_1' as before. During phase 2 (figure 2.2b), the opamp is put in a negative feedback configuration, forcing node x to zero (virtual ground). Because the input is also ground, there is no charge storage on C_s, and all the charge is transferred to C_f. Thus, a *voltage* gain of C_s/C_f is achieved. Analytically, charge on node x is conserved, so $q = q'$:

$$q = q' \tag{2.12}$$

$$-C_s V_i = C_f(0 - V_o) \tag{2.13}$$

$$\frac{V_o}{V_i} = \frac{C_s}{C_f} \tag{2.14}$$

If we consider the input V_i as a discrete-time sequence $V_i(n) = V_i(nT)$, where T is the sampling period, then the output is

$$V_o(n) = \frac{C_s}{C_f} V_i(n-1). \tag{2.15}$$

This equation reflects the one period latency of this discrete-time circuit.

Because this circuit incorporates an opamp, it has the same limitations as the sample-and-hold circuit in figure 2.3. In addition, the exact gain of the stage is de-

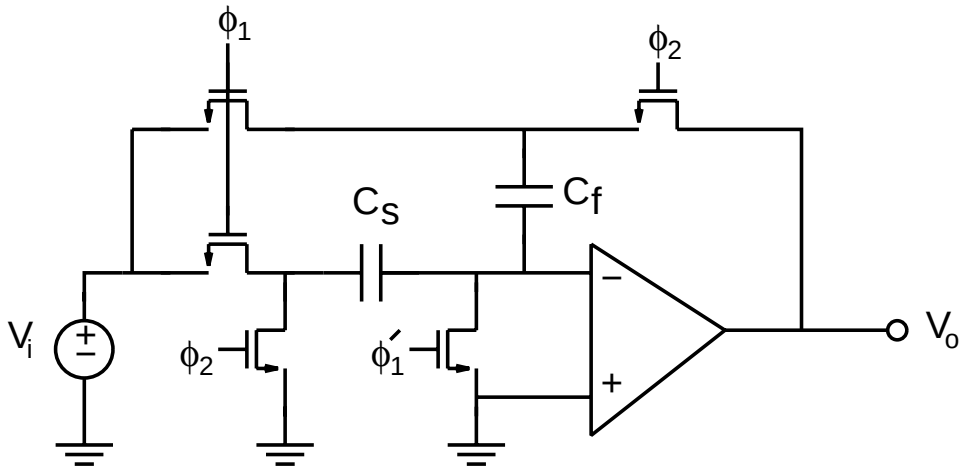


Figure 2.5 Improved settling speed gain stage

pendent on the capacitor matching of C_s and C_f. For example if $C_s = C_f + \Delta C$ then:

$$V_o(n) = \frac{C_s}{C_f} V_i(n-1) + \frac{\Delta C}{C_f} V_i(n-1) \tag{2.16}$$

In high-resolution applications, the second term can represent a significant error. Similarly if the capacitors have a voltage-dependent value, the gain will be distorted (section 6.4.2).

Figure 2.5 shows another gain stage configuration that uses both the feedback (C_f) and sampling (C_s) capacitors to sample the input. This configuration has the advantage that the opamp settling is inherently faster for a given stage gain [46] due a larger feedback factor.

The transfer function of this stage is:

$$V_o(n) = \left(1 + \frac{C_s}{C_f}\right) V_i(n-1) \tag{2.17}$$

Notice the gain is larger for the same capacitor loading. This effect can be significant for low-gain stages of 2 or 3 for example.

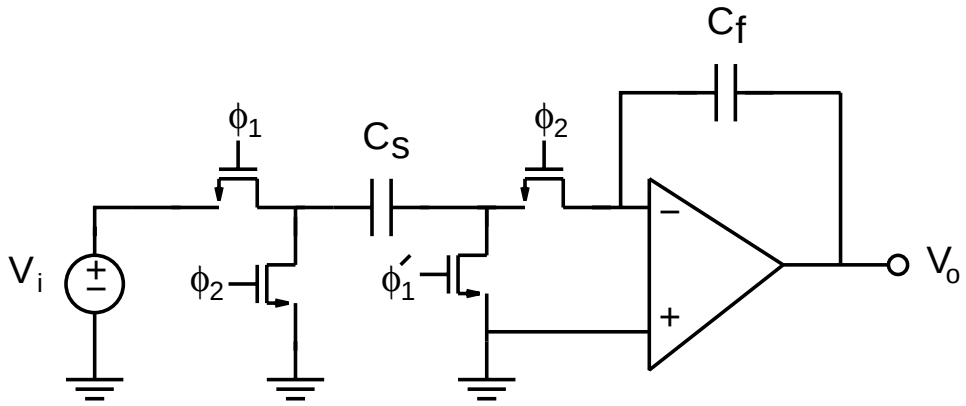


Figure 2.6 Switched-capacitor integrator

2.3 Integrator

Another modification of the basic sample-and-hold circuit yields a switched-capacitor integrator. Integrators are used throughout switched-capacitor filters (section 3.1) and sigma-delta modulators (section 3.2). Figure 2.6 shows a switched-capacitor integrator. For simplicity the single-ended version is shown, but these results apply to a differential implementation as well.

The same analysis used for the gain stage (above) can be used for the integrator. Unlike the gain stage, the feedback capacitor C_f is not reset each cycle. Therefore, C_f accumulates previous sampled values:

$$V_o(n) = V_o(n-1) + \frac{C_s}{C_f} V_i(n-1) \tag{2.18}$$

The integrator has the same performance limitations as the gain stage.

2.4 Comparator

Comparators are not strictly considered switched-capacitor elements. They, however, often employ switched-capacitor techniques and are used in switched-capacitor applications such as pipeline analog-to-digital converters and sigma-delta analog-to-digital converters. Figure 2.7 shows a comparator suitable for use in a two-phase, switched-capacitor circuit. For simplicity, a single-ended version is shown,

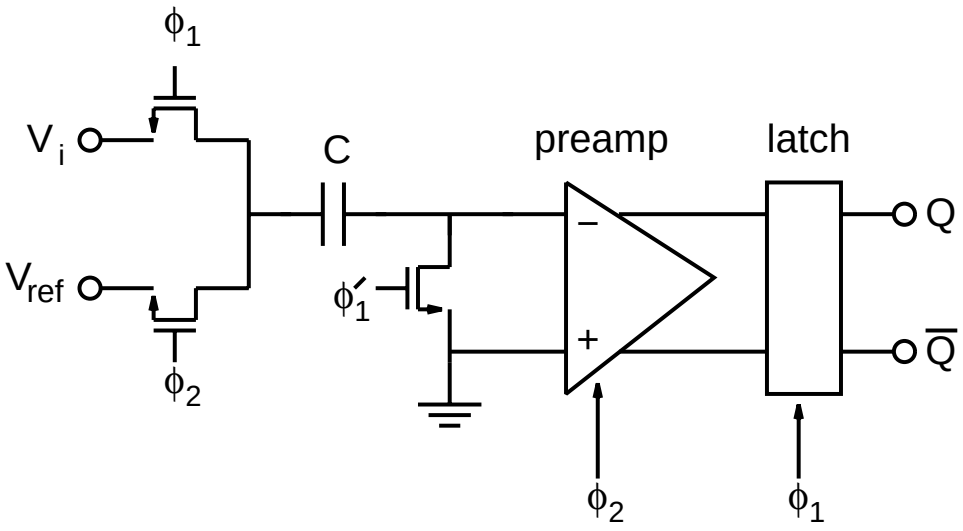


Figure 2.7 Switched-capacitor comparator

but this circuit can be extended to a fully differential implementation.

During phase 1 the input signal V_i is sampled across capacitor C. Again, bottom plate sampling is employed using the early clock phase ϕ_1'. During phase 2, the reference voltage V_{ref} is applied to the left side of the capacitor. The voltage difference $(V_{ref} - V_i)$ appears at the input of the pre-amplifier. The pre-amplifier amplifies this difference and applies it to the input of a regenerative latch. At the end of phase 2 the pre-amplifier outputs are disconnected from the input. At the beginning of phase 1 of the next cycle, the latch is strobed, creating digital logic levels at the output.

Any offset voltage of the pre-amplifier is directly referred to the input of the comparator. The potentially large offset of the latch is divided by the small-signal gain of the pre-amplifier when referred to the input. Multiple pre-amplifiers can be cascaded to further reduce the effective latch offset at the expense of power consumption. If a low offset is required, auto-zero techniques can be employed in the pre-amplifier(s) [3, 65, 22, 66].

The speed of the comparator is determined by the regenerative time constant of the latch. Consider the representative latch shown in figure 2.8. It consists of two inverters or amplifiers in a positive feedback loop, which are capable of

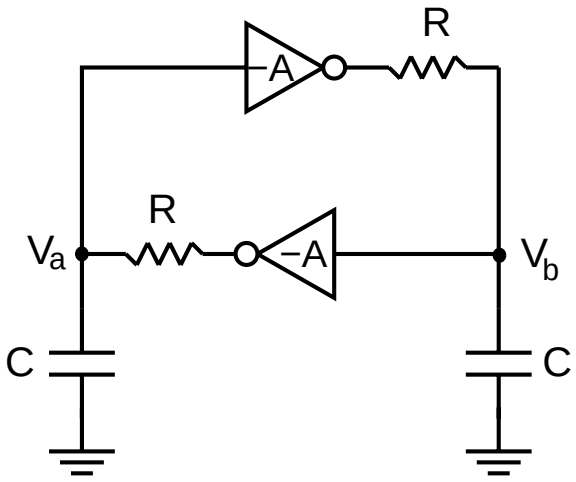


Figure 2.8 Regenerative latch time constant

amplifying a small difference in V_a and V_b to full logic levels. The time required for this amplification is dependent on the initial difference. If the initial difference between V_a and V_b when the latch is strobed is V_0, and the desired voltage difference is V_f, then the time required is [75, 66]

$$t_{comp} = \frac{\tau}{A-1} \ln\left(\frac{V_f}{V_0}\right) \tag{2.19}$$

where $\tau = RC$ and A is the small-signal gain of the inverters. Thus, for arbitrarily small inputs, the amplification time is arbitrarily long or meta-stable. If the input to the comparator is random, then there is a finite probability that the comparator will not be able to render a decision in a given time period. If the time given to make a decision is T, and the input is uniformly distributed from $[-V_f, V_f]$ then probability that the comparator will not amplify to full output levels is [75, 66]

$$P(t_{comp} > T) = \exp\left(\frac{-(A-1)T}{\tau}\right). \tag{2.20}$$

This result is *independent* of thermal noise and offsets. Therefore, if $P << 1$ then the mean time to failure (MTF) is given by

$$\mathrm{MTF} \approx \frac{1}{NfP} \tag{2.21}$$

where N are the number of concurrently operating comparators in the system and f is the frequency of comparisons per second. In a real design, τ and T must be chosen such that the mean time to failure is sufficiently long, such as the lifetime of the system (e.g. 20 years).

Chapter 3

Switched-Capacitor Applications

FROM THE building blocks described in chapter 2, larger applications, such as filters, sigma-delta analog-to-digital converters, pipeline analog-to-digital converters, and digital-to-analog converters can be built. This chapter gives a brief overview of how these representative applications incorporate switched-capacitor blocks. It is not intended as a rigorous and detailed analysis; a more rigorous analysis can be found in the references.

3.1 Filters

Discrete-time filters are the most common application of switched-capacitor circuits. Switched-capacitor filters were conceptualized in the early 1970's [26, 33] and soon after implemented in MOS integrated circuit technology [83, 34, 1, 11]. They remain prevalent in mixed-signal interfaces today. The following section gives a brief overview of the operation of switched-capacitor filters, their capabilities, and performance limitations. A more detailed tutorial can be found in [30, 31].

Before the advent of active elements, filters were implemented with passive elements, such as inductors, capacitors, and resistors. Practical inductors, however, are typically lossy, limiting the attainable selectivity. Furthermore, at low frequencies, the required size and weight of the inductors becomes large. On the other hand, very high Q can be attained with practical capacitors. Active RC filters were developed to exploit this fact and use active integrators to emulate RLC networks.

Given a continuous-time transfer function, an active RC realization can be created by mapping it onto active RC integrators. For example, a very common

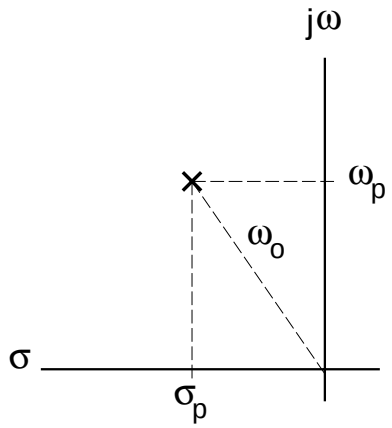


Figure 3.1 Poles positions of biquad in s-plane

function is the biquad (two zeros and two poles).

$$\frac{V_{out}}{V_{in}(s)} = -\frac{K_2 s^2 + K_1 s + K_0}{s^2 + \frac{\omega_o}{Q} s + \omega^2} \tag{3.1}$$

The poles of this transfer function are shown in figure 3.1.

$$\omega_o \triangleq |s_p| = \sqrt{\sigma_p^2 + \omega_p^2} \tag{3.2}$$

$$Q \triangleq \frac{|s_p|}{2|\sigma_p|} \tag{3.3}$$

The transfer function can then be (non-uniquely) algebraically partitioned into the following form:

$$V_{out} = -\frac{1}{s}(a_1 V_x + a_2 s V_y + ...) \tag{3.4}$$

Notice that V_{out} is expressed as the output of a $\frac{1}{s}$ integrator. The terms V_x and V_y are intermediate signals containing either V_{in}, V_{out}, or the output of another $\frac{1}{s}$ integrator. It is sometimes helpful to represent the partition schematically as shown in figure 3.2. Each arc in the graph is labeled with a weighting factor. Expressions of this form map directly onto active RC integrators as shown in figure 3.3. The capacitor is set to unity value. In the actual realization, once the in-

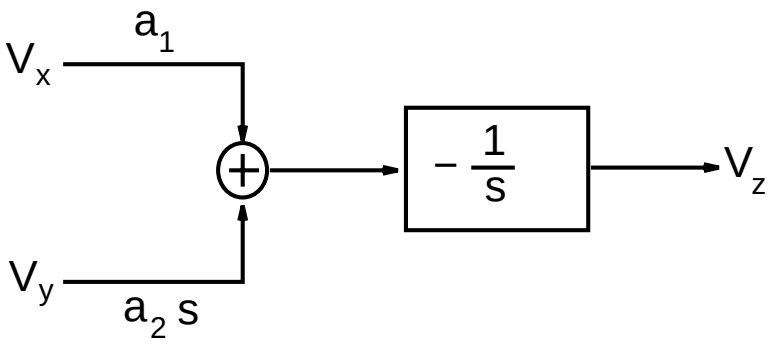


Figure 3.2 Flow diagram of RC integrator

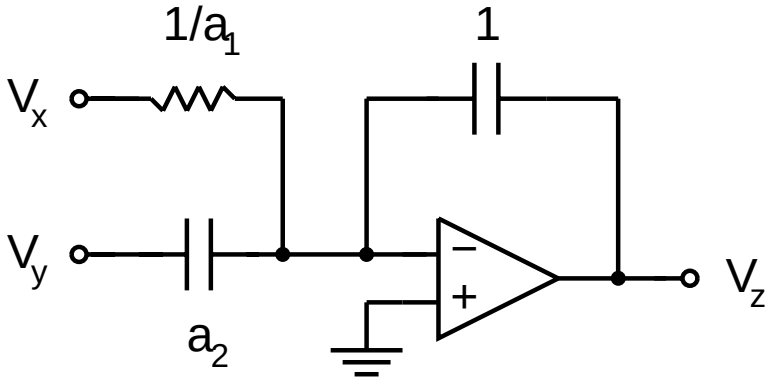


Figure 3.3 Equivalent RC integrator

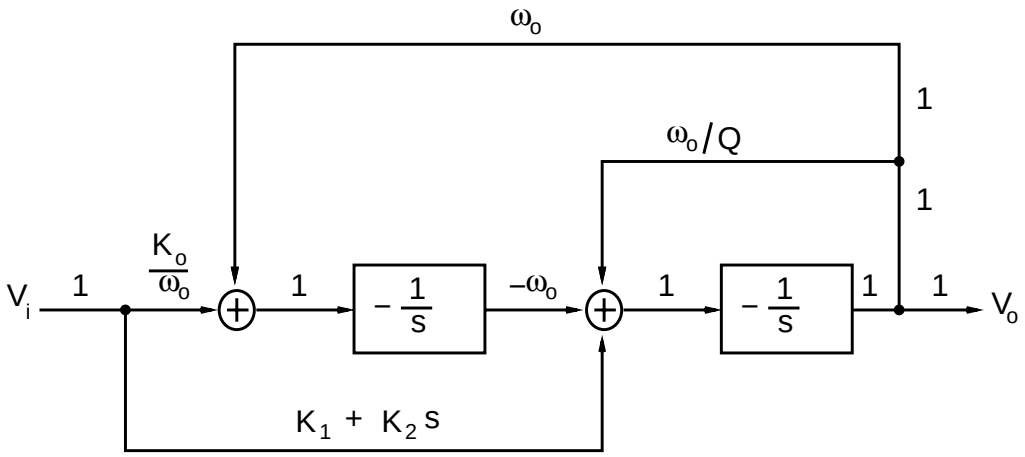


Figure 3.4 Flow diagram of low-Q biquad

tegrating capacitor value is chosen all other component values are scaled by the same factor.

One partition for the biquad transfer function is shown in figure 3.1. This partition of the transfer function is not unique. The exact partition chosen is determined by implementation considerations such as sensitivity to component variation, capacitance spread (C_{max}/C_{min}), and pole Q [31].

Using the equivalent RC integrators from figure 3.3, this flow diagram can be directly mapped to an RC equivalent implementation as shown in figure 3.5. Ideally, the transfer function of this circuit exactly matches equation 3.1. Real integrated circuit implementations, however, suffer from non-ideal components. While better than 1% component matching can be achieved on-chip, the absolute tolerance on resistors and capacitors is typically 10%. This leads to poor control over the actual placement of poles and zeros and errors in the frequency response. Furthermore, resistors are difficult to implement in integrated technologies. The area is usually large and the linearity is poor which causes signal distortion.

Alternatively, discrete-time, switched-capacitor integrators can be used as described in chapter 2. As a first approximation, the switched capacitor at the input of the integrator can be modeled as a continuous-time resistor. Consider the switched capacitor shown in figure 3.6. If the switches are operated on a two-phase clock as shown, a charge $\Delta q = C(V_1 - V_2)$ will flow through the capaci-

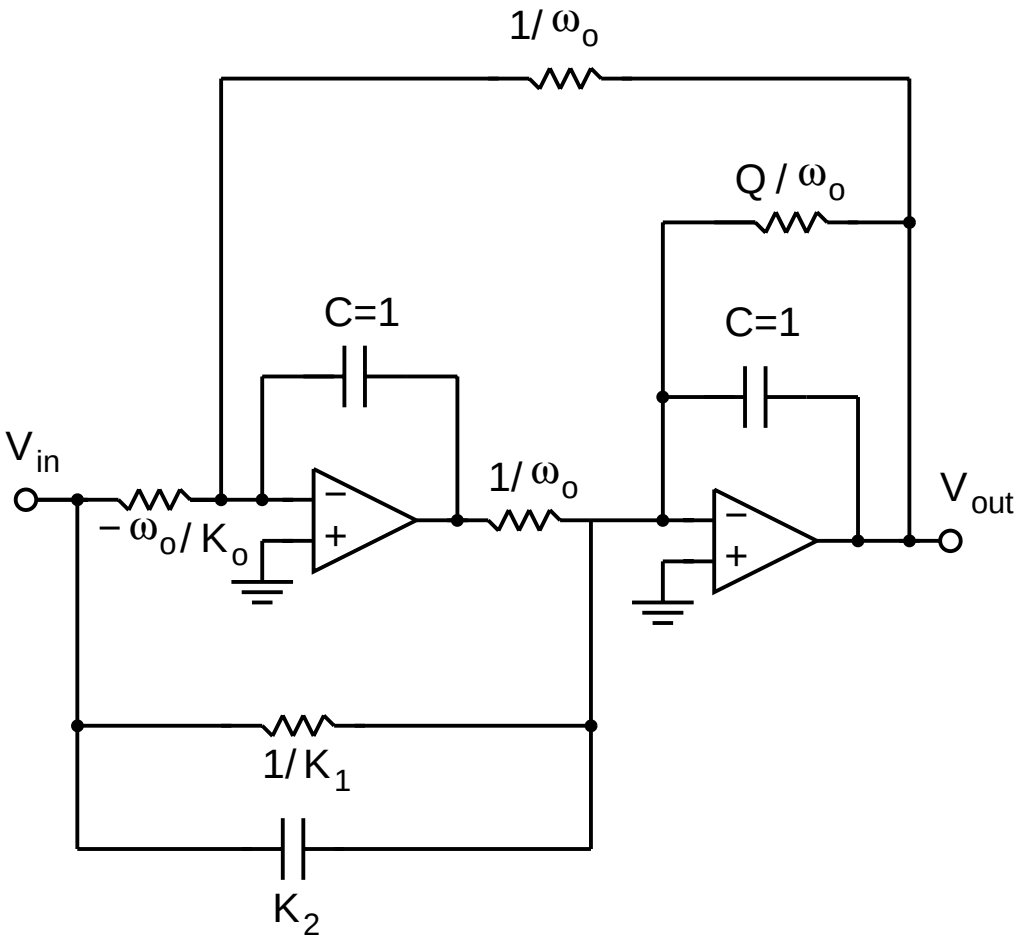


Figure 3.5 Active-RC equivalent of low-Q biquad

tor. If the clock period is T, then the average current will be

$$i_{av} = \frac{\Delta q}{T} \tag{3.5}$$

$$= \frac{V_1 - V_2}{T/C} \tag{3.6}$$

T/C can then be thought of as the average resistance. Thus, the resistor values in figure 3.5 can now be mapped directly to capacitance values, clocked at a given frequency $1/T$. Note that a negative value of resistance in figure 3.5 corresponds to using the clock phases shown in parentheses in figure 3.6. Intuitively, a switched capacitor will better approximate a continuous-time resistor if the switching frequency is large compared to the time constant of the circuit. Analytically, this approximation is equivalent to:

$$z \rightarrow e^{sT} \approx 1 + sT$$

This is the mapping of the discrete-time, z-domain to the continuous-time, s-domain. The approximation holds for $s = j\omega, |\omega| \ll 1/T$ as expected. The resulting biquad is shown in figure 3.7.

Finally, if a more accurate design is required, the exact z-domain transfer function of the circuit in figure 3.7 can be determined from the time-domain difference equations. The result is a biquadratic z-domain transfer function, which can be mapped to its continuous-time equivalent using the bilinear transformation [60] for example.

For higher-order filters, a cascade of biquadratic filters can be difficult to implement due to the high sensitivity to component values. The ladder filter topology overcomes many of these problems, and can be found in the references [61, 39, 30].

The performance of switched-capacitor filters is limited by the accuracy of the three basic components–switches, capacitors, and opamps.

The linearity of the input sampling switches is a significant contribution to the overall distortion. For MOS switches, the device conductance varies with input

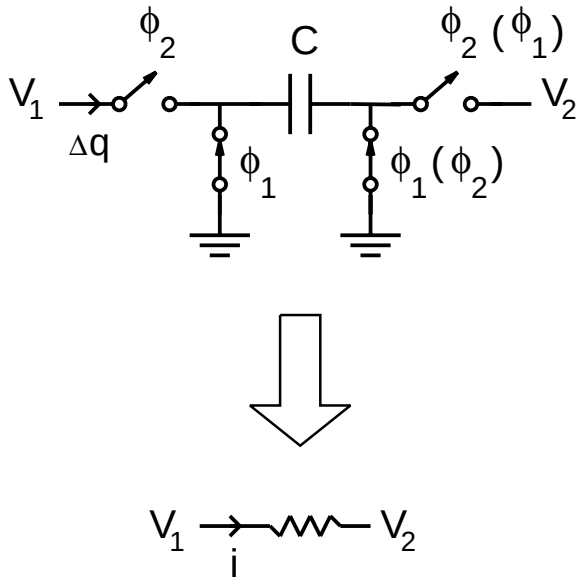


Figure 3.6 Switched capacitor

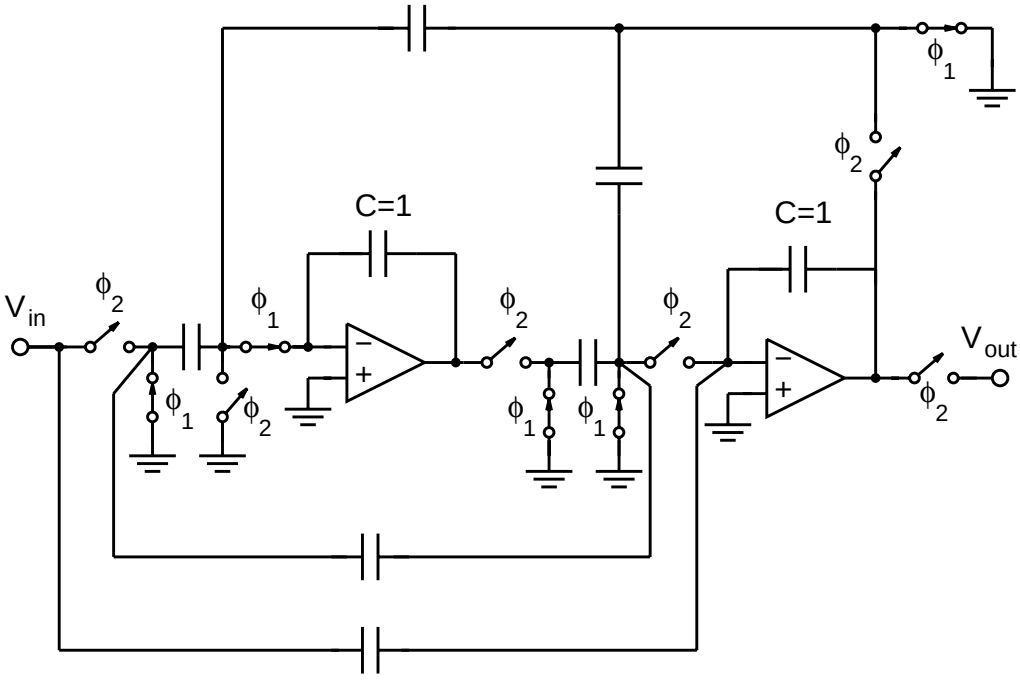


Figure 3.7 Switched-capacitor equivalent of low-Q biquad

Band edge	Dynamic range	Technology	Reference
20 MHz	76 dB	GaAs	[32]
10.7 MHz	68 dB	BiCMOS	[56]
10.7 MHz	42 dB	CMOS	[72]
6 kHz	92 dB	CMOS	[9]

Table 3.1 Switched-capacitor filter performance

signal level, creating a signal-dependent time constant in the sampling network. The distortion becomes more pronounced as the input signal bandwidth becomes comparable to the sampling network bandwidth.

If diffusion layers are used to implement the capacitors, the capacitance may have voltage-dependent characteristics. This effect also contributes to the overall distortion. If poly-poly or metal-metal capacitors are used, this effect is usually negligible.

The signal-to-noise ratio (SNR) is limited by kT/C thermal noise generated by the sampling switches and opamp transistor thermal noise. The capacitor values must be carefully chosen, so that the desired SNR can be achieved.

Finally, the accurate placement of poles and zeros in the frequency domain depends on the capacitor matching, opamp DC gain and settling time. Capacitor ratios determine the effective integrator time constants, so although absolute capacitance is not critical, relative matching is essential. 0.1% matching can typically be obtained in contemporary CMOS technologies. The finite DC gain of the opamp lowers the Q factor in the frequency response from its ideal value. The opamp settling time also limits the sampling rate of the overall filter. If the opamp cannot settle to sufficient accuracy in each clock period, the poles and zeros will shift in frequency domain of the overall filter response.

The performance of some reported switched-capacitor filters is shown in table 3.1.

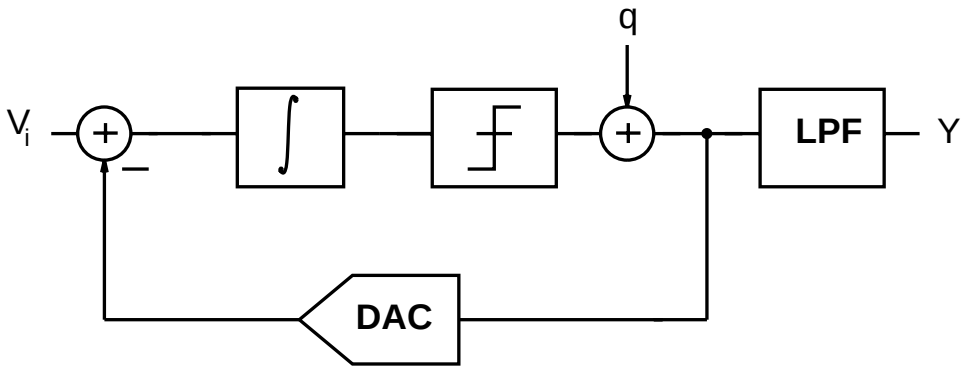


Figure 3.8 Sigma-delta analog-to-digital converter block diagram

3.2 Sigma-delta analog-to-digital converters

Another important switched-capacitor application is the sigma-delta analog-to-digital converter [5, 8]. Sigma-delta converters achieve high resolution by pushing the quantization noise to high frequency and removing it with a digital filter.

Figure 3.8 shows the block diagram of a first-order, sigma-delta ADC. Intuitively, the negative feedback loop causes the output the DAC to on average equal the input voltage V_i. Therefore the output bits of the quantizer are a rough, low-frequency representation of the analog input V_i. If the signal bandwidth is sufficiently smaller than the sampling frequency then the high frequency quantization noise can then be digitally separated and removed with a low-pass filter (LPF) to yield the final digital output signal Y.

