

Copyright © 1993, by the author(s).
All rights reserved.

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

**OPTIMUM PARTITIONING OF ANALOG AND
DIGITAL CIRCUITRY IN MIXED-SIGNAL
CIRCUITS FOR SIGNAL PROCESSING**

by

Ken A. Nishimura

Memorandum No. UCB/ERL M93/67

26 July 1993

**OPTIMUM PARTITIONING OF ANALOG AND
DIGITAL CIRCUITRY IN MIXED-SIGNAL
CIRCUITS FOR SIGNAL PROCESSING**

Copyright © 1993

by

Ken A. Nishimura

Memorandum No. UCB/ERL M93/67

26 July 1993

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Acknowledgments

First and foremost, I appreciate the guidance and support given by Professor Paul R. Gray, who even during his tenure as Chairman of the Department, found time to give me excellent advice in all aspects of my stay here at the University. I also would like to thank Professor Jan M. Rabaey and Professor Robert G. Meyer for their help in my research and academic pursuits.

My stay here at Berkeley would have been interminable had it not been for the support given to me by my fellow graduate students. Tim Hu, Weijie Yun, Robert Neff, Dave Cline and Cormac Conroy all gave me vital assistance when it was needed. Special thanks goes to (now Professor) Greg Uehara — his insight into system and circuit design was instrumental in the successful completion of this project. Thanks goes to Eric Boskin for his help in device characterization. Brian Richards and Ken Lutz provided needed support to test the prototype chips.

Fellow graduate students not only provided academic assistance, but also provided opportunities for much needed extra-curricular activities. The “lunch gang” — Henry C., Chris, Eric, Mark, and Jeff — helped break the monotony of many days of work. Conversations with Sherry and Henry S. helped pass the time more easily.

Special thanks to my parents and sister, for they provided the support throughout my life so that I could be here to write this thesis.

This work was supported by the NSF under contract number MIP-8911017. I was supported by the Fannie and John Hertz Foundation during the entirety of my graduate studies. Their financial support was invaluable in providing a secure environment for conducting my studies.

Finally, I thank my girlfriend Judy for her unending patience and understanding. Contrary to popular belief, this thesis was finished sooner rather than later as a result of her positive encouragement, support and love.

Table of Contents

CHAPTER 1

Introduction

1.1	Background and Motivation.....	1
1.2	Thesis Organization.....	2

CHAPTER 2

Analog vs. Digital Implementations

2.1	Introduction.....	5
2.2	Continuous vs. Discrete Time.....	6
	2.2.1 Errors in the Sampling Process.....	8
	2.2.1.1 Sample Acquisition Delay.....	8
	2.2.1.2 Finite Track Mode Bandwidth.....	9
	2.2.1.3 Aperture Delay.....	10
	2.2.1.4 Clock Feedthrough.....	10
	2.2.2 Thermal Noise of Sampling.....	11
	2.2.2.1 Oversampling.....	12
	2.2.3 Maximum Sampling Rate.....	13
2.3	Analog-Digital Conversion.....	14
2.4	Analog vs. Digital Implementations.....	16
	2.4.1 Determination of Circuit Building Blocks.....	16
	2.4.2 Switched-Capacitor Charge Transfer Circuits.....	17
	2.4.3 Fundamental Limits of Traditional Switched Capacitor Integrators.....	18
	2.4.4 Practical Power Constrained Switched Capacitor Charge Transfer Circuits.....	22
	2.4.4.1 Subthreshold Operation of Charge Transfer Circuit.....	24
	2.4.4.2 Optimal Design of Switched-Capacitor Charge Transfer Circuits.....	28
	2.4.4.3 Minimum Achievable Area of Switched Capacitor Charge Transfer Circuits ...	35
	2.4.5 Digital Equivalents of Charge Transfer Circuits.....	36
	2.4.5.1 Power Characteristics of Digital Circuits.....	37
	2.4.5.2 Digital Circuits to Realize the MAC Function.....	37
	2.4.5.3 Lower Limits of Power Consumption for Digital Circuits.....	38
	2.4.5.4 Power Consumption of Practical Digital MAC Circuits.....	39
	2.4.5.5 Area of Practical Digital MAC Circuits.....	40
2.5	Conclusions.....	42
2.6	References for Figure 2.5.....	44
2.7	References.....	44

CHAPTER 3

Overview of NTSC System

3.1	Introduction.....	47
3.2	Raster Scanning of an Image	47
3.2.1	Spectral Analysis of the Raster Scanned Image	49
3.3	Principles of Color Video.....	50
3.3.1	Additive Color Theory.....	50
3.3.2	Color Vision Characteristics of the Human Eye.....	52
3.4	The NTSC Color Video Signal	52
3.4.1	Methods of Transmitting Color Information	52
3.4.2	The NTSC Color Encoding System.....	54
3.5	PAL and SECAM Color Encoding Methods	56
3.5.1	PAL	57
3.5.2	SECAM	58
3.6	Decoding NTSC Signals.....	58
3.6.1	Colorburst Extraction	58
3.6.2	Chrominance demodulation.....	59
3.6.3	De-matrixing of the chrominance signal	60
3.7	Y/C Separation.....	60
3.7.1	Effects of Incomplete Separation.....	60
3.7.2	Bandpass separation	61
3.7.3	Comb Filters for Y/C Separation.....	61
3.7.3.1	2H Comb.Filters.....	63
3.7.3.2	Adaptive Comb Filters	64
3.7.4	Delay Elements for Comb Filters	65
3.7.4.1	Bulk Delay Devices	65
3.7.4.2	CCD Delay Lines	66
3.7.4.3	Digital Delay Lines	66
3.7.4.4	Analog Line Memories	67
3.8	References.....	67

CHAPTER 4

Prototype Comb Filter

4.1	Introduction.....	69
4.2	System-Level Design Considerations	69
4.3	System Overview	70
4.3.1	Operational Overview.....	71
4.4	Description of Silicon Technology	72
4.5	Architecture of the Analog-RAM.....	73
4.5.1	Storing Information onto the Analog RAM	74
4.5.2	Reading Stored Values from the Analog-RAM.....	75
4.5.3	Overall Read/Write Architecture of Analog-RAM	75
4.6	Topology of the Storage Array	77

4.6.1	Contents of the Cell	78
4.6.2	Two-Dimensional Array	79
4.6.3	Single Switch Storage Cell	79
4.6.4	Cell Addressing.....	80
4.6.5	Sizing of Storage Capacitor	83
4.6.5.1	Thermal Noise	84
4.6.5.2	Matching Requirements	84
4.6.5.3	Clock Feedthrough Rejection.....	85
4.6.5.4	Loop Stability and Dynamics.....	86
4.6.6	Select Switch Sizing	90
4.6.7	Cell Layout	90
4.6.8	Row Multiplexers	92
4.6.8.1	Mux Switch Design.....	92
4.6.8.2	Multiplexer Logic.....	94
4.6.8.3	Auxiliary Output for Shunt Capacitor.....	94
4.6.8.4	Read Switch Delay Circuit.....	94
4.6.8.5	Layout Considerations for the Multiplexers.....	95
4.6.9	Assembly of the Storage Array.....	95
4.7	Design of the OTA	96
4.7.1	Introduction.....	96
4.7.2	Amplifier Specifications	97
4.7.2.1	Effect of Finite Amplifier Gain.....	97
4.7.2.2	Secondary Requirements.....	98
4.7.3	Basic Single-Stage Amplifiers.....	98
4.7.3.1	Unfolded "Telescopic" Cascode OTA	98
4.7.3.2	Folded Cascode OTA.....	100
4.7.3.3	Class AB OTAs and Dynamic Biasing	101
4.7.4	Two Stage OTAs.....	102
4.7.5	Single Stage Amplifier with Preamp Stage.....	103
4.7.5.1	Low-Gain Wide-Bandwidth Preamp.....	104
4.7.5.2	Complete OTA Circuit.....	105
4.7.5.3	Layout Considerations.....	108
4.8	Bias Generator.....	109
4.8.1	Introduction.....	109
4.8.2	Reference Voltage Generation	109
4.8.3	Output Common Mode Voltage Generator.....	111
4.9	Auxiliary Feedback Capacitor Stage.....	112
4.9.1	Introduction.....	112
4.9.2	Write Cycle Auxiliary Capacitor	112
4.9.3	Effect of Auxiliary Capacitor.....	114
4.10	S/H and Summer Stages.....	115
4.10.1	Introduction.....	115
4.10.2	Input S/H Stage.....	115
4.10.3	Intermediate S/H Stage	117
4.10.4	Output Scaling and Summing Stage.....	117
4.10.4.1	Effect of Mismatch in Scaling and Non-Unity Gain in Delay Line	118
4.10.4.2	Compensation of Mismatch in Coefficients	120
4.11	Output Buffers.....	122

4.12	Clock and Address Generation	124
4.12.1	Introduction	124
4.12.2	Analog Clock Generation	125
4.12.3	Cell Selection and Address Generation	126
4.12.3.1	Horizontal Shift Register	127
4.12.3.2	Vertical Shift Register	129
4.12.3.3	Row Increment Logic	130
4.12.3.4	Row Reset Lockout Logic	131
4.12.3.5	Power On Reset	132
4.12.3.6	Digital Clock Generation	132
4.13	Circuit Layout	133
4.14	References	135

CHAPTER 5

Experimental Results

5.1	Introduction	137
5.2	Test Fixture	137
5.3	Measured Performance	141
5.3.1	Comb Notch Depth	142
5.3.2	Composite Channel Frequency Response	143
5.3.3	Power Consumption	143
5.3.4	Dynamic Range (Random Noise)	144
5.3.5	Fixed Pattern Noise (220.3 KHz)	144
5.3.6	Full Scale Input and Linearity	144
5.3.7	Differential Gain and Phase	145
5.3.8	Active Area	146
5.3.8.1	Storage Array Area	146
5.3.8.2	Address Generators	146
5.3.8.3	Analog Read/Write Amplifiers	147
5.3.8.4	Clock Generator and Reset Circuitry	147
5.4	Functional Performance	147
5.5	Visual Inspection	148
5.6	Additional Tests	150
5.6.1	Operation with Reduced Supply Voltage	150
5.6.2	Operation with Faster Clocks	151
5.7	Future Modifications to Prototype Circuit	151
5.8	Conclusion	152
5.9	References	152

CHAPTER 6

Conclusions

6.1	Summary of Research Results	153
6.2	Projected Performance in Scaled Technologies	154

Abstract

**Optimum Partitioning of Analog and Digital Circuitry in
Mixed-Signal Circuits for Signal Processing**

by

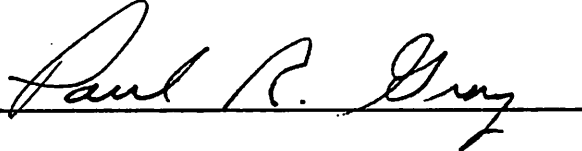
Ken A. Nishimura**Doctor of Philosophy in Engineering-Electrical Engineering
and Computer Sciences****University of California at Berkeley****Professor Paul R. Gray, Chair**

Advances in digital signal processing (DSP) technologies have resulted in an increased proportion of signal processing tasks being performed in the digital domain. However, increased interest in low-power circuitry and economic factors have placed pressure to minimize power dissipation and silicon area in such circuits. An examination of the relative strengths and weaknesses of analog versus digital circuits is made in this dissertation. Comparisons of power dissipation and silicon area based on fundamental limits and practical considerations as a function of signal bandwidth and dynamic range are made. The final objective is to determine the range of frequencies and dynamic range for which analog processing is more efficient than digital processing.

A monolithic analog video comb filter has been fabricated in 1.2- μm CMOS technology to demonstrate the area and power advantages of analog processing for video-rate signals. This chip, which dissipates 170 mW and consumes 11.7 mm², requires only a single $4 f_{sc}$ clock and reference current and no adjustments. The chip which uses a fully differential architecture, achieves a dynamic range of 51 dB and a comb notch depth of > 28 dB. Fixed pattern noise is less than 55 dB below full scale. Circuit techniques to mitigate the effects of large parasitic capacitances are introduced.

This dissertation shows that considering only the power and area of the actual processing circuitry, signal processing tasks with modest (< 60dB) dynamic range requirements are more effi-

ciently undertaken with analog processing compared to equivalent digital processing techniques. This result is derived from an examination of fundamental limits and demonstrated using numerical examples representative of a 1 μm technology with a 3.3V supply. Specifically, sampled-data analog processing of NTSC video signals is achievable using standard CMOS technologies, allowing the use of such techniques within a larger mixed-signal integrated circuit.

A handwritten signature in black ink, reading "Paul R. Gray", is written over a solid horizontal line.

Paul R. Gray, Chairman of Committee

CHAPTER 1

Introduction

1.1 Background and Motivation

The signals present within traditional analog circuitry are continuous valued, continuous time physical quantities such as voltage and current. In signal processing systems, such signals are often replicas of real physical quantities. For example, in a voiceband circuit, the voltage at a circuit node may represent the sound pressure level of speech. Because of this property, signal processing with analog circuitry is conceptually straightforward. However, due to non-idealities present in any real system, signals within analog circuitry is susceptible to degradation, especially by additive noise. Meanwhile, improvements in integrated circuit technologies have allowed the production of low-cost, high-density digital circuits. This in turn, coupled with analog-to-digital (A/D) conversion, which transforms an analog signal into a discrete valued, discrete time representation, has brought about a viable alternative to analog signal processing. Once in the digital domain, the signals are immune to factors which would degrade analog signals — this results, for example, in the increased fidelity of digitally recorded audio media over analog phonographs

Recent advances in A/D technology, coupled with advances in digital microprocessor design have allowed digital signal processing (DSP) to take over many signal processing tasks formally performed using analog techniques. Because the ease of testability, software programmability, relative ease of digital design and the advantages of digital representation of signals, the proportion of signal processing performed using DSP continues to increase.