Stated more analytically, the transfer function from V_i to Y is a pure delay. If the quantizer is modeled as additive white quantization noise, the converter can be treated as a linear system. Now the transfer function of the quantization noise q to the output Y is a high-pass. Thus the quantization noise at low-frequencies is suppressed as shown in figure 3.9. The remaining high-frequency quantization noise can then be removed with a the digital low-pass filter. The benefits increase as the sampling rate is increased relative to the input signal bandwidth. This ratio is called the oversampling ratio. The ideal output SNR increases 9dB per octave in oversampling ratio. Thus, sigma-delta converters can achieve very high resolution for small signal bandwidths, such as audio applications. By using a high

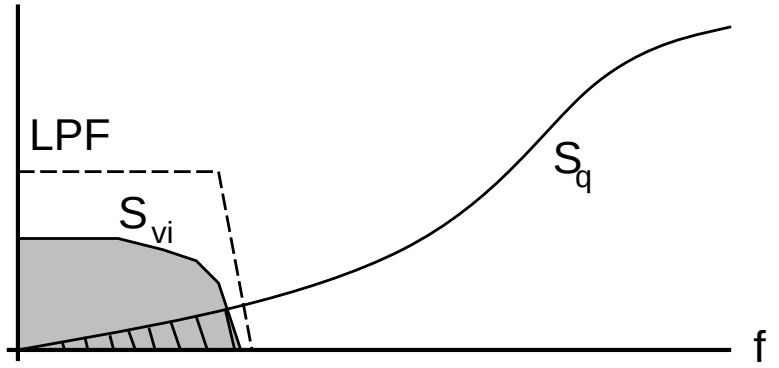


Figure 3.9 Power spectral density of sigma-delta modulator output

oversampling ratio, an inherently linear, single-bit quantizer can be used which is beneficial because it does not require any precision, matched components. Finally a high oversampling ratio is also an advantage because it relaxes the roll-off characteristics of the anti-alias filter preceding the ADC.

Furthermore, more integrators may be added to increase the noise-shaping characteristics of the system. This dramatically reduces the in-band quantization noise. Adding too many integrators, however, introduces loop stability problems.

Sigma-delta applications usually make heavy use of switched-capacitor circuits. Switched-capacitor integrators usually used in the forward path, and switched-capacitor comparators and DACs are also used in the loop. It is the settling time of the integrators, however, that typically limits the sampling rate of the system.

3.3 Pipeline Analog-to-digital converters

Pipeline analog-to-digital converters use a technique similar to digital circuit pipelining to trade latency for throughput [45, 15]. In a pipeline converter only a few bits are resolved at a time. This approach increases the throughput and reduces the required number of comparators compared to a flash or half-flash converter.

Figure 3.10 shows the block diagram of a pipeline ADC. The pipeline converter consists of N cascaded stages. Each stage samples an analog input and does a course B-bit quantization with a sub-ADC. Using a DAC, the quantization

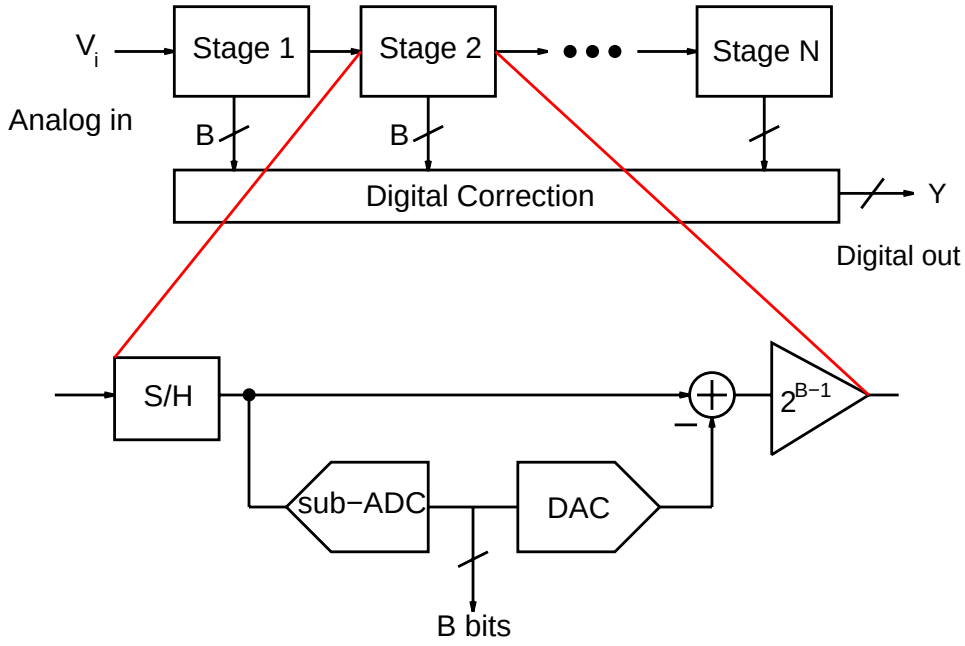


Figure 3.10 Block diagram of a pipelined ADC

error can then be determined by subtracting the quantized value from the analog input. This error is then amplified by a precision gain of 2^{B-1}. The resulting full-scale residue signal is further resolved by the remaining stages.

Each stage is typically implemented with switched-capacitor circuits. The sub-ADC can be constructed from switched-capacitor comparators, and the DAC output voltage can be generated capacitively as described in the following section. The sample-and-hold/gain stages from chapter 2 can be used to generate the output residue. Again, it is the settling time of the gain stage opamp that limits the converter throughput. Also, because a precision, inter-stage gain is required, the capacitors must match to the same accuracy as the desired resolution of the overall converter. There, however, is a large body of work devoted to the self-calibration of pipeline converters [49].

3.4 Capacitor digital-to-analog converters

Using the principle of charge division, an array of capacitors can be used to perform digital-to-analog conversion [66]. Unlike resistor-string DACs, a switched capacitor array does not consume any static power. Furthermore, the charge domain nature of the capacitor DAC complements switched-capacitor integrators in many applications such as sigma-delta modulators and pipeline ADCs.

Figure 3.11 shows one implementation of a capacitor DAC. During the reset phase, the all switches to ground are closed and the others are opened. The reset switch is first opened, then a digital codeword $d_N \cdots d_3 d_2 d_1$ $(d_k \epsilon [0,1])$ sets the switch positions. The output voltage will then be

$$V_o = \frac{\sum_{i=1}^{N} d_i C_i}{\sum_{i=1}^{N} C_i}$$

If the codeword is binary weighted then the capacitors C_k must also be binary weighted. If the DAC must drive a large capacitive or resistive load, a switched-capacitor buffer can be placed at the output. The will add static power consumption.

The linearity of this type of DAC is limited by the matching of the capacitors in the array. The rate at which it can clocked is limited by the RC settling time constants of the switching network.

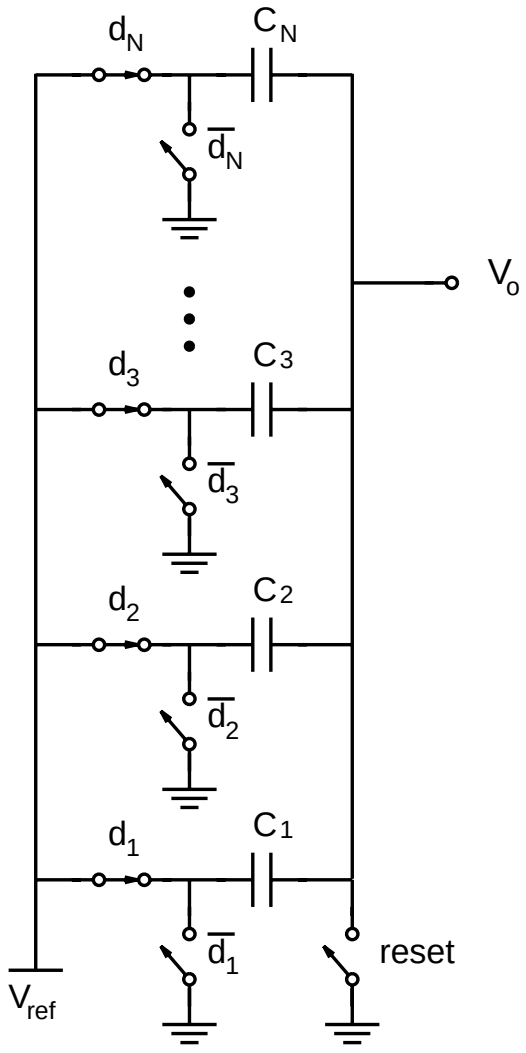


Figure 3.11 Switched-capacitor DAC

Chapter 4

CMOS Technology Scaling

THE miniaturization of complementary-metal-oxide-semiconductor (CMOS) technology has enabled dramatic increases in integrated circuit performance over the years. In the 1960's when integrated circuits came into production, the minimum feature sizes were greater than $10\mu m$. Contemporary fabrication facilities of the late 1990's have $0.35\mu m$ and smaller sizes in production. Improvements in fabrication technology have also greatly increased functional production yield to allow larger silicon dies to be economically manufactured. The result has been an increase in the number of transistors and functions per chip, circuit speed, and a reduction in power consumed per transistor. Early integrated circuits consisted of only a few transistors. Contemporary microprocessors contain several million transistors clocking at several hundred megahertz.

These advances in CMOS circuit performance have been enabled by the miniaturization of the MOSFET. Clearly, as the device becomes smaller and wafer yield increases, more functional devices can be fabricated on the same die. As device dimensions shrink, the parasitic capacitances also tend to decrease. Furthermore, the reduction in channel length and gate oxide increase the current density, achieving more current drive in the same area. The combined effect is a reduction in propagation delay allowing higher throughput and clock rates for digital circuits.

Although there are many potential problems, this trend of improved performance is expected to continue. Table 4.1 shows one outline of expected advances in CMOS technology over the next decade [70]. The rate of advances may slow, but it is thought that there are still many more performance gains that can be achieved.

This chapter is an overview of the goals of scaling CMOS technology and fo-

Product shipment	1999	2001	2003	2006	2009	2012
DRAM bits/chip	1.07G	1.7G	4.29G	17.2G	68.7G	275G
μP transistors/chip	21M	40M	76M	200M	520M	1.4B
DRAM area (mm^2)	400	445	560	790	1120	1580
MPU area (mm^2)	340	385	430	520	620	750
External clock (MHz)	1200	1400	1600	2000	2500	3000
Gate length (μm)	0.14	0.12	0.1	0.07	0.05	0.035
V_{dd} (V)	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6

Table 4.1 SIA Technology Roadmap 1997

cuses on the need for voltage supply scaling. It provides motivation for why low-voltage operation is an important issue for future, and this may affect analog, switched-capacitor circuits.

4.1 CMOS Scaling

The goals of CMOS scaling are to increase the speed and density of the transistors. This allows the design of faster circuits with more functionality or value for a given area of silicon. An increase in speed requires a higher current density per unit width of transistor. The load capacitance per unit transistor width has historically remained constant [37], therefore, an increase in current density enables faster charging and discharging of this capacitance and lower propagation delays. The increase in current density simultaneously translates to higher device density per unit area of silicon.

Examination of one model for the drain current of a saturated MOSFET, it can be seen that the critical dimensions for increasing current density are channel length L and oxide thickness t_{ox}. In equation 4.2, W is the device width, v_{sat} is the carrier saturation velocity, and $C_{ox} = \epsilon_{ox}/t_{ox}$ is the gate oxide capacitance per unit area.

$$I_{dsat} = W v_{sat} C_{ox} \frac{(V_{gs} - V_t)^2}{(V_{gs} - V_t) + \varepsilon_{sat} L} \tag{4.1}$$

$$= W v_{sat} C_{ox} (V_{gs} - V_t - V_{dsat}) \quad (4.2)$$

$$V_{dsat} \triangleq \frac{\varepsilon_{sat} L (V_{gs} - V_t)}{\varepsilon_{sat} L + (V_{gs} - V_t)} \quad (4.3)$$

It should be noted, however, that the benefits of channel length scaling diminish as the drain current becomes dominated by velocity saturation.

Furthermore, the device dimensions, such as oxide thickness t_{ox} and junction depth X_j, and doping N_A must also be scaled to combat short-channel effects, such as drain-induced barrier lowering (DIBL). Otherwise, the device will not behave like a transistor; namely a device whose current is controlled by the gate. Several scaling models [35, 6] express the minimum achievable channel length as a function of t_{ox}, X_j, W_D (depletion width) as shown in equations 4.4 and 4.5. These dimensions must all be scaled together.

$$L_{min} \propto t_{ox} \cdot X_j^{1/3} \quad (4.4)$$

$$L_{min} \propto t_{ox}^{1/3} \cdot X_j^{1/3} W_D^{2/3} \quad (4.5)$$

From a speed performance perspective, as high an operating voltage as possible is desired because this maximizes current density. However, the demand for low-power circuits, and the unavoidable issue of device reliability require that the voltage supply also be scaled down with device dimensions as shown in table 4.1. From this general trend, it is clear that circuits will need to operate at 1.5 V and below within a decade [70, 18]. With each new generation of CMOS technology, the applied voltages will need to be proportionately scaled to maintain power density and reliable operation.

4.2 Voltage scaling for low power

The popularity of portable applications, such as cellular phones, laptop computers, and digital assistants, has created a strong demand for low-power (long battery life) integrated circuits. The power in digital circuits is a strong function of the supply voltage.

$$P = CV_{dd}^2 f + I_{leak} V_{dd}$$

This equation illustrates that the power P has two components–dynamic switching energy and standby leakage current. The former has a quadratic dependence on the supply voltage (C is the load capacitance, and f is the average switching frequency). This energy can be substantially reduced by modest reductions in the supply voltage. The latter component becomes substantial for large chips with millions of transistors but is a weaker function of the supply voltage. This constraint also puts a lower bound on V_t scaling as discussed below in section 4.4.

For a technology optimized for low-power operation, the goal is to maintain constant power density across scaling. In other words, more functions can be performed for the same power, or less power with be consumed for the same number of functions. In this scenario, a lower supply voltage is chosen at a modest sacrifice in speed. Furthermore, circuit architectures incorporating parallelism have been developed to allow voltage scaling within a given technology without a loss in speed performance for certain applications [13]. The alternative scenario is to maximize performance at the expense of increasing power density. In this case, the power supply is chosen as large as possible within leakage and reliability limits (section 4.3).

To manage leakage currents, the voltage supply must be scaled. There are three major current leakage phenomenon in short channel MOS devices: drain-induced barrier lowering, punch-through, and gate-induced drain leakage. All contribute to current leakage from the drain when the device is in the off state. This leakage impacts the power for die with a large number of devices as discussed above. For DRAM circuits, it also impairs the ability to store charge for long periods of time. These leakage phenomenon place an upper-bound on the supply voltage.

For short channel devices, the electric field from the drain begins to have a significant influence on the channel charge. This phenomenon is called drain-induced barrier lowering (DIBL). The effective threshold voltage drops as a function of the drain-to-source voltage V_{ds} and the extent of the drain depletion region into the channel [37].

$$\Delta V_t \approx V_{ds} e^{-L/l_1} \tag{4.6}$$

$$l_1 \approx 0.1(X_j t_{ox} X_{dep}^2)^{1/3} \propto X_J^{1/3} t_{ox} \tag{4.7}$$

Equation 4.7 further emphasizes the need for a thin gate oxide to maintain strong control over the channel. DIBL further exacerbates the problem of limited V_t scalability (section 4.4).

Punch-through or sub-surface DIBL is another source of current leakage that occurs even when $V_{gs} = 0$. For sufficiently large V_{ds}, the drain and source depletion regions will begin to merge below the channel. This occurs because the increasing reverse bias on the drain enlarges its depletion width, and the bulk is less heavily doped below the surface due to the V_t implant. When the two depletion regions touch, the potential barrier for diffusion carriers is greatly reduced. This gives rise to a current component that is exponentially dependent on the drain-to-source voltage. Thus, to maintain an acceptable amount of leakage the magnitude of V_{ds} should be limited as as described by equation 4.8. where N_{sub} is the substrate doping concentration. The circuit designer can increase the punch-through voltage by increasing the channel length as shown in equation 4.8.

$$V_{ds} < V_p \propto \frac{N_{sub} L^3}{X_j + 3t_{ox}}. \tag{4.8}$$

The last major source of leakage is gate-induced drain leakage (GIDL). GIDL occurs when the gate is grounded (device cutoff) and the drain is at a high voltage. Thus, for an NMOS device, a depletion region forms in the n+ drain. If the drain voltage is high enough, an inversion layer of holes will begin to form. Because the gate overlaps both the n+ drain and the p- channel, however, the holes flow into the p- channel which is at a lower potential. It is theorized that the holes are not generated thermally, but by electrons tunneling across the oxide via band-to-band tunneling. This leakage current puts a lower bound on the off-state leakage in a device.

Equation 4.9 [37] shows the maximum allowed voltage across the gate and drain, where E_{gidl} is electric field in the oxide that induces tunneling (typically

4MV/cm), 1.2V is the bandgap voltage, and V_{FB} is the MOSFET flat-band voltage (approximately 0V for n+ poly over n+ drain and 1.1V for n+ poly over p+ drain).

$$V_{dg} < E_{gidl} \cdot t_{ox} + 1.2V - V_{FB}. \tag{4.9}$$

4.3 Voltage scaling for reliability

The voltage limitations of the technology dictate that in mixed-signal applications, the integrated analog circuits operate at the same low voltage as the digital circuitry. Furthermore, in achieving low-voltage operation, the reliability constraints of the technology must not be violated. Therefore it is important to thoroughly understand the reliability issues and how not to over stress CMOS devices.

Most critical are oxide breakdown and hot-electron effects which can cause CMOS circuit failure or a degradation in performance. A CMOS technology is typically designed such that these failure modes occur at a similar stress level which is an important factor in determining the rated supply voltage. Therefore, if the following device voltages are kept within the rated supply voltage, a long circuit lifetime can be assured with high confidence.

4.3.1 Gate oxide breakdown

A thin gate oxide is desirable because it increases the current density of the device and allows the gate to control the channel charge effectively. If other device dimensions are scaled without also scaling the gate oxide, the device experiences various short-channel effects such as drain-induced barrier lowering as discussed above. If, however, the oxide thickness is scaled and the applied voltage is not also proportionately scaled, then the electric field in the oxide increases. A sufficiently large electric field will cause the oxide to instantaneously break down or to break down over time (called time-dependent dielectric breakdown–TDDB).

Oxide failures have been experimentally found to occur in three distinct groups [82]. The first occurs under very weak electric fields typically less than 1MV/cm. It

is thought that these cases have gross defects, such as pin holes that immediately conduct. This is typically the statistically least common failure. The second group breaks down under moderate fields from 3-7MV/cm. This is the most common failure mode and the level of fields typical in integrated circuits. In this case, it is thought that some minor defects exist that assist the breakdown process. Potential defects could be contaminants in the oxide, surface roughness, or variations in oxide thickness. The last group occurs above 8MV/cm. This is called intrinsic or defect free breakdown.

The first group is relatively rare and can be screened out with burn-in testing. The second group with minor defects is the largest concern. Oxide breakdown can also occur at lower field strengths over a longer period of time. Thus, breakdown becomes a long-term reliability issue. As the total oxide area per die decreases, however, the probability of oxide defects also decreases. Thus, a small number of devices can be subjected to larger voltages without statistically reducing the overall circuit lifetime. For a small oxide area, the probability of being defect free is high and can approach intrinsic break down fields of 8MV/cm.

The exact physical mechanism that causes oxide breakdown is not fully known. There are, however, several plausible theories. Most theories agree that electrons tunnel into the oxide conduction band via Fowler-Nordheim tunneling as shown in the energy-band diagram in figure 4.1. This process involves electrons at equilibrium or “cold” electrons, unlike “hot” electron injection described below. Once these electrons reach the conduction band of the oxide, the electric field accelerates them toward the gate.

Some theories propose that these energetic electrons then generate electron-hole pairs [20, 41, 14]. A fraction of these “hot holes” then can be trapped in oxide traps. This charge locally increases the electric field and in turn increases the local tunneling current. If enough positive charge is accumulated, the tunneling barrier is reduced sufficiently to let current to freely flow and the oxide has been broken down.

An alternate theory proposes that the energetic electrons collide with the crystal lattice at the gate-SiO_2 interface and break Si-O bonds, forming defects [79].

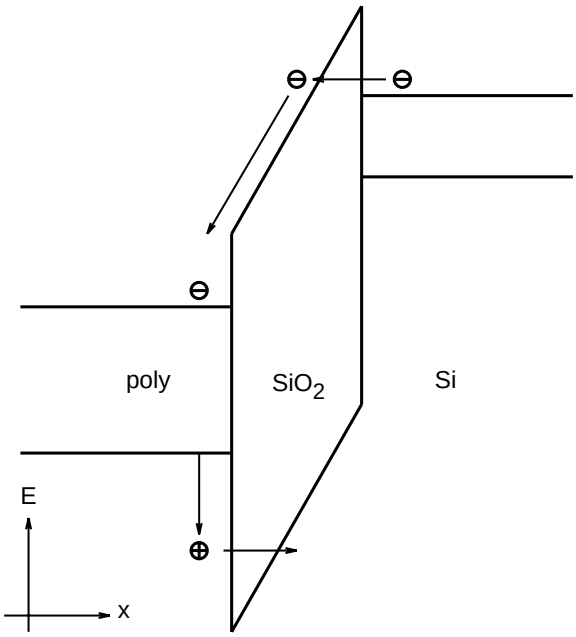


Figure 4.1 Fowler-Nordheim tunneling

The positively charged defects locally attract more electrons that deepen the damage into the oxide. Eventually a conductive path is formed through the oxide.

Both theories support the empirical findings that the time to oxide failure is a function of applied voltage, time duration, and defect density. One quantitative model for the lifetime t_{BD} of gate oxide relates these three parameters [54].

$$1 = \frac{1}{\tau_0} \int_0^{t_{BD}} \exp\left(-\frac{GX_{eff}}{V_{ox}(t)}\right) dt \tag{4.10}$$

τ_0 and G are constants, X_{eff} is the effective oxide thickness due to defects, and $V_{ox}(t)$ is the time-dependent voltage across the oxide. Equation 4.10 shows that it is the accumulation of stress over time that determines breakdown. Thus, short AC or transient stress is less harmful than DC stress. This equation can be simplified for the case when V_{ox} is constant. Equation 4.12 shows that under such DC stress, the lifetime of gate oxide is exponentially dependent on the field in the oxide. E_{bd} is the electric field in the oxide that causes breakdown. For circuit lifetime of 30 years, E_{bd} is typically 5MV/cm [55]. This value allows for sta-

tistical defects. For defect-free oxide (i.e. small area) the intrinsic breakdown is approximately 7-8MV/cm [35]. t_{ox} is the gate oxide thickness. As the oxide voltage V_{ox} increases beyond E_{bd}, the lifetime of the oxide decreases exponentially where $\tau_0(T) \approx 10^{-11}$s and $G(T) \approx 350$MV/cm for $T = 300$K.

$$V_{ox} < E_{bd} \cdot t_{ox} \tag{4.11}$$

$$t_{BD} = \tau_0(T) e^{G(T) t_{ox}/V_{ox}} \tag{4.12}$$

Thus, as oxide thickness t_{ox} is scaled to improve device performance, the voltage supply must also be scale to maintain the same device lifetime. For small sub-sections of circuits, such as the small, analog portion of a large, mixed-signal die, however, it may be possible to allow higher oxide fields because the probability of oxide defects is lower for a small area. Yield Y (probability of a defect in a given area) is an exponential function of defect density D and area A [44] as shown in equation 4.13. Thus, oxide lifetime is a function of the magnitude of the voltage across the oxide, the duty cycle of voltage stress, and the area under stress.

$$Y = e^{-DA} \tag{4.13}$$

4.3.2 Hot-electron effects

As the channel length of the device becomes shorter, if the drain-to-source voltage is not proportionately reduced, the electric field along the channel increases. If the peak electric field is sufficiently large, electrons in the channel are accelerated to the point where they cause damage that degrades device performance over time or causes instantaneous breakdown.

Figure 4.2 illustrates how energetic "hot" electrons are generated in an NMOS device. The device gate G must be biased above threshold, such that a conductive channel of electrons exists. Then when a large V_{DS} is applied, 1. electrons are accelerated by the large lateral electric field. 2. These electrons strike the crystal

lattice in the drain depletion region and create electron-hole pairs through impact ionization. 3. If the generated electron has more than approximately 1.5eV it can tunnel into the oxide. This accumulation of trapped charge will lower saturation current I_{dsat}, cause V_t drift, lower the linear region transconductance, and degrade the sub-threshold slope S_t. 4. The generated holes will either be collected by the substrate or the source. In both cases, resistive drops in the substrate tend to forward bias the source-bulk junction. 5. This causes minority carrier electrons to be injected into the bulk via lateral BJT action. These carriers then cause more impact ionization, which closes a positive feedback loop. If a sufficient number of electrons are injected, the device will experience snapback breakdown or latch-up rendering it in operable and large current will flow into the substrate.

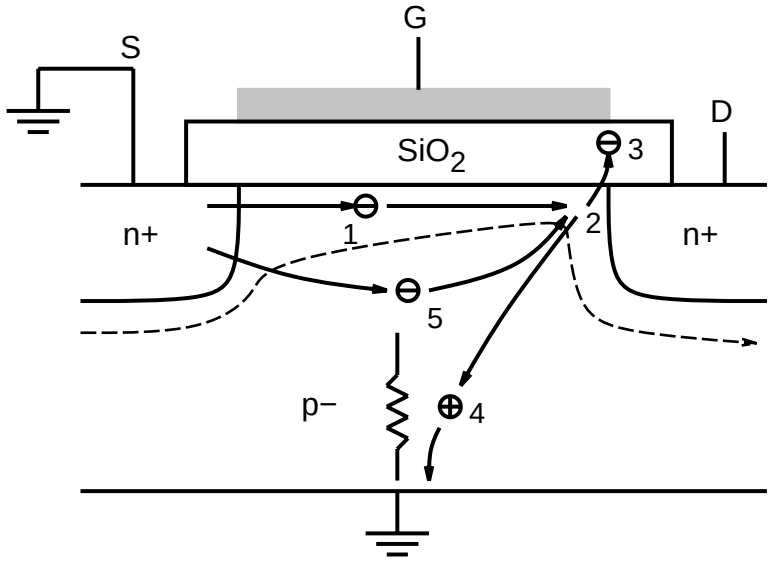


Figure 4.2 Hot electron effects

Hot electron generation is a strong function of the maximum lateral electric field. This value can be analytically modeled by equation 4.14 [12, 42]. This model shows that hot-electron stress is worst for large V_{DS} with short channel length and low $(V_{GS} - V_t)$. l_{LDD} is the effective length of the lightly doped region. l_{LDD} can be increased to allow larger applied voltages. The increased series

resistance, however, degrades the device's current carrying capability. Therefore, the trade-off between reliability and speed, implies there is an optimal choice of V_{dd} that maximizes speed given an maximum allowed ε_{max}.

$$\varepsilon_{max} = \frac{(V_{DS} - V_{DSAT})}{l} \tag{4.14}$$

$$l \approx 0.22 t_{ox}^{1/3} X_j^{1/3} + l_{LDD} \tag{4.15}$$

$$V_{DSAT} \triangleq \frac{\varepsilon_{sat} L (V_{GS} - V_t)}{\varepsilon_{sat} L + (V_{GS} - V_t)} \tag{4.16}$$

Similar to oxide stress, low levels of hot-electron stress can accumulate over time to lead to time-dependent degradation. DC stress is the most damaging. AC stress can be de-rated as a function of the duty cycle. Empirical simulations show, for a typical inverter, the combination of V_g and V_d gives an effective hot-carrier stress given by equations 4.17, 4.18. T is the period, t_r is the rise time of the output, and t_f is the fall time of the output.

$$\text{NMOS: } \frac{\text{AC lifetime}}{\text{DC lifetime}} = \frac{4T}{t_r} \tag{4.17}$$

$$\text{PMOS: } \frac{\text{AC lifetime}}{\text{DC lifetime}} = \frac{10T}{t_f} \tag{4.18}$$

Hot carrier protection lightly doped drain (LDD) structures permit higher voltage operation, but add series drain resistance which degrades speed. Therefore, an optimal $V_d d$ exists that maximizes speed for a given a fixed reliability level.

Thus, the choice of power supply voltage for a CMOS technology is a trade-off of speed, power, and reliability. For speed, generally the highest voltage allowed by leakage and reliability considerations is desired. A large supply voltage maximizes the drain saturation current, which allows faster dis/charging of capacitive loads. Oxide TDDB and hot-electron damage, however, set the voltage upper-bound. For the latter, LDD design can extend the allowable voltage before series

resistance begins to defeat the speed benefits. If low-power operation is desired, the voltage supply is scaled more aggressively. This has the benefit of greatly reducing dynamic switching energy CV_{dd}^2, reducing voltage dependent leakage currents, and allowing more aggressive scaling of the device dimensions that improve performance (oxide thickness, channel length). The trade-off is a reduction in speed gains with each technology generation. Most likely, the historical trend of doubling circuit speed every two technology generations will probably slow.

4.4 Fundamental scaling limits

Certainly many technological advances in semiconductor manufacturing need to be made to realize the circuit performance shown above in table 4.1 and beyond. Moreover, there are fundamental limits to the scaling of CMOS devices as they are presently used and implemented. It is unclear if solutions can be found to these problems.

For oxide thin oxides less than 6nm the phenomenon of direct tunneling begins to occur. The rate of tunneling is larger than predicted by Fowler-Nordheim tunneling. For these thin oxides there is significant probability of an electron tunneling directly across the bandgap of SiO_2 without entering the oxide conduction band. For oxides thinner than 3nm, the rate of tunneling reaches a critical point where the charges cannot be replenished by equilibrium thermal generation. Thus, it becomes difficult to form the inversion layer necessary for transistor operation. This limit may be the lower bound of oxide thicknesses.

Scaling the voltage supply also requires scaling the threshold voltage in order to increase the drain saturation current. Otherwise, scaling the device does not result in an increase in speed performance. There are fundamental limitations, however, to how small V_t can be made. A non-zero V_t allows the sub-threshold, off-state, drain leakage current to be sufficiently small. The amount of leakage current that exists when the gate is grounded depends on the sub-threshold swing S_t of the device. S_t is defined as the reduction in V_{gs} that results in a 10x decrease in the sub-threshold current. The fundamental lower bound of S_t is $2.3kT/q$ or about 60mV at room temperature. Imperfections in a typical device cause S_t to

be 80-100mV. An approximate model for the sub-threshold current at zero gate bias is given by equation 4.4[64].

$$I_{leak} = 10\frac{\mu\text{A}}{\mu\text{m}} W 10^{-V_t/95\text{mV}}$$

As an example a VLSI circuit with $10^8\mu$m of transistor width will have 100mA of leakage current with a threshold voltage of 0.38V. Thus, threshold voltages lower than 0.3V for VLSI are problematic. This problem is exacerbated by short-channel roll-off due to DIBL and punch-through for short channel lengths.

The value of V_t also becomes more difficult to control for smaller devices due to the statistical variation of the number of dopant atoms in channel region. Even if absolute V_t variation remains constant, the variation as a fraction of the gate bias $V_{gs} - V_t$ increases as V_{dd} is scaled down. It has been shown that the statistical variation of digital circuit propagation delay due to this effect increases dramatically below 1V for a sub-0.5μm CMOS process [74]. This problem makes it difficult to achieve reasonable performance yield. V_t variation between devices on the same die also presents a matching problem for precision analog circuits.

4.5 Analog circuit integration

It is clear that for reasons of low power and reliability, the operating voltage for CMOS technology will be greatly reduced in the future. For digital circuits, it is clear there will be an increase in performance through device scaling. For mixed-signal applications, with integrated analog circuitry, the reduction in supply voltage presents new challenges (chapter 5). Analog designers are accustomed to having 3V to 5V to work with. With supplies of 1.5V and lower in the near future, new circuit techniques will need to be developed. The alternative is to add special processing, such as thick oxides and multiple threshold voltages to allow analog sections to run at higher voltages. Such special processing, however, detracts from the cost benefits of mixed-signal integration and would preferably be avoided. It is always desirable to be as technology-independent as possible for maximum flexibility and lowest cost. In the worst case, separate technologies

could be used in a multi-chip module, or separate packages. This lower integration solution, however, would greatly increase the overall cost of manufacture and testing.

Most technologies are designed with a single voltage V_{dd} in mind. Clearly if all no absolute node voltage exceeds V_{dd} reliable operation can be assured with high confidence. This absolute restriction, however, may be overly conservative. Consider the mechanisms of device degradation discussed above–DIBL, punch-through, GIDL, TDDB, and CHE. In steady-state, it is only the *relative* terminal voltages V_{gs}, V_{gd}, and V_{ds} that are critical as shown in figure 4.3. If these critical terminal voltages are kept within the rated operating voltage V_{dd} of the technology, the relevant electric fields in the device (which are defined by relative potentials) will not over-stress the device. Long-term, reliable operation can still be maintained.

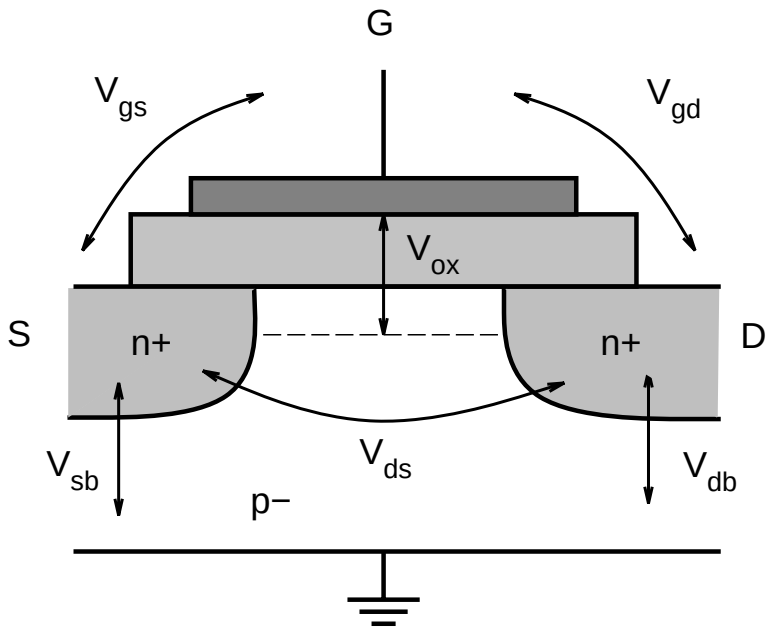


Figure 4.3 Relative potential determines reliability

The fact that reliability is determined by relative potential and not referenced to an absolute voltage such as ground, can be exploited by certain analog circuits. For example, the absolute V_g referenced to ground may exceed the rated

V_{dd} if $V_{gs} < V_{dd}$ is maintained. This fact has been exploited in implementing the low-voltage MOS switch described in chapter 5. Care must be taken, however, that the source-to-substrate and drain-to-substrate junctions do not exceed reverse breakdown voltages. These voltages are referenced to absolute ground (assuming a grounded substrate). This reverse breakdown, however, is typically much larger than the supply because the substrate is doped much less than the drain and source diffusions.

Chapter 5

Low-voltage Circuit Design

THE RELIABILITY constraints of scaled CMOS technology require low-voltage operation as discussed in chapter 4. Furthermore, chapter 4 discussed the types of device stress and modes of operation that are allowable. These new constraints impact how circuits are implemented for analog, switched-capacitor applications. This chapter discusses the impact of operating at low voltage and specific implementations of low-voltage, switched-capacitor building blocks.

5.1 Low-voltage, switched-capacitor design issues

The reduction in supply voltage introduces several factors that complicate the design of low-voltage, analog circuits. This discussion focuses specifically on CMOS switched-capacitor circuits.

As the supply voltage is scaled down, the voltage available to represent the signal is reduced; therefore dynamic range becomes an important issue. In order to maintain the same dynamic range on a lower supply voltage, the thermal noise in the circuit must also be proportionately reduced. There, however, exists a trade-off between noise and power consumption. Because of this strong trade-off, it will be shown that under certain conditions, the power consumption will actually increase as the supply voltage is decreased.

Consider the typical switched-capacitor circuit shown in figure 5.1. It consists of a class A, operational transconductance amplifier (OTA) operating from a supply voltage V_{dd} with a static bias current I. The OTA is configured with capacitive negative feedback and drives a fixed capacitive load. For high-resolution applications, the dynamic range of such a circuit is often limited by thermal noise. If the circuit is then optimized for minimum power, it can be shown that the power

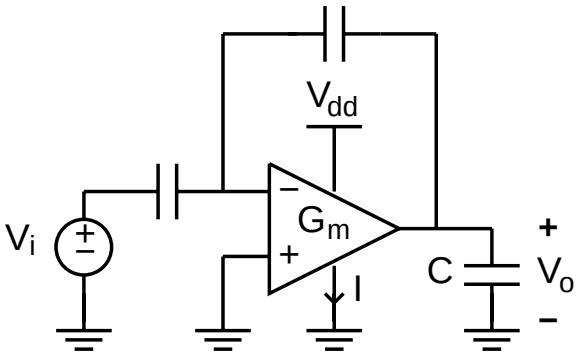


Figure 5.1 Typical class A, switched-capacitor circuit

will tend to increase as the supply voltage is lowered [10]. This result can be derived from a few simplifying assumptions. First, the power in the circuit is the static bias current times the voltage supply.

$$P \propto I \cdot V_{dd} \tag{5.1}$$

Second, if the OTA can be modeled as a single transistor, then the bias current is proportional to the transconductance times the gate over-drive. There exists an optimal value of $(V_{gs} - V_t)$ that minimizes the power consumption. If this $(V_{gs} - V_t)$ is too small, then the f_T (unity current gain frequency) of the device will be too small. In that case, either it is infeasible to meet the specified speed, or the device must be prohibitively large to make the load insignificant compared to the intrinsic device parasitics. Such a large device consumes excessive power. If the value of $(V_{gs} - V_t)$ is too large, then the I/g_m ratio is too large and the device consumes more power than is necessary to meet the desired speed (i.e. it is over-designed). Thus, for minimum power, there exists an optimal value $V_{gt}(opt)$ that is a function of the technology and load.

$$I \propto g_m \cdot (V_{gs} - V_t) = g_m \cdot V_{gt}(opt) \tag{5.2}$$

Third, the closed loop bandwidth of the circuit must be high enough to achieve the desired settling accuracy at the given sampling rate f_s. For simplicity, the loading of capacitive feedback network is not included in the bandwidth expression. In

fact, there exists an optimum for minimum power that makes the feedback loading a certain fraction of the output load C.

$$\frac{g_m}{C} \propto f_s \tag{5.3}$$

Finally, the dynamic range (DR) is proportional to the signal swing, αV_{dd} (where $0 < \alpha < 1$), squared over the sampled kT/C thermal noise. α represents what fraction of the available voltage supply is being utilized.

$$\text{DR} \propto (\alpha V_{dd})^2 / \frac{kT}{C} \tag{5.4}$$

From these assumptions, it follows that for a given technology, dynamic range, and sampling rate, the power will be inversely proportional to the supply voltage V_{dd} (eq. 5.5). Furthermore, the power is inversely proportional to the square of fractional signal swing α. Although the supply voltage is a fixed constraint of the technology which cannot be modified, the circuit designer can choose α. Therefore, to minimize power consumption, it is important to use circuits that maximize the available signal swing, α. As the device technology improves, (shorter channel length) the power-optimal value of V_{gt} will decrease which will tend to mitigate the trend of an increase in power consumption. For sufficiently small V_{gt}, however, the I/g_m ratio will approach a constant due to sub-threshold conduction. Therefore, power consumption will most likely be flat or tend to increase.

$$\Rightarrow P \propto kT \cdot \text{DR} \cdot \left(\frac{V_{gt}(opt)}{\alpha^2 V_{dd}} \right) \cdot f_s \tag{5.5}$$

If the settling time of the circuit is dominated by slew-rate, the above analysis can be modified by substituting equation 5.3 with equation 5.6 below.

$$\frac{I}{C(\alpha V_{dd})} \propto f_s \tag{5.6}$$

This modifies the final result, such that the power is independent of the power supply voltage, but still inversely proportional to the fractional signal swing α.

$$P \propto \frac{kT \cdot \mathrm{DR} \cdot f_s}{\alpha} \tag{5.7}$$

Threshold voltage variation also becomes a larger problem at low voltage. As device geometries become smaller and the threshold V_t is scaled, the variation in device V_t becomes more difficult to control. As discussed in chapter 4, this creates matching problems in precision analog circuits. Precision devices may be required to use larger than minimum geometries and therefore fail to benefit from device scaling. Another solution is to use offset cancellation techniques or offset-insensitive architectures, such as the low bits-per-stage pipeline ADC.

Another critical problem in designing switched-capacitor circuits on a low-voltage supply is the difficulty of implementing MOS switches. Typically in a switched-capacitor circuit an analog input signal, V_i, is sampled through a MOS switch or transmission gate as shown in figure 5.2 (see also chapter 2). Ideally the switch in the on-state acts as a fixed linear conductance g_{ds}. In practice the conductance of the switch varies with the signal voltage as shown in the right half of figure 5.2. Plotted in the figure is the switch conductance versus input signal V_i for three different supply voltages. The dashed line shows the individual conductances of the NMOS and PMOS devices, and the solid line shows the effective parallel conductance. In the top case, V_{dd} is much larger than the sum of the two threshold voltages, V_{tn} and V_{tp}. In this case, it easy to achieve a large on-conductance from rail to rail for V_i. In the middle case, V_{dd} is comparable to the sum of the threshold voltages, and there is a substantial drop in conductance when V_i approaches $V_{dd}/2$. Finally, in the bottom case where V_{dd} is less than the sum of the two threshold voltages, there is a large range of V_i for which the switch will not conduct. Previous work [15, 80] has addressed this problem by using voltage boosting circuits that subject devices to large terminal voltages. This technique, however, introduces potential long-term reliability oxide problems. Section 5.2 introduces an alternate approach to this problem.

Finally, operational amplifiers become more difficult to implement at low voltage. Section 5.3 describes a folded-cascode, common-source cascade that achieves

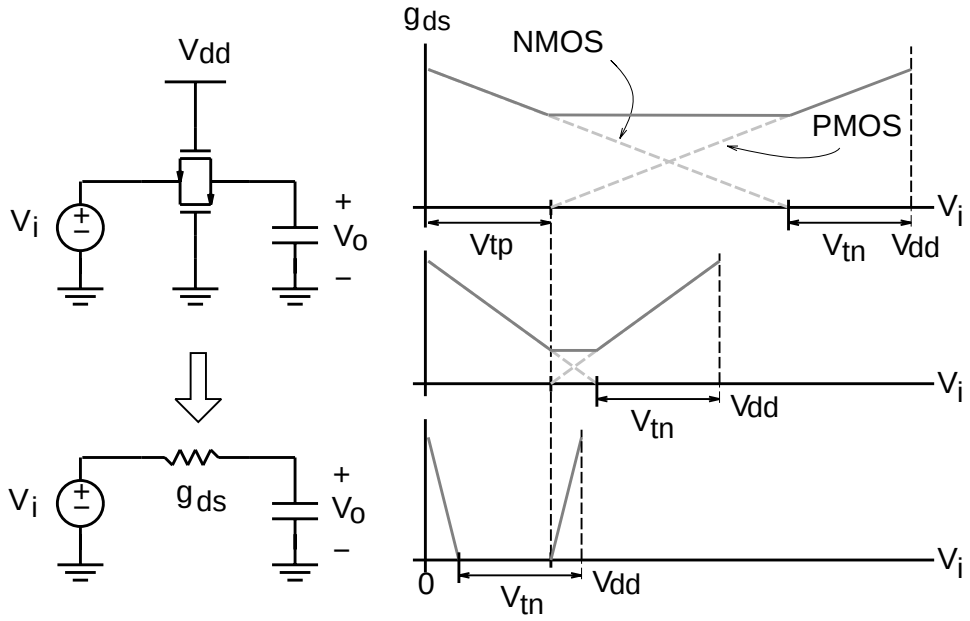


Figure 5.2 Conductance of MOS switches

gain and bandwidth sufficient for moderate resolution, video-rate applications.

The following sections describe low-voltage implementations of these critical switched-capacitor building blocks.

5.2 Reliable, high-swing MOS switch

As previously discussed, transmission gates cannot be directly realized on a supply voltage below the sum of the two threshold voltages. Therefore, an alternate approach is required. Earlier bootstrap implementations [15] resulted in (relative terminal) voltage stress exceeding the supply by a large margin. Another approach has been to use switched-opamps that can be turned on and off [17, 4]. The clock rates of these circuits has been limited well below video rates. In this work, a bootstrapped switch was designed to observe device reliability considerations.

5.2.1 Operation

This switch is conceptually a single NMOS transistor as shown in figure 5.3. In the "off" state, the gate is grounded and the device is cutoff. In the "on" state, a constant voltage of V_{dd} is applied across the gate to source terminals, and a low on-resistance is established from drain to source independent of the input signal. Although the absolute voltage applied to the gate may exceed V_{dd} for a positive input signal, none of the relative terminal-to-terminal device voltages exceed V_{dd}. In chapter 4 it was shown that this will not degrade the lifetime of the device.

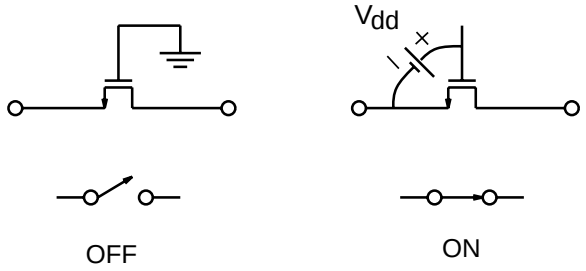


Figure 5.3 Bootstrapped MOS switch

The switch operates on a two phase clock as shown in figure 5.4. During the "off" phase on the left, the switch is turned off by grounding the gate. Simultaneously, the capacitor, which acts as the battery, is charged to the supply voltage. During the "on" phase the capacitor is then switched across the gate and source terminals of the switching device.

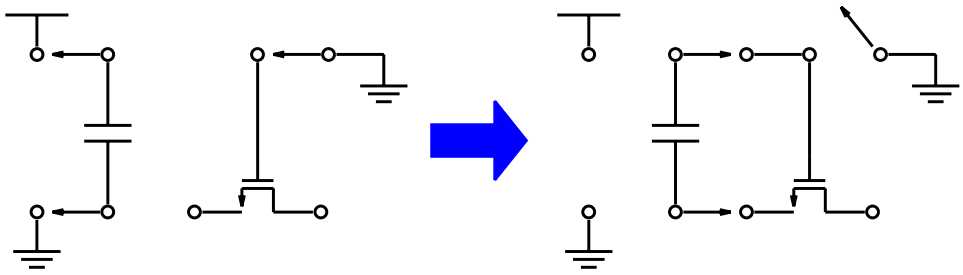


Figure 5.4 Operation of bootstrap switch

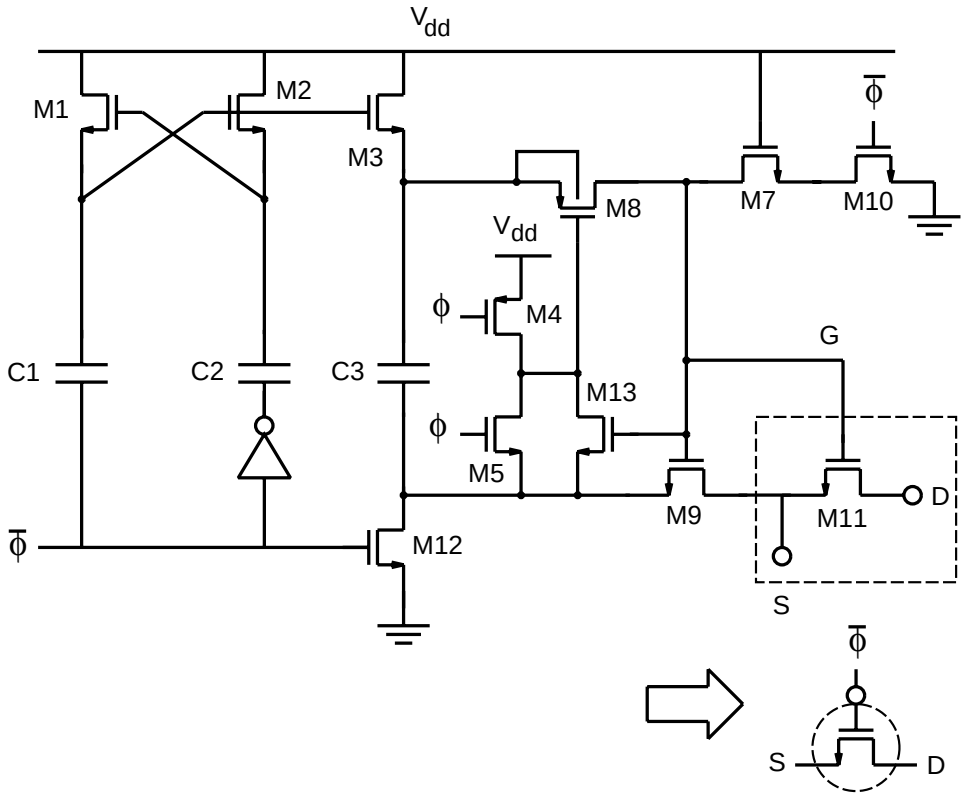


Figure 5.5 Bootstrap circuit and switching device

Figure 5.5 shows the actual bootstrap circuit. It operates on a single phase clock ϕ that turns the switch M11 on and off. During the off phase, ϕ is low. Devices M7 and M10 discharge the gate of M11 to ground. At the same time, V_{dd} is applied across capacitor C3 by M3 and M12. This capacitor will act as the battery across the gate and source during the on phase. M8 and M9 isolate the switch from C3 while it is charging. When ϕ goes high, M5 pulls down the gate of M8, allowing charge from the battery capacitor C3 to flow onto the gate G. This turns on both M9 and M11. M9 enables the gate G to track the input voltage S shifted by V_{dd}, keeping the gate-source voltage constant regardless of the input signal. For example, if the source S is at V_{dd}, then gate G is at $2V_{dd}$, however, $V_{gs} = V_{dd}$. Because the body (nwell) of M8 is tied its the source, latch-up is suppressed. Note that node S is best driven by a low-impedance due to the added

capacitance at this node. This circuit will be represented later by the symbol at the bottom the figure. Node S is indicated by the arrow, and the bubble on the gate indicates the inverted clock input.

Two devices in figure 5.5 are not functionally necessary but improve the circuit reliability. Device M7 reduces the V_{ds} and V_{gd} experienced by device M10 when $\phi = 0$. The channel length of M7 can be increased to further improve its punch-through voltage. Device M13 ensures that V_{gs8} does not exceed V_{dd}.

M1, M2, C1, and C2 form a clock multiplier [15] that enables M3 to unidirectionally charge C3 during the off phase. This entire circuit was carefully designed such that no device experiences a relative terminal voltage greater than V_{dd}. This circuit is similar to a previous low-distortion sampling switch approaches [28, 19, 59, 68, 7] that provide a constant V_{gs} across the switching device. In this case, however, there is the added constraint of device reliability.

Figure 5.6 shows the conceptual output waveforms of the bootstrap circuit. When the switch is on, its gate voltage, V_g, is greater than the analog input signal, V_i, by a fixed difference of V_{dd}. This ensures the switch is operated in a manner consistent with the reliability constraints. Because the switch V_{gs} is relatively independent of the signal, rail-to-rail signals can be used which is important in minimizing power consumption as discussed in section 5.1. Furthermore, the switch linearity is also improved and signal-dependent charge injection is reduced. Variations in on-resistance due to body effect, however, are not eliminated.

5.2.2 Design guidelines

Although it is difficult to derive an analytic set of design equations for this circuit, some general design guidelines are given. The capacitor values should be chosen as small as possible for area considerations but large enough to sufficiently charge the load to the desired voltage levels. The device sizes should be chosen to create sufficiently fast rise and fall times at the load. The load consists of the gate capacitance of the switching device M11 and any parasitic capacitance due to the interconnect between the bootstrap circuit and the switching device M11. Therefore, it is desirable to minimize the distance between the bootstrap circuit

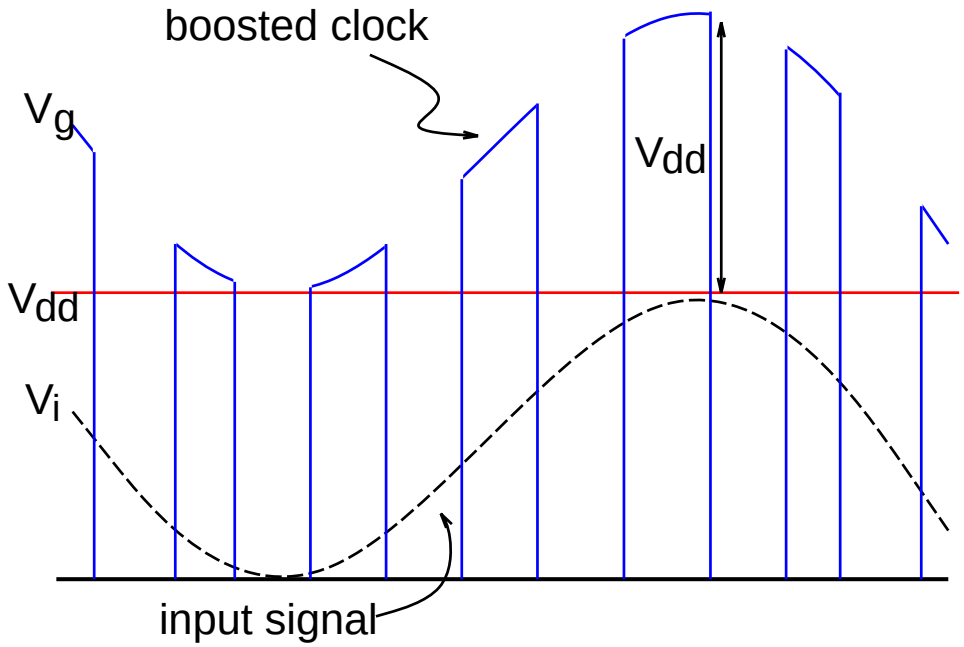


Figure 5.6 Conceptual bootstrap circuit output

and the switch in the layout.

Once the load is known, the other device sizes can be chosen. First, C3 must be sufficiently large to supply charge to the gate of the switching device in addition to all parasitic capacitances in the charging path. Otherwise, charge-sharing will significantly reduce the boosted voltage according to equation 5.8, where C_p is the total parasitic capacitance connected to the top plate of C3 while it is across the main switching device M11.

$$V_g = V_i + \frac{C3}{C3 + C_p} V_{dd} \tag{5.8}$$

Several bootstrap circuits were designed for the pipeline ADC prototype described in chapter 7. In this design, typical values of C3 were 0.5 pF to 1.8 pF which was approximately six times C_p. These capacitors were implemented with poly over n-diffusion layers with approximately 2 fF/μm^2.

M3 and M12 need to be large enough to charge C3 to V_{dd} during the reset phase. Because C3 is not fully discharged during the "on" phase, however, the voltage across C3 does not change too much. C1 then needs to be chosen large enough so that the boosted voltage at the gate of M3 is sufficient to turn M3 on (approximately $2V_{dd}$). Similarly C2 needs to be large enough to boost the gate of M1 to turn it on. These device sizes do not directly affect the rise and fall time at the load, however.

M8 and M9 are critical to the rise time of the voltage at the load. The (W/L) ratio should be increased until the rise time begins to decrease due to self-loading. M7 and M10 alone are responsible for the fall time and should be sized appropriately.

Due to the auxiliary devices, significant parasitic capacitance is introduced at node "S" particularly if there is a large bottom-plate parasitic associated with floating capacitor C3. Therefore, it is best to drive this node with low-impedance output. This, circuit, however, can still be used on the transfer switch of an integrator which sees a high impedance. If a fully-differential integrator is used, the

charge sharing caused by the parasitic capacitance creates only a common-mode error. If there is a mismatch in the parasitic capacitances of the two differential halves, a DC offset is added to the integrator, which is indistinguishable from an opamp offset. The DC offset is proportional to the amount of mismatch in the two halves. Certain DC offsets in switched-capacitor filters can corrected at the system level [51, 81].

5.2.3 Layout considerations

The reliability of this circuit can be further improved by carefully laying out some of the critical devices. Although the relative voltages between gate, source, and drain do not exceed V_{dd}, the drain-to-substrate and source-to-substrate voltages of some devices exceeds V_{dd} (assuming an nwell process). Devices M1, M2, M3, and M7 are subjected to this large voltage. Typically a CMOS technology is designed such that the reverse breakdown of a stand alone n+/p- junction is approximately $3V_{dd}$ [36]. This voltage called BV_{dss} is tested under the condition that the gate and source are grounded. In a NMOS transistor, however, a n+/p+ junction is formed between the n+ drain (or source) and the p+ channel-stop implant. The break-down voltage is further reduced when it is under a thin oxide as shown in figure 5.7. The right-hand side shows a cross-section of the transistor through plane AB. The break-down voltage in this region is typically designed to $1.7V_{dd}$.

Using a circular drain or "race track" layout, the p+ channel stop can be removed around the drain to add another 1-2V to the break-down voltage. Figure 5.8 shows the layout of a circular drain device.

The lightly-doped drain region can also be extended to further increase the drain breakdown voltage. By using the PMOS photo-resist mask already present in most processes, this LDD region can be extended into the drain as shown in figure 5.9. This step will increase the series drain resistance but will also increase the drain break-down 1-2V. By combining the circular drain layout and the extended LDD, BV_{dss} can typically be increased 2-4V [36].

Thus, for improved reliability, the sources of devices M1, M2, M3 and the drain of M7 should be laid out circularly with extended LDD regions. Finally,

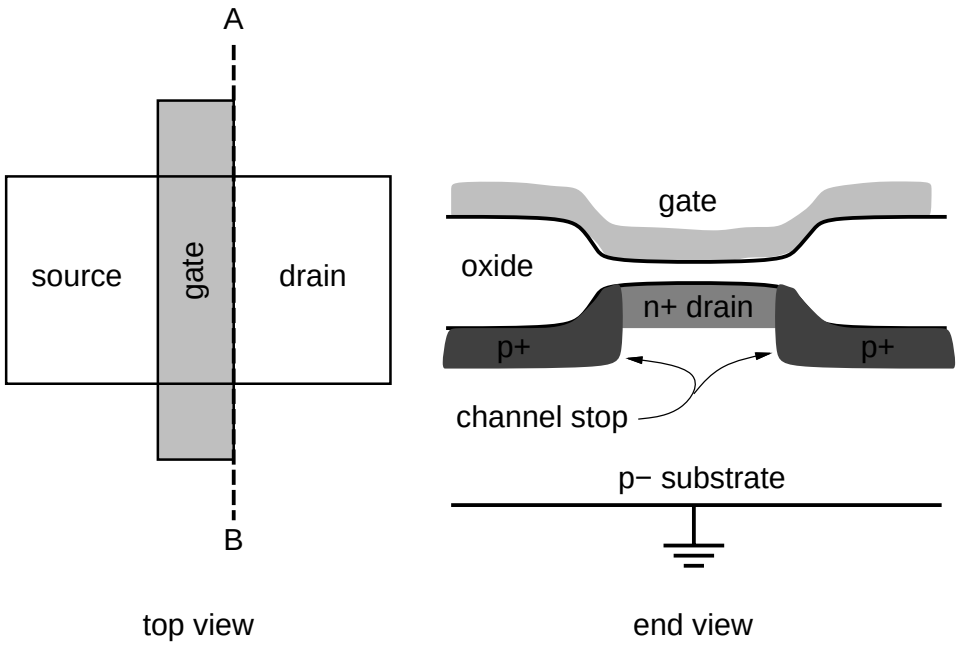


Figure 5.7 Weakest region of drain reverse break-down

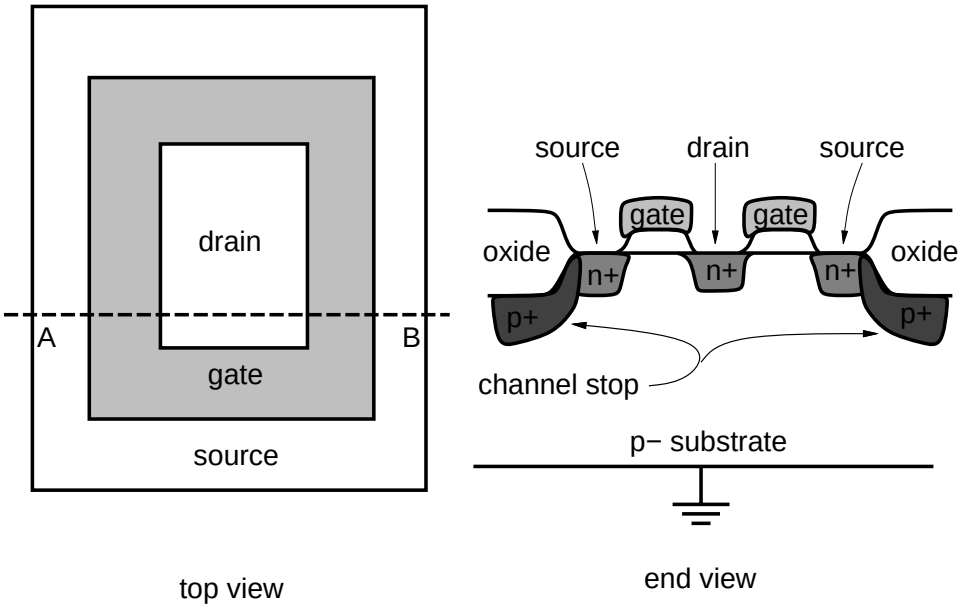
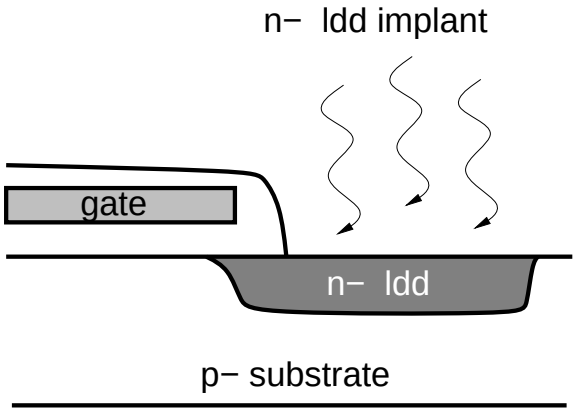
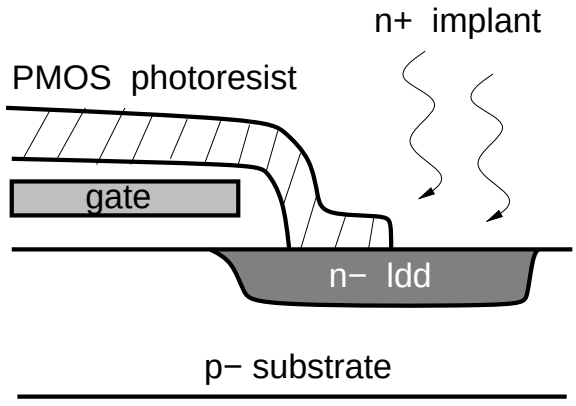


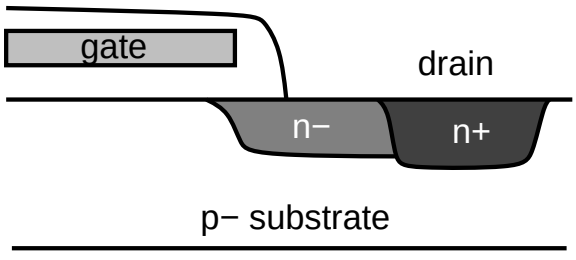
Figure 5.8 Circular drain layout



(a) lightly doped drain implant



(b) heavily doped drain implant



(c) improved break-down drain structure

Figure 5.9 Processing steps for improved break-down drain

the "off" drain-to-source voltage of M7 can exceed V_{dd} introducing a potential punch-through problem. If, however, the channel length of this device is increased (typically 1.5xL_{min}) this punch-through voltage can be significantly beyond the supply voltage.