While the advantages of using digital technology are fairly clear, the resulting solutions, when integrated on silicon, are not necessarily optimal with respect to certain key parameters such as power consumption and silicon die area. The increased interest in portable equipment places renewed emphasis on low-power consumption, while economic factors always favor a smaller silicon die area. As a result, a determination of the relative strengths and weaknesses of analog vs. digital circuits as a function of signal bandwidth and dynamic range will allow designers to reach a more optimal system architecture when making a choice between analog and digital processing. This determination will be made on the basis of fundamental physical limits taking into account practical limitations that are associated with the production of real components.

Recent interest in multimedia applications in the computing environment have created a need for signal processing circuits for video signals. For reasons outlined above, the trend has been to digitize the video signal and relegate subsequent processing to DSP circuitry. Although this results in a workable solution, based on the analysis conducted in this thesis, analog circuitry should provide a more efficient solution for a certain class of video processing tasks. One such task is that of comb filtering, an advanced method of luminance-chrominance (Y/C) separation used to decode NTSC video signals for use within a computing environment.

Advances in circuit design techniques makes it possible to overcome many of the non-idealities which affect analog signal processing. Continued research in this area can result in increased use of analog processing with associated reduction in power dissipation and silicon area. The use of standard CMOS technologies allows high levels of integration with the potential of combining both analog and digital processing on a single chip for maximum efficiency.

1.2 Thesis Organization

An examination of the relative strengths and weaknesses of analog vs. digital circuitry is presented in this thesis. Based on the results, a prototype video processing circuit has been fabricated and tested, and the results reported. Chapter 2 presents the analysis of analog and digital circuitry and attempts to determine preferred regions of operation for both types of circuits. Chapter 3 gives a brief background of the NTSC video standard, with emphasis on concepts necessary to understand the function of the prototype circuit. Chapter 4 describes the design of the prototype circuit. Discussions of the effect of non-idealities such as noise, amplifier settling

behavior, and parasitic elements are included. Trade-offs in the design process and circuit solutions are examined. Experimental results from the prototype converters are presented in Chapter 5. A conclusion and a summary of research results are given in Chapter 6.

Introduction

The following text is a very faint and illegible scan of an introduction section. It appears to contain several lines of text, but the characters are too light and blurry to be transcribed accurately. The text likely discusses the purpose and scope of the document.

CHAPTER 2

Analog vs. Digital Implementations

2.1 Introduction

The recent advent of digital signal processing (DSP) combined with the explosive increase in computational power has resulted in a marked shift away from analog signal systems to those that subject the incoming signal to a data converter and proceed in the digital domain thereafter. In addition to the rapidly dropping cost of such alternatives, convenience and ease of implementation contributes to the popularity of digital processing. The design process for digital circuits continues to be more efficient than that of analog circuits, and is helped by a plethora of automatic synthesis tools, while analog circuits still make use of large amounts of hand design and layout. In addition, the ease with which digital circuits can be programmed allows the use of adaptive algorithms and software control. Finally, digital circuits are easier to test than similar analog circuits, reducing production costs. As a result, digital implementations are becoming increasingly popular, even though they may not be optimal when other metrics such as elegance, power consumption and chip area are considered.

This chapter will investigate the limitations of analog signal processing, and compare it to equivalent digital implementations, mainly on the basis of chip area and power consumption. Theoretical limits based on thermal noise and quantum limits will be explored, followed by more practical implementations. Throughout, the effect of scaled technologies will play an important part. In end, an attempt to fundamental limits that govern analog and digital techniques as a function of speed and required dynamic range will be made.

Analog signal processors will always be present, even in the limit of complete superiority of digital methods strictly due to the nature of the real world. Measurable physical quanta are all analog quantities, though more often than not, the result is expressed digitally. Even though many forms of data transmission and storage are “digital,” the actual transmitted or stored quantities are electric or magnetic fields, which are inherently analog. Figure 2.1 illustrates one view of analog signal processors in the mixed-signal (analog and digital) world [1]. In the center lies the digital VLSI signal processing system (DSP), which is surrounded by the shell of analog signal processors; the outside represents the real world. The limits of the diagram range from no digital processing (i.e. pure analog system), to that of a thin eggshell consisting strictly of data converters (A/D and D/A blocks). The challenge is to determine the optimum thickness of the shell and what to place within the shell and what to relegate to DSP.

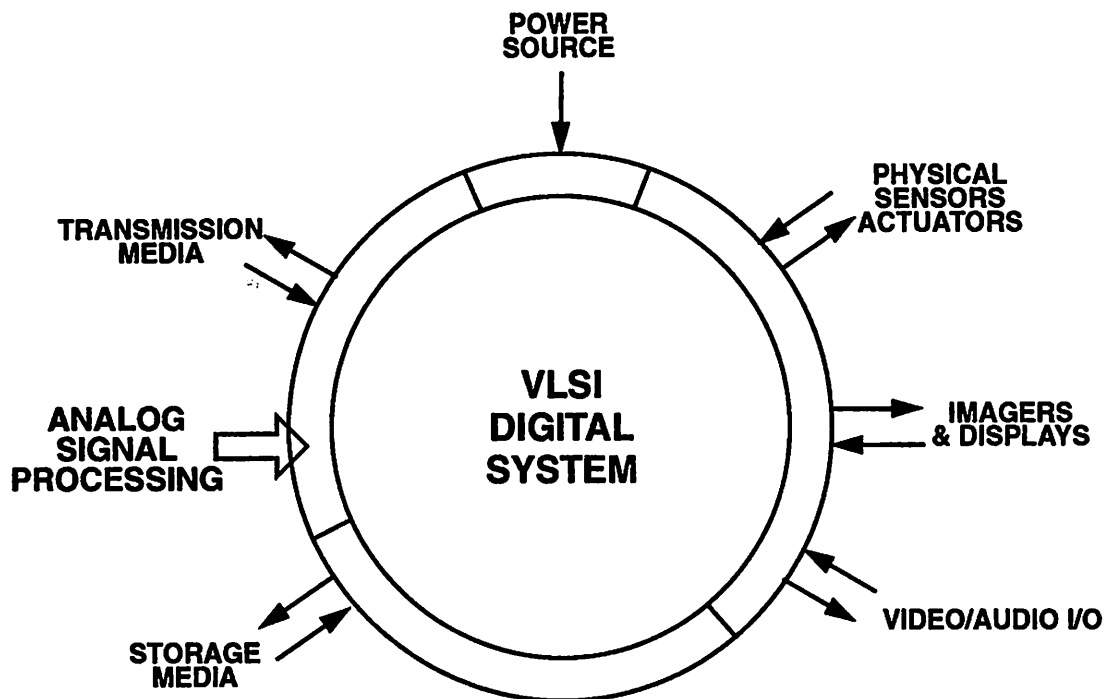


Figure 2.1 Role of Analog Processors in a Mixed-Signal Environment

2.2 Continuous vs. Discrete Time

Digital signal processors are discrete-time systems due to the inherent clocked nature of digital computation units. On the other hand, analog processors have a choice between continuous

and discrete time implementations. The choice between the two is largely one of implementation, as a system that is designed in continuous time can usually be implemented in discrete time and vice versa. Discrete time processing is very popular in monolithic analog processors because the desired circuit function can be obtained without using precision devices, which are difficult to obtain in the integrated circuit environment. However, all discrete time systems are limited by the Nyquist criterion, which limits the signal bandwidth to one-half the sampling rate [2]. Thus, the first separating factor to be considered is whether the signal frequency is low enough to be sampled and acted upon.

To begin, assume that the signal processing units are infinitely fast, and the sampling operation is the limiting factor. As the quantity being sampled is analog-continuous time, the only possible solution is that of an analog sampling system. Most discrete time systems expect that the sampling operation be performed as a sample and hold (S/H), where the input is sampled at regular intervals and the value held until the next sample (Figure 2.2). This S/H is almost universally performed using a switch and a storage element. In the CMOS environment, a MOS switch

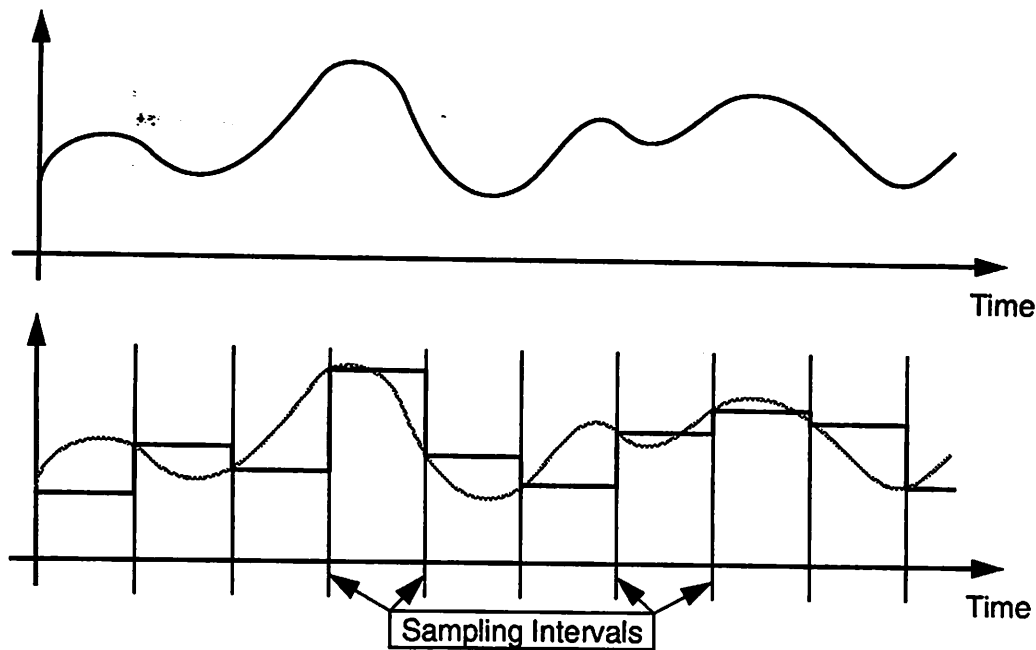


Figure 2.2 Sample-and-Hold Operation: Input (top), Sampled Output (bottom)

connected to a sampling capacitor is used to realize the S/H stage. If the sampling clock is perfectly timed and infinitely fast, in line with the infinitely fast processor assumption, the speed

limitation of this rudimentary S/H will be governed by the speed of the MOS switch. Figure 2.3 shows a simple implementation of such a circuit.

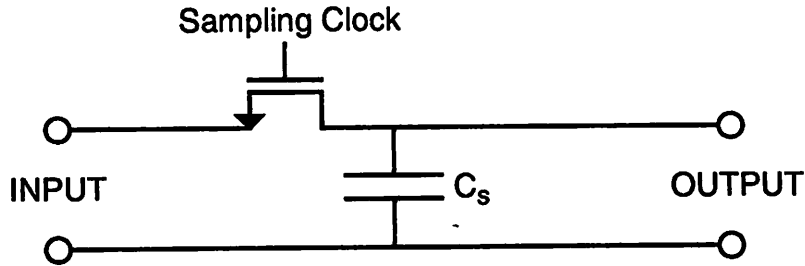


Figure 2.3 MOS Sample and Hold

A closer investigation of the actual sampling process shows that there is a finite time between the sampling clock and the time when the held output represents the value of the input at the sampling instant. Moreover, there is an additive noise and an offset contribution due to the sampling switch itself. These errors present themselves as limitations to processing using discrete time.

2.2.1 Errors in the Sampling Process

Expansion of the waveforms shown in Figure 2.2 will reveal non-idealities in the sampling process. The sampling waveform shown in Figure 2.4 outlines the most significant errors which affect the sampling operation. The output of the S/H circuit attempts to track the input during the period the switch is closed, and will attempt to hold the exact value of the input the instant the switch is opened. However, due to the finite resistance of the switch when turned on, and the nature of the MOS switch, errors such as acquisition time, track mode bandwidth limitations, aperture delay, and clock feedthrough are introduced.

2.2.1.1 Sample Acquisition Delay

Figure 2.4 shows that the output of the S/H does not instantaneously track the input after the switch has been closed. Due to the finite resistance of the switch in its on state, the sampling capacitor forms a low-pass filter network. Since in general the previously held value is different than the current input, the output of the S/H circuit will follow a step response of a single-pole system. This response, which has an exponentially decreasing error with time constant $R_{sw}C_s$, sets a minimum time between the start of the acquisition cycle and the next hold cycle, and hence

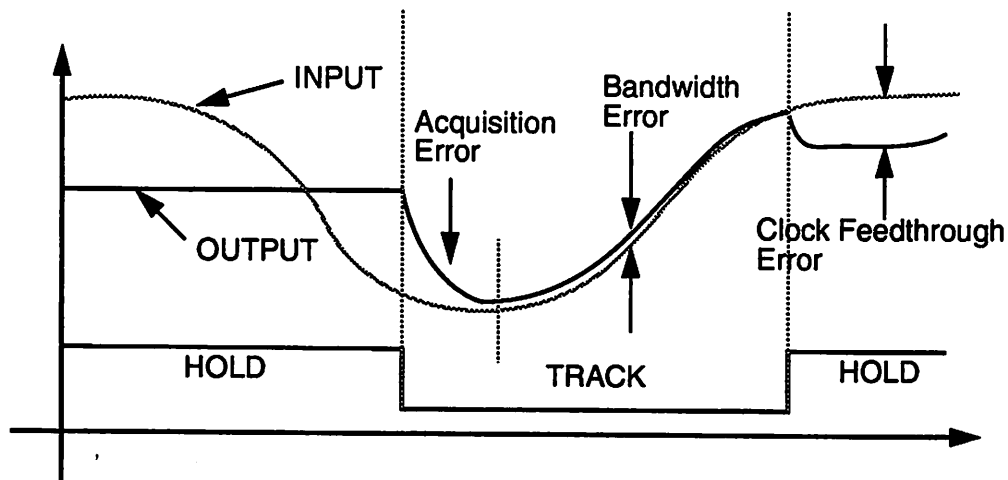


Figure 2.4 MOS Sample and Hold Errors

the maximum sampling rate [3]. For the acquisition error to be less than $1/2$ LSB of a N -bit equivalent system, the minimum acquisition of settling time is $\ln(2)NR_{sw}C_s$.