One potential transient reliability problem exists for this circuit. If the rise time of the voltage at the gate is too fast, a large voltage could exist across the oxide of the switching device before a channel is formed to equalize the potential between source and drain. Consider the case where the switching device's source is driven by a low-impedance voltage V_{dd} and the drain is attached to a large capacitor discharged to ground. As the switching device turns on, a voltage of approximately $2V_{dd}$ will be generated on the gate. Before a channel is formed and the capacitor is charged to V_{dd}, an excessive voltage greater than V_{dd} may exist across the gate to drain terminals. This effect could create an oxide reliability problem. One solution would be to reduce the rise time by decreasing the W/L of M9 and or M8. It should also be noted that the lifetime of gate oxide is roughly inversely proportional to the voltage duty cycle [54]. Therefore, transient stress is less harmful than DC stress. Furthermore, this type of stress was not seen in the simulations of the ADC prototype described in chapter 7. A more thorough investigation of these transient effects using a reliability simulator such as the Berkeley Reliability Tool (BERT) [38] would be useful.

5.3 Opamp

Opamps become more difficult to properly bias on a low supply voltage. It is the opamp biasing that often limits the minimum supply voltage for switched-capacitor circuits. Figure 5.10 shows the practical minimum supply voltage for a class A, CMOS amplifier. At a minimum, the supply must be able to support a common source device driven by another common source stage. Therefore, the minimum supply voltage is $V_t + 2V_{dsat}$. This value can be thought of as a benchmark for low-voltage opamps. A good low-voltage opamp candidate will approach this value.

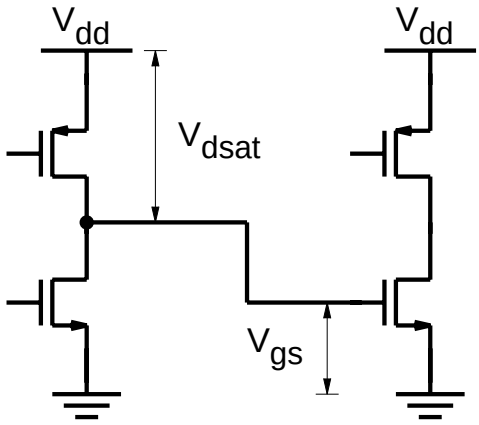


Figure 5.10 Minimum practical supply voltage

$$V_{dd} < V_{gs} + V_{dsat} \quad (5.9)$$

$$< V_t + 2V_{dsat} \quad (5.10)$$

5.3.1 Application

In switched-capacitor applications, opamps are configured either as integrators or gain-stages as described in chapter 2. This discussion will focus on design for gain-stages. Recall that a gain-stage operates on a two-phase clock. These two phases are shown in figure 5.3.1. During phase 1 the opamp is inactive and auxiliary tasks such as common-mode feedback reset or input offset cancellation can be performed. During phase 2 the opamp is actively amplifying a signal and generating a low-impedance output to drive other blocks. Although this discussion describes a single-ended version, it is completely applicable to the fully differential case.

Typically, the capacitors are determined first by system considerations, such as kT/C thermal noise, and the opamp is designed around this load. The gain, settling time, noise, offset required of the opamp are also determined by overall system specifications as discussed in chapter 6. Once these specifications are determined, an opamp topology can be chosen.

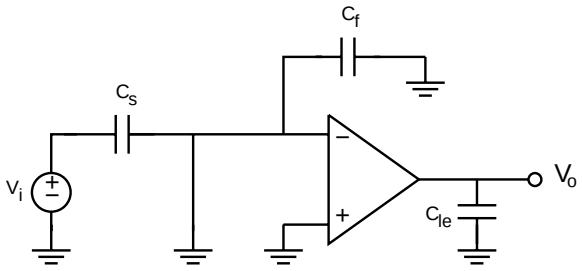


Figure 5.3.1a: Phase 1

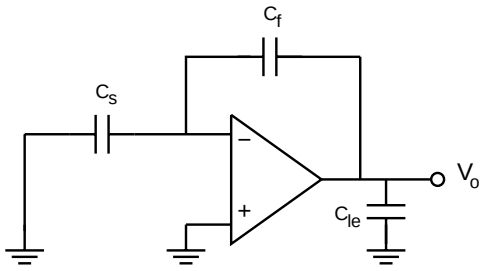


Figure 5.3.1b: Phase 2

5.3.2 Topology

The best topology for an opamp is highly dependent on the desired performance. As an example of a low-voltage opamp applicable to video-rate, moderate resolution, switch-capacitor applications, the design of a specific opamp is described. This opamp was used in the pipeline ADC prototype described in chapter 7. For this application, the opamp DC gain must be greater than 60 dB, settle to 0.1% accuracy in less than one-half clock cycle (35 ns), and operate on a 1.5V supply.

Figure 5.11 shows the topology chosen for this opamp design. This two-stage, fully differential amplifier consists of a folded-cascode first stage followed by a common-source second stage. The common-source second stage increases the DC gain by an order of magnitude and maximizes the output signal swing for a given voltage supply. As discussed in section 5.1, this is important in reducing the power consumption. Cascode compensation [2, 67, 62] was used to improve the bandwidth over conventional Miller compensation. A high bandwidth was necessary to achieve fast linear settling. Switched-capacitor common-mode feedback was employed to stabilize the common-mode output voltage. The following sections outline a design method for this opamp topology once the specifications of settling time, gain, noise, and offset are given.

5.3.3 Biasing

The first step in the design was to choose a set of feasible bias voltages for all the devices. This step is particularly important in the case of low-voltage design. In this case the limitations of the opamp bias circuit determined several of the

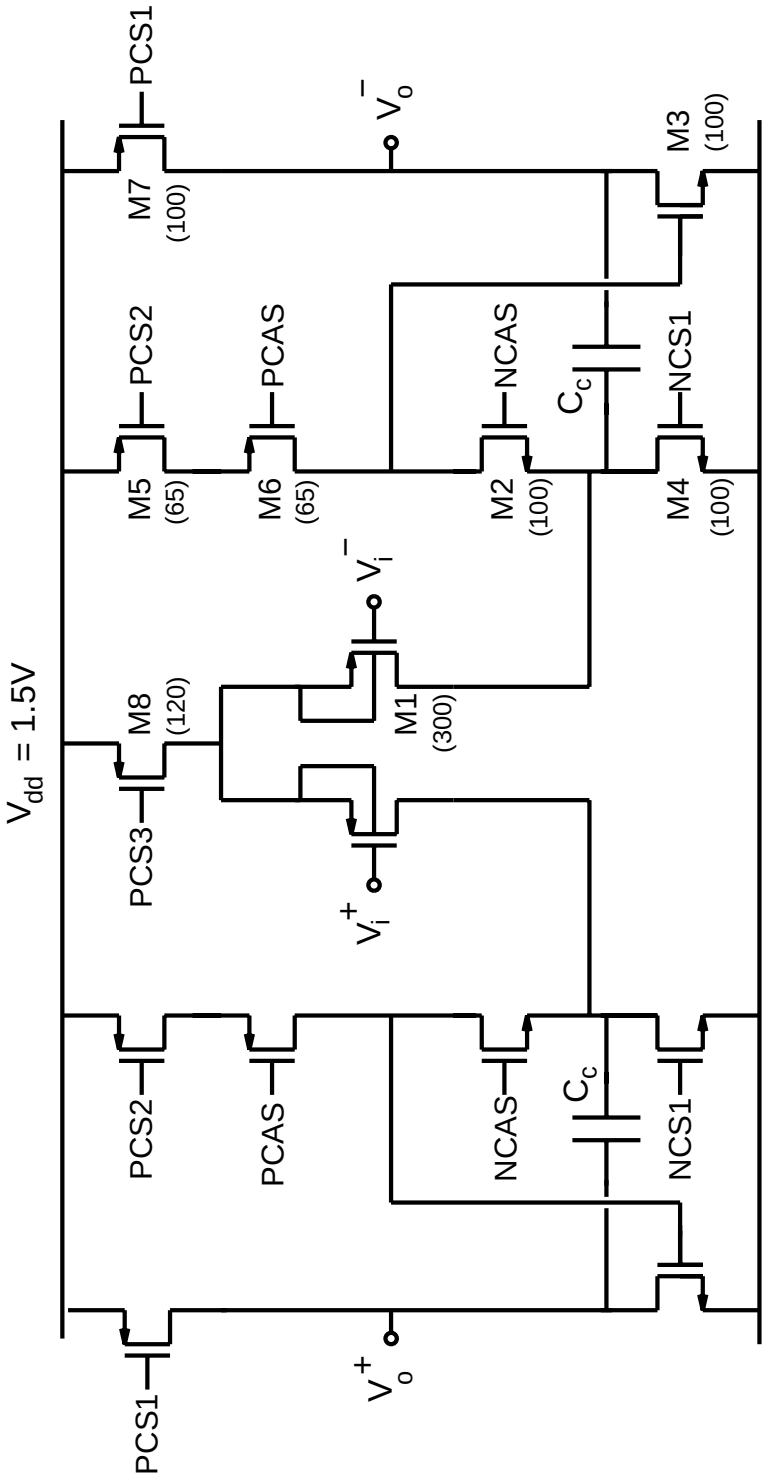


Figure 5.11 Low-voltage opamp

voltages. This circuit is shown in figure 5.12. To bias common-source devices, such as M3 (fig. 5.11), simple single transistor current mirrors were used, such as M11 (fig 5.12). For common-source devices whose drain is connected to the source of a cascode device, such as M4, M5, M8, a configuration such as M12 and M13 is used to generate the gate bias voltage. M12 is essentially a diode-connected device except that M13 mimics the cascode device (M6) so that the drain-to-source voltage of in-circuit device (M5) and the bias device (M12) are approximately equal. To first order this eliminates current mismatch due to finite output impedance (Early effect). To bias the cascode devices, M2, M6, a stack of devices driven by a current source was used as such as M14 and M15. The PMOS version is described; the NMOS version is similar. This scheme provides a high-swing cascode bias [43, 49]. M14 operates in the triode region, M15 is a diode-connected device operating in the saturation region, and M19 acts as a current source. M14 is sized to create a V_{ds} sufficient to keep M5 and M6 in the saturation region. M15 and M13 have the same current density as M6, therefore $V_{gs15} = V_{gs13} = V_{gs6}$. Thus, the minimum operating supply voltage is:

$$V_{dd} > V_{dsat5} + V_{gs15} + V_{dsat19} \tag{5.11}$$

$$> 3V_{dsat} + V_t \tag{5.12}$$

This particular path sets the minimum supply voltage for the opamp. Due to the limited operating voltage of 1.5 V, several devices were biased at moderate to weak inversion. In this design, the worst case threshold voltages were $V_{tn} \leq 650$ mV and $V_{tp} \leq 900$ mV. Furthermore to ensure that all devices are operating in saturation under all conditions, we require $V_{ds} \geq V_{dsat} + 200$ mV. At the output of the opamp a single-ended swing of 800mV was chosen. Therefore, M3 and M7 must still be in saturation when the output swings within 350 mV of the supply rails. From these specifications, the $(V_{gs} - V_t)$ values of all devices were chosen as shown in parentheses in millivolts. The current levels in the bias circuits were chosen such that each voltage is generated with a reasonably low output impedance less than 2kΩ. This allows the bias levels to be quickly restored after

a perturbance in the bias voltages.

Once a viable set of bias voltages has been chosen, the current density of each device (I/W) is fixed. This reduces the number of independent design parameters that need to be chosen. The bias current I, transconductance g_m, and parasitic capacitance C_{gs}, C_{gd}, C_{db}, C_{sb}, are all dependent functions of the device width W. Conversely any of these variables determines the other. By fixing the $(V_{gs} - V_t)$ some design flexibility is lost, however, in the low-voltage case, there is little flexibility in any case. If necessary, these voltages can be adjusted later in the design process.

Some care should be taken to note that V_t mismatch between the bias devices and the slave devices causes a larger current error for lower values of $V_{gs} - V_t$. It is the fractional ratio of the threshold error over the gate over-drive that determines current error.

$$\frac{\Delta I_D}{I_D} = \frac{\Delta(W/L)}{(W/L)} - 2\frac{\Delta V_t}{V_{GS} - V_t}$$

5.3.4 Linear Settling time

Unlike continuous-time applications where a particular frequency response is desired, settling time is the relevant specification for discrete-time applications. Because the settling time of the opamp limits the clock frequency of switched-capacitor circuits, this is often a critical specification.

Design equations for the linear settling time of this circuit can be derived from a small-signal analysis. Typically in feedback opamp design, an open-loop approach is used to design for a target phase margin. In a two-pole system, this phase margin directly translates to a unique, time-domain step-response and small-signal, settling time. This is the case when traditional Miller compensation is used [29]. Using cascode compensation, however, creates an inherently higher bandwidth, three-pole system. In this case, there is no direct relationship between phase margin, step-response, and settling time. Therefore, traditional open-loop analysis does not lead to design equations for the settling time. There is a significant trade-off in ease of design and bandwidth when choosing between the

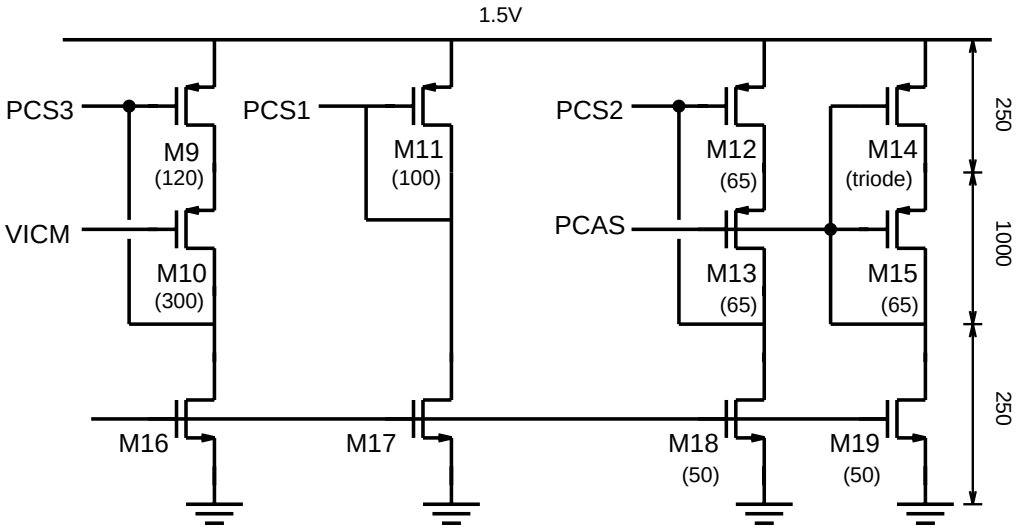


Figure 5.12 Opamp PMOS bias circuit

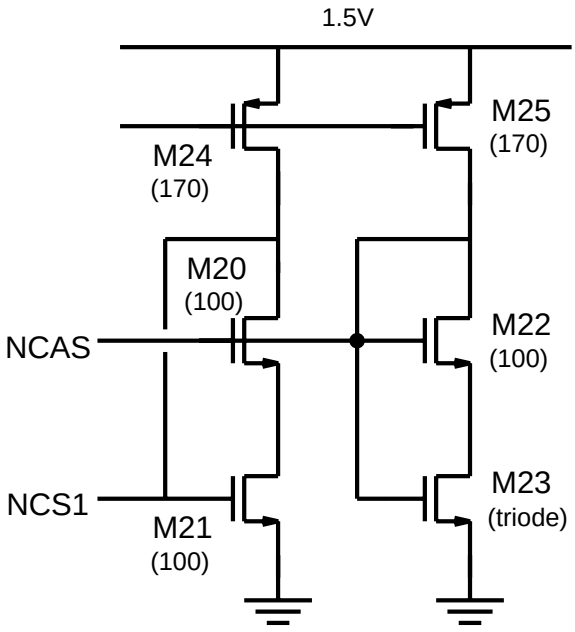


Figure 5.13 Opamp NMOS bias circuit

two compensation schemes. Previous work [2, 67, 62] has shown some open-loop design equations for this type of opamp, but little analysis has been done on closed-loop settling time. Therefore, an alternate closed-loop approach was taken here.

The design approach taken is similar to previous work [16, 25]. Unfortunately, the author has not seen a satisfactory set of simplified design equations for this compensation topology. The following is a somewhat complex, but functional design method. First a closed-loop, small-signal analysis is used to derive a relationship between the closed-loop poles (and zeros) and the small-signal parameters of devices in the signal path. Second, the desired pole positions are chosen that achieve the desired settling time with the desired stability. The result is a non-linear system of equations with the three signal-path device sizes and the compensation capacitance as free variables. Using the computer program MATLAB [52], a constrained numerical optimization was performed on this system of equations with opamp power as the minimized cost function. Final verification and tuning of the design was done using HSPICE. Unfortunately, this approach does little to lend intuitive insight in the design process.

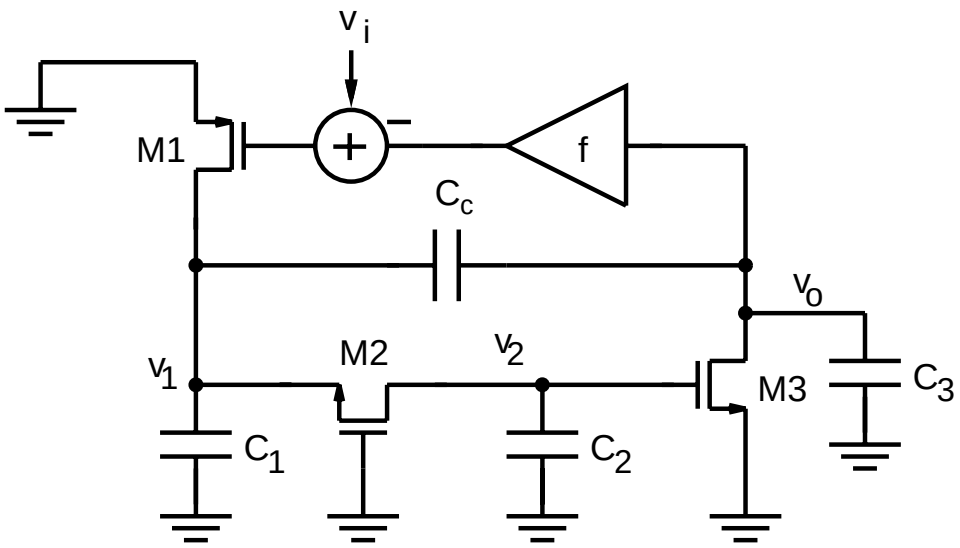


Figure 5.14 Small-signal equivalent circuit of gain-stage

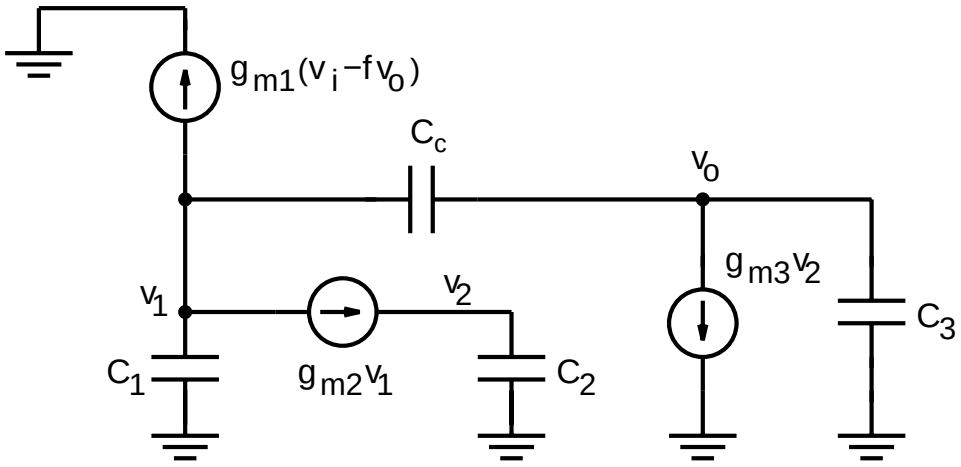


Figure 5.15 Small-signal equivalent circuit of a gain-stage

Figure 5.14 shows the small-signal equivalent circuit of the closed-loop gain-stage (fig. 5.3.1b) including external load, sampling, and feedback capacitors. In this circuit, the capacitive feedback has been lumped into a single feedback factor

$$f = \frac{C_f}{C_f + C_s + C_{ip}},$$

where C_{ip} is the input capacitance of the opamp. This circuit can be further simplified by replacing each transistor with its small-signal equivalent transconductance generator (fig. 5.15). The parasitic and external capacitances have been lumped into three capacitors C_1, C_2, C_3. As an approximation, the output impedance of each transistor is assumed to be infinite. This has the effect of pushing low-frequency poles to DC and slightly increasing the bandwidth of high-frequency poles [16]. The resulting simplified small-signal equivalent is shown in figure 5.15. From here, a straightforward nodal analysis yields the following small-signal equations:

$$0 = g_{m1}(v_i - fv_o) + sC_1v_1 + sC_c(v_1 - v_o) + g_{m2}v_1 \quad (5.13)$$

$$0 = -g_{m2}v_1 + sC_2v_2 \quad (5.14)$$

$$0 = sC_c(v_o - v_1) + g_{m3}v_2 + sC_3v_o \quad (5.15)$$

This linear system of equations can be solved to yield the closed loop transfer function.

$$A(s) = \frac{v_o(s)}{v_i(s)} = \frac{1}{f} \frac{1 - s^2 \frac{C_c C_2}{g_{m2} g_{m3}}}{1 + s \frac{C_c}{f g_{m1}} + s^2 \frac{C_2}{g_{m3}} \left(\frac{C_c + C_3}{f g_{m1}} - \frac{C_c}{g_{m2}} \right) + s^3 \frac{C_2 (C_c C_1 + C_c C_3 + C_1 C_3)}{f g_{m1} g_{m2} g_{m3}}}$$

By inspection the transfer function has two zeros and three poles of the form in equation 5.3.4, which is shown graphically in figure 5.16

$$A(s) = \frac{1}{f} \frac{(1 + s/z_o)(1 - s/z_o)}{(1 + s/\omega_{cl})(1 + 2\zeta s/\omega_n + s^2/\omega_n^2)}$$

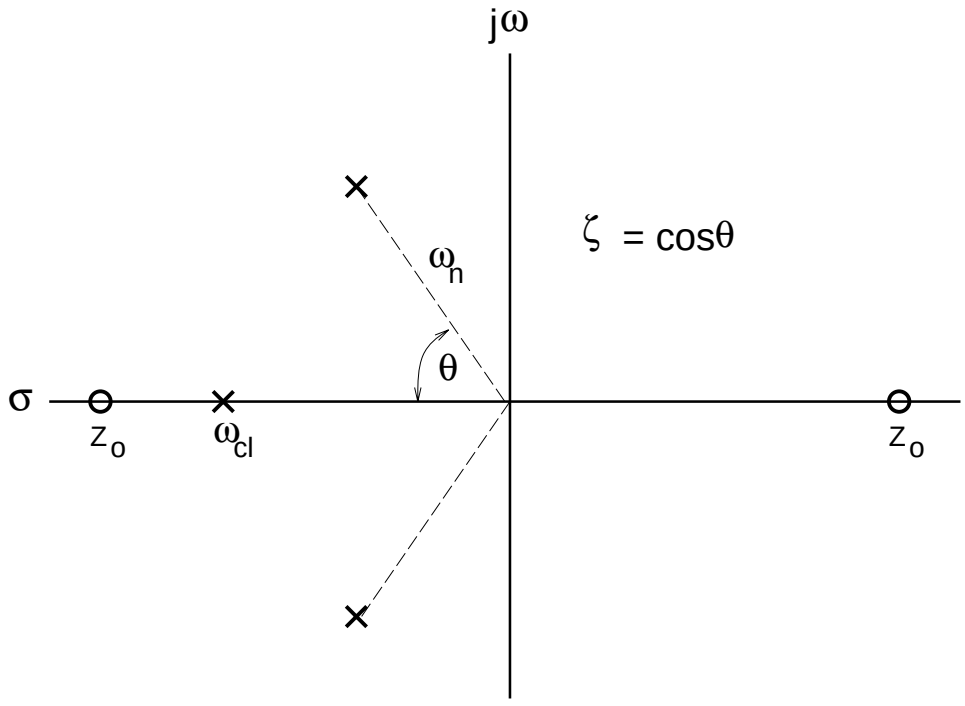


Figure 5.16 Closed-loop poles and zeroes

By equating coefficients of s, the following system of non-linear equations relates the circuit small signal parameters to the placement of the closed-loop poles and zeroes [16].

$$\pm\sqrt{\frac{g_{m2}g_{m3}}{C_cC_2}} = z_o \tag{5.16}$$

$$\frac{fg_{m1}}{C_c} = \frac{\omega_{cl}\omega_n}{2\zeta\omega_{cl}+\omega_n} \tag{5.17}$$

$$\frac{fg_{m1}g_{m2}g_{m3}}{C_2(g_{m2}(C_c+C_3)-fg_{m1}C_c)} = \frac{\omega_{cl}\omega_n^2}{\omega_{cl}+2\zeta\omega_n} \tag{5.18}$$

$$\frac{fg_{m1}g_{m2}g_{m3}}{C_2(C_cC_1+C_cC_3+C_1C_3)} = \omega_{cl}\omega_n^2 \tag{5.19}$$

Values for ω_n, ζ, ω_{cl} must now be chosen. The pole parameters uniquely determine the small-signal step-response as derived in [25]. The goal is choose a set of poles that minimize the settling time for a given bandwidth. Because it is a linear system, the bandwidth can simply be scaled to achieve the desired settling time. Given a fixed load, amplifier power tends to be proportional to bandwidth. Therefore, this approach will lead to a design that meets the settling time requirement with low power consumption.

The zeros z_o are typically sufficiently higher in frequency than the poles such that they do not significantly affect the step-response. The step-response is important because it determines the settling time in a switched-capacitor circuit. The step-response is the superposition of a decaying exponential due to the real pole at ω_{cl} and the tuned response of the complex poles defined by ω_n and ζ. Intuitively, if the real pole is too low in frequency, then the step response will be slow. If the real pole at ω_{cl} is too high, the step response will be fast but may require bandwidth (power) or be unrealizable. The same is true of the complex poles at ω_n. These poles, however, have the added damping parameter ζ. Like all second order systems, depending on ζ, the step-response can be under-damped $(\zeta < 1)$, critically damped $(\zeta = 1)$ or over-damped $(\zeta > 1)$. Depending on the allowed settling error there is a compromise between rise time and error due to overshoot and ringing. As the error tolerance becomes tighter, the optimal response approaches a critically damped response. For looser specifications ζ can be somewhat less than one.

Because of the potentially transcendental nature of the step-response, an analytic expression for the settling time cannot be generally derived as a function of the pole locations. Therefore, optimal placement of the poles cannot be solved for analytically. A reasonable pole configuration, however, can be found empirically using numerical methods.

Certainly a conservative choice would be to select $\omega_{cl} < \omega_n$ and $\zeta \geq 1$. Under these conditions, the step-response will be over-damped and the settling time will be insensitive to fabrication-induced variations in the small-signal circuit parameters. The trade-off is an increase in power consumption compared to a more aggressive, inherently faster configuration that incorporates some overshoot. For this design, 10-bit settling was required ($< 0.1\%$ error), and $a = \omega_{cl}/\omega_n = 0.9$ and $\zeta = 0.8$ were chosen. This results in the approximate relationship $\omega_n \approx 14/t_s$. Figure 5.17 shows the ideal 0.05% settling time of a three-pole system normalized to the shortest settling time versus the damping factor ζ for several values of $a = \omega_{cl}/\omega_n$. From the graph it can be seen that by decreasing the ratio of ω_{cl} to ω_n the optimum is broadened, but the step response is slower. Increasing the ratio beyond 0.9, however, does not increase the speed.

Furthermore, figure 5.18 shows the amplifier power increases with that ratio. Therefore, choosing $a = 0.9$ is good compromise between speed and power. For the damping factor ζ, the range of 0.7 to 0.9 achieves the best speed. Thus for a desired settling time of less than 14 ns requires $\omega_n > 1$ Grad/s (160 MHz). In simulation, the measured phase margin was 67 degrees.

Once the constants in the right-hand side of eqs. 5.16–5.19 have been chosen, a numerical solver can be used to find the opamp small signal-parameters g_{m1}, g_{m2}, g_{m3}, and C_c. Because the transistor $(V_{gs} - V_t)$ values have already been chosen, these four parameters uniquely define the opamp design including all device sizes. In this case, MATLAB [52] was used to find a set of device sizes that satisfy eqs. 5.16–5.19 with the minimum power consumption. From this basic design, the device sizes can be adjusted using HSPICE to verify the settling time. Table 5.3.4 shows the resulting settling time variations over process measured from HSPICE simulation with target specifications of $a = 0.9$ and $\zeta = 0.8$. The

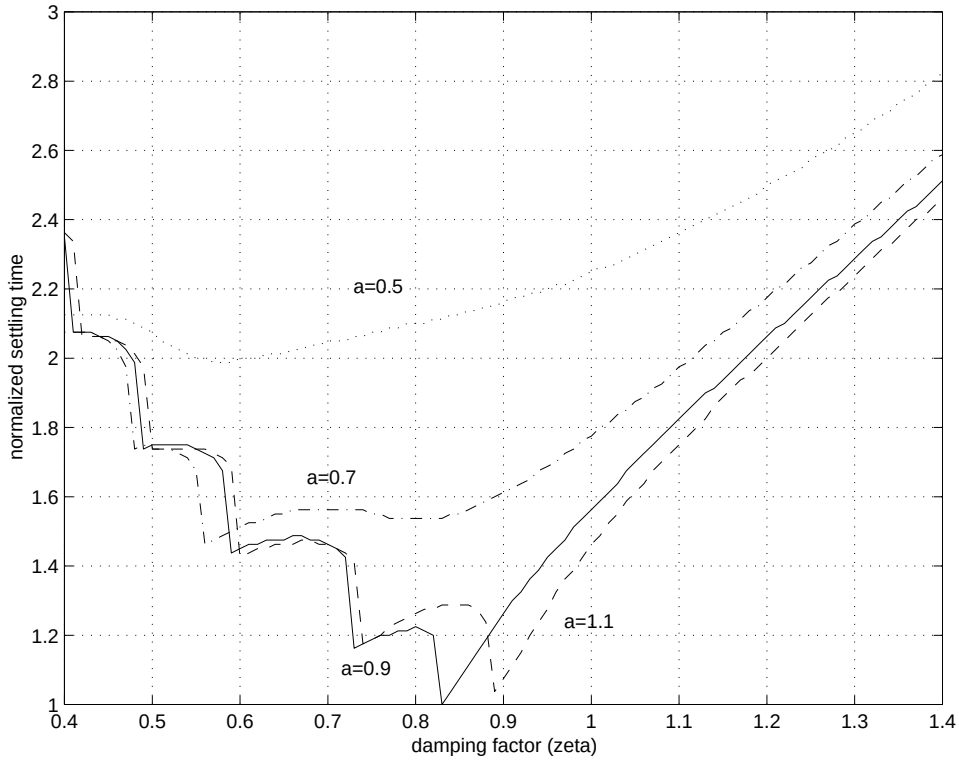


Figure 5.17 Normalized settling time vs. damping factor ζ

opamp configuration for the measurement is shown in figure 5.3.1, and the input was a full scale step.

5.3.5 Slew rate

Typically for high-bandwidth opamps, the slew rate scales with the bandwidth. Therefore, the fraction of the settling time spent in the slew-limited regime is small. Because this is a two-stage amplifier, there are two different slew rates. The lesser of the two will limit the overall rate of voltage of change at the output. Consider first the output node of the opamp. If the source of M2 is considered a virtual ground, then the maximum (positive) rate of change at the drain of M7 is

$$\text{SR1} = \frac{I_7}{C_c + C_{out} + (1-f)C_f},$$

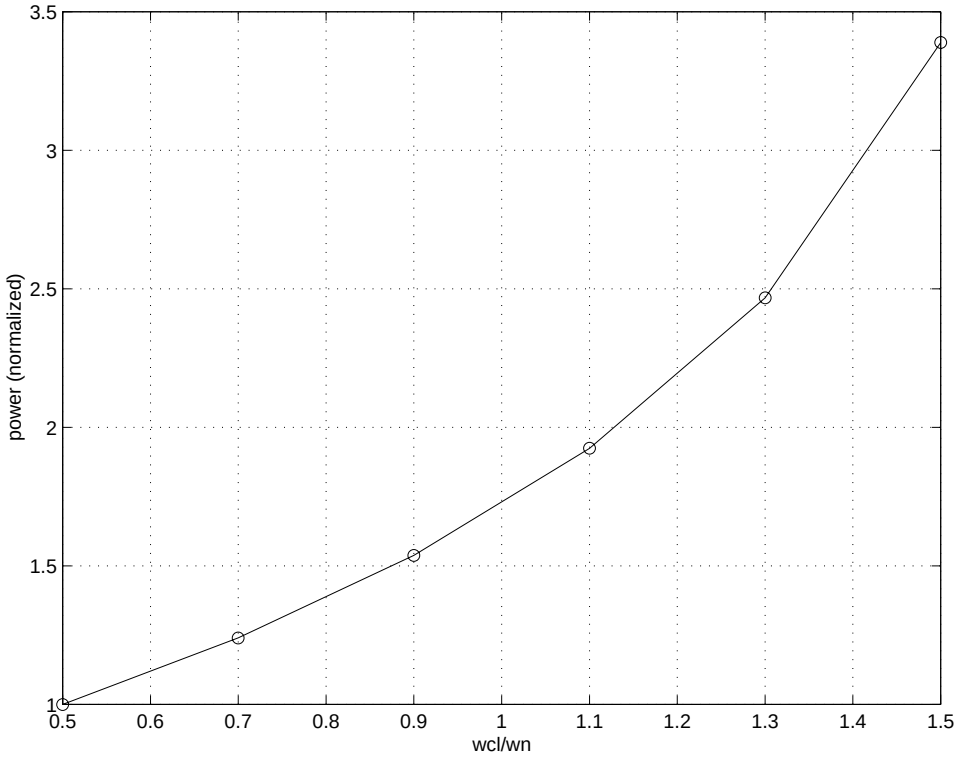


Figure 5.18 Normalized power vs. $a = \omega_{cl}/\omega_n$

where C_{out} is the total parasitic and external capacitance at the output. The source of M2, however, will only remain a virtual ground if M1 can supply sufficient charge to C_c to support the voltage change across C_c during a change in voltage at the output. Otherwise, this node will move and potentially cause M4 to leave saturation or M2 to cutoff. Therefore,

$$\text{SR2} = \frac{I_8}{C_c}.$$

For SR1, the slew rate may be improved by choosing a larger $(V_{gs} - V_t)$ for M1. This will raise the I/g_m ratio, which increases I_1 while keeping g_{m1} fixed. Similarly for SR2, the $(V_{gs} - V_t)$ for M3 may be increased.

NMOS	PMOS	t_s (0.05%)
slow	slow	17.1 ns
slow	fast	17.0 ns
nom	nom	12.9 ns
fast	slow	15.3 ns
fast	fast	11.6 ns

Table 5.1 Simulated settling time skew

5.3.6 Noise

For switched-capacitor applications, the opamp output noise directly contributes to the output signal in the hold phase such as shown in figure 5.3.1b. Therefore, this is typically the relevant configuration to compute the output noise. For simplicity, consider the single-ended noise due only to M1, M4 and M5. The cascode devices M6 and M2 do not significantly contribute to the noise; neither do the output devices M3 and M7 (because the other noise sources see a much larger gain to the output). Furthermore assume the noise due to M4 and M5 can be input referred. This assumes the transfer functions to the output are approximately the same as from M1. The output noise is then

$$\frac{\overline{v_o^2}}{\Delta f} \approx A^2(s) 4kT \frac{2}{3} \frac{1}{g_{m1}} \left(1 + \frac{g_{m4}}{g_{m1}} + \frac{g_{m5}}{g_{m1}} \right),$$

where $A^2(s)$ is the continuous-time, closed loop gain from the input of the opamp to the output. In this case,

$$A(s) = \frac{1}{f} \frac{(1 + s/z_o)(1 - s/z_o)}{(1 + s/\omega_{cl})(1 + 2\zeta s/\omega_n + s^2/\omega_n^2)}$$

To obtain the rms output noise voltage, this expression is then integrated over frequency for $\omega = [0, \infty]$. If it is assumed that there are three coincident real poles at frequency B (Hz), the equivalent noise bandwidth is $\frac{3}{16}\pi B$. Thus,

$$\overline{v_o^2} \approx \left(\frac{1}{f} \right)^2 4kT \frac{2}{3} \frac{1}{g_{m1}} \left(1 + \frac{g_{m4}}{g_{m1}} + \frac{g_{m5}}{g_{m1}} \right) \frac{3}{16} \pi B,$$

Because the opamp is fully differential, this singled-ended expression must be multiplied by 2. A precise expression for the output noise has been derived from a similar opamp in [25].

5.3.7 DC gain

A straightforward, small-signal analysis of figure 5.14 yields the open-loop DC gain of the opamp. This topology typically has a minimum gain of 1000, but it is dependent on bias and technology. Lower current levels and longer channel lengths tend to increase the overall DC gain.

$$A_1 \approx \frac{(g_{m1}r_{o1})(g_{m2}r_{o2})(g_{m5}r_{o5})r_{o4}r_{o6}}{(g_{m5}r_{o5})r_{o4}r_{o6} + (g_{m2}r_{o2})r_{o1}r_{o4} + (g_{m5}r_{o5})r_{o1}r_{o6}} \quad (5.20)$$

$$A_2 = (g_{m3}r_{o3})\frac{r_{o7}}{r_{o3} + r_{o7}} \quad (5.21)$$

$$\frac{v_o}{v_i} = A_1 \cdot A_2 \quad (5.22)$$

5.3.8 Common-mode feedback

Figure 5.19 shows the switched-capacitor common-mode feedback used in the opamp. This type of circuit is commonly used in fully-differential opamps. It operates on a 50% duty cycle. During phase ϕ_1, the output of the opamp is clamped to the desired common-mode output voltage VCMO by switches M29 and M30. The output of the first stage is also clamped by M31 and M32. This speeds the recovery of opamp during the next phase. The common-mode sensing capacitors C_{cm} are also reset by M28. During the next phase all clamped nodes are released. The capacitors C_{cm} sense the difference between the actual common-mode output and the desired common-mode output. The differential pair M26 and M27 convert this voltage difference to a current that is added or subtracted to the source of M1. This closes a negative feedback loop that stabilizes the common-mode output voltage around the desired common-mode voltage. The clamping switch gates are driven by the high-swing bootstrap circuit described above.

The benefit of injecting a single, common-mode current to the source of M1 is that no differential noise is added. One caveat of this approach, however, is potential large signal instability. The feedback loop passes through the input devices M1 and M33. If the input common-mode level is too high for some reason and M1 and M33 are cut off, then the output common-mode feedback loop is inoperable. Furthermore, the global feedback around the opamp formed by C_s and C_f (shown in the dashed lines) forms a common-mode positive feedback loop, which can potentially exacerbate the problem. An alternative approach is to inject the correction current at the drains of M1 and M33, which bypasses M1 and M33. Of course this then adds differential noise since there are two injection points.

The sampling and feedback capacitor sizes are determined by kT/C thermal noise constraints of the application (figure 5.3.1). For example in the pipeline ADC (chapter 7) the external capacitors were $C_s = C_f = 585$fF. Based on layout extraction the external loading driven by the opamp is approximately 2.3 pF. Based on HSPICE simulation, the approximate power consumption of the first stage opamp was 1.9 mW. Finally, although the opamp DC gain and settling time cannot be directly measured, the measured overall converter performance shows that the opamp DC gain exceeds 60 dB and the 0.1% settling time is less than 35 ns (at 14.3MS/s). Due to relaxed accuracy constraints in later pipeline stages, the sampling and feedback capacitors were scaled down to help reduce power consumption [15]. There are a total of 8 opamps in the ADC; the last stage of the pipeline does not need to generate a residue and does not require an opamp.

5.4 Comparator

The sub-ADC in each pipeline stage consists of two fully differential comparators, as shown in figure 5.20. In the 1.5-bit-per stage architecture, the sub-ADC thresholds are $+V_{ref}/4$ and $-V_{ref}/4$, where the ADC input range is $-V_{ref}$ to $+V_{ref}$ differential. The switched-capacitor comparator operates on a two phase non-overlapping clock. The differencing network samples V_{ref} during phase ϕ_2 onto capacitor C, while the input at capacitor $3C$ is shorted giving differential zero. During phase ϕ_1, the input signal V_i is applied at the inputs of both capac-

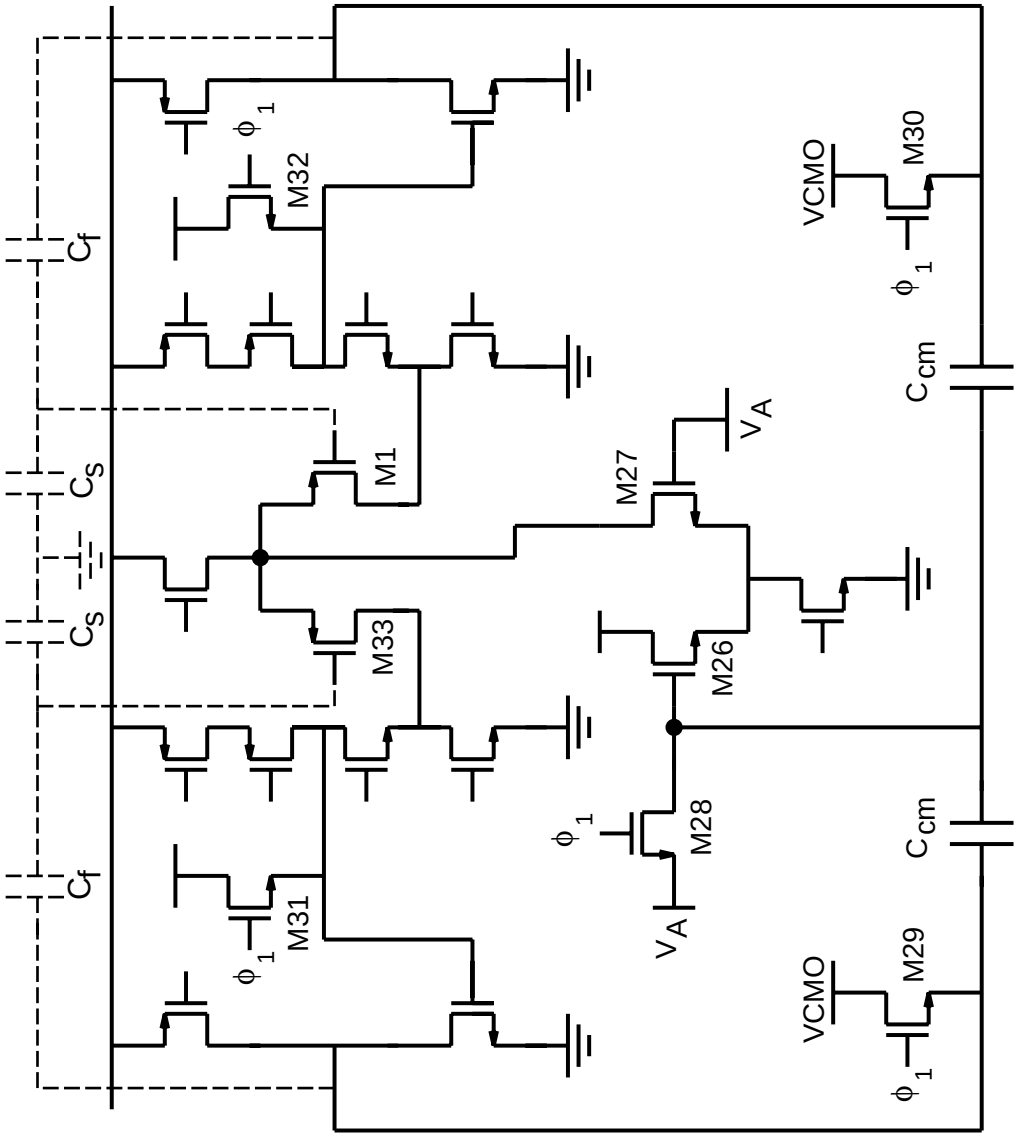


Figure 5.19 Common-mode feedback circuit

itors, causing a differential voltage proportional to $V_i - V_{ref}/4$ to appear at the input of the comparator preamp. At the end of phase ϕ_1 ($\overline{\phi_1}$ high) the regenerative flip-flop is latched to make the comparison and produce digital levels at the output V_o. Based on matching and common-mode charge injection errors, C was chosen to be near minimum size, approximately 40fF.

The pre-amp and latching circuit is shown in figure 5.21. Due to digital correction, a comparator error of $V_{ref}/4$ (200 mV) can be tolerated. With such a large allowable offset, a fully dynamic comparator is often used in order to reduce static power consumption [15]. At this low voltage, however, there are significant meta-stability problems with a fully dynamic comparator. Therefore, a class AB approach was used. During phase ϕ_1 the input ($V_i^+ - V_i^-$) is amplified by the input transistors, M1 and M2, which are connected to PMOS triode load devices, M3 and M4. During phase $\overline{\phi_1}$, the input is disconnected and the the NMOS flip-flop regenerates the voltage difference. Digital inverters buffer the outputs and restore full logic levels.

5.4.1 Offset

The DC offset of the comparator due to random device mismatch can be broken down into three contributions. Consider first mismatch between M1 and M2.

$$V_{os12} = \Delta V_{t12} + \left(\frac{\Delta\beta}{\beta}\right)_{12}\left(\frac{V_{gs}-V_t}{2}\right)_{12} \tag{5.23}$$

$$\Delta V_{t12} \triangleq V_{t1} - V_{t2} \tag{5.24}$$

$$\Delta\beta_{12} \triangleq \left(\mu C_{ox}\frac{W}{L}\right)_1 - \left(\mu C_{ox}\frac{W}{L}\right)_2 \tag{5.25}$$

The mismatch between devices M5 and M6 can be input referred.

$$V_{os56} = \frac{1}{g_{m12}}\left(g_{m56}\Delta V_{t56} + \left(\frac{\Delta\beta}{\beta}\right)_{56} I_{56}\right)$$

Load devices M3 and M4 also contribute to the offset.

$$V_{os34} = \frac{g_{m34}}{g_{m12}}\left(\Delta V_{t78} + \left(\frac{\Delta\beta}{\beta}\right)_{78}\left(\frac{V_{gs}-V_t}{2}\right)_{78}\right)$$

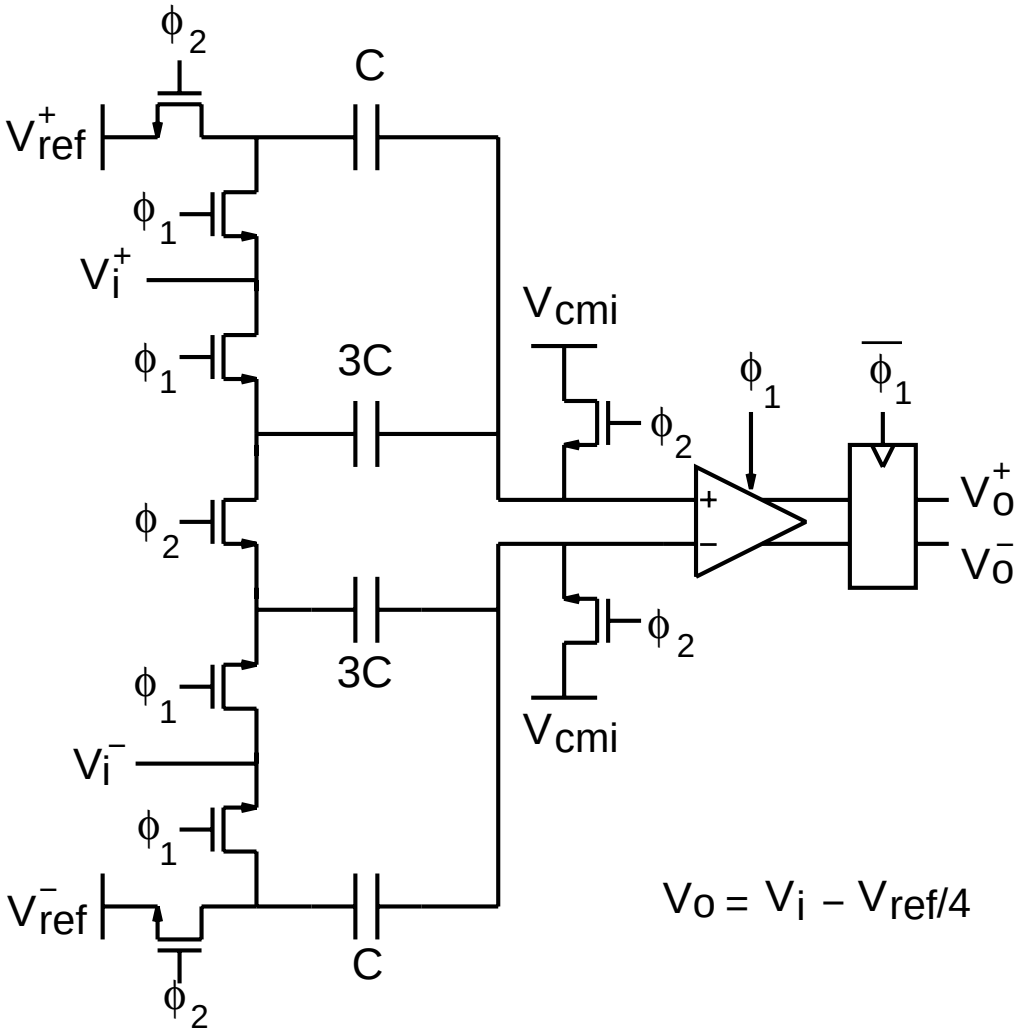


Figure 5.20 Differential comparator

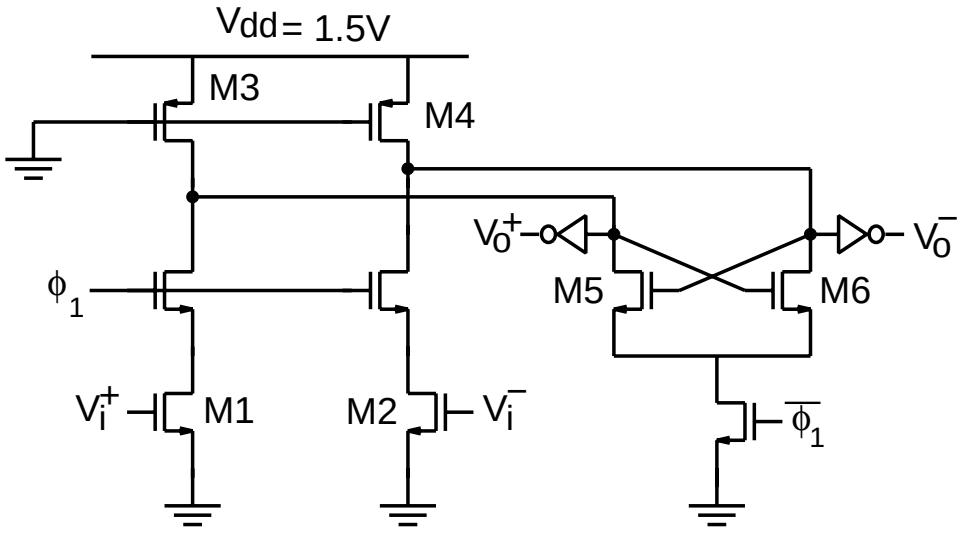


Figure 5.21 Comparator preamp and latch

If the offset are assumed to be small, then they add linearly to give a total input-referred offset.

$$V_{osT} = V_{os12} + V_{os34} + V_{os56}$$

The random mismatch between two devices can be characterized by a normal distribution that is a function of technology, device area, and distance between the two devices [63].

$$\sigma^2_{\Delta V_t} = \frac{A^2_{VTO}}{WL} + S^2_{VTO}D^2 \tag{5.26}$$

$$\frac{\sigma^2_{\Delta\beta}}{\beta^2} = \frac{A^2_{\beta}}{WL} + S^2_{\beta}D^2 \tag{5.27}$$

A_{VTO}, S_{VTO}, A_β, S_β are technology fitting parameters, WL is the device area, and D is the distance between the devices. Data from [63] gives typical values for the fitting parameters from a 2.5μm CMOS technology shown in table 5.4.1.

As these numbers show, the device mismatch is typically dominated by the ΔV_t mismatch. The $\Delta\beta$ mismatch can be neglected especially when the devices are physically close. If the mismatch contributions are assumed to be statistically

	NMOS	PMOS	
A_{VTO}	30	35	mV·μm
S_{VTO}	4	4	μV/μm
A_β	2.3	3.2	%·μm
S_β	2	2	$10^{-6}/\mu$m

Table 5.2 Mismatch data

independent, the total variance $\sigma^2(V_{osT})$ is the linear sum of the individual variances. Using this data, the offset for the comparator in the prototype was calculated to be $\sigma(V_{osT}) = 32$mV, or a worst-case 3σ offset of 96mV. Due to digital correction, an offset of 200mV can be tolerated in the pipeline ADC prototype.

5.4.2 Meta-stability

The relevant specification for comparators is not comparison time, but mean time to metastability. That is the average time between events when the comparator has not made a decision within the allotted time. This is a probabilistic event because the difference between the input signal and the reference is a random variable. The smaller the difference, the longer the required decision time which can approach infinity.

A typical regenerative (positive-feedback) comparator or flip-flop can be modeled by a two-stage, positive feedback loop. The loop consists of two ideal amplifiers of DC gain A and a time constant τ (see section 2.4). For this type of comparator, if the difference between the input signal and the reference is assumed to be a uniform across the input range, the probability of a metastable event is

$$P(t > T) = \exp\left(-\frac{A-1}{\tau}T\right),$$

,

where t is the actual comparison time, T is the allotted time [75]. Note that while a meta-stable state is always possible, simply waiting more time constants greatly reduces the probability. If you have a collection of N such comparators all clocking at a frequency f_s, then the mean time to failure (MTF)

$$\text{MTF} \approx \frac{\exp\left(\frac{A-1}{\tau}T\right)}{Nf_s}$$

as long as $P(t > T) \ll 1$. This is actually a lower bound for MTF. Typically MTF should be greater than the lifetime of the device. It should be noted that thermal noise does not affect MTF because the noise is equally likely to put the comparator into a meta-stable state as it is to take it out of a meta-stable state [75].

For the comparator shown in figure 5.21 the quantity $(A-1)/\tau$ is approximately

$$\frac{A-1}{\tau} \approx \frac{C_T}{g_{m5}}, \tag{5.28}$$

where C_T is the total parasitic capacitance at the drain of M5 or M6. The comparator latches during the clock non-overlap period, giving approximately 4 ns to make a decision. The regenerative time constant of the latch was chosen to make the system (18 comparators) mean time to meta-stability on the order of years at 14.3MS/s.

5.4.3 Pre-amplifier bandwidth

If a sample-and-hold circuit is not used in front of the comparator, the pre-amp bandwidth response needs to be greater than the input signal. Otherwise, high-frequency components in the input signal may go undetected.

For the first pipeline stage, the bandwidth of the pre-amplifier must exceed the bandwidth of the input signal being digitized, which is 7.15 MHz for the prototype. The static power consumption of the pre-amplifier is approximately 200μW based on HSPICE simulation.

Chapter 6

Pipeline ADC Architecture

THERE ARE a wide variety of analog-to-digital converter architectures. Each has its strengths and weaknesses that make it appropriate for a given set of specifications, such as speed, resolution, power, latency, and area. Such architectures include flash, two-step, interpolating and folding, pipeline, successive approximation, parallel, and others. A survey of these architectures is beyond the scope of this work, however, such surveys can be found in [65, 16, 15]. Here the focus is on the implementation of a video-rate, pipeline ADC that demonstrates the low-voltage circuit techniques described in chapter 5.

6.1 Pipeline ADC architecture

A generic pipeline ADC consists of N cascaded stages, each resolving B bits as shown in figure 6.1. Within each stage, the analog input is first sampled and held. Then it is coarsely quantized by a sub-ADC to resolve B bits. Then using a DAC, the quantized value is subtracted from original input signal to yield the quantization error. The quantization error is then restored to the original full-scale range by an amplifier of gain 2^B. The resulting residue signal is then applied to the next pipeline stage for further quantization on the next clock cycle. Due to the sample-and-hold nature of the pipeline, each stage works concurrently to achieve high throughput. The sample rate is only limited by the time it takes one stage to resolve B bits. There is a trade off in latency, however. The fully resolved $N \cdot B$ bits of resolution per sample experiences a delay N clock cycles from the sampling instant to full quantization. Therefore, the pipeline may be inappropriate for applications where latency is not acceptable.

The are several advantages to this type of architecture. As more bits of reso-

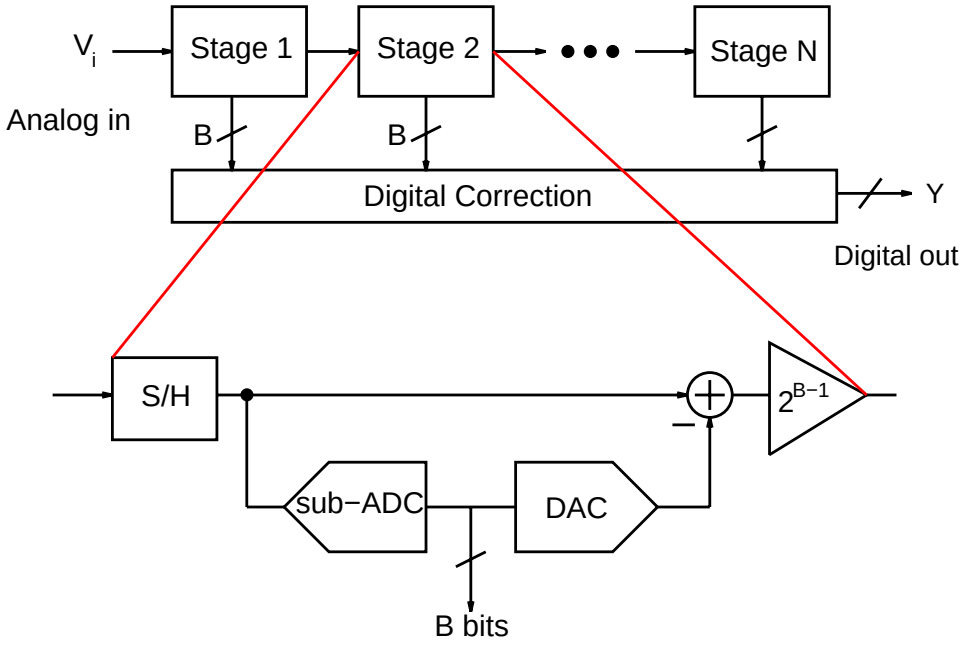


Figure 6.1 B-bit/stage Pipeline ADC

lution are added, the hardware and area requirements grow linearly since more stages are simply added. Using digital correction techniques, the accuracy requirements of the sub-ADCs are greatly relaxed allowing low-power comparators. On the other hand, however, because a precision sample-and-hold is required between each stage, fast-settling opamps are required which limit the throughput and increase power consumption. For same reason, high accuracy capacitor matching is required, but these errors can be self-calibrated with trimming [48], digital compensation [40], or capacitor averaging [71].