2.2.1.2 Finite Track Mode Bandwidth

Subsequent to the settling of the acquisition transient, the S/H circuit enters the track mode, where the output attempts to follow the input. However, the same low-pass response which caused a non-zero acquisition delay also results in a finite tracking error. Here, the circuit acts as a simple single-pole low-pass filter. For most reasonable circuits, satisfying the acquisition time requirements results in a minimal impact on the amplitude of the tracking output. However, there is a finite time delay, and the phase shift can cause unacceptable errors. Moreover, because this system is not “constant-delay,” there is a frequency dependent phase shift. Phase modulated systems such as NTSC composite signals will be affected by this error. The phase shift is a given frequency is:

$$\phi = -\tan^{-1}\left(\frac{f}{2\pi R_{sw}C_s}\right) \quad (2.1)$$

Because the acquisition time requirements are typically more stringent than those imposed by these phase shifts, the $R_{sw}C_s$ time constant is usually much higher than the signal bandwidth of interest. As a result, the phase shift is nearly linear over the range of interest, and is usually not a problem. Both this error and acquisition time errors can be reduced by making the sampling switch larger, thereby reducing R_{sw} . However, this strategy comes at the expense of heavier clock

loads and larger clock feedthrough errors, which forces a compromise in selecting the size of these MOS switches.

2.2.1.3 Aperture Delay

Because of the finite switching time of the MOS device, there is a delay between the edge of the gate drive and the actual hold instant. This delay is known as the aperture delay, and results in a small error between the input at the instant of the sampling clock and the output of the S/H circuit in excess of the above errors. More serious, however, is a variation in this delay, called aperture jitter. This results in a random error being introduced into the signal. Assuming that the input is a sinusoid, the aperture jitter is most pronounced at the zero crossing of the sinusoid. Again, for a 1/2 LSB error for a N -bit equivalent system, the jitter must be less than $\frac{1}{2\pi f 2^N}$. This source of error becomes a serious issue in multi-phase clocked sampling systems, where the skew in the sampling clocks results in a relatively large time error.

2.2.1.4 Clock Feedthrough

At the instant the switch is opened, there is a small step in the output of the S/H circuit which causes an error. There are two main sources of these errors, commonly lumped as clock feedthrough, gate overlap capacitance and channel charge injection. Every MOS switch has a small parasitic capacitance between the gate the source-drain terminals. As a result, when the gate is switched from the on to the off state, a capacitive divider between the gate overlap capacitance and the sampling capacitor (plus parasitics) is formed. Since the gate typically swings the entire supply voltage, this error can be substantial for small signal voltages even with a small overlap capacitance. Channel charge injection is more serious and is the major source of this error. By nature, the MOS switch has a collection of free charge in the channel while the switch is on with a magnitude $Q_{ch} = C_{ox}WL(V_{GS}-V_T)$. When the switch is turned off, this charge must go somewhere, and a portion inevitably is injected into the sampling capacitor. The exact proportion of the charge which flows into C_s is not deterministic, but is affected by the speed of the sampling clock and the relative impedances seen by the source and drain terminals [4].

A substantial problem with clock feedthrough is that the magnitude of the error for a given circuit is dependent on the voltage across the switch because both V_{GS} and V_T are a function of the signal voltage. Thus, a simple S/H circuit of Figure 2.3 will result in a clock feedthrough error which is signal dependent. Therefore, the error will manifest itself as harmonic distortion, which is highly undesirable. To eliminate this, techniques such as *bottom-plate* sampling [5]

which use a switch with constant potential across the device for sampling are used to remove this source of harmonic distortion. The resultant error is now independent of the signal, and adds a small constant pedestal to the output. The magnitude of this error is proportional to the size of the switch, and sets a practical limit on the switch size. Thus, as mentioned earlier, a compromise is required between the reducing the switch resistance and reducing the clock feedthrough error. A fully differential system aids greatly in removing the effects of this error for the pedestal is a common-mode error which can be removed unless the pedestal is so large that it drives the subsequent stage into saturation.

2.2.2 Thermal Noise of Sampling

In addition to the deterministic errors introduced by the sampling process, random noise is a substantial contributor to signal degradation in the sampling process. Consider the simple S/H stage shown in Figure 2.3. During the period when the switch is on, the MOS switch can be replaced by a finite resistance due to the channel present. The resultant circuit is a simple single-pole RC circuit. The resistance introduces a thermal noise, whose noise power density is the familiar $4kTR\Delta f$. However, the RC combination acts as a low pass filter which limits the bandwidth of interest. It can be shown that an equivalent noise-bandwidth can be found for such a circuit, and is equal to $1/(4RC)$. Thus, the noise power at the output terminals is simply kT/C . Note that this is independent of the size of the switch — an increase in the channel resistance results in a compensating reduction of the noise bandwidth.

If the switch is now turned off to take a sample, the value of the noise at the instant the switch turns off is held on the capacitor. A simplistic view of showing that the value of the noise at the instant is kT/C is to consider the instant of time the switch is turning off. Just before the switch opens, the channel resistance will be approaching infinity. However, as shown above, the noise power remains constant at kT/C , such that when the switch opens, that noise will be held on the capacitor. Another approach is to make use of the fact that one is sampling a signal which happens to be the thermal noise of the channel of the switch filtered by the sampling capacitor. The noise is spread over a bandwidth of $1/(4R_{sw}C_s)$, with a spectral density near DC of $2kTR_{sw}$, assuming a two sided spectrum. Assuming that the switch is closed for a period many times the time constant of the filter (a condition necessary for settling), the thermal noise is substantially undersampled. This results in aliasing, where the noise components from the switch are repli-

cated, shifted in frequency and summed. Because the noise is white Gaussian noise, each component is uncorrelated and the sum is easily found to be kT/C , with a spectral density of $kT/f_s C$ [5].

This so called “ kT/C noise” is a fundamental limit of analog signal processing systems [6]. Although it is not an issue with systems that have moderate dynamic range requirements, the noise voltage is a function of $\sqrt{\frac{kT}{C}}$, and the capacitor size must be quadrupled for a doubling of dynamic range. Thus, it becomes a severe limitation when systems with a large dynamic range is desired. For so called “Nyquist” sampled systems, those whose sampling rates are roughly double the maximum signal frequency, achieving resolutions in excess of 100 dB is very difficult, and the resulting circuit consumes large amounts of silicon area and power due to the size of the capacitors in the circuit.

2.2.2.1 Oversampling

The fact that the thermal noise associated with sampled data systems is uniformly distributed between DC and $f_s/2$ allows for a technique known as oversampling to reduce the effective noise floor. Simply, if the entire system is run faster than required to meet the Nyquist criterion, then the noise floor in the signal band will be reduced. Of course, the benefits of oversampling require that the signal bandwidth be limited to the frequency of interest, which can be a daunting task in and of itself. The most popular example of this technique is that of oversampled A/D converters, commonly known as Σ - Δ converters. These data converters have been made with dynamic ranges in excess of 120 dB [7], yet use reasonably small capacitors. Because of the high degree of oversampling required to noise shape the quantization noise, typically in the 100 to 200 range, the noise density in the signal bandwidth is reduced by the same factor, allowing the use of a capacitor that is $1/m$ the size that would be required in a Nyquist sampled system, where m is the oversampling ratio. The output of the Σ - Δ converter is a bitstream which is then digitally filtered to remove out of band components. Thus, in addition to removing the quantization noise which has been shifted into the higher frequencies, the bulk of the thermal noise which is distributed between DC and half the oversampled sampling rate is removed.

Use of oversampling techniques in systems where both input and output are in the analog domain is more difficult because of the high degree of filtering required. The low-pass filters in Σ - Δ converters are typically multi-stage high-order FIR filters with hundreds of taps. Realizing this type of transfer function in the sampled data analog domain without introducing additional baseband thermal noise is difficult, if not impossible. Thus, the use of oversampling is limited to

special cases such as Σ - Δ converters. Finally, oversampling places a stricter requirement on the input S/H stage, which places an ultimate limit in the use of this technique.

2.2.3 Maximum Sampling Rate

The sources of errors discussed in the previous sections can be combined to obtain a practical upper limit on the sampling rate of a MOS switch based system. On one hand is the inherent $R_{sw}C_s$ time constant which can be improved by making the capacitor smaller or the switch bigger, while the clock feedthrough error is improved by making the switch smaller and the capacitor bigger. Finally, kT/C noise, which is independent of the switch size but is improved with a larger C_s factors into the sampling rate limitation.

The maximum sampling rate itself can be increased without limit by making the capacitor very small with a large switch; however, the errors would badly corrupt the signal making the circuit unusable. A more meaningful metric is that of the sampling rate-dynamic range product, or the throughput of the system. It is common to measure dynamic range as N -bit equivalent, even for pure analog systems to allow easy comparison to digital systems, and this notation will be used throughout this dissertation. For most purposes, the dynamic range of a N -bit system measured in dB is just $6.02N$ dB. Deriving the maximum sampling rate begins with the selection of the minimum sampling capacitor size by requiring the kT/C noise to be less than 1/2LSB of the full scale voltage:

$$C_s = \frac{kT2^{2N}}{A^2} \quad (2.2)$$

where A is the full scale signal voltage. Then, by requiring that the channel charge be a certain fraction of the full scale charge in the sampling capacitor, the maximum switch size can be determined. A correction factor, ρ , is used to determine the degree to which the differential nature of the circuit removes the clock feedthrough error:

$$C_{ox} \frac{W}{L} = \frac{2^{N+1}kT}{\rho L^2 (V_{GS} - V_T) A} \quad (2.3)$$

Finally, the minimum sampling period is determined by taking into account the acquisition time produced by the finite switch resistance:

$$t = \frac{(\ln 2) N \rho L^2 2^{N-1}}{\mu A} \quad (2.4)$$

As an example, in a typical 1 μ m technology with a 2 volt swing and 90% cancellation of clock feedthrough, a 12-bit system will have a maximum sampling rate of about 60 MHz. Thus, the above described limitations can be a substantial bottleneck in the sampling process. Achieving higher speeds is usually accomplished by reducing the effect of the clock feedthrough. Careful use of bottom plate sampling techniques can make the clock feedthrough substantially signal independent. If the feedthrough is totally signal independent, the result is a DC offset, which can be removed by self-calibration techniques. However, at high frequencies approaching the inverse of the transit time of the MOS switch, the amount of charge in the channel will no longer be constant. This will lead to signal and signal-slope dependent components of clock feedthrough which cannot be cancelled. Thus, while the cancellation factor can be substantially improved by careful arrangement of switches, the presence of non-cancellable components of clock feedthrough will continue to limit the performance of high speed, high accuracy S/H stages [5].

Given the fact that one is able to sample an input signal with sufficient accuracy and speed as given in the above section, there are many advantages to signal processing in discrete time. In the analog domain, discrete time processing allows for the use of switched capacitor circuits, which are first-order insensitive to the absolute value of the circuit components without complex circuitry to track the value. Unlike continuous time circuits, switched-capacitor circuits make use of ratios of component values, which in a monolithic process can be maintained to very close tolerances. Moreover, discrete time systems are invariant to path delay mismatches as long as all circuit branches are allowed to settle before reclocking occurs. In continuous time circuits, care must be taken to insure that the signal paths are carefully matched when phase variations would cause a degradation of circuit performance. Finally, discrete time processing is a requirement of digital signal processing systems. Therefore, the remainder of this dissertation will concentrate on discrete time implementations of analog and digital signal processing techniques.

2.3 Analog-Digital Conversion

An obvious requirement of digital signal processing is the conversion of the input signal which is an analog sampled data value. There are extensive references regarding the architectures, trade-offs, limitations, implementations and performance of A/D converters. In conducting an analysis of the relative merits of analog vs. digital based processing, it is only important to

explore the limits of A/D conversion. A closed form analysis is intractable — empirical data gathered from various works however, gives a fairly consistent limit of the A/D process. A plot of A/D throughput (Figure 2.5), based on performance reported at the ISSCC, on a log-scale with speed on the x -axis and dynamic range on the y -axis results in a roughly straight line. Operation

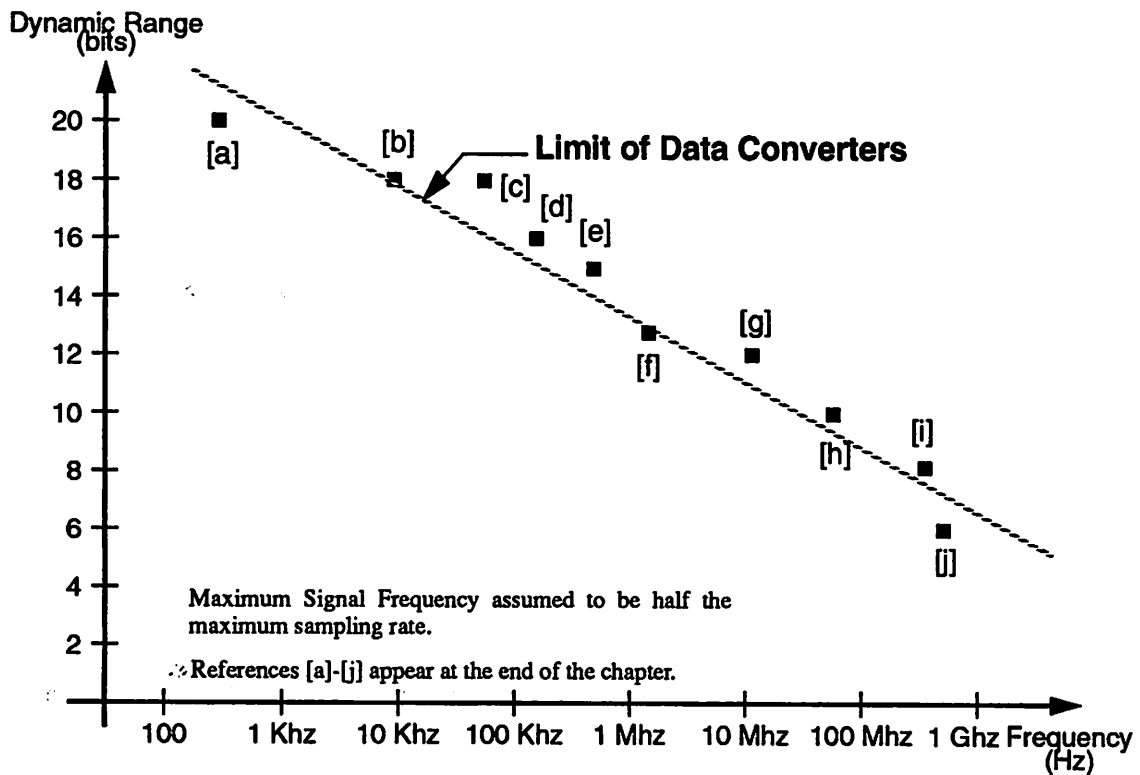


Figure 2.5 Current Limits of A/D Converters

to the right of the dashed line is relegated to the analog domain strictly due to the lack of data converters. This establishes one limit of digital signal processing. This line is slowly moving towards the right, as improved technology and circuit design techniques allow for faster and more accurate data converters. However, the A/D bottleneck will remain a barrier to digital processing for the time being. Moreover, the data points shown in Figure 2.5 does not take into account the technology used, nor does it factor in power dissipation or chip area. In many cases, the power dissipation, chip area or the requirement of specialized technologies can make the digital implementation unattractive, though possible with currently available technology.