6.2 1.5-bit/stage architecture

The number of bits per stage has a large impact on the speed, power, and accuracy requirements of each stage. Therefore the best choice is dependent on the overall ADC specifications. For fewer number of bits per stage, the sub-ADC comparator requirements are more relaxed, and the inherent speed of each stage is faster. The latter occurs because the inter-stage gain is lower allowing higher speed due

to the fundamental gain-bandwidth tradeoff of amplifiers. However, more stages are required if there are fewer bits per stage. Furthermore, the noise and gain errors of the later stages contribute more to the overall converter inaccuracy because of the low inter-stage gain. Thus, high-speed, low-resolution specifications favor a low number of bits per stage, where low-speed, high-resolution specifications tend to favor higher number of bits per stage. A more detailed analysis can be found in [47].

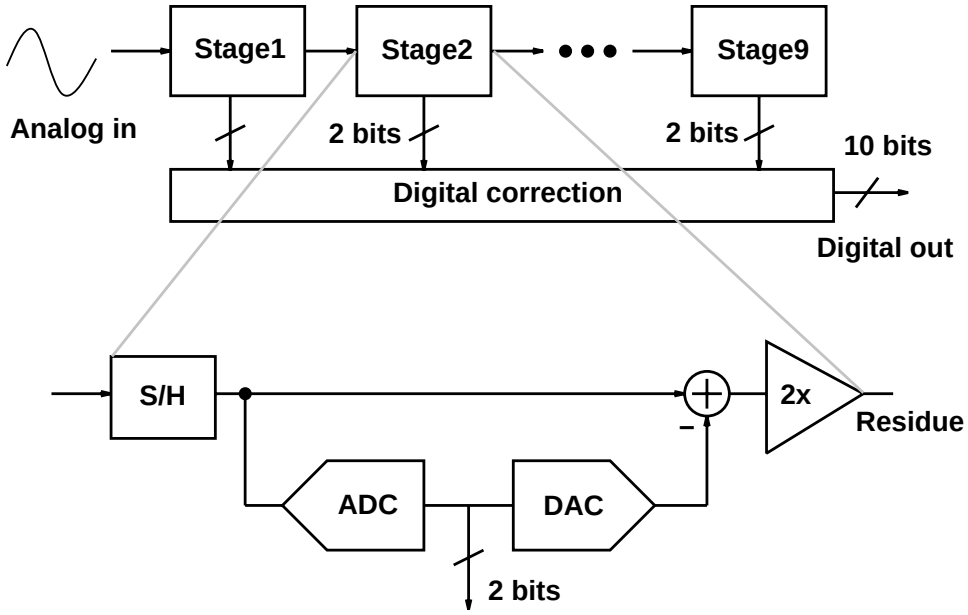


Figure 6.2 Pipeline ADC 1.5-bit/stage architecture

The ADC prototype uses a pipeline 1.5-bit/stage architecture [57, 46, 53] with 9 stages as shown in figure 6.2. Each stage resolves two bits with a sub-ADC, subtracts this value from its input and amplifies the resulting residue by a gain of two. The input signal ranges from $-V_{ref}$ to $+V_{ref}$, and the sub-ADC has thresholds at $+V_{ref}/4$ and $-V_{ref}/4$. The DAC levels are $-V_{ref}, 0, +V_{ref}$. Therefore, the residue transfer function is

$$V_o = \begin{cases} 2V_i - V_{ref} & \text{if } V_i > V_{ref}/4 & d = 2\ (10)_2 \\ 2V_i & \text{if } -V_{ref}/4 \leq V_i \leq +V_{ref}/4 & d = 1\ (01)_2 \\ 2V_i + V_{ref} & \text{if } V_i < -V_{ref}/4 & d = 0\ (00)_2 \end{cases}$$

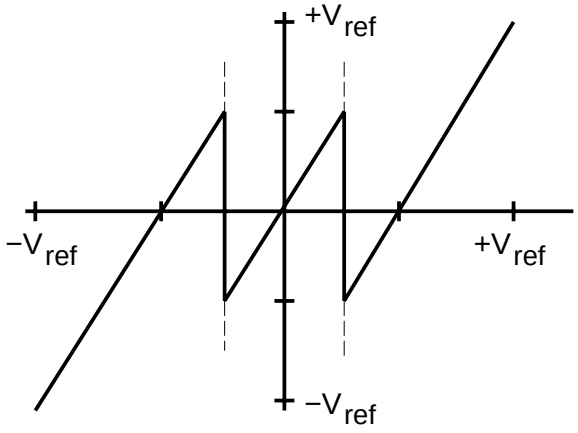


Figure 6.3 1.5 bit/stage residue transfer function

d is the output bit code for that stage. The transfer function is shown graphically in figure 6.3 The gain is lower (2 instead of 4) and there are more stages (9 instead of 5) than shown in figure 6.1 because digital correction [45, 15] was used. By reducing the inter-stage gain and introducing redundant bits, the accuracy requirements on the sub-ADCs are greatly reduced. In this case, a maximum offset of $V_{ref}/4$ can be tolerated before bit errors occur. A total of 18 bits (2 from each of 9 stages) are generated and combined using digital correction to yield 10 effective bits at the output. The bits are combined as follows:

$$d_{out} = \sum_{i=1}^{N} 2^{N+1-i} d_i$$

where N is total number of stages (9) and d_i is the output code for the ith stage. The Nth stage cannot be digitally corrected (since there are no following stages) and should have standard thresholds of

$$\begin{array}{ll} \text{if } V_i > V_{ref}/2 & d = 3(11)_2 \\ \text{if } 0 < V_i < V_{ref}/2 & d = 2(10)_2 \\ \text{if } -V_{ref}/2 < V_i < 0 & d = 1(01)_2 \\ \text{if } V_i < -V_{ref}/2 & d = 0(00)_2 \end{array}$$

If instead the digitally corrected offsets are used, the top code $1111\cdots$ will be missing which is typically not critical.

This architecture has been shown to be effective in achieving high throughput at low power [15, 50]. The low number of bits per stage coupled with digital correction relaxes the constraints on comparator offset voltage and DC opamp gain.

6.3 1.5 bit/stage implementation

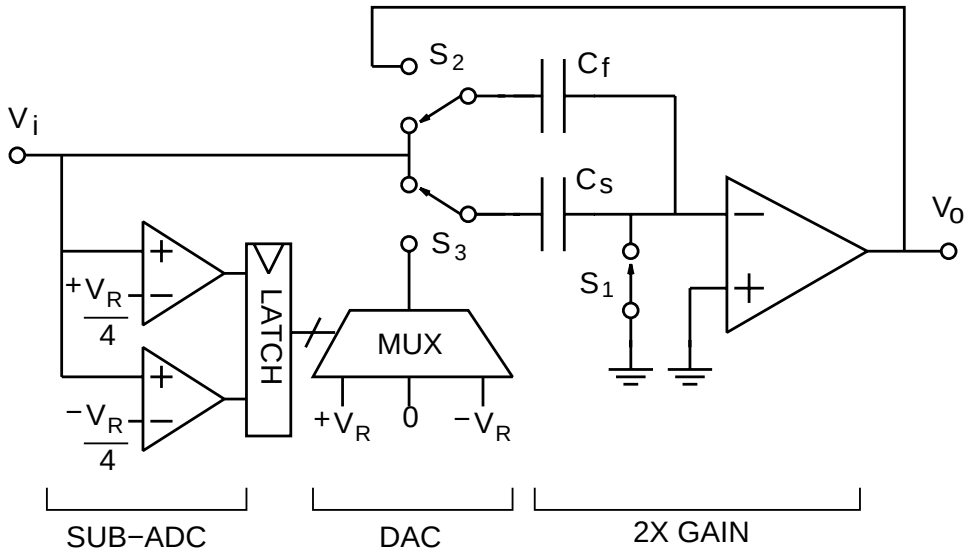


Figure 6.4 Switched-capacitor implementation of each pipeline stage

The implementation of each pipeline stage is shown in figure 6.4. Although a single-ended configuration is shown for simplicity, the actual implementation was fully differential. A common, switched-capacitor implementation [15] was chosen which operates on a two-phase clock. During the first phase, the input signal, V_i, is applied to the input of the sub-ADC, which has thresholds at $+V_{ref}/4$ and $-V_{ref}/4$. The input signal ranges from $-V_{ref}$ to $+V_{ref}$ (differential). Simultaneously, V_i is applied to sampling capacitors C_s and C_f. At the end of the first clock phase, V_i is sampled across C_s and C_f, and the output of the sub-ADC is latched. During the second clock phase, C_f closes a negative feedback loop around the opamp, while the top plate of C_s is switched to the DAC output. This configuration generates the stage residue at V_o. The output of the sub-ADC is

used to select the DAC output voltage, V_{dac}, through an analog multiplexor. V_{dac} is capacitively subtracted from the residue such that:

$$V_o = \begin{cases} \left(1 + \frac{C_s}{C_f}\right) V_i - \frac{C_s}{C_f} V_{ref} & \text{if } V_i > V_{ref}/4 \\ \left(1 + \frac{C_s}{C_f}\right) V_i & \text{if } -V_{ref}/4 \leq V_i \leq +V_{ref}/4 \\ \left(1 + \frac{C_s}{C_f}\right) V_i + \frac{C_s}{C_f} V_{ref} & \text{if } V_i < -V_{ref}/4 \end{cases}$$

In the 1.5bit/stage architecture $C_s = C_f$ is chosen to give a gain of two in the transfer function. The DAC levels can be generated from a single differential reference because the negative reference and differential zero are readily available by using the reverse polarity or shorting the outputs together respectively.

A precision interstage gain is required to achieve the desired overall ADC linearity. Because the capacitor ratio C_s/C_f determines this interstage gain, capacitor matching is critical. In the prototype, a capacitor trim array was used to ensure better than 0.1% matching. This array could be set either manually or automatically with a self-calibration scheme [48]. Also, the DC opamp gain must be sufficiently large (> 60 dB) to reduce finite gain error. Finally, the opamp must settle to better than 0.1% accuracy in one clock phase (one half-cycle). It is this settling time that limits the overall pipeline throughput.

6.4 Pipeline stage accuracy requirements

To achieve the desired resolution, linearity, and signal-to-noise ratio, each stage must be designed such that non-ideal effects do not excessively degrade the overall performance. Capacitor linearity and matching, opamp gain and settling, and thermal noise are all critical to pipeline ADC performance.

6.4.1 Capacitor matching

If the capacitors C_s and C_f are not equal, then an error proportional to the mismatch is generated in the residue output. Assume the capacitors differ by ΔC, then we can define

$$\Delta C \triangleq C_f - C_s \tag{6.1}$$

$$C \triangleq \frac{C_f + C_s}{2} \tag{6.2}$$

$$\Rightarrow C_s = C + \frac{\Delta C}{2} \tag{6.3}$$

$$C_f = C - \frac{\Delta C}{2} \tag{6.4}$$

$$\frac{C_s}{C_f} = \frac{C + \frac{\Delta C}{2}}{C - \frac{\Delta C}{2}} \approx 1 + \frac{\Delta C}{C} \tag{6.5}$$

The approximation holds if $|\Delta C/C| \ll 1$. Therefore, the new residue transfer function becomes:

$$V_o' \approx \left(2 + \frac{\Delta C}{C}\right) V_i \pm \left(1 + \frac{\Delta C}{C}\right) V_{ref} \tag{6.6}$$

$$\tag{6.7}$$

Due to the finite resolution of the lithographic process, capacitor mismatch is due mainly to variations at the edges of the capacitor plates. Therefore, capacitors with large area to perimeter ratios will tend to have better matching. Variations in oxide thickness between the capacitor plates also affect the matching, but to a lesser degree (especially for small, adjacent capacitors). The standard deviation of the fractional matching error between two adjacent square capacitors can be modeled as

$$\sigma_{\Delta C/C} = \frac{A_C}{S} \tag{6.8}$$

where S is one side of the capacitor in μm. The value of A_C is technology dependent, but can typically vary between $2 - 5\%\mu$m. For example, if A_C is $5\%\mu$m, then two adjacent, $15\mu\text{m} \times 15\mu\text{m}$ capacitors will match to better than 1% with 99.7% probability.

6.4.2 Capacitor linearity

If the sampling and feedback capacitors, C_s and C_f respectively, are implemented with diffusion layers, the capacitance is usually a function of the applied voltage.

Therefore, the charge transferred in a gain stage is no longer a linear function of the input voltage. This distortion degrades the linearity of the overall pipeline ADC.

If the value of each capacitor, C, can be approximated as a quadratic function of applied voltage V,

$$C(V) = C_0(1 + \alpha_1 V + \alpha_2 V^2)$$

then the resulting residue transfer function is altered. If the stage is assumed to be implemented as fully differential, then the new transfer function becomes [84]

$$V_o' \approx 2V_i - \frac{\alpha_2}{2} V_i^3 \pm V_{ref}.$$

Therefore, the error voltage is given by $\frac{\alpha_2}{2} V_i^3$. Typically for poly-poly or metal-metal capacitors this error is negligible.

6.4.3 Opamp DC gain

Because the DC gain of the opamp is finite, a gain error is introduced in the residue transfer function. If the actual gain of the opamp is a, then the residue transfer function becomes:

$$V_o' = \left(\frac{1}{1 + af}\right)(2V_i \pm V_{ref}),$$

where f is the feedback factor and C_{IP} is the input capacitance of the opamp,

$$f = \frac{C_f}{C_f + C_s + C_{IP}}.$$

This can also be expressed as

$$V_o' \approx \left(1 + \frac{\Delta G}{G}\right)(2V_i \pm V_{ref}) \tag{6.9}$$

$$\frac{\Delta G}{G} = -\frac{1}{af}. \tag{6.10}$$

6.4.4 Opamp settling

Because the opamp has finite bandwidth, the output takes time to settle to its final value. In the simplest case, if the opamp can be characterized by a single-pole time constant τ, then the output at the end of the settling period T is given by:

$$V_o' = \left(1 - e^{-t/\tau}\right)(2V_i \pm V_{ref}) \tag{6.11}$$

$$= \left(1 + \frac{\Delta G}{G}\right)(2V_i \pm V_{ref}) \tag{6.12}$$

$$\frac{\Delta G}{G} = -e^{-T/\tau}. \tag{6.13}$$

Typically most amplifiers have multiple poles and cannot be characterized by such a simple function. In this case, more empirical methods must be applied. The settling time T is less than one-half the ADC sampling period since the circuit operates on a two-phase clock. Some time must also be allocated to rising and falling edges of the clock as well as the non-overlap interval.

6.4.5 Thermal noise

There are two sources of thermal noise in this circuit. One is the thermal noise of the non-zero resistance switches. The other is the noise from the active transistors inside the opamp. Therefore [15],

$$\overline{v_n}^2 = \frac{1}{f}\frac{kT}{C_f} + \overline{v_{opamp}}^2.$$

$\overline{v_{opamp}}^2$ is the variance at the output due to noise from the opamp (e.g. see section 5.3.6). This value is dependent on the opamp topology and bias.

6.4.6 Error tolerances

If all of the above errors are sufficiently small, they can be linearly summed together to give the total error. The allowed total error per stage is now calculated. Recall, the implementation shown in figure 6.4 has the ideal transfer function

$$V_o = \left(1 + \frac{C_s}{C_f}\right) V_i \pm \frac{C_s}{C_f} V_{ref}.$$

The actual transfer function includes sources of error given by

$$V_o' = \left(1 + \frac{\Delta G}{G}\right) \left(\left(2 + \frac{\Delta C}{C}\right) V_i \pm \left(1 + \frac{\Delta C}{C}\right) V_{ref}\right) + v_n \quad (6.14)$$

$$\approx \left(1 + \frac{\Delta G}{G} + \frac{1}{2}\frac{\Delta C}{C}\right) 2V_i \pm \left(1 + \frac{\Delta G}{G} + \frac{\Delta C}{C}\right) V_{ref} + v_n. \quad (6.15)$$

$\Delta G/G$ is the combined interstage gain error caused by opamp finite gain, settling, and distortion. $\Delta C/C$ is the fractional difference between capacitors C_s and C_f. v_n is the output-referred thermal noise. For now, this term will be ignored and discussed below. A bit error results if the error in this residue signal is larger than 0.5 LSB of the following stages. The largest voltage error will result if $V_i \geq V_{ref}/4$. Note that in this case, the DAC output is also V_{ref} which is subtracted in the transfer function.

$$V_\varepsilon \leq (V_o' - V_o)|_{V_i = V_{ref}/4} \quad (6.16)$$

$$\leq \left(\frac{\Delta G}{G} + \frac{1}{2}\frac{\Delta C}{C}\right) 2\frac{V_{ref}}{4} - \left(\frac{\Delta G}{G} + \frac{\Delta C}{C}\right) V_{ref} \quad (6.17)$$

$$\leq \left(-\frac{1}{2}\frac{\Delta G}{G} - \frac{3}{4}\frac{\Delta C}{C}\right) V_{ref} \quad (6.18)$$

$$\leq \left(\frac{1}{2}\left|\frac{\Delta G}{G}\right| + \frac{3}{4}\left|\frac{\Delta C}{C}\right|\right) V_{ref} \quad (6.19)$$

Thus, if the overall converter has a resolution of N bits, then the first stage output error cannot cause a bit error greater than 0.5 LSB at the $N-1$ bit level. Intuitively, since one effective bit is resolved at each stage, the first stage is followed by a $N-1$ bit ADC.

$$V_\varepsilon < \frac{1}{2} \cdot \frac{2V_{ref}}{2^{N-1}} \quad (6.20)$$

$$\left(\frac{1}{2}\left|\frac{\Delta G}{G}\right| + \frac{3}{4}\left|\frac{\Delta C}{C}\right|\right) V_{ref} \quad < \quad \frac{1}{2} \cdot \frac{2V_{ref}}{2^{N-1}} \tag{6.21}$$

$$\Rightarrow \frac{1}{2}\left|\frac{\Delta G}{G}\right| + \frac{3}{4}\left|\frac{\Delta C}{C}\right| \quad < \quad \frac{1}{2^{N-1}} \tag{6.22}$$

Using this logic, the same argument can be applied to each stage. Therefore, in general,

$$\frac{1}{2}\left|\frac{\Delta G}{G}\right|_i + \frac{3}{4}\left|\frac{\Delta C}{C}\right|_i < \frac{1}{2^{N-i}} \quad (i = 1, 2, ..., N-2) \tag{6.23}$$

Notice i only goes up to $N-2$ because the last stage does not require a residue amplifier. If these conditions are met, the worst case differential non-linearity (DNL) will be less than 0.5 LSB at the N bit level [45, 49]. The integral non-linearity, however, accumulates from stage to stage. Therefore, equation 6.23 alone does not guarantee the INL will be less than 0.5 LSB at the N bit level. Unlike DNL, the INL accumulates from stage to stage. Due to the non-linear nature of residue transfer function, it is difficult to analyze. Based on empirical behavior simulations, the worst-case INL accumulates approximately as follows:

$$\text{INL} \leq \sum_{i=1}^{N-2} \frac{1}{2^{i+1}} \left(\frac{1}{2}\left(\frac{\Delta G}{G}\right)_i + \frac{3}{4}\left(\frac{\Delta C}{C}\right)_i\right) \tag{6.24}$$

Equation 6.24 is expressed as a fractional value. To convert to LSB at the N bit level, multiply the result by 2^N. In the worst case, the errors all have the same sign and accumulate in the same direction. Typically the errors have a more random nature and perhaps could be considered to add in a root-mean-square manner. Thus, to achieve an INL less than 0.5 LSB, equation 6.23 is a necessary but not sufficient condition.

The random thermal noise v_n will degrade the converter signal-to-noise ratio (SNR). The noise can be considered a small-signal; this allows the total input-referred noise to be calculated by linearly adding the noise contributions of all the stages. The noise contribution from each stage is divided by the total gain

from that point to the ADC input. Therefore, for an inter-stage gain of 2, the total input-referred noise is

$$\overline{v_s}^2 = \frac{\overline{v_1}^2}{2^2} + \frac{\overline{v_2}^2}{2^4} + \frac{\overline{v_3}^2}{2^6} + \cdots + \frac{\overline{v_N}^2}{2^{2N}}.$$

$\overline{v_k}^2$ is the output-referred thermal noise variance for stage k. An ideal ADC has a peak SNR (SNR with largest amplitude input applied) that is limited only by quantization noise, which can be approximated as

$$\text{SNR}_{dB} < 6N + 1.76 \text{ dB},$$

where N is converter resolution in bits. If thermal noise is added, the peak SNR becomes

$$\text{SNR}_{dB} = -10 \log_{10} \left(\frac{2}{3} \frac{1}{2^{2N}} + 2 \frac{\overline{v_s}^2}{V_{ref}^2} \right).$$

The amount of tolerable thermal noise is dependent on the application. However, 6 dB degradation of the SNR translates to a least significant bit that is essentially random. Typically, 1-2 dB degradation of SNR is designed for in the error budget. Equivalently,

$$\sqrt{\overline{v_s}^2} < \frac{1}{6} \frac{2V_{ref}}{2^N} = \frac{1}{6}\text{LSB} \tag{6.25}$$

6.4.7 Design example

Using the above equations we can allocate the sources of error and specify the required performance of each of the pipeline stages. As a specific example, assume the goal is to design a 10-bit ADC ($N = 10$) with INL less than 1 LSB and DNL less than 0.5 LSB. The clock frequency of ADC is 14MS/s. The input signal swing is 2V peak-to-peak.

For the first stage ($i = 1$) we will allocate errors equally to gain errors and capacitor mismatch, $\Delta G/G \leq 2^{-10}$ and a worst case $\Delta C/C \leq 2^{-10}$. Two untrimmed capacitors as discussed in section 6.4.1, would need to be approximately $150\mu\text{m} \times 150\mu\text{m}$ to have a worst-case matching of less than 0.1%. Therefore, a calibration

scheme such as capacitor trimming (section 7.5) or digital-domain calibration is typically used to guarantee precision matching and allow the use of smaller capacitors (as dictated by kT/C noise). Notice this allocation satisfies equation 6.23

$$\frac{1}{2}|2^{-10}| + \frac{3}{4}|2^{-10}| < \frac{1}{2^9} \tag{6.26}$$

$$0.00122 < 0.00195 \tag{6.27}$$

The gain error $\Delta G/G$ then must be further broken down into DC gain error (eq. 6.10) and settling error (eq. 6.13). If we allocate these equally, then 2^{-11} is allowed for each. From equation 6.10 if we assume the feedback factor of the gain stage is $1/2$, then the DC gain of the must opamp must be at least 4096. If we assume the clock frequency is 14MS/s (T=70ns) with a two-phase clock, then the opamp must settle to better than 2^{-11} within 35ns.

Allocating the noise for minimum power is a non-trivial issue and is discussed in detail in [16, 15]. For simplicity here, we will allocate 50% of the total input-referred noise power to the first stage, 25% for the second stage, 12.5% for the third stage and so on. If we use equation 6.25, then the total allowed input-referred rms noise is $2V/2^{10}/6 = 325\mu$V. Therefore, the first stage must have an input-referred rms noise of less than 230μV.

We can continue this process for each stage, taking care to observe equations 6.23 and 6.24.

Chapter 7

Prototype Implementation

AS A DEMONSTRATION of the low-voltage circuit techniques described in the previous chapters, a 1.5V, 10-bit, 14.3MS/s analog-to-digital converter prototype was implemented in a 0.5 μm CMOS technology. This chapter discusses specific implementation details including schematics and layout issues.

7.1 Technology

The prototype was fabricated using Hewlett Packard's 0.5μm CMOS14 process through MOSIS. The process has standard threshold voltage levels of 0.7 V and 0.9 V for NMOS and PMOS devices respectively. The maximum rated voltage supply is 3.3 V, however, the prototype was designed for reliable operation at 1.5 V. Three levels of metal interconnect and one layer of poly were used. Linear capacitors were implemented using poly over n+ diffusion in a n-well. The capacitance of these capacitors was approximately 2.3 fF/μm^2. The substrate consisted of low-resistance p+ with an epitaxial layer of p- on top. The die was packaged in a 68-pin J-LDCC ceramic package with a 0.265"x0.265" square cavity from Spectrum Semiconductor, Inc.

7.2 Layout

A die photograph of the prototype ADC is shown in figure 7.1. The differential input enters the chip from the top, third and fourth pad from the left side. The visible probe pads are for test devices used in process characterization. The eight opamp stages are clearly visible in two rows of four. In between the two rows are the charge pumps and routing of the digital signals such as clocks and data outputs. Using this arrangement there is a minimum of analog and digital signal

lines crossing.

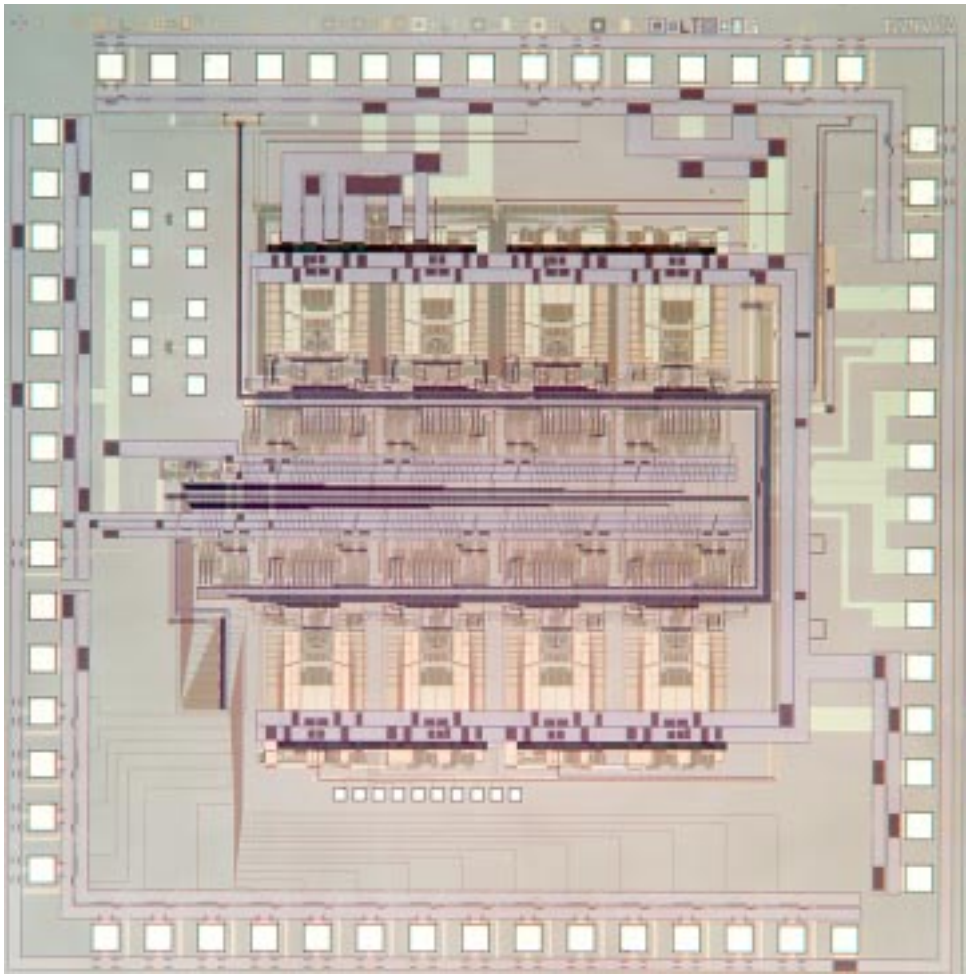


Figure 7.1 Photomicrograph of prototype ADC

In mixed-signal chips, it is often unclear what the best strategy is for minimizing the impact of noise coupling from the digital circuitry to the sensitive analog circuitry via the common substrate. The most effective way to reduce substrate noise is to create a low-impedance path from the p+ substrate to ground (or the lowest potential in the chip). Typically, however, the back side of the die is oxidized by exposure to air which increases its resistance. Thus even if the die is conductively attached to a package with a grounded cavity, the resistance to the substrate is high. If cost permits, the backside of the die can be back-lapped

(ground down), gold-coated, and conductively attached to package. The prototype die, however, was not back-lapped.

In this layout, the following approach was taken. Separate supply rails were used for the digital and analog signals, which were named VDDD, GNDD, VDDA, and GNDA respectively. Because an n-well process was used, the digital and analog PMOS transistors were naturally isolated by separate wells. The NMOS transistors, however, interact via the common, low-resistance p+ substrate.

The p+ substrate has the advantage that it makes it difficult to create latch-up, which is critical for digital circuits. It has the disadvantage of creating a low-resistance path for the coupling of undesired signals. Because noise travels almost exclusively in the p+ region, traditional isolation using grounded n-well guard rings to collect noise is not effective [73, 27].

For the analog NMOS transistors, it is important that the source-to-body voltage is constant. Otherwise, if these voltages move relative to each other, the drain current is modulated through the body effect. Therefore, it is important to locally have a low-resistance path from body to source. In the layout, a p+ substrate ring was placed around each NMOS analog transistor. This ring was then contacted to GNDA, which is the same potential as the source for common-source devices. This helps keep the potential of the source and the body the same. For cascode NMOS devices, this arrangement helps reduce fluctuations on the body terminal, but it cannot guarantee that the source and body will move together (since the source is not at ground potential). For cascode devices, however, the relationship between drain current and V_{sb} is much weaker due to source degeneration. Care should be taken to use enough of these GNDA substrate contacts to make the source-to-body path low resistance. Excessive contacts, however, will act as noise receptors and unnecessarily couple in noise from the substrate [73]. In the prototype, a separate substrate pin PSUB was used to set the substrate potential. Contacts between PSUB and the substrate were made liberally in the digital areas to provide a low-impedance path to ground for substrate noise. PSUB and GNDD are were disjoint on-chip. Off-chip, GNDA, GNDD, and PSUB were connected together. An alternate approach would be to tie PSUB and GNDD together on-

chip if it is preferable to have fewer pin-outs. VDDA and VDDD were generated by separate but equal voltage regulators to allow the supply currents to be measured independently.

All analog paths were differential to increase the rejection of common mode noise, such as substrate noise and supply voltage fluctuations. Furthermore, analog signal paths were shielded from the substrate by an n-well tied to a separate SHIELD pin tied to VDDA off-chip. VDDA was used instead of GND to increase the back-bias of the n-well thereby reducing the coupling capacitance to the substrate.

7.3 Master bias

Each opamp requires a set of bias voltages that is supplied by a bias generator. There is one bias generator shared between every two pipeline stages for a total of four. To minimize pin-out, each generator is driven by an input current that is slaved from a single master bias current shown in figure 7.2.