2.4 Analog vs. Digital Implementations

In the region of operation underneath the dotted line in Figure 2.5, both analog and digital implementations of the signal processing task are achievable. The question therefore is which is more attractive. First, a metric of measuring the relative merits of each technique is required. For most cases, a cost function taking into account power dissipation and chip area seems to suffice. There is a direct relationship between chip area and the final cost of the device due to both fabrication throughput and yield issues, while power dissipation is becoming more important as higher levels of integration are achieved. Moreover, the current trend towards portable equipment provides a serious motivation to reduce power dissipation.

Obtaining an all-encompassing solution to the analog vs. digital question is almost impossible due to the enormous variation in signal processing tasks. Analysis based on very simple assumptions have been undertaken with plausible results [8], but the applicability to more complex circuits is uncertain. The following sections will explore the limitations of both analog and digital circuits in more detail in an attempt to provide a more extensive solution to this problem.

2.4.1 Determination of Circuit Building Blocks

Prior to making any computations about analog vs. digital circuits, it is necessary to produce a set of circuits with which to make comparisons. The realm of signal processing circuitry is extensive, and to attempt to find a circuit that represents all types of signal processing while remaining simple enough to be tractable is futile. Abstracting the types of signal processing usually performed by monolithic circuits shows that there are two or three main classes of circuits, such as filters, timing recovery loops and data converters. Of these, filters are of great interest not only because of their ubiquitousness, but also because they are easily realizable in both analog and digital domains. Moreover, in many cases, the choice to include more than the A/D converter in the analog signal path is equivalent to placing a filter immediately prior to the A/D rather than after conversion to the digital domain. For this reason, a filter will be used as a model building block to determine the applicability of analog vs. digital techniques.

Switched capacitor low-pass filters and their digital equivalents both make use of scaling accumulators as building blocks for filters. In switched capacitor circuits, these accumulators are often sampled data integrators, made to realize a $1/s$ transfer function of the frequency range of interest [9], and are constructed using an amplifier in the switched capacitor domain, or a regis-

ter, multiplier and an accumulator in the digital domain. While integrators make a good metric circuit for comparing the relative merits of analog and digital implementations, not all integrated filters make use of integrators as building blocks, as is the case with transversal filters. As a result, the switched capacitor charge transfer circuit and its digital equivalent, the multiply accumulator stage will be used as a metric for the purposes of this comparison.

2.4.2 Switched-Capacitor Charge Transfer Circuits

The switched capacitor integrator shown in Figure 2.6 in reality functions as a charge-transfer circuit, for it transfers the charge stored in C_S to C_I . The integrator function results from the fact that C_I , if not reset after each clock cycle, continues to accumulate charge from previous samples. While IIR filters and switched capacitor representations of ladder filters use the charge transfer circuit as an integrator, circuits such as transversal filters use the circuit strictly as a method of scaling and moving data, and not as an integrator to provide a $1/s$ transfer function. The circuit in Figure 2.6 also demonstrates “bottom-plate” sampling to minimize signal-depen-

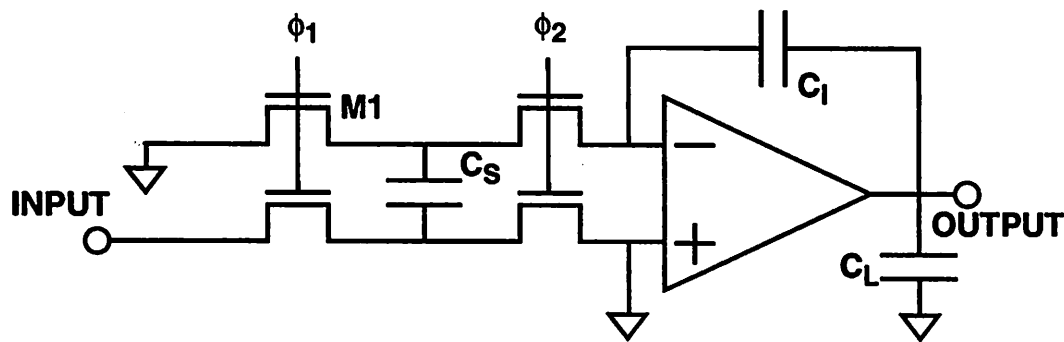


Figure 2.6 Switched-Capacitor Integrator

dent clock feedthrough. The input sample is taken when ϕ_1 goes low — in this scheme, M1 is clocked slightly earlier so that it becomes the device which isolates the top plate of C_S , thereby taking the sample. Because the source voltage of M1 is independent of the input signal, the magnitude of the clock feedthrough is constant. Charge transfer takes place during ϕ_2 , where the charge in C_S is transferred to C_I by the action of the amplifier. The ratio of C_S to C_I determines the magnitude of the voltage step at the output as a function of the input, and is nominally $-(C_S/C_I)$ assuming an ideal amplifier.

The characteristics and limitations of the S/H stage have already been discussed — the remaining performance limiting portion of the circuit is the amplifier whose role is to provide an active means to transfer charge from C_S to C_I . Of course, an ideal amplifier with infinite gain, zero input current and infinite speed is desired. With these attributes, it is easy to see that all the charge in C_S is transferred to C_I . Deviations from ideality will affect the operation of the charge transfer circuit. The requirement that the amplifier have a DC input impedance approaching infinity is especially strict to avoid leakage of charge from the sampling capacitor, C_S . A non-infinite gain results in incomplete charge transfer that affects the final transfer function. The speed, accuracy and power consumption of the circuit are all intertwined. For example, the feedback capacitor, C_I , combines with the sampling capacitor C_S , the input capacitance of the amplifier and any parasitics to form a feedback network. This, along with the open loop response of the amplifier forms the closed loop response which dictates the maximum speed and accuracy of the circuit. Because the load is effectively capacitive, the speed of the circuit is, in general, related to the amount of current available for charging and discharging the capacitors. Hence, higher speed for a given circuit topology implies higher power consumption.

2.4.3 Fundamental Limits of Traditional Switched Capacitor Integrators

Traditional switched capacitor low pass filters utilize charge transfer circuits configured as integrators with a very large ratio of C_I to C_S ($C_I \gg C_S$) [10]. Because of the large capacitor ratio, the unity gain frequency of the integrator is much lower than the clock frequency:

$$f_{\text{unity}} = \frac{C_S f_{\text{clock}}}{2\pi C_I} \quad (2.5)$$

Such conditions are ideal when switched capacitor filters are used as anti-alias filters as might be found in a voiceband PCM codec. However, the very fact that there is a high degree of oversampling implies that there is more work being performed by the circuit for a given data throughput than in an equivalent Nyquist sampled circuit. Put another way, the technology of the circuit is not being used to its fullest extent. Thus, high-speed circuits which attempt to extract the maximum throughput available for a given technology will tend not to use circuits with high oversampling ratios. In these cases, the assumption of $C_I \gg C_S$ which is traditionally associated with switched capacitor circuits is invalid.

However, because of the pervasiveness of the traditional switched capacitor low pass filter, it is useful to review the performance limitations of such a circuit. Assume that an ideal technology is available, and it is desired to compute the absolute minimum power consumption and silicon area required to effect the integrator function. To make the analysis tractable, some assumptions must be made regarding the architecture of the integrator. First, under these conditions, allow the assumption that the integrating capacitor is much larger than the sampling capacitor. Second, the amplifier is assumed to have infinite gain and zero noise, consumes no static power and swings over the entire supply voltage (0 to V_{DD}). Under these conditions, the power necessary to integrate a sinusoid waveform of amplitude V_i is given by [11]:

$$P = \frac{2}{\pi} V_i V_{DD} C_S f_{\text{clock}} \quad (2.6)$$

which has the familiar CV^2f relationship. The power consumption can also be expressed as a function of the output, and assuming that the input is driven such that the output is at the verge of clipping,

$$P = 2C_I V_{DD}^2 f \quad (2.7)$$

where f is now the signal frequency; because the integrator “gain” is a function of input frequency, the clock frequency is implicitly a function of input frequency for a constant output. This implies that the power can be reduced to arbitrarily low values while maintaining the desired transfer function strictly by reducing the value of C_I , and scaling C_S accordingly. However, this disregards the noise contribution of the S/H stage in front of the amplifier. Shrinking C_I and C_S to arbitrarily small values will result in excessive noise at the output of the integrator thereby degrading the dynamic range of the circuit. Recall the earlier discussion of kT/C noise of MOS sample and hold circuits where it was shown that the noise contributed by the sampling switch M1 in Figure 2.6 contributes a noise of variance kT/C_S distributed over the bandwidth from DC to $f_{\text{clock}}/2$. The resultant noise at the output can be computed by multiplying this noise by the integrated magnitude squared of the integrator transfer function [12]:

$$\bar{n}_o^2 = \frac{kT}{C_S} \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 |H(e^{-j\omega})|^2 d\omega = \frac{1}{2\pi} \frac{kT}{C_I} \frac{B_N}{f_{\text{unity}}} \quad (2.8)$$

where B_N given by:

$$B_N = \frac{f_{\text{clock}}}{2\pi} \int_{-\pi}^{\pi} H(e^{j\omega}) H(e^{-j\omega}) d\omega \quad (2.9)$$

represents the equivalent noise bandwidth from the integrator input to the output, given that $H(e^{j\omega})$ represents the transfer function from the input to the output of the circuit. The effective noise at the output is actually twice that given in (2.8) because of the contribution of the second switch. This assumes that the noise generators are uncorrelated and an infinite bandwidth of the op-amp. The maximum signal power, s_o^2 , is assumed to be $\frac{V_{DD}^2}{8}$ as the output swings to the rails. Thus, the voltage signal-to-noise ratio is given by:

$$\text{SNR} = \sqrt{\frac{s_o^2}{n_o^2}} = \sqrt{\frac{\pi V_{DD}^2 C_I f_{\text{unity}}}{8kTB_N}} = \sqrt{\frac{\pi U_{\text{max}} f_{\text{unity}}}{4kTB_N}} \quad (2.10)$$

where U_{max} is the maximum energy ($C_I V_{DD}^2/2$) that can be stored in the integrating capacitor. Thus, as expected, the SNR is strictly a function of the ratio of energy in the circuit to the thermal energy, with an allowance for the relative bandwidth of the circuit as set by the particular feedback network [Castello]. It can be shown that in the case of a simple single-pole response, (2.10) reduces to:

$$\text{SNR} = \sqrt{\frac{V_{DD}^2 C_I}{8kT}} \quad (2.11)$$

The minimum area required to implement this switched capacitor integrator is a function of both capacitor area and amplifier area. In the technological limit, it can be argued that the amplifier scales to the point that its area is negligible, and hence the area is strictly given by the size of C_I . The size of C_I is driven by the need to store a given amount of energy to obtain a desired SNR, for it was shown in (2.10) and (2.11) that the SNR is a function of the ratio of energy stored in the capacitor to the thermal energy kT . The relationship of the capacitance to area is well known:

$$C = \frac{\epsilon A}{d} = \frac{\epsilon_{\text{ox}} A E_{\text{max}}}{V_{DD}} \quad (2.12)$$

where E_{max} is the maximum electric field sustainable by the dielectric with relative permittivity ϵ_{ox} , with area A . Thus, the following relationships

$$\text{SNR} = \sqrt{\frac{\pi V_{DD} \epsilon_{ox} E_{max} A f_{unity}}{2kTB_N}} = \sqrt{\frac{\pi P f_{unity}}{16kTB_N f}} \quad (2.13)$$

outline the theoretical minimum requirements for power and area as a function of the desired SNR.

The statement above that the area of the amplifier is negligible is based on the assumption that the technology would eventually improve to the point that device sizes would become negligible. A verification of this assumption proves interesting, and results in an upper speed bound where this assumption becomes false. For simplicity, let the amplifier be a single transistor, whose task is to sink or source current out of C_I . It is desired to compute the minimum W and L of the device to determine its area. The minimum L is dictated by the material breakdown properties of silicon. As a very rough approximation, let $L_{min} = V_{DD}/E_{maxSi}$. (Note that E_{maxSi} is distinct from the E_{max} used earlier as the latter refers to oxide, not bulk silicon.) Under conditions of small L , it is assumed that the device is velocity saturated, and hence the current can be approximated by [13]:

$$I_D = WC_{ox} v_{sat} V_{GS} - V_T - V_{Dsat} \sim WC_{ox} v_{sat} V_{DD} \quad (2.14)$$

where the second approximation is made under the assumption that V_T and V_{Dsat} can be scaled, and that the entire supply voltage is available for gate drive, an albeit very generous assumption. The switched capacitor integrator will be called upon to transfer the maximum charge, $Q_{max} = V_{DD}C_S$, when the input voltage is equal to the supply voltage. For steady-state operation, to avoid clipping the output, this implies that the input frequency is f_{unity} . Furthermore, assume that this charge must be transferred within half the clock cycle, allowing slew limiting by the amplifier. This determines the peak current that needs to be supplied by the amplifier:

$$I_{max} = 2Q_{max} f_{clock} = 4\pi f_{unity} C_I V_{DD} \quad (2.15)$$

Combining (2.14) and (2.15) gives an expression for the minimum width of device:

$$W_{min} = \frac{4\pi f_{unity} C_I V_{DD}}{\epsilon_{ox} E_{max} v_{sat}} \quad (2.16)$$

where v_{sat} is the saturation velocity of electrons in silicon ($\sim 10^7$ cm/s). Combining the expressions for W_{min} and L_{min} results in the expression for the minimum area of the amplifier:

$$A_{\min} = W_{\min} L_{\min} = \frac{4\pi f_{\text{unity}} C_1 V_{\text{DD}}^2}{\epsilon_{\text{ox}} E_{\text{max}} v_{\text{sat}} E_{\text{maxSi}}} = \frac{C_1 V_{\text{DD}}}{\epsilon_{\text{ox}} E_{\text{max}}} \left[1 + \frac{4\pi V_{\text{DD}} f_{\text{unity}}}{v_{\text{sat}} E_{\text{maxSi}}} \right] \quad (2.17)$$

The factor $(C_1 V_{\text{DD}} / \epsilon_{\text{ox}} E_{\text{max}})$ is the minimum area for the integrating capacitor from (2.12). Thus, for conditions such that $4\pi V_{\text{DD}} f_{\text{unity}} \ll v_{\text{sat}} E_{\text{maxSi}}$, the assumption that the integrator area is dictated by the capacitor area holds. Otherwise, the amplifier area becomes significant, and cannot be neglected. Note that the right hand side of the equation is strictly a function of the material property of silicon, and is therefore not likely to be enhanced by improvements in technology. For typical values, the $v_{\text{sat}} E_{\text{maxSi}}$ product is about 10^{12} V/s, implying that the inequality holds for most all frequencies of interest. It should be noted, however, that the assumptions made to derive this result are extremely generous, such as negating the effects of V_T and V_{Dsat} , allowing the device to operate at the threshold of breakdown, etc. For real devices, the above result may have to be derated by a factor of 100. As speeds of circuits are increasing with improvements in technology, the frequency where the area of the amplifier becomes significant within the context of this analysis is rapidly being approached, and the result that the amplifier area is negligible is no longer true.

The theoretical limits for the minimum power and area of a switched capacitor charge transfer circuit have been discussed above. Unfortunately, while they give a lower bound and a goal for circuit designers, it is highly unlikely that the limits will be approached due to practical constraints of realizable circuits. Moreover, as they are based on the assumption of $C_I \gg C_S$ and the traditional oversampled LPF architecture, they are not a good basis for comparison. Therefore, it is important to investigate the effect of the constraints of practical implementations combined with more aggressive architectures in order to derive a more practical solution to the analog vs. digital question.

2.4.4 Practical Power Constrained Switched Capacitor Charge Transfer Circuits

In this section, the assumption that $C_I \gg C_S$ will not be made, and that the limitations inherent in moving charge from C_S to C_I be investigated. Moreover, practical limitations will be considered to allow for a more meaningful result to power constrained switched-capacitor integrators will be investigated. The goal is to provide maximum throughput (SNR and speed) while consuming the minimum power. Referring to Figure 2.6 and Figure 2.7, the speed of such a circuit is governed by the forward transfer function of the amplifier and the feedback circuit. Assuming a single-pole response for the amplifier, the settling time of the circuit is given by:

$$\tau = \frac{C_{LT}}{g_m} \times \frac{1}{f} = \frac{\left(C_L + \frac{C_I (C_S + C_P)}{C_I + C_S + C_P} \right)}{g_m} \times \frac{C_S + C_I + C_P}{C_I} \quad (2.18)$$

where g_m refers to the forward transconductance of the amplifier, C_{LT} to the total load seen by the amplifier. C_L is the load at the output of the amplifier, and C_P to the total parasitics at the amplifier summing node, including the input capacitance of the amplifier. f refers to the “feed-

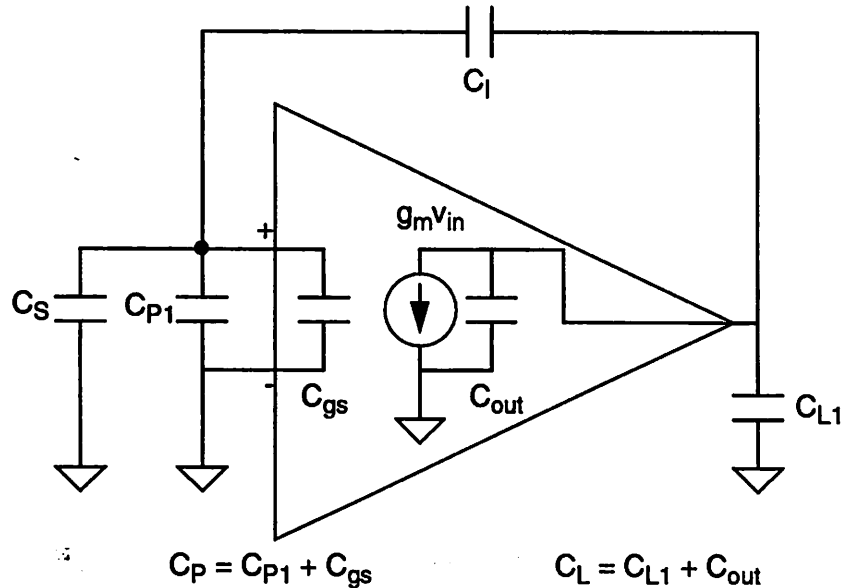


Figure 2.7 Amplifier Model for Equation 2.18

back factor,” or the fraction of the output signal available to the summing node to provide feedback. For the purposes of the remainder of this analysis, it will be assumed that the integrating capacitor, C_I , and the source capacitor, C_S , are of equal value. This is distinctly different from the assumptions of the earlier section. However, it more closely resembles the situation found in high-speed Nyquist rate sampled data signal processors. For example, the prototype circuit described in the later chapters makes use of roughly equal values of C_I and C_S . Under these conditions, and assuming that the circuit is loaded with another identical stage such that $C_L = C_S$, then (2.18) simplifies to:

$$\tau \approx \frac{3C_S + 2C_P}{g_m} \quad (2.19)$$

Thus, the speed becomes a function of g_m , and C_S , where the required size of C_S is dictated by dynamic range requirements, and the parasitic loading presented by the amplifier itself. Thus, in order to maintain a minimum SNR, the speed of the circuit is governed primarily by the amplifier g_m , which in most all cases is a function of the power dissipated by the amplifier. In the case of single-stage amplifiers, the g_m available is that of the inverting device, typically a common source MOS device. Under normal operating conditions, the g_m of the amplifier is simply:

$$g_m = \sqrt{2k' \frac{W}{L} I_D}. \quad (2.20)$$

where $k' = \mu C_{ox}$. The square-root dependence of g_m as a function of I_D suggests that the speed per unit of bias current increases with decreasing bias currents. That is, the device is most “efficient” at low bias currents. Moreover, the gain of most CMOS amplifiers increases as the bias current is reduced because the incremental output resistance is inversely proportional to I_D , making it easier to meet the gain requirements to achieve adequate charge transfer. Thus, reducing the drain current to infinitesimally small values appears to be strategically advantageous for ultra-low power operation.

2.4.4.1 Subthreshold Operation of Charge Transfer Circuit

Below a certain current density, the MOSFET no longer remains in the “saturated” region of operation as implied by (2.20), but enters the subthreshold region. It has been suggested [11],[14] that operation in the subthreshold (weak-inversion) regime provides the most efficient use of power for switched capacitor circuits. In this region of operation, the MOSFET behaves very similarly to a BJT, in that the drain current becomes exponentially dependent on the gate voltage, and to first degree independent of V_{DS} provided $V_{DS} > 3kT/q$. Under these condition, the transconductance of the device becomes a linear function of the drain current and is given by:

$$g_m \approx \frac{qI_D}{mkT} \quad (2.21)$$

where m is the subthreshold slope multiplication factor. (m is a function of the applied substrate bias, gate oxide thickness and substrate doping, and is usually between 1 and 2).

Assuming linear settling, the number of time constants required to settle to a given accuracy is given by $\ln(SNR)$, and is usually between 5 and 10 for most circuits. Furthermore, assuming

that half the clock period is available to settle, $\tau_{min} = (1/2 \ln(SNR)) / f_{clock}$. This gives an expression for the minimum g_m required:

$$g_{m_{min}} = 2 \ln(SNR) \times (2C_S + C_P) f_{clock} \approx 4 \ln(SNR) C_S f_{clock} \quad (2.22)$$

which implies that the minimum current necessary for proper operation is:

$$I_{D_{min}} = \frac{4kT \ln(SNR) C_S f_{clock}}{q} \quad (2.23)$$

implying the minimum power is:

$$P_{min} = \frac{4kT \ln(SNR) V_{DD} C_S f_{clock}}{q} \quad (2.24)$$

under the assumption that C_P is negligible to C_S . C_P is primarily the input capacitance of the active device; the effect of C_P can be made negligibly small by making the device smaller while keeping the drain current constant. However, this approach implies that the current density increases as C_P is made smaller for a given g_m . Thus, at some point, the original assumption that the device is in subthreshold operation and (2.21) applies becomes false. This limitation can be found by examining the expression for the drain current of a subthreshold device [15]:

$$I_D = \frac{W}{L} I'_M e^{\frac{q(V_{GS} - V_M)}{mkT}} \left(1 - e^{-\frac{qV_{DS}}{kT}} \right) \quad (2.25)$$

where

$$I'_M = \mu C_{ox} \left(\frac{kT}{q} \right)^2 \frac{\gamma}{2\sqrt{1.5\phi_f + V_{SB}}} \quad (2.26)$$

and V_M is the maximum V_{GS} allowable while maintaining subthreshold operation, γ is the body effect coefficient, and ϕ_f is the Fermi potential of silicon. Clearly, the maximum allowable drain current in subthreshold is when $V_{GS} = V_M$; by assuming that $V_D \gg (kT/q)$, and $V_{SB} = 0$, the expression for the maximum drain current becomes, with appropriate substitutions:

$$I_{D_{max-subthreshold}} = \frac{\mu W}{L} \left(\frac{kT}{q} \right)^2 \sqrt{\frac{q\epsilon_s N_A}{3\phi_f}} \quad (2.27)$$

where N_A is the substrate doping and ϵ_s is the permittivity of silicon. Note that this quantity is composed strictly of physical constants, the substrate doping, mobility and geometry. Thus, for

reasonable values of N_A and μ , the maximum drain current available in subthreshold is about $(3W/L) \times 10^{-9} A$.

The actual value of C_P when the device is operated in the subthreshold region is difficult to determine accurately. The worst case value is WLC_{ox} , which assumes that there is a conducting sheet immediately underneath the gate extending completely from the source to the drain. In reality, however, the actual capacitance seen at the gate is a fraction of WLC_{ox} ; as the device moves from strong inversion to subthreshold, the value of C_P falls from the standard $2/3WLC_{ox}$ to a smaller value. The value of C_P in weak inversion is dominated by the parasitic capacitances and the capacitance from the gate to the bulk (assuming the technology has a grounded substrate). The value for the gate-bulk capacitance is given by: [15]

$$C_{gb} = C_{ox} \frac{\gamma}{2\sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}}} \quad (2.28)$$

and is typically between 0.3 and 0.5 C_{ox} [16].

The maximum value of I_D sets a severe restriction on the available speed of the device, and thus, for circuits which operate above a few megahertz, the approach of utilizing subthreshold devices becomes unreasonable. Increasing the g_m of the device by making it larger forces the input capacitance of the amplifier to rise, thereby lowering the feedback factor, which in turn will increase the loop settling time. As with circuits operating in the conventional regime, in the limit of large devices, the loss in feedback factor directly offsets the increase in g_m , resulting in no net gain. For large devices, the settling time is given by (C_P/g_m) , which when the subthreshold expression for g_m is substituted and assuming that C_P is half WLC_{ox} , becomes:

$$\tau_{\text{big-device}} = \frac{WLC_{ox}}{2g_m} \approx \frac{m \frac{kT}{2q} L^2 C_{ox}}{3 \times 10^{-9}} \quad (2.29)$$

which, for a value of $L=1\mu m$, $C_{ox}=2fF/\mu m^2$, $m=1.5$, becomes 5.18×10^{-8} . This implies, for an 8-bit equivalent system, the maximum clock frequency is 6.44 MHz, which allows for voiceband signal processing circuits. Note that (2.29) is the inverse of the expression for the f_T of the device, implying that this represents the speed limit of subthreshold devices.

Examining the optimum device sizing for operation in subthreshold reveals that the circuit is most efficient when the amplifier input capacitance is made as small as possible with respect to the rest of the capacitances in the circuit. This is distinctly different than for operation in the normal regime as will be shown in the next section. Thus, maximum speed is obtained at the edge of subthreshold operation. Under the assumptions that led to (2.19), the circuit speed is dependent on the size of C_S as dictated by the required system SNR, and can approach, but not exceed that given in (2.29).