The master current I_{master} is generated off-chip. In a commercial chip, this would typically be generated on-chip with a self-biasing current source, such as a bandgap reference. Similarly, the opamp common-mode output voltage VCMO is set off-chip and buffered on-chip with a simple source-follower. This also could be generated on-chip with a resistor string. Its exact value is not critical, but should be close to midway between the GNDA and VDDA. It is important to use high-output-impedance current sources wherever possible to reduce errors in the current values due to differences in drain voltages. Typically, cascode current sources are preferable. On a low supply voltage, however, this is not possible, so long channel devices were used.

The alternative to routing bias *currents* is to route bias *voltages* across the chip as shown in figure 7.3. Routing bias voltages locally is not a problem. Routing over long distances, however, can generate errors in the resulting bias current. If the distance between the master device M1 and the slave device M2 is too great, then the resistance R in the supply line can be significant. This drop reduces V_{GS2} and creates a current error in the slave device.

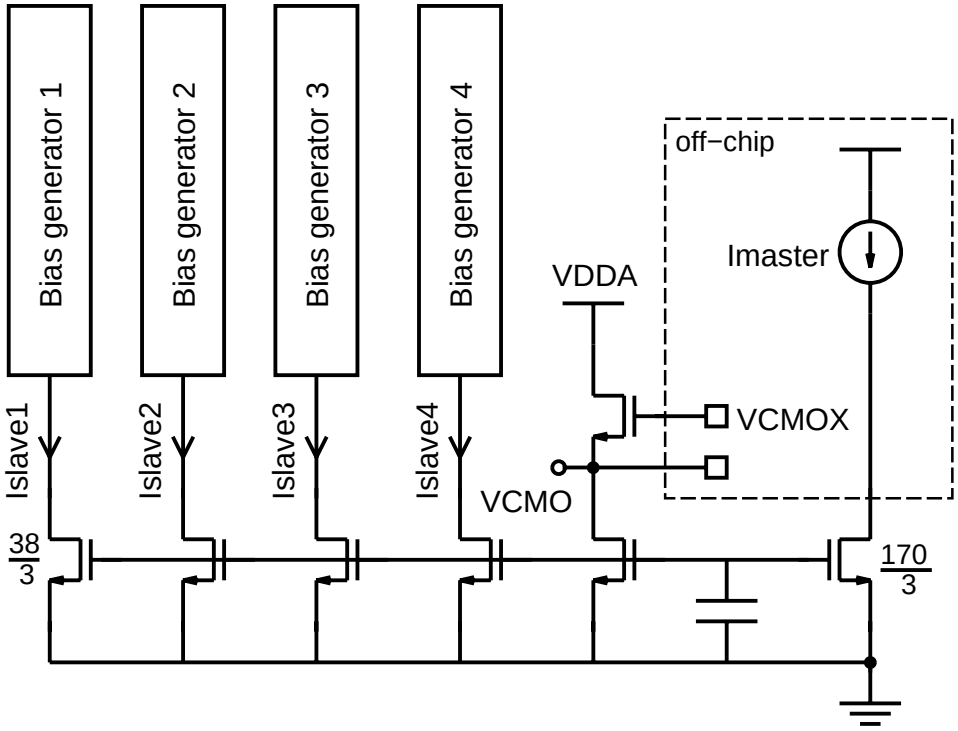


Figure 7.2 Master bias circuit

$$V_{GS2} = V_{GS1} - I_2 R \Rightarrow I_2 \neq I_1$$

When bias *currents* are routed, global resistances in the supply lines do not affect the final current in the slave devices. Locally, diode-connected devices can convert the current back into a bias voltage.

7.4 Clock generator

All the pipeline stages operate on a two-phase, non-overlapping clock. All the odd stages sample during phase ϕ_2 and present a valid residue output to the next stage during phase ϕ_1. All even stages work on the opposite phases, so that all stages operate concurrently. The commonly used circuit shown in figure 7.4 was used in the prototype. An external 50% duty-cycle reference clock drives the in-

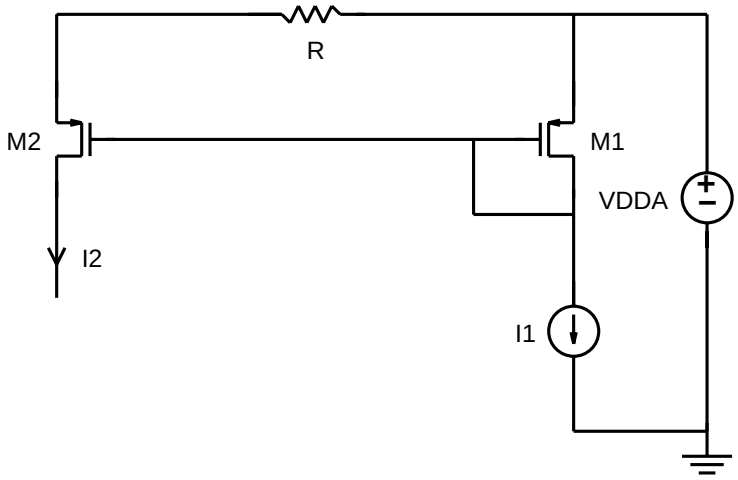


Figure 7.3 Problem with routing a bias voltage

put `clkin`. In the prototype, the inverted phases are buffered and routed across the chip. Figure 7.5 shows the clock waveforms. For simplicity the non-inverted phases are shown.

At the top of the figure the reference clock `clkin` is shown with a period of T and rise and fall times of `tr` and `tf` respectively. The duration of the dependent phases is a function of the propagation delays of the various gates in the clock generator. By adjusting these delays, the designer can allocate the time spent in each of the phases. `ts1` is the opamp settling time for pipeline stages that generate an output during phase ϕ_1. `ts2` is the opamp settling time for pipeline stages that generate an output during phase ϕ_2. Typically these are designed to be as equal as possible. `tlag` is the time between the early clock ϕ_1' and the regular clock ϕ_1. This delay ensures that the bottom plate switch of the sample and hold is opened first to reduce signal-dependent charge injection. `tnov` is the non-overlap interval during which neither phase is active. For proper operation of the two-phase circuits this overlap must be non-zero. Furthermore, in a pipeline ADC, this time is used for the sub-ADCs to digitize the sample and select the cor-

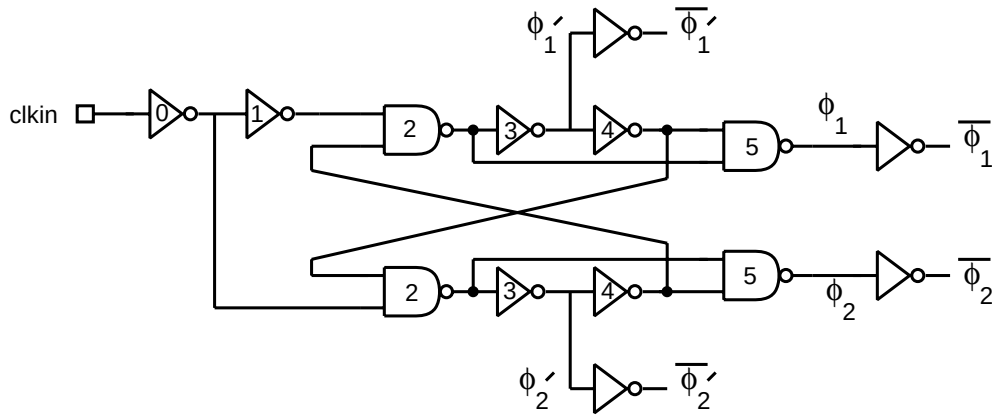


Figure 7.4 Non-overlapping two-phase clock generator

`ts1`	$\frac{T}{2} - t_r - t_2 - t_3 - t_4 + t_1 + t_f$
`ts2`	$\frac{T}{2} - t_f - t_1 - t_2 - t_3 - t_4 + t_r$
`tlag`	$t_4 + t_5$
`tnov`	$\min(t_2, t_2 + t_3 - t_5)$

Table 7.1 Clock generator timing

rect DAC level. If this interval is too small, the probability of meta-stability will increase. Table 7.1 gives the dependencies of the clock intervals. In the table, t_n represents the propagation delay of gate n.

7.5 Capacitor trimming

Capacitors in the gain stages were trimmed using the circuit shown in figure 7.6. By setting the switches to ground, no capacitance is added between points A and B. The array does, however, add some parasitic capacitance to ground. By selectively putting switches in the opposite position, a very small amount of capacitance can be added between points A and B. In the prototype, the unit capacitor C was set to 10 fF, which allowed a trim range of 0-6 fF in 0.4 fF steps.

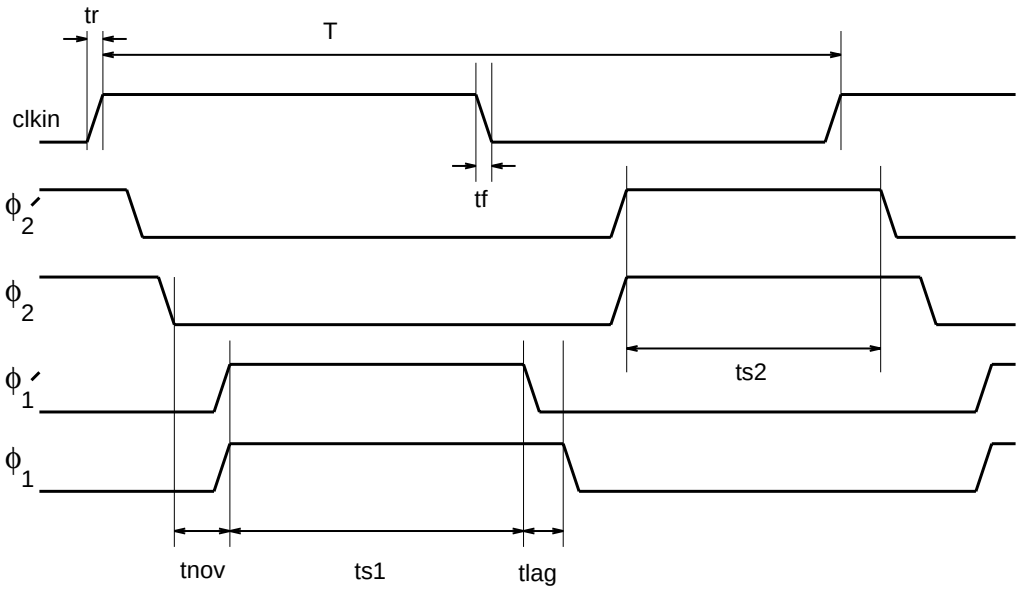


Figure 7.5 Clock generator timing diagram

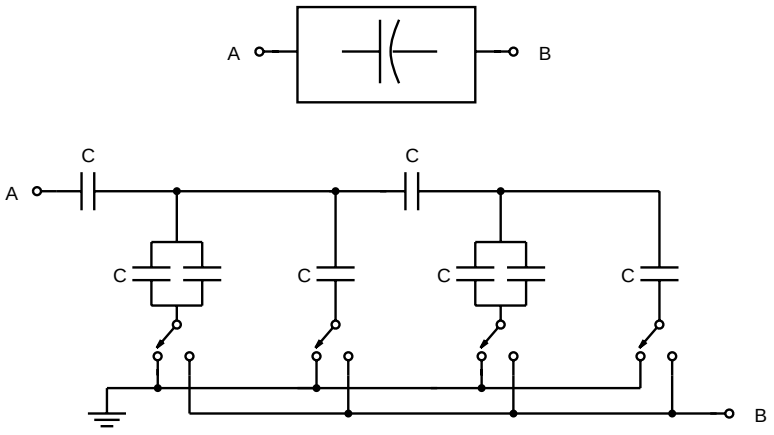


Figure 7.6 Capacitor trimming circuit

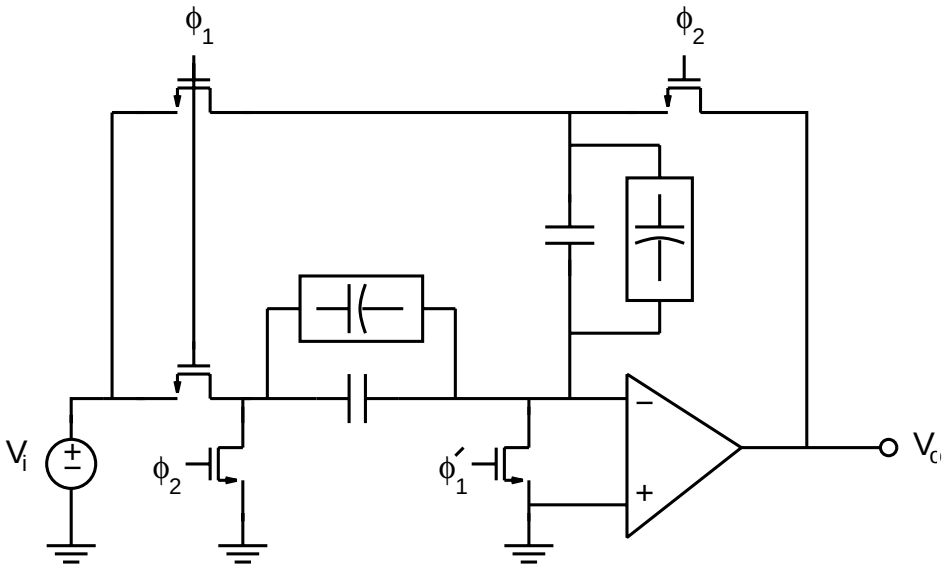


Figure 7.7 Trim applied to gain stage

Figure 7.7 shows the trim circuit applied to a gain stage. Notice that the bottom plate (node B) of the trim array is placed on the summing node. Because the summing node (or common-mode input) voltage was set to GND, the switches in the trim array can be operated on a 1.5V supply without extra boosting circuitry. By placing the bottom plate on the summing node, however, two trim arrays are needed—one for each capacitor. This allows trim to be added to the appropriate capacitor to achieve the correct ratio. If the opposite configuration were possible, where the top plate (node A) is connected to the summing node, only one array is necessary. In this configuration the array switches could be used to add trim to either capacitor as necessary. A shift register was used to digitally store the trim settings. In the prototype, there was a 2dB degradation in SNDR without using the trim.

7.6 Gain stage

The fully-differential implementation of the gain stage is shown in figure 7.8. Gate boosted switches are indicated with a dashed circle as defined previously in figure 5.5. Recall, extra parasitic capacitance is incurred at the source terminal of the switch. The bottom-plate sampling switches do not need gate boost because they only pass the GND potential. The opamp was described in detail in section 5.3. The clock phases shown in figure 7.8 correspond to sampling on ϕ_1 and holding on ϕ_2.

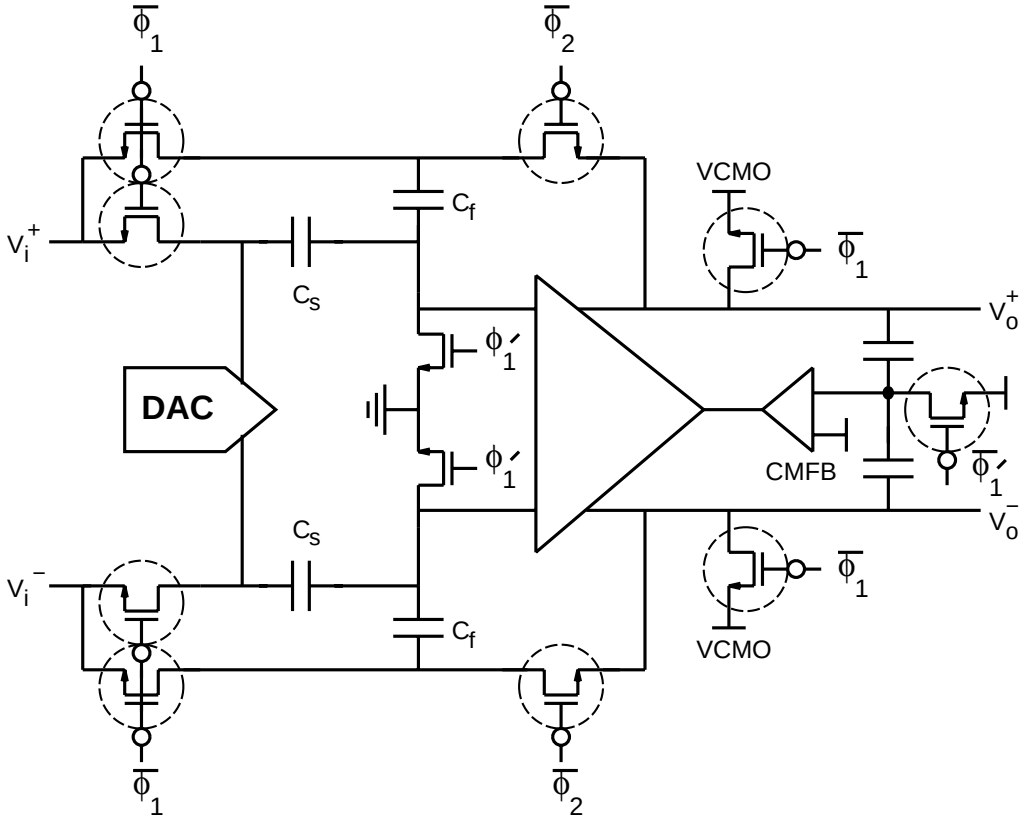


Figure 7.8 Differential gain-stage

7.7 Sub-ADC/DAC

The sub-ADC and DAC within each pipeline stage is shown in figure 7.9. The two differential comparators threshold the input at $-V_{ref}/4$ and $+V_{ref}/4$. The result is then decoded by some digital logic which selects one of three values, $-V_{ref}$, 0, or $+V_{ref}$ for the differential DAC output. The output is also gated by the hold clock ϕ_2, such that the DAC output presents a high-impedance until the gain stage is in the hold phase. The data bits MSB and LSB are then fed to set of latches. These latches align all the bits from all the stages in time. Because one sample takes 9 clock cycles (for 9 stages) to propagate through the entire pipeline, the first stage bits are delayed by 8 clock cycles, and the last stage has no delay. The resulting 18 bits were digitally corrected by a simple add and shift operation to yield 10 effective bits.

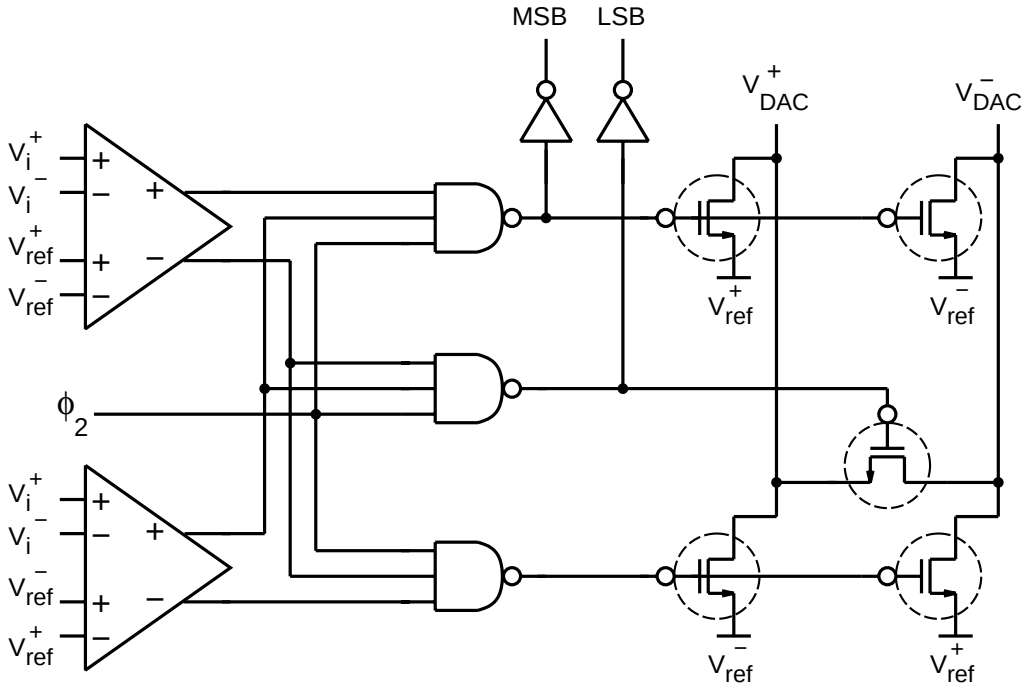


Figure 7.9 Sub-ADC and DAC

Chapter 8

Experimental Results

FOLLOWING the design and fabrication of the prototype, the ADC was characterized for linearity, signal-to-noise ratio, and power. Reliability characterization, however, was not performed.

8.1 Test setup

The basic setup for the experimental testing is shown in figure 8.1. A single-frequency, sinusoidal signal, V_s is generated by a Krohn-Hite 4400A oscillator and applied at the first to a narrow, band-pass filter to remove any harmonic distortion and extraneous noise, and then to the test board. The signal is connected via 50Ω coaxial cables to minimized external interference. On the test circuitboard, the single-ended signal is converted to a balanced, differential signal using a transformer. The transformer was a PSCJ-2-2 from Mini Circuits. The common-mode voltage of the test signal going into the ADC is set through 50Ω mmatching resistors connected to a voltage reference. This common-mode voltage can be adjusted using a potentiometer, but is nominally at half the supply or 0.75 V. This voltage is bypassed to the board ground with a 10μF ccapacitor.

A 50% duty-cycle clock is generated by a Hewlett Packard HP8131A pulse generator. Both ADC prototype and the HP16500B logic analyzer use this clock. The logic analyzer stores all the digital output codes in uncorrected format, 18 bits wide. The data is then down-loaded to a computer where the digital correction (simple shift and add) is done in software. The correction can easily be implemented on-chip, but was done off-chip for simplicity and diagnostic feedback.

The circuit board has four layers. The top was used mainly for analog signal traces. The second layer was a common ground plane. The third layer was broken

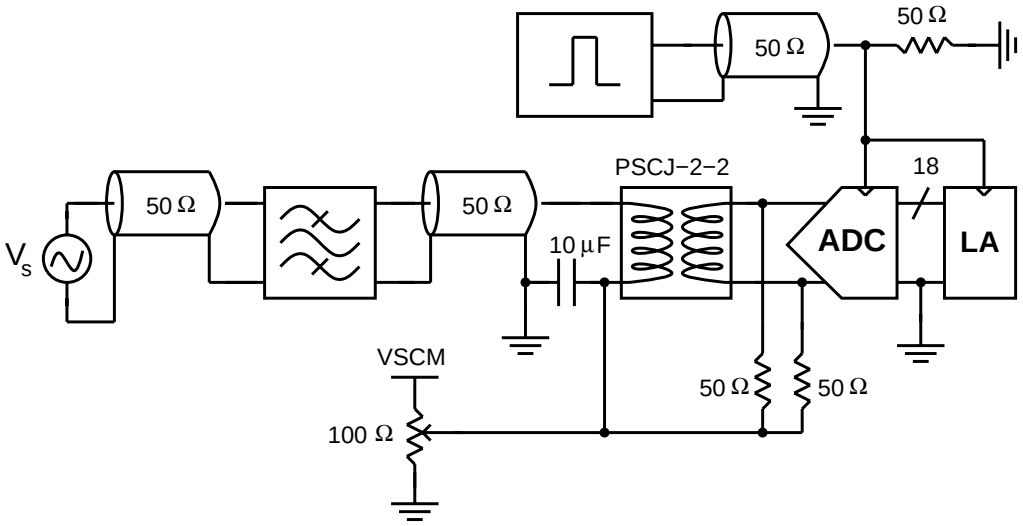


Figure 8.1 Test setup (ADC=device under test, LA=logic analyzer)

into separate VDDA and VDDD power planes. The bottom side was used for control traces.

All voltage references were generated using a National Semiconductor LM317 regulator as shown in figure 8.2. This includes all voltage supplies, VDDA, VDDD, V_{ref}, VCMOX, VSCM. The output voltage, VOUT, can be adjusted by selecting R1 and R2.

$$V_{out} = 1.25\text{V}\left(1 + \frac{R2}{R1}\right) + I_{ADJ}R2$$

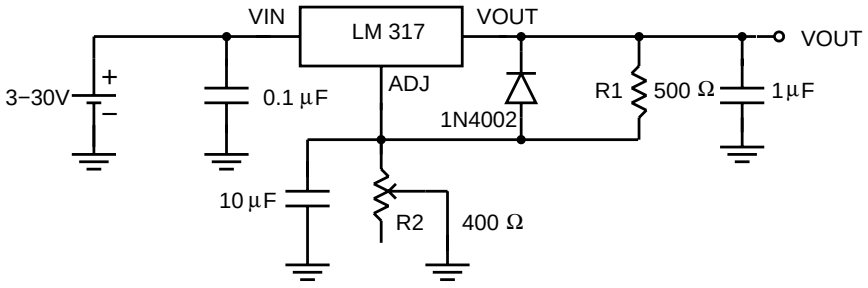


Figure 8.2 Voltage reference generation

The master current is generated using a National Semiconductor LM334 current source as shown in figure 8.3. The current is set by choosing R_{SET}.

$$I_{OUT} \approx \frac{(227\mu\mathrm{V/K})T}{R_{SET}} = \frac{68\mathrm{mV}}{R_{SET}}$$

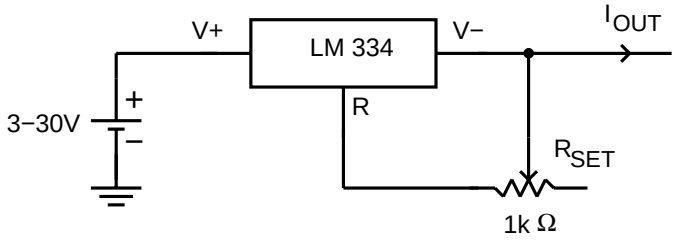


Figure 8.3 Current reference generation

8.2 Dynamic linearity and noise performance

The linearity and noise performance of the converter can be characterized by a signal-to-noise-plus-distortion (SNDR) measurement. The SNDR is defined as the ratio of signal power to all other noise and harmonic power in the digital output stream. This characteristic determines the smallest signal that can be detected in the presence of noise, and the largest signal that does not overload the converter. The peak SNDR is highest achievable SNDR for a given converter, which usually occurs for an input signal near full scale. Ideally, if a B-bit converter has no distortion and is noiseless, the peak SNDR is given by:

$$\mathrm{SNR}_{dB} < 6N + 1.76 \text{ dB},$$

The dynamic linearity of the ADC was measured by analyzing a Fast-Fourier Transform (FFT) of the output codes for the single input tone. 64k codes were used to compute the FFT using a Blackman windowing algorithm. The peak SNDR for a 100 kHz sine wave input was measured at 58.5 dB with a clock frequency of 14.3 MHz. For a 4 MHz sine wave input the peak SNDR was 53.7 dB. Figure 8.5 shows the SNDR versus input level at a clock frequency of 14.3 MHz.

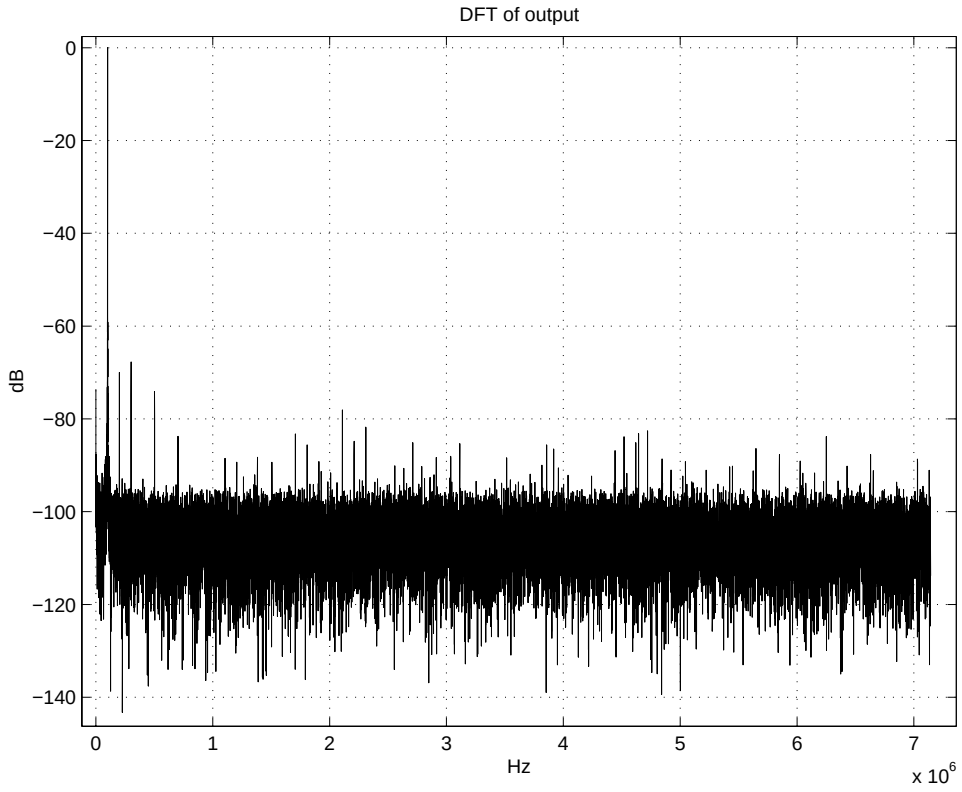


Figure 8.4 FFT of output codes for 100 kHz input

8.3 Static linearity

The static linearity characterizes the DC transfer function of the converter from the analog to the digital domain as the input is swept from $-V_{ref}$ to $+V_{ref}$. Integral non-linearity (INL) is defined as the maximum deviation of the transfer characteristic from a straight line fit. A low INL is important if large-signal linearity. Differential non-linearity (DNL) is defined by measuring the increase in input voltage it takes to transition between output codes. Ideally, this interval is 1LSB. DNL is defined as the actual interval minus 1LSB. A low DNL is important for small-signal linearity.