As a point of reference, if the areas of the device and integrating capacitor are made equal, Castello [11] finds that maximum unity gain frequency of a traditional integrator ($C_S \ll C_I$) is:

$$f_{\text{unity}} = \left(\frac{kT}{q}\right) \frac{\mu}{8\pi\beta \ln(\text{SNR}) qmL^2} \quad (2.30)$$

where β is the ratio of the device area to the gate area. Under typical assumptions for a $1\mu\text{m}$ technology, (2.30) implies a f_{unity} of 2 MHz. It is important to note however, that this computation does not take into consideration the parasitic loading effect of the device input capacitance, and hence is extremely generous. Nevertheless, it confirms the earlier finding that operation in the subthreshold region, while power efficient, is appropriate for circuits with signal bandwidths much less than 0.1% of the inherent device f_T .

The L^2 dependence on the time constant suggests that continued improvements in technology will allow the use of subthreshold circuits well into the hundred megahertz range with $0.1\mu\text{m}$ line widths. However, it is important to note that the above analysis neglects many second order effects which will limit the usefulness of such circuits. First and foremost is the grossly simplified amplifier model used. No practical circuit can operate with a single device amplifier — at a minimum, some type of load device must be present, which will add parasitics. Second, the minimum gain requirement has been completely ignored. Lastly, but perhaps most importantly, the issue of subthreshold swing, m , and the inability to turn off ultra-short devices with small gate swings affects the operation of such hypothetical circuits.

Increasing the throughput of systems designed around subthreshold devices will most likely utilize parallel approaches where the low power consumption of these circuits outweigh the added area penalty of parallel circuits. However, for high-speed applications where the desired

clock rate is orders of magnitude higher than those achievable through subthreshold circuits, operation of the MOSFET in the conventional “saturated” regime is required.

2.4.4.2 Optimal Design of Switched-Capacitor Charge Transfer Circuits

Assuming that utilizing massively parallel circuits operating in the subthreshold region is prohibitively expensive in silicon area limits one to utilizing MOS devices operating in the conventional “saturated” regime. Thus, the desired goal is to determine the minimum power and area in order to achieve a certain throughput without resorting to parallel approaches. Assuming that the devices are operating in the conventional “saturated” regime, the equations of interest are:

$$I_D = \frac{\mu C_{ox} W}{2L} (V_{GS} - V_T)^2 \quad (2.31)$$

$$g_m = \sqrt{2\mu C_{ox} \frac{W}{L} I_D} = \mu C_{ox} \frac{W}{L} (V_{GS} - V_T) \quad (2.32)$$

$$C_{gs} = \frac{2}{3} WLC_{ox} \quad (2.33)$$

Note that (2.32) and (2.33) imply that the maximum f_T of the device is a function of gate overdrive voltage, and is not limited as in the subthreshold case. However, there are practical limits such as gate oxide breakdown and mobility degradation that prevent excessive the attainment of extremely high values of f_T .

$$f_T = \frac{3}{2L^2} \mu (V_{GS} - V_T) \quad (2.34)$$

What (2.34) does imply, however, is that the amplifier can be made arbitrarily fast through increased gate overdrive, at the expense of increased power dissipation. Determination of a power and area metric for an integrator of Figure 2.6 requires that the size of the inverting device be determined. A means of sizing the device for greatest efficiency, or minimizing the settling time as a function of device size for some fixed technology and drain current is required. Again, examining the circuit of Figure 2.6, and once again assuming that $C_S = C_J$, recall that the settling time of such a circuit is given by:

$$\tau = \frac{\alpha WLC_{ox} + C_S}{\sqrt{2\mu C_{ox} \frac{W}{L} I_D}} \quad (2.35)$$

where α is usually $\frac{2}{3}$. Minimizing τ as a function of W follows as:

$$\frac{d\tau}{dW} = \frac{\alpha LC_{ox} \sqrt{2\mu C_{ox} \frac{W}{L} I_D} - \mu C_{ox} \frac{1}{L} I_D \times \frac{\alpha W L C_{ox} + C_S}{\sqrt{2\mu C_{ox} \frac{W}{L} I_D}}}{2\mu C_{ox} \frac{W}{L} I_D} = 0 \quad (2.36)$$

which implies that

$$2W L C_{ox} \alpha = W L C_{ox} \alpha + C_S \quad (2.37)$$

which, of course, gives the familiar result of $C_{gs} = C_S$. Thus, to maximize speed for a given power dissipation level, the optimum sizing of the inverting device is to equate the gate capacitance with the sampling capacitance. If one inserts a load capacitance at the output of the amplifier equivalent to C_S , thus simulating cascading of stages, then the results change slightly so that $C_{gs} = 2C_S$. Assuming that the C_S and C_{gs} are sized appropriately, using the earlier definition of $\tau = 1/(2f_{clock} \ln(SNR))$, the maximum clock rate becomes a function of the drain current, and hence power dissipation:

$$f_{clock} = \frac{1}{2 \ln(SNR) \alpha L} \sqrt{\frac{\mu I_D}{2W L C_{ox}}} \quad (2.38)$$

while the area consumed by the amplifier is some constant times the input device area.

A means of determining a reasonable estimate of power and area requires that certain assumptions be made regarding the technology. It was argued earlier that the active device can be scaled to the material limits of silicon, and extremely small devices used. However, with practical circuits, those limits are highly unlikely to be observed. First, an adequate supply voltage must be used. The current standard of 5 volts is slowly giving way to 3.3 volt circuits so that the analog circuitry will remain compatible with digital technologies. However, moving to lower voltages is proving difficult, as the mainstay of switched capacitor circuitry, the MOS switch, suffers greatly from reduced voltage. As the supply voltage is reduced, the device threshold voltage becomes a larger proportion of the supply voltage, resulting in a two-fold reduction in gate overdrive voltage, $(V_{GS} - V_T)$. As a result, the "on-state" resistance of the switches increases resulting in lowered performance of the circuits. Reduction of V_T is possible to some extent, but effects associated with the short channels of today's technologies results in enhanced subthresh-

old swing [13]. As a result, a minimum V_T of about 500mV is needed to insure that the device can be turned off and prevent charge leakage in a switched capacitor circuit. Thus, it is unlikely that supply voltages will be reduced much below 3 volts in the near future. Assuming a 3 volt technology allows one to set the minimum channel length and gate oxide thickness, C_{ox} . Both parameters are limited by material constants, namely hot electron and breakdown, and are unlikely to be improved upon significantly. Minimum channel lengths of 0.5 μm and gate oxides of 100Å are required to withstand 3 volt potentials that may arise in generalized circuits [13]. The mobility of silicon under such conditions is subject to numerous scattering and high-field effects, and is unlikely to exceed commonly found low-field values of 500 cm^2/Vs . These values, when substituted into (2.38) result in clock rates in excess of 5 GHz, implying that between the upper limit of subthreshold operation, and multi-GHz frequencies, switched capacitor MOS technology utilizing devices in the “saturated” regime is applicable.

Unlike switched capacitor integrators, where the dynamic range is ultimately a function of the integrating capacitor because of bandlimiting of noise caused by the integrator transfer function, the switched capacitor charge transfer circuit, in general, transfers all of the noise sampled on C_S to the output. Thus, the dynamic range of the switched capacitor charge transfer circuit is governed by C_S , not C_f . Thus, the design of an optimal circuit is fairly straightforward:

- Determine the system level parameters, in particular, the signal swing and clock rate
- Determine the minimum value of C_S such that the dynamic range requirements due to kT/C noise are met.
- Size the device such that C_{gs} is equal to twice the value of C_S .
- Determine the minimum drain current, I_D , such that the clock rate requirements are met.
- In the case of a class A amplifier, verify that the computed drain current allows for linear or near linear settling. Otherwise, insure that the amplifier has the capability to supply the peak current necessary for slewing, while maintaining a minimum g_m for settling.

For purposes of this computation, let β be the signal swing as a fraction of the supply voltage. Then, the minimum sampling capacitor, C_S is given by the SNR requirements:

$$C_s = \frac{kT(\text{SNR})^2}{(\beta V_{DD})^2} \quad (2.39)$$

Substituting this into (2.38) and solving for the minimum drain current, which shall be denoted I_{D1} , yields:

$$I_{D1} = \frac{kT\alpha}{\mu} \times \left(\frac{4 \ln(\text{SNR}) L f_{\text{clk}} \text{SNR}}{\beta V_{DD}} \right)^2 \quad (2.40)$$

This current, I_{D1} , represents the minimum drain current necessary to provide the forward g_m in the amplifier to insure settling in the time allowed. It assumes, *a priori*, that there the loop settles linearly, with no slew rate limiting. As will be seen shortly, this is a fairly poor assumption, unless class AB or class B amplifiers are used to insure that the peak charging current is provided. Otherwise, the minimum bias current required to settle is subject to a more stringent requirement, in that the circuit must not be forced into slew limiting. Recall that for single-pole settling, the maximum rate of change occurs at the beginning of the settling cycle. The maximum current, therefore, required to charge C_I occurs at $t = 0$, and is given as:

$$I_{o-\text{max}} = \frac{C_I \beta V_{DD}}{\tau} \quad (2.41)$$

Substituting the relationship between τ and f_{clk} , and assuming that $C_I = C_S$, gives another expression for the minimum drain current (this assumes that the maximum charging current is equal to the drain current, correct for class A circuits), denoted I_{D2} :

$$I_{D2} = \frac{2f_{\text{clk}} \ln(\text{SNR}) kT}{\beta V_{DD}} (\text{SNR})^2 \quad (2.42)$$

For most combinations of speed and dynamic range, $I_{D2} > I_{D1}$; the exact criteria is given by:

$$f_{\text{clk}} < \frac{\mu \beta V_{DD}}{8\alpha \ln(\text{SNR}) L^2} \quad (2.43)$$

For clock frequencies below that given in (2.43), the circuit is slew limited, and the larger drain current given in (2.42) must be supplied to prevent non-linear effects from adversely affecting the expected settling time. In the rare event that the clock frequency is greater than that given in (2.43), usually the result of high dynamic range requirements, the current necessary to prevent slewing is insufficient to provide the transconductance needed to settle the circuit in a timely fashion.

As the criteria in (2.43) is unlikely to be met for most circuits — note that for $1\mu\text{m}$ technologies, the critical frequency from (2.43) near 1GHz, and increases with the square of the scaling factor — slew rate limiting is an important factor in amplifier design. However, it is important to note that the peak current necessary to avoid slewing need only be supplied for a small fraction

of the settling cycle. With a class A amplifier, where the maximum output current is equal to the quiescent bias current, this results in an inherent inefficiency, for during the majority of the settling cycle, the current flowing into the capacitor is only a fraction of the bias current. Although the power needed to charge the capacitors is only $C_I(\beta V_{DD})^2 f_{clk}$, the power consumed by the circuit is the product of the appropriate I_D (usually I_{D2}) and V_{DD} .

As alluded to earlier, the use of class AB or B amplifiers, where the circuit is able to supply output currents many times greater than the quiescent bias current, allows the overall efficiency to be boosted considerably. In this case, the peak charging current necessary to avoid slewing, I_{D2} , is supplied for an instant at the beginning of each charging cycle. As the capacitor charges, the current supplied by the amplifier falls accordingly, limited only the minimum g_m requirements necessary to insure a fast enough settling dictated by linear network analysis, which is simply I_{D1} . As a result, the power dissipated by the circuit more closely resembles the power consumed by the capacitors, and is given by:

$$C_I (\beta V_{DD})^2 f_{clk} + I_{D1} V_{DD} \quad (2.44)$$

As I_{D1} in many cases is much less than I_{D2} , the use of class AB or B amplifiers can lead to much higher efficiencies, in that it can approach the theoretical minimum power of $C_I(\beta V_{DD})^2 f_{clk}$. Moreover, as the power in (2.44) assumes that the signal swings full scale at all times, it represents the maximum dissipation for these circumstances. If the signal being processed is voice for example, the average power dissipated in the capacitor is much less than $C_I(\beta V_{DD})^2 f_{clk}$.

The choice of whether to use a class A amplifier and pay the penalty in efficiency or to use a class AB/B amplifier is dependent on the level of circuit complexity and speed desired. Class A amplifiers are inherently fast, as they can be very simple — in this analysis, we assume a single device inverting stage — and can approach f_T in the usable gain-bandwidth. Class AB/B amplifiers are inherently more complex, requiring extra devices to form the circuitry necessary to supply the peak charging currents. Because of this added complexity, the inherent speed of such a circuit for a given technology is less than that for a class A amplifier. In addition, the added circuit elements increase the amplifier area required. Thus, for situations where speed and small area are important, class A amplifiers offer a better solution. Situations with large parasitic output capacitances, which adversely impact the slewing requirements of the circuit while leaving the loop feedback factor unchanged may save considerable amounts of power through the use of

class AB/B amplifiers. To provide a point of reference, Table 2-1 shows the power required for the switched capacitor charge transfer circuit for various speeds and dynamic ranges for both class A and class AB/B circuits. ($L = 1 \mu\text{m}$, $\beta = 0.33$, $\mu = 500 \text{ cm}^2/\text{Vs}$, $V_{DD} = 3\text{V}$)

TABLE 2-1 Theoretical Power Dissipation for SC Charge Transfer Circuit

Clock Freq.(MHz)/ Dynamic Range (Bits)	Class A (W)	Class AB/B (W)
10 MHz 8 Bits	9.03×10^{-8}	3.249×10^{-9}
10 MHz 12 Bits	3.47×10^{-5}	1.003×10^{-6}
10 MHz 16 Bits	1.18×10^{-2}	3.180×10^{-4}
100 MHz 8 Bits	9.03×10^{-7}	8.055×10^{-8}
100 MHz 12 Bits	3.47×10^{-4}	3.773×10^{-5}
100 MHz 16 Bits	0.118	1.578×10^{-2}
1 GHz 8 Bits	9.03×10^{-6}	5.612×10^{-6}
1 GHz 12 Bits	3.47×10^{-3}	3.147×10^{-3}
1 GHz 16 Bits	1.418	1.418

It should be noted that for very low resolutions, the required value of C_S from (2.39) can become unmanageably small, in that fabrication of such a small capacitor may not be possible, or the quality of the capacitances may not be adequate for the task. In this event, the value of C_S must be made artificially large, with a concurrent increase in current and power. For example, a 4-bit system requires capacitances on the order of 1aF, which is clearly not realizable in technologies available at the present time. Below about the 12-bit level, the size of the capacitances required to avoid degradation due to kT/C noise is very small, and the bias currents and power necessary for circuit operation are dominated by other factors such as parasitics and matching requirements. Figure 2.8 shows the relationship between power required per circuit as a function of the clock rate (f_{clk}) and the dynamic range (SNR) expressed in bits, assuming $\mu=500 \text{ cm}^2/\text{Vs}$, $\beta = 0.33$, and $V_{DD} = 3\text{V}$ for a class A amplifier. The straight curves on a log plot show a near square-law relationship between power and dynamic range, demonstrating the severe penalty imposed by large dynamic range requirements.

One final factor needs to be considered. At very high clock frequencies, the displacement current necessary to charge the capacitor can be quite high as shown in (2.42). The above computations are made based on the "optimal" sizing of devices from (2.36). At a high enough fre-

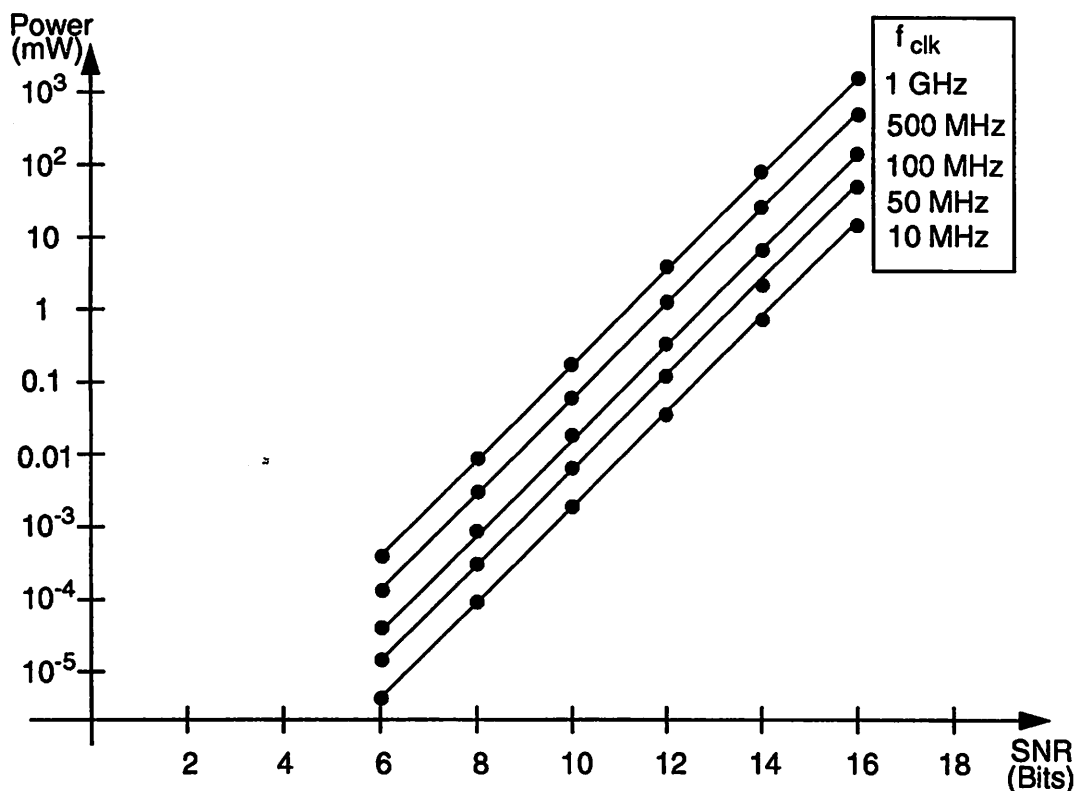


Figure 2.8 Power for SC Charge Transfer Circuit vs. SNR and Clock Freq. ($L=1\mu\text{m}$)

quency, the appropriately sized device is not capable of providing the current necessary with a reasonable amount of gate drive. A reasonable assumption of available quiescent gate drive is the signal swing, βV_{DD} , which provides a maximum current of:

$$I_{D3} = \frac{\mu C_{ox} W}{2L} (\beta V_{DD})^2 = \frac{\mu C_s}{\alpha L^2} (\beta V_{DD})^2 \quad (2.45)$$

In the event that $I_{D3} < I_{D2}$, then for class A amplifiers, the device size will have to be increased to meet the slew rate requirements. Of course, this upsets the original assumption that $C_{gs} = 2C_S$. Thus, the device and bias currents will need to be increased beyond what is expected to compensate for the additional load of the device. This problem is most severe with high clock rates and long channel lengths. For the example presented here, the restriction of (2.45) results in a slightly increased power requirement at $f_{clk} = 1\text{GHz}$. Finally, because the inherent f_T of a MOS device is given by:

$$f_T = \frac{\mu (V_{GS} - V_T)}{2\pi\alpha L^2} \quad (2.46)$$

there is a minimum gate overdrive voltage to achieve a given clock rate. It is entirely possible that under conditions of high dynamic range with a long minimum channel length, and low clock rate, the f_T of the device may be insufficient to settle the loop, and additional current with concurrent penalties in power and area will be required. However, for the examples given in this section, this is not an issue.

2.4.4.3 Minimum Achievable Area of Switched Capacitor Charge Transfer Circuits

A lower limit for the area required by the circuit described in the previous section can be approximated under the assumption that $C_{gs} = 2C_S$. Although for most all situations, the drain current required is dictated by displacement current and not the transconductance requirements, the above approximation is still useful as a reference point. The minimum required area of the circuit is that consumed by the integrating and sampling capacitors, plus the device (amplifier) area. For simplicity, it will be assumed that the dielectric available for the device gate oxide and the capacitors are equal. Then, the gate area of the device is just $(C_{gs}/\alpha C_{ox})$ or $(2C_S/\alpha C_{ox})$. The area of the two capacitors, is just $(2C_S/C_{ox})$. The device area is usually assumed to be some multiple of the gate area, so the device area can be expressed as $(2\delta C_S/C_{ox})$, thereby making the total circuit area $([1+\delta][2C_S/C_{ox}])$.

The area of the SC charge transfer circuit can then be plotted using (2.39) assuming some characteristics of a typical $1\mu\text{m}$ technology, with $\delta = 5$ and $C_{ox} = 1.5\text{fF}/\mu\text{m}^2$, and $\beta V_{DD} = 1$. It is clear that the area figures suggested by Figure 2.9 are in most cases unreasonable given the assumption of a $1\mu\text{m}$ technology. However, what the figure does show is that for lower resolution systems, the area requirements imposed by the inherent limits of analog technology are negligible, and will be completely swamped out by practical considerations. One final fact to keep in mind is that the amplifier was modelled as a single transistor. In all practical circuits, there will be at least one other device to act as a load, plus bias circuits. Moreover, metal routing can consume a large fraction of the total area. These additional components will add to the δ factor used above. The critical fact to note is that there is a square law relationship between circuit area and desired SNR, making analog circuits very expensive in both power and area in instances where a large dynamic range is required.

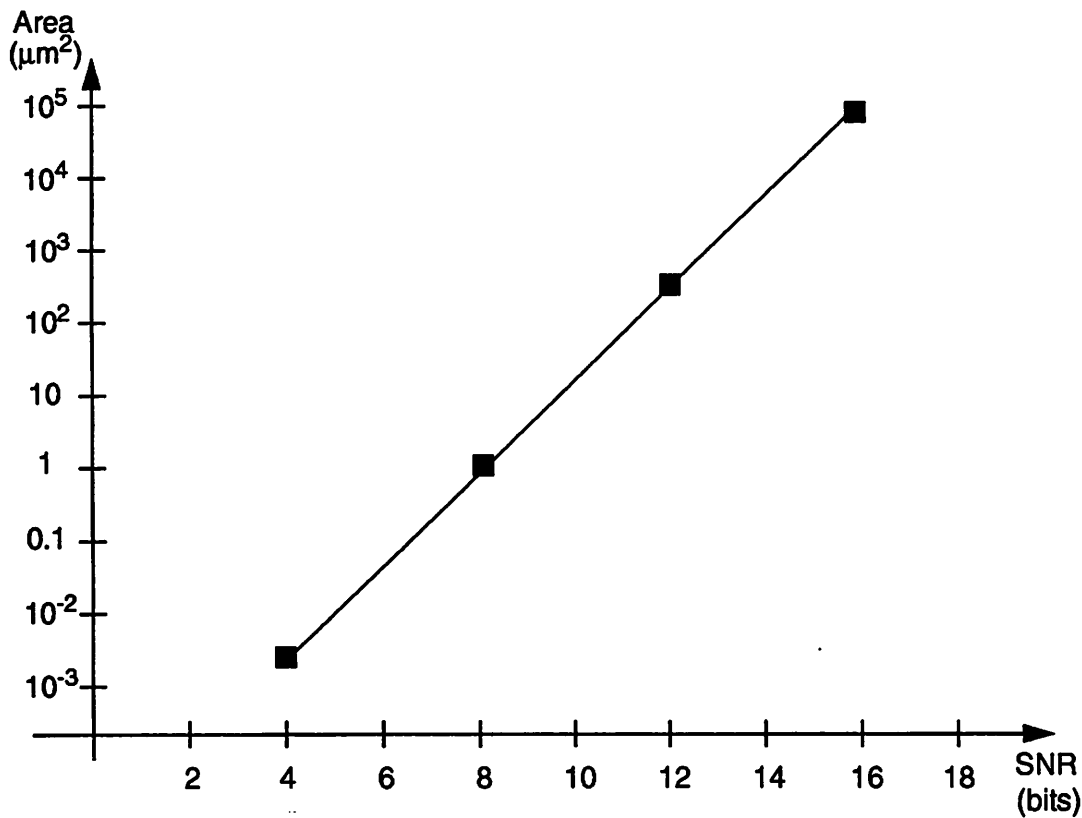


Figure 2.9 Area of SC Charge Transfer Circuit vs. SNR ($C_{ox}=1.5\text{fF}/\mu\text{m}^2$)

2.4.5 Digital Equivalents of Charge Transfer Circuits

The switched capacitor charge transfer circuit provides two functions, that of scaling an input and that of accumulating data. The digital equivalent that most closely mimics these functions is the multiply-accumulate (MAC) stage. In fact, many digital signal processor implementations use large numbers of MAC stages to perform filtering functions [17]. Thus, it is of value to compare the power and area consumed by a digital circuit which performs the MAC function.

Unlike the analog switched capacitor implementation of this function, there are a large number of different realizations of the MAC function in the digital domain, which makes it difficult to make a direct comparison between analog and digital implementations. However, using some basic assumptions, a trend will appear, which will provide the key information to determine the applicability of analog vs. digital implementations.

2.4.5.1 Power Characteristics of Digital Circuits

A digital circuit can be viewed as a collection of nodes with mechanisms to alter the logic level at the nodes as a function of the logic levels at some other nodes. Information is processed by changing the logic level at appropriate nodes. Therefore, the power consumed by a digital circuit is fairly easy to compute in that it is roughly the product of the rate of nodal transitions and the energy per nodal transition. Each node has associated with it a capacitance due to the logic gates connected to the node (input gate capacitance) plus any parasitics. Because there are only two logic levels, the energy to effect a nodal transition is just CV_{DD}^2 , where C is the total capacitance at the node. Assuming that the circuit consumes no static power and all current consumed goes to charging capacitances, the power of a digital circuit is proportional to the clock rate, and the familiar $CV_{DD}^2f_{clk}$ expression results. For modern CMOS circuits, the assumption that there is zero static power is acceptable, as junction leakages are orders of magnitude less than the dynamic power. The second assumption, that there is no "short-circuit" current from V_{DD} to ground during a logic transition is not as clear. This short circuit current occurs when both devices are turned on at the same time. In a CMOS circuit, this occurs during the transition where the input to the logic gate is swinging between V_{in} and V_{tp} . The magnitude of this current, and its effect on the assumption that the power of a digital circuit is given by $CV_{DD}^2f_{clk}$ has been studied [18]. For a well designed circuit, the error is found to be about 10 percent, so that the power consumed in a digital circuit is approximated by $1.1CV_{DD}^2f_{clk}$.

2.4.5.2 Digital Circuits to Realize the MAC Function

The digital multiply accumulate stage can be realized by using a combination of a multiplier, an adder and a register. While this provides the functionality required to emulate the switched capacitor charge transfer circuit, the resultant digital circuit will be quite large due to the amount of circuitry required by a full digital multiplier. Fairness requires that techniques to avoid the use of a full digital multiplier be made in this comparison. One of the advantages afforded by use of a full digital multiplier is the ability to change the multiplicand quickly and easily. In the switched capacitor circuit, the multiplicand is set by a ratio of capacitors, and hence is not easily changed. Techniques have been developed to allow switched capacitor circuits to change the multiplicand, but these tend to have a small range of adjustment, and add considerably to the complexity of the circuit.

If the digital equivalent of the switched capacitor circuit does not need to support changing multiplicands, or equivalently, is a fixed coefficient system, then a technique known as Canonical Signed Digital representation of numbers allows multiplication without the use of multipliers [19],[20]. This technique, which makes use of the fact that binary numbers can be represented by a summation of powers of two, allows fixed coefficient multiplies with only a small number of shift and adds. Typically, an 8-bit multiply can be performed with three 8-bit adds and a 8-bit shift register. The use of this technique greatly reduces the power and area necessary to implement the multiply accumulate function, and more closely replicates the functionality of the analog switched capacitor circuit described earlier.