The static linearity of the ADC was measured using a code density test [21]. The input signal was a low-frequency (100 kHz) tone, and a continuous sample of

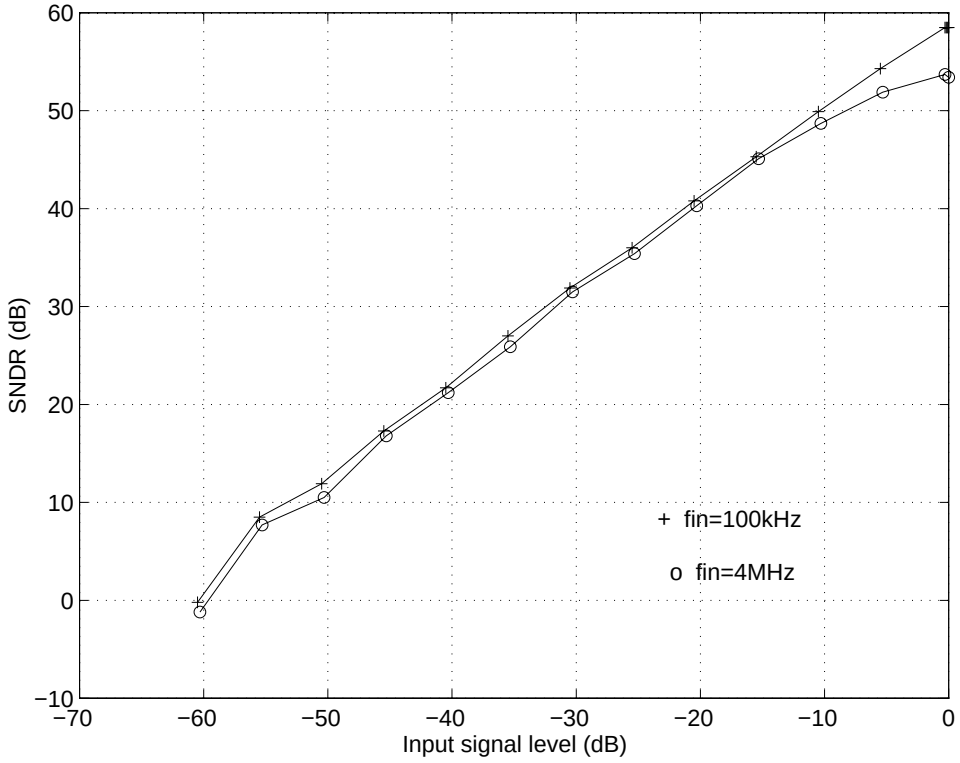


Figure 8.5 Output SNDR vs input signal level

over 1.04×10^6 output codes were analyzed. This implies that the measurement is accurate to within ± 0.1 LSB with confidence of greater than 99%. In general to compute the number of samples N required to measure the static linearity to within $\pm\beta$ LSB at the n bit level is given by:

$$N \geq \frac{Z^2_{\alpha/2}\pi 2^{n-1}}{\beta^2}$$

N	=	number of samples
β	=	maximum absolute error in LSB
n	=	number of bits in ADC
α	=	probability that error exceeds β
Z_x	=	$Z : F(Z) = 1 - x$
$F(Z)$	=	$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2} dt$

For example, $n = 10$ bits, $\beta = 0.1$ LSB, $\alpha = 0.01$, ($Z_{0.005} = 2.58$) gives:

$$N \geq \frac{(2.58)^2 \pi 2^9}{(0.1)^2} = 1.07 \times 10^6$$

The maximum differential non-linearity was measured at 0.5 LSB, and the maximum integral non-linearity using a least-squares fit was measured at 0.7 LSB. Figure 8.6 shows the DNL and INL versus output code. The regular variations in the INL characteristic are likely due to finite DC opamp gain error in the first three stages of the pipeline.

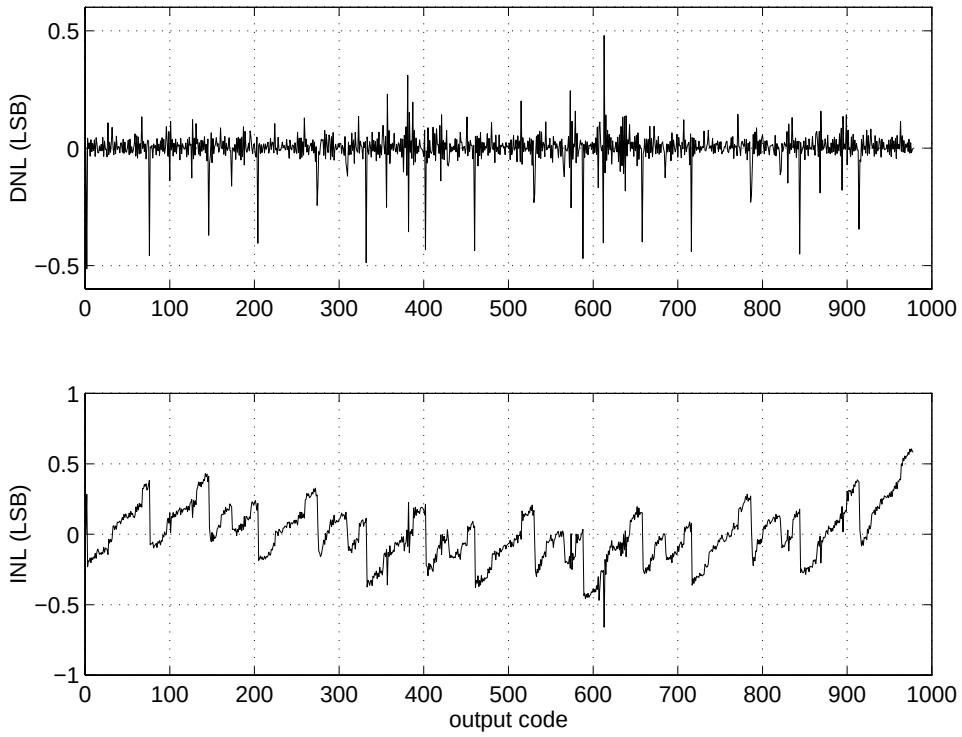


Figure 8.6 INL and DNL versus output code

8.4 Summary

The power consumption at a clock frequency of 14.3 MHz was 36 mW from a 1.5 V supply. Table 8.1 summarizes the measured results.

Table 8.1 Measured A/D performance at 25^o C

V_{dd}	1.5 V
Technology	0.6 μm CMOS $V_{tn} = 0.7V$, $V_{tp} = 0.9V$
Resolution	10 bits
Conversion rate	14.3 MS/s
Active area	2.3 x 2.5 mm^2
Input range	± 800 mV differential
Power dissipation	36 mW (no pad drivers)
DNL INL	0.5 LSB 0.7 LSB
SNDR	58.5 dB (f_{in} = 100 kHz)

Chapter 9

Conclusion

IT IS CLEAR that the operating voltages for CMOS technology will be reduced much faster than historical trends. The main drivers of this trend are the need for long-term product reliability and low-power, digital operation. Because integrated analog circuitry is becoming a smaller, but fundamentally necessary, portion of die area, it becomes more difficult to justify modifying the technology for analog needs. Therefore, the analog circuits must be modified to operate on low voltage. Furthermore, it is important that the MOS devices are not unduly stressed in the process, which would degrade product lifetime.

Fundamentally, operating switched-capacitor circuits on a lower supply voltage will tend to increase power consumption. In order to maintain the same dynamic range on a lower supply voltage requires a quadratic increase in sampling capacitance to reduce thermal noise. The required increase in bias current to maintain circuit bandwidth results in a net increase in the overall power consumption. Improvements in device f_T will mitigate this trend, but there is a fundamental tendency to increase power consumption. Furthermore, from a practical viewpoint, reduced voltage operation tends to increase the circuit complexity with increases power.

An examination of the breakdown and degradation phenomenon in a MOS device has shown that *relative* terminal potential determines device lifetime. Absolute potential is less critical. For switched-capacitor circuits, this implies that certain, restricted forms of bootstrapping for switches will not degrade lifetime. In particular, by referencing bootstrapped clocks to the input signal, switches can be operated in a reliable manner. Furthermore, for small areas, such as the analog portion of a large, mixed-signal application, the stress requirements can be somewhat relaxed without significantly degrading yield.

The specific research contributions of this work include (1) identifying the MOS device reliability issues that are relevant to switched-capacitor circuits, (2) introduction of a new bootstrap technique for operating MOS transmission gates on a low voltage supply without significantly degrading device lifetime, (3) development of low-voltage opamp design techniques. Low-voltage design techniques for the switched-capacitor building blocks have been demonstrated enabling the implementation of larger applications such as sample-and-holds, filters, and data converters. In particular, a 1.5 V, 10-bit, 14.3 MS/s, 36 mW pipeline analog-to-digital converter was implemented in a 0.6μm CMOS technology. It demonstrates that video-rate analog signal processing can be achieved at low voltage without requiring special enhancements to CMOS technology.

In the future, both digital and analog circuits will clearly need a finite amount of voltage to represent signals with. Furthermore, a minimum amount of margin is fundamentally necessary for noise margin. Thus, what is the minimum voltage required to operate these circuits? Due to current leakage considerations, a threshold voltage less than 0.4V is unlikely. To maintain circuit speed some margin of over-drive voltage is required beyond that. This is true for both digital and analog circuits as current is a function of $V_{gs} - V_t$. Therefore, it is possible that both circuit types can scale together. Due to the unscalability of V_t, however, it is unclear if voltage supplies will drop below 0.6V.

An interesting continuation of this work would be actual reliability testing of the bootstrapping techniques proposed. To first order, the principle of operation is sound, but actual verification would prove the concept. The experimental prototype was fully characterized for performance at 1.5V operation, however, lifetime extrapolation of the device was beyond the scope of the project. This could be done both with a reliability simulator of the bootstrap circuit, such as the Berkeley Reliability Tool [38], and with the actual accelerated stressing of a statistically significantly large population of test devices.

Bibliography

[1] —-, “MOS sampled data recursive filters using switched capacitor integrators,” *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 600-608, Dec. 1977.

[2] B. K. Ahuja, “An Improved Frequency Compensation Technique for CMOS Operational Amplifiers,” *IEEE J. Solid-State Cicuits*, vol. SC-18, no. 6, pp. 629-33, December 1983.

[3] D. J. Allstot, “A Precision Variable Supply CMOS Comparator,” *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 412-416, Dec. 1975.

[4] A. Baschirotto, R. Castello, “A 1-V 1.8-MHz CMOS switched-opamp SC filter with rail-to-rail output swing,” *IEEE J. Solid-State Circuits*, vol. 32, no. 12, Dec. 1997, pp. 1979-86.

[5] B. E. Boser, B. A. Wooley, “The Design of Sigma-Delta Modulation Analog-to-Digital Converters,” *IEEE J. Solid-State Circuits*, vol. SC-23, pp. 1298-1308, Dec. 1988.

[6] J. R. Brews, et al., “Generalized Guide to MOSFET Miniaturization,” *IEEE Electron Dev. Letts.*, EDL-1, p. 2, January 1980.

[7] T. L. Brooks, D.H. Robertson, D. F. Kelly, A. Del Muro, “A 16b Sigma Delta pipeline ADC with 2.5 MHz output data-rate,” *1997 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, pp. 208-9, San Francisco, 1997.

[8] J. C. Candy, G. C. Temes, *Oversampling Delta-Sigma Data Converters*, IEEE Press, New York, 1992.

[9] R. Castello, P. R. Gray, “A High-Performance Micropower Switched-Capacitor Filter,” *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp. 1122-1132, Dec. 1985.

[10] R. Castello, F. Montecchi, F. Rezzi, A. Baschirotto, “Low-Voltage Analog Filters,” *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, vol. 42, no. 11, pp. 827-40, Nov. 1995.

[11] J. T. Caves, M. A. Copeland, C. F. Rahim, S. D. Rosenbaum, “Sampled analog filtering using switched capacitors as resistor equivalents,” *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 592-599, Dec. 1977.

[12] T. Y. Chan, P. K. Ko, C. Hu, "Dependence of channel electric field on device scaling," *IEEE Electron Dev. Lett.*, EDL-6, p. 551, October 1985.

[13] A. P. Chandrakasan, "Low Power Digital CMOS Design," PhD Thesis, University of California, Berkeley, 1994. Available as ERL M94/65.

[14] I. C. Chen, S. Holland, C. Hu, "Electrical breakdown of thin gate and tunneling oxides," *IEEE Trans. Electron Dev.*, p. 413, Feb. 1985.

[15] T. Cho, *Low-power Low-voltage Analog-to-digital Converstion Techniques Using Pipelined Architectures*, PhD Thesis, University of California, Berkeley, 1995. Available as UCB/ERL M95/23.
tt http://kabuki.eecs.berkeley.edu/~tcho/Thesis1.pdf

[16] D. Cline, *Noise, Speed, and Power Trade-offs in Pipelined Analog to Digital Converters*, PhD Thesis, University of California, Berkeley, 1995. Available as UCB/ERL M95/94.

[17] J. Crols, M. Steyaert, "Switched-opamp: an approach to realize full CMOS switched-capacitor circuits at very low power supply voltages," *IEEE J. Solid-State Circuits*, vol. 29, no. 8, Aug. 1994, pp. 936-42.

[18] B. Davari, R. Dennard, G. Shahidi, "CMOS Scaling for High Performance and Low Power–The Next Ten Years," *Proceedings of the IEEE*, vol. 83, no. 4, pp. 595-606, April 1995.

[19] M. de Wit, "Sample and Hold Circuit and Methods," United States Patent, 5,170,075, December 8, 1992.

[20] T. H. Distefano, M. Shatzkes, "Impact ionization model for dielectric instability and breakdown," *Appl. Phys. Lett.*, vol. 25, no. 12, p. 685, Dec. 1974.

[21] J. Doernberg, H-S. Lee, D. A. Hodges, "Full-Speed Testing of A/D Converters," *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, Dec. 1984.

[22] J. Doernberg, P. R. Gray, D. A. Hodges, "A 10-Bit 5-Msample/s CMOS Two-Step Flash ADC," *IEEE J. Solid-State Circuits*, vol. SC-24, pp. 241-249.

[23] Stephen Anthony Edwards, *The Specification and Execution of Synchronous Reactive Systems*, PhD thesis, University of California, Berkeley, 1997. Available as UCB/ERL M97/31.
`http://ptolemy.eecs.berkeley.edu/papers/97/sedwardsThesis/`

[24] C. Enz, G. Temes, "Circuit Techniques for Reducing the Effects of Op-Amp Imperfections: Autozeroing, Correlated Double Sampling, and Chopper Stabilization," *Proceedings of the IEEE*, Vol. 84, No. 11, Nov. 1996.

[25] A. R. Feldman, "High-Speed, Low-Power Sigma-Delta Modulators for RF Baseband Channel Applications," PhD thesis, University of California, Berkeley, 1997. Available as UCB/ERL M97/62. `http://kabuki.eecs.berkeley.edu/~arnold/phd.pdf`

[26] D. L. Fried, "Analog sample-data filters," *IEEE J. Solid-State Circuits*, vol. SC-7, no. 4, pp. 302-4, Aug. 1972.

[27] R. Gharpurey, "Modeling and Analysis of Substrate Coupling in Integrated Circuits," PhD Thesis, University of California, Berkeley, 1995. Available as UCB/ERL M95/47.

[28] J. L. Gorecki, "Dynamic Input Sampling Switch for CDACs," United States Patent 5,084,634, January, 28, 1992.

[29] P. R. Gray, R. G. Meyer, *Analysis and Design of Analog Integrated Circuits–3rd Ed.*, John Wiley & Sons, Inc., 1993.

[30] R. Gregorian, K. W. Martin, G. C. Temes, "Switched-Capacitor Circuit Design," *Proc. IEEE*, vol. 71, no. 8, pp. 941-966, Aug. 1983.

[31] R. Gregorian, G. C. Temes, *Analog MOS Integrated Circuits for Signal Processing*, John Wiley & Sons, Inc., 1986.

[32] D. G. Haigh, C. Toumazou, S. J. Harrold, K. Steptoe, J. I. Sewell, R. B. Bayruns, "Design Optimization and Testing of a GaAs Switched-Capacitor Filter," *IEEE Trans. Circuits and Systems*, vol. 38, no. 8, pp. 825-37, Aug. 1991.

[33] K. Hirano, S. Nishimura, "Active RC filters containing periodically operated switches," *IEEE Trans. Circuit Theory*, vol. CT-19, May 1972.

[34] B. J. Hosticka, R. W. Brodersen, P. R. Gray, "MOS sampled data recursive filters using state variable techniques," *Proc. Int. Symp. on Circuits and Systems*, Phoenix, pp. 525-529, April 1977.

[35] C. Hu, "Gate Oxide Scaling Limits and Projection," *IEDM Technical Digest*, 1996 IEEE International Electron Devices Meeting, San Francisco, pp. 319-22, 1996.

[36] C. Hu, private communication, May, 1996.

[37] C. Hu, "Ultra-large-scale integration device scaling and reliability," *J. Vac. Sci. Technol. B*, vol. 12, no. 6, pp. 3237-41, Nov/Dec 1994.

[38] C. Hu, "IC reliability simulation," *IEEE J. Solid-State Circuits*, pp. 241-246, March 1992.

[39] G. M. Jacobs, D. J. Allstot, R. W. Brodersen, P. R. Gray, "Design Techniques for MOS Switched Capacitor Ladder Filters," *IEEE Trans. Circuits Syst.*, vol. CAS-25, no. 12, pp. 1014-1021, Dec. 1978.

[40] A. N. Karanicolas, H-S. Lee, K. L. Bacrania, "A 15-b 1-Msample/s Digitally Self-Calibrated Pipeline ADC," *IEEE J. Solid-State Circuits*, vol. 28, no. 12, pp. 1207-1215, Dec. 1993.

[41] N. Klein, P. Solomon, "Current runaway in insulators affected by impact ionization and recombination," *J. Appl. Phys.*, vol. 47, no. 10, p. 4364, 1976.

[42] P. K. Ko, R. S. Muller, C. Hu, "A Unified Model for Hot-Electron Currents in MOSFETs," *Tech. Dig. IEDM*, p. 600, 1981.

[43] C. A. Laber, C. F. Rahim, S. F. Dreyer, G. T. Uehara, P. T. Kwok, P. R. Gray, "Design Considerations for a High-Performance 3-μm CMOS Analog Standard-Cell Library," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 2, pp. 181-9, April 1987.

[44] J. C. Lee, I-C. Chen, C. Hu, "Modeling and Characterization of Gate Oxide Reliability," *IEEE Trans. Electron. Dev.*, vol. 35, no. 12, Dec. 1988.

[45] S. Lewis, *Video-rate Analog-to-digital Conversion Using Pipelined Architectures*, PhD Thesis, University of California, Berkeley, UCB/ERL 87/90, 1987.

[46] S. Lewis, et al, "10b 20Msample/s analog-to-digital converter," *IEEE J. Solid-State Circuits*, vol. 27, pp. 351-358, March 1992.

[47] S. H. Lewis, "Optimizing the stage resolution in pipelined, multistage, analog-to-digital converters for video-rate applications," *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 8, pp. 516-23, Aug. 1992.

[48] Y-M. Lin, B. Kim, P. R. Gray, "A 13-b, 2.5-MHz Self-Calibrated Pipelined A/D Converter in 3-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 628-36, April 1991.

[49] Y. M. Lin, *Performance Limitations on High-Resolution Video-rate Analog-to-Digital Interfaces*, PhD Thesis, University of California, Berkeley, UCB/ERL M90/55, 1990.

[50] F. Maloberti, F. Francesconi, P. Malcovati, O. J. A. P. Nys, "Design Considerations on Low-Voltage Low-Power Data Converters," *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, vol. 42, no. 11, pp. 853-63, November 1995.

[51] D. G. Marsh, B. K. Ahuja, T. Misawa, M. R. Dwarakanath, P. E. Fleisher, and V. R. Saari, "A single-chip CMOS PCM CODEC with filters," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 308-315, Aug. 1981.

[52] The Mathworks, Inc., 24 Prime Park Way, Natick, Mass., 01760-1500. `info@mathworks.com`.

[53] T. Matsuura, et. al., "A 92mW, 10b, 15MHz low-power CMOS ADC using analog double-sampled pipelining scheme," *Symposium on VLSI Circuits Dig. Tech. Papers*, pp. 98-99, June 1992.

[54] E. R. Minami, S. B. Kuusinen, E. Rosenbaum, P. K. Ko, C. Hu, "Circuit-Level Simulation of TDDB Failure in Digital CMOS Circuits," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 3, August 1995.

[55] R. Moazzami, C. Hu, "Projecting oxide reliability and optimizing burn-in," *IEEE Trans. Electron Devices*, vol. 37, pp. 1643-50, July 1990.

[56] A. Nagari, A. Baschirotto, F. Montecchi, R. Castello, "A 10.7-MHz BiCMOS high-Q double-sampled SC bandpass filter," *IEEE J. Solid-State Circuits*, vol. 32, no. 10, pp. 1491-8, Oct. 1997.

[57] K. Nakamura, et. al., "A 85mW, 10bit 40Ms/s ADC with decimated parallel architecture," *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 23.1.1-23.1.14, May 1994.

[58] K. Nakamura, M. Hotta, L. R. Carley, D. J. Allsot, "An 85 mW, 10 b, 40 Msample/s CMOS parallel-pipelined ADC," *IEEE J. Solid-State Circuits*, vol. 30, no. 3, pp. 173-83, March 1995.

[59] J. R. Naylor, M. A. Shill, "Bootstrapped FET Sampling Switch," United States Patent 5,172,019, December, 15, 1992.

[60] A. V. Oppenheim, R. W. Schafer, *Discrete-time Signal Processing*, Prentice-Hall, Inc., New Jersey, 1989.

[61] H. J. Orchard, "Inductorless Filters," *Electron. Lett.*, vol.2, pp.224-225, June 1966.

[62] G. Palmisano, G. Palumbo, "A Compensation Strategy for Two-Stage CMOS Opamps Based on Current Buffer," *IEEE Trans. Circuits and Systems–I: Fund. Theory and Applications*, vol. 44, no. 3, March 1997.

[63] M. J. M. Pelgrom, A. C. J. Duinmaijer, A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5 , pp. 1433-1439, Oct. 1989.

[64] J. Rabaey, et al., *Low Power Design Methodologies*, Kluwer Publishing, 1996.

[65] B. Razavi, B. A. Wooley, "Design Techniques for High-Speed High-Resolution Comparators," *IEEE J. Solid-State Circuits*, vol. SC-27, pp. 1916-1926, Dec. 1992.

[66] B. Razavi, *Principles of Data Conversion System Design*, IEEE Press, 1995.

[67] D. B. Ribner, M. A. Copeland, "Design Techniques for Cascoded CMOS Op Amps with Improved PSRR and Common-Mode Input Range," *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, pp. 919-25, December 1984.

[68] D. J. Sauer, "Constant Impedance Sampling Switch for an Analog to Digital Converter," United States Patent, 5,500,612, March, 19, 1996.

[69] B. J. Sheu, C. Hu, "Switch-Induced Error Voltage on a Switched-Capacitor," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 519-525, April 1984.

[70] Semiconductor Industry Association, "The National Technology Roadmap for Semiconductors," 1997. Available from `http://www.sematech.org/`.

[71] B-S. Song, M. F. Tompsett, K. R. Laksmikumar, "A 12-bit 1-Msample/s Capacitor-Averaging Pipelined A/D Converter," *IEEE J. Solid-State Circuits*, vol. SC-23, pp. 1324-1333, Dec. 1988.

[72] B-S. Song, "A 10.7-MHz switched-capacitor bandpass filter," *IEEE J. Solid-State Circuits*, vol. 24, no. 2, pp. 320-4, April 1989.

[73] D. K. Su, M. J. Loinaz, S. Masui, B. A. Wooley, "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits," *IEEE J. Solid-State Circuits*, vol. 28, no. 4, April 1993.

[74] S-W. Sun, P. G. Y. Tsui, "Limitation of CMOS Supply-Voltage Scaling by MOSFET Threshold-Voltage Variation," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, Aug. 1995.

[75] H. J. M. Veendrick, "The Behavior of Flip-Flops Used as Synchronizers and Prediction of Their Failure Rate," *IEEE J. Solid-State Circuits*, vol. SC-15, no. 2, April 1980.

[76] G. Wegmann, E. A. Vittoz, F. Rahali, "Charge Injection in Analog MOS Switches," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 1091-1097, Dec. 1987.

[77] W. B. Wilson, et al., "Measurement and Modeling of Charge Feedthrough in N-Channel MOS Analog Switches," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1206-1213, Dec. 1985.

[78] S. Wolf, *Silicon Processing for the VLSI Era, Vol. 3 - The Submicron MOSFET*, Lattice Press, Sunset Beach, California, 1995.

[79] D. R. Wolters, A. T. A. Zeegers-van Duijnhoven, "Breakdown of thin dielectrics," *Ext. Abs. Mtg. of Electrochem. Soc.*, p. 272, spring 1990.

[80] J-T. Wu, Y-H. Chang, K-L Chang, "1.2V CMOS Switched-Capacitor Circuits," *1996 IEEE Solid-State Circuits Conference Digest of Technical Papers*, San Francisco, pp. 388-9, February 1996.

[81] K. Yamakido, T. Suzuki, H. Shirasu, M. Tanaka, K. Yasunari, J. Sakaguchi, and S. Hagawara, "A single-chip CMOS filter/CODEC," *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 302-307, Aug. 1981.

[82] K. Yamabe, K. Taniguchi, "Time-dependent dielectric breakdown of thin thermally grown SiO_2 films," *IEEE Trans. Electron Dev.*, ED-32, p. 423, 1985.

[83] I. A. Young, P. R. Gray, D. A. Hodges, "Analog NMOS sampled data recursive filters," *Dig. Int. Solid-State Circuits Conf.*, Philadelphia, pp. 156-157, Feb. 1977.

[84] P. C-W. Yu, "Low-Power Design Techniques for Pipelined Analog-to-Digital Converters," Ph.D. Thesis, Massachusetts Institute of Technology, 1996.

Index

ADC, *see* analog-to-digital
analog-to-digital, 1
    flash, 26
    pipeline, 26, 81
    sigma-delta, 25
anti-alias filter, 4, 26
aperture error, *see* jitter
applications, 1, 17
architecture, 74, 81, 83
auto-zero, 8, 14

BERT, 58
biasing, 97
bilinear transformation, 22
binary weighted, 28
biquad, 18
body effect, 52
bottom-plate sampling, **7**
breakdown
    oxide, 35
    time-dependent dielectric, 35

capacitor, 94
    feedback, 66, 87
    integrating, 20
    linearity, 24, 87
    matching, 12, 20, 24, 27, 28, 82, 86, **86**
    parasitic, 8, 30, 55, 63
    sampling, 6, 66, 87
    trim, 86, **100**
cascode, 62, 72
channel length, 30
charge injection, 5, 52, 99
CHE, *see* hot-electron effects
circuit-board, 105
clock
    fall time, 89
    generator, **98**
    rise time, 89
    two-phase, 7, 13, 50, 59, 74, 85, 89
clock feed-through, 5
CMOS, **30**
    current, 31
    forecast, 30
    reliability, **35**, 43
    scaling, 31
    scaling limits, 41
codec, 1
common-mode feedback, 10, 59, 60, **73**
common-mode input, 102
common-mode output, 97
comparator, **13**, 27, **74**
    dynamic, 76
    mean time to failure, 15
    meta-stability, 15
    offset, 14, 79
    speed, 14
components, 20
cost, 1
current density, 32

DAC, 104, *see* digital-to-analog
damping, 68
depletion region, 34
DIBL, *see* drain-induced barrier lowering
differential non-linearity, 91, 110
digital correction, 76, 84, 104
digital-to-analog, 2, 26
    capacitor array, 28
discrete-time, 22
disk drive, 2
distortion, 20, 22, 90, 105
    sample-and-hold, *see* sample-and-hold
DNL, *see* differential non-linearity

drain-induced barrier lowering, 33
dynamic range, 45, 47, 107

electric field, 39

feedback factor, 8, 12, 66, 88
FFT, 107
filter, 1, **17**
    active RC, 17
    ladder, 22
    tutorial, 17
flow graph, 18

gain error, 8, 88, 90
gain stage, **10**, 27, 103
gate oxide, 30, 31, 35
    breakdown, 36
    defects, 37
gate-induced drain leakage, 34
GIDL, *see* gate-induced drain leakage
guard rings, 96

Hewlett Packard, 94
hot-electron effects, 38

impact ionization, 39
INL, *see* integral non-linearity
integral non-linearity, 110
integral non-linearity, 91
integration, 42
integrator, **13**, 54
    active RC, 18
interfaces, 1

jitter, **6**

latch, 14, 76
    time constant, 14
latch-up, 39, 51, 96
latency, 81
layout, 55, 94
leakage current, 33
lightly doped drain, 39, 40
lightly-doped, 55
linearity, 28, 86
    dynamic, 107
    static, 108
low power, 32, 33

matching
    capacitor, 82
    device, **78**
mean time to failure, *see* comparator
meta-stability, 76, 79, 100, *see* comparator

noise, 89, 91
    bandwidth, 72
    kT/C, 10, 59, 74
    opamp, **72**
    quantization, 25, 92
    thermal, 45

offset, 8, 55, 84
    cancellation, 48, 59
    comparator, **76**, *see* comparator
    sample-and-hold, 6
opamp, 8
    bandwidth, 8, 46, 60
    biasing, **60**
    gain, 60, **73**, 85, 88
    noise, **72**
    phase margin, 63
    power, 74
    settling time, 26, 27, 60, **63**, 68, 86, 89
    slew rate, 70
    topology, 60
open-loop, 63
output impedance, 62
oversampling, 25
oxide, *see* gate oxide

package, 94
pipeline
    errors, 89
poles, 65–67, 72, 89
power, 45, 47, 74, 85, 110
pre-amplifier, 14, 76
    bandwidth, 80
propagation delay, 100
prototype, 94
punch-through, 34

quantization error, 26
quantization noise, 92

residue, 27, 81, 83
resistor, 20

reverse breakdown, 44, 55
rise time, 54

sample-and-hold, **4**
  bottom-plate, 103
  distortion, 5
  offset, 6
  top-plate, 4
sampling frequency, 22, 25, 46
self-calibration, 27, 82, 86
Semiconductor Industry Association, 30
settling error, 89, 90
settling time, 74
signal-to-noise ratio, 6, 24, 91, 107
  ideal, 92
slew rate, 70
small-signal, 63, 66
SNR, *see* signal-to-noise ratio
step-response, 68
stress, 49
sub-ADC, 27, 74, 81, 83, 85, 104
sub-threshold, 41
  slope, 39
substrate contacts, 96
substrate coupling, 95
switch, 48
  high-swing, 49
  layout, 55
  low-distortion, 52
switched-capacitor, 4, 20

TDDB, *see* breakdown
technology, 94
testing, 105
thermal noise, 6, 91
threshold voltage, 33, 48, 94
  matching, 63
  scaling, 41
  variation, 42
tolerance, 89
transconductance, 46, 63
transient stress, 37, 40, 58
transmission gate, *see* switch
tunneling, 34, 36
  direct, 41
  Fowler-Nordheim, 36

voltage scaling
  low power, 32
  reliability, 35

wireless, 2

yield, 30, 38

zeros, 65, 67, 68

Colophon

This document was typeset with Times Roman and Courier using LaTeX2$_\epsilon$ on a Hewlett Packard workstation. Text editing was done using Emacs 20.2.2 and AUCTeX 9.0v. The figures were drawn using `idraw`. The LaTeX2$_\epsilon$ format was given to me by Stephen Edwards [23].