2.4.5.3 Lower Limits of Power Consumption for Digital Circuits

As stated earlier, the power consumption of a digital circuit is given by $1.1CV_{DD}^2f_{clk}$. In theory, as technology improves, both the capacitance driven and the supply voltage can be reduced so that power is minimized. However, as supply voltages are reduced, the large noise margin associated with digital CMOS circuits becomes smaller, and factors usually associated with analog circuits become a concern. For example, with small supply voltages and small devices, which imply small capacitances, the kT/C noise can cause false data to propagate through the circuit. As is known from digital communication theory, the probability of a bit error increases exponentially with falling signal to noise ratios. The bit error rate for a single valued digital signal such as that found in digital circuits can be approximated by [21]:

$$\text{BER} = Q\left(\frac{\left(\frac{V_{DD}}{2}\right)}{\sqrt{\frac{kT}{C}}}\right) = \frac{1}{2}\text{erfc}\left[\frac{V_{DD}}{2\sqrt{2\frac{kT}{C}}}\right] \quad (2.47)$$

assuming that the noise margin is half the supply voltage. The desired BER for digital signal processing systems is on the order of 10^{-12} to 10^{-15} to insure that the system is capable of many hours of error free operation. For a BER of 5×10^{-13} , the minimum gate capacitance required is on the order of 1aF for a V_{DD} of 1V. Clearly, operating at this limit implies scaling by a factor of 100 from today's technologies, perhaps with a line width of $0.01\mu\text{m}$. Thus, for the near future, it is unlikely that kT/C noise is an issue for digital circuits, even when scaled.

A circuit making use of the limit derived above, gates with a capacitance of 1aF and a supply voltage of 1V, has logic gates that operate with 1aJ of energy per transition. The thermal energy

is kT , or $4.143 \times 10^{-21} \text{J}$ at 300K. Thus, at this limit, the circuit is operating at roughly 200 times the thermal noise floor. This compares well with earlier figures which placed the limit between 100 and 200 times kT [22],[23]. Ripple carry adders can be fabricated with about 7 “inverter-equivalent gates” per bit, implying that an 8-bit adder will consume about 50 aJ per operation. Thus, using CSD representations of numbers, and assuming 4 gates per bit for registers, the power required for an 8-bit digital MAC using this limiting technology is about 180 aJ of energy. Therefore, 100 MHz operation implies a power dissipation of 18 nW at for a 1V supply voltage. Scaling this figure to a 3V supply voltage for comparison with the previous section yields a power dissipation of 162 nW. This figure compares well with the theoretical power dissipation for an equivalent analog circuit using a 1 μm gate length. In fact, the power for the digital circuit exceeds that for the analog circuit if a class AB/B circuit is used. Computation of the minimum analog power (class AB/B) with a 0.01 μm gate results in a power of 27 nW for $V_{DD} = 3\text{V}$. This shows that if higher supply voltages are maintained, analog circuitry can be more efficient than digital technologies even in the limit of scaled technologies. Only when the supply voltage is substantially reduced does the inherent swing limitation of analog circuitry tilt the power advantage shift to the digital arena.

2.4.5.4 Power Consumption of Practical Digital MAC Circuits

While the above comparison based on theoretical limits proves interesting, it is more useful to investigate the relative merits of both implementations for technologies currently available and those projects to be available in the near future. It was shown above that the power of digital circuits is simply the CV^2f power, and is not subject to limitations of scaling due to noise for all current and near future technologies. As such, determination of power is equivalent to determining the actual circuit used to implement a CSD based multiply accumulate stage. Digital circuitry offers even more alternatives than analog circuitry in the exact implementation of a function — analysis of each possible circuit design for an adder would be intractable. Thus, a “generic” implementation of the basic building blocks necessary will be used for this comparison. The LAGER [24] logic design package contains a large collection of building blocks which allow the construction of a wide range of digital circuits. The speed, power, and area figures used are those based on LAGER designs scaled for a 1 μm technology.

The choice of the adder topology is most critical, as the maximum throughput of the MAC circuit will be determined by the speed of the adder. Relatively low resolution systems are best

suited with ripple carry adders as the delay is tolerable — 5 to 10 ns for 8 bits — but as the delay increases linearly with word length, become too slow except for systems with low clock rates. When speed is required in combination with long word lengths, the carry select or carry lookahead structure is required, as the delay of the adder is only a weak function of the word length. Of course, the added speed of these techniques comes at the penalty of area and power, with the area of the carry lookahead adder increasing quasi-exponentially with increasing word length. The exact boundary between ripple carry and carry lookahead strategies is heavily technology dependent, but is in the neighborhood of 8 to 10 bits [25].

Determining the power required to perform the MAC task as a function of word length and clock rate requires that certain assumptions be made:

- Use of the CSD technique of multiplication results in $\frac{N}{3}$ *N-bit* adds in lieu of an *N-bit* x *N-bit* multiplication.
- Allow use of ripple carry adders for clock rates up to 200 MHz for 8-bits, decreasing to 100 MHz for 16-bits. Above this, carry lookahead or carry select adders are required.
- To realize an output per clock cycle, pipeline registers are required.
- Ripple carry adders consume 9 pJ per 8-bit section, while carry lookahead adders consume 18 pJ per 8-bit section; pipeline registers consume 3 pJ per 8-bit section. ($V_{DD}=3V$)

with these assumptions, a graph similar to Figure 2.8 can be developed for the digital MAC circuit. Upon examination of Figure 2.10 in comparison with Figure 2.8, it is obvious that the square-law relationship of the power of an analog circuit as a function of dynamic range becomes a big penalty with resolutions in excess of 12 to 14 bits. Thus, signal processing of high resolution signals is best done in the digital domain, provided an accurate data converter exists.

2.4.5.5 Area of Practical Digital MAC Circuits

Much as in the previous section, a set of assumptions can be made to provide a quasi-quantitative comparison of the area required by analog and digital realizations of the circuit function. Before delving into the numerical comparison, it is interesting to explore the theoretical limitations of the area of a digital MAC circuit. Unlike the analog case, where the kT/C noise required that a certain minimum capacitor size be maintained, it was shown in the previous section that for gate capacitances in excess of 1aF, the digital circuit itself is immune from the effects of noise due to small feature sizes. In order for thermal noise to become an issue with digital circuits, the technology would have to scale such that $L_{\min} = 10\text{\AA}$. With this hypothetical technol-

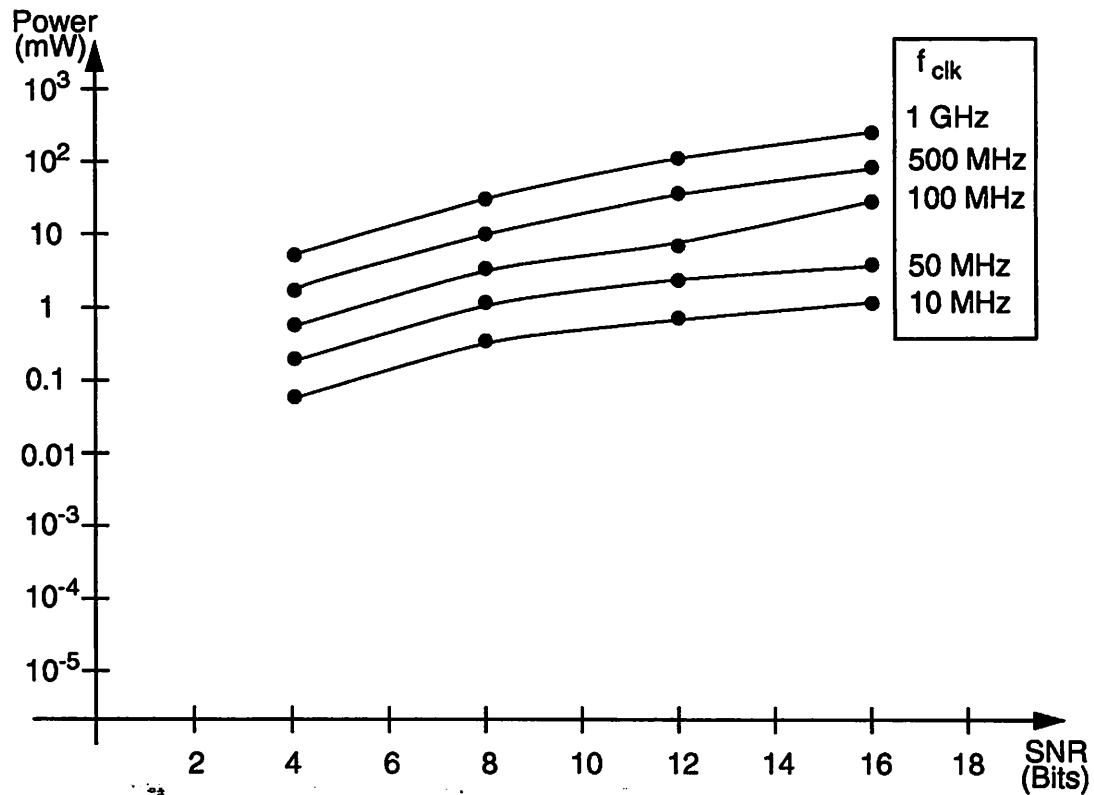


Figure 2.10 Power of Digital MAC circuit vs. SNR and Clock Freq. ($L=1\mu\text{m}$)

ogy, the area for an 8-bit ripple carry adder would be on the order of $2\mu\text{m}^2$, assuming that all dimensions scaled with gate length. As a result, an 8-bit digital MAC stage would take somewhere on the order of 6 to $10\mu\text{m}^2$ of silicon area; this clearly assumes the most aggressive scaling possible, and is unlikely to be realized. However, it does set some theoretical lower bound on the area required.

Quantitative figures of area required to implement the adder and register functions are once again obtained from the LAGER library. While the absolute numbers are very technology dependent, the trends — area as a function of word length and topology — are more important in determining the relative merits of analog vs. digital implementations. Similar numbers are found by Rabaey [26] which shows a linear dependence on area with word length with the exception of lookahead designs, where the area increases quasi-exponentially. The same assumptions relating to the adder style used for the power computations will be applied for the area calculations. Ripple carry adders are assumed to consume $2568\mu\text{m}^2$ per bit, carry select adders consuming 3810

μm^2 per bit, and shifters consuming $1224 \mu\text{m}^2$ per bit, and pipeline registers at $1792 \mu\text{m}^2$ per bit. As seen in Figure 2.11, there is only a weak dependence in the area required on the speed of the circuit, with a linear increase in area as function of word length. This is in stark contrast with the analog case, where area increases exponentially with the word length. As such, it is clear that analog techniques have an area advantage at low resolutions, while at higher resolutions, digital implementations are more area efficient.

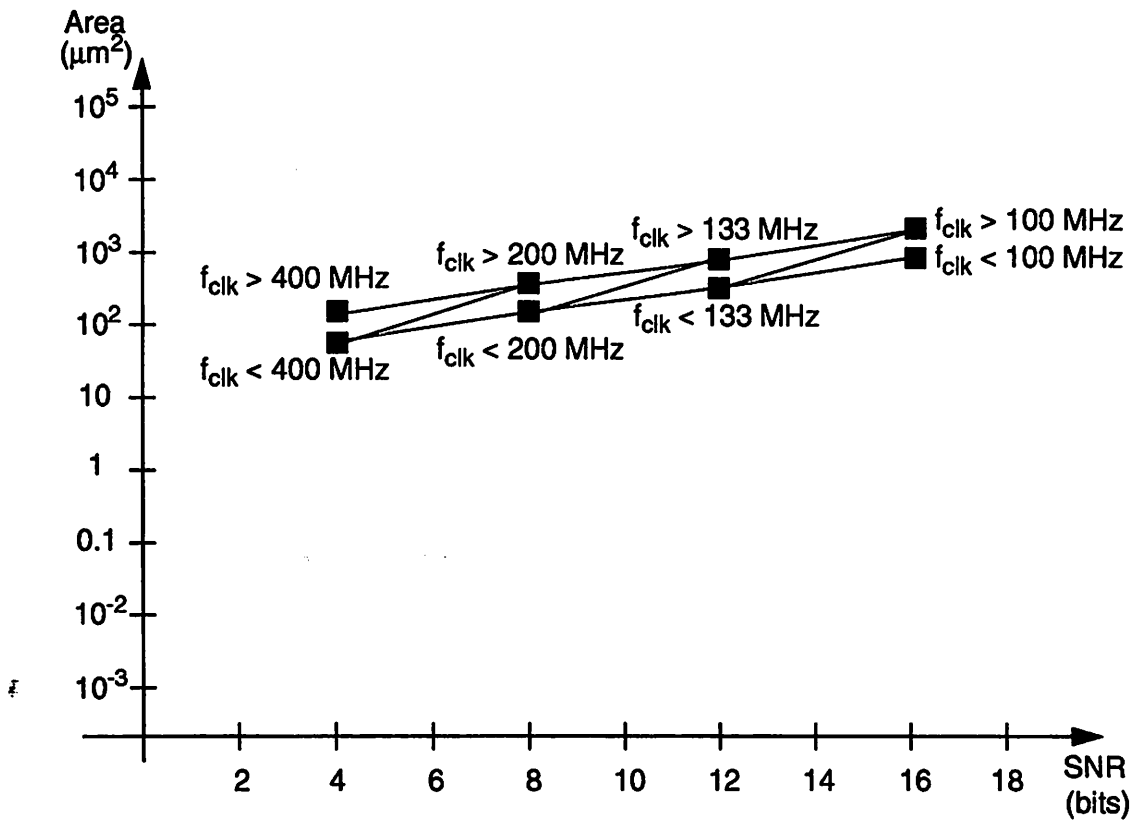


Figure 2.11 Area of Digital MAC Circuit vs. SNR and Clock Freq. ($L=1\mu\text{m}$)

2.5 Conclusions

Through superimposition of Figure 2.8 and Figure 2.10, and Figure 2.9 and Figure 2.11, there are areas where analog is superior to digital and vice versa. It is meaningful to take into account the information shown in Figure 2.5 and present a combination graph which attempts to delineate areas of preferential operation for both analog and digital circuits. Figure 2.12 shows that digital processing is suited best for high precision (high SNR) applications, provided that a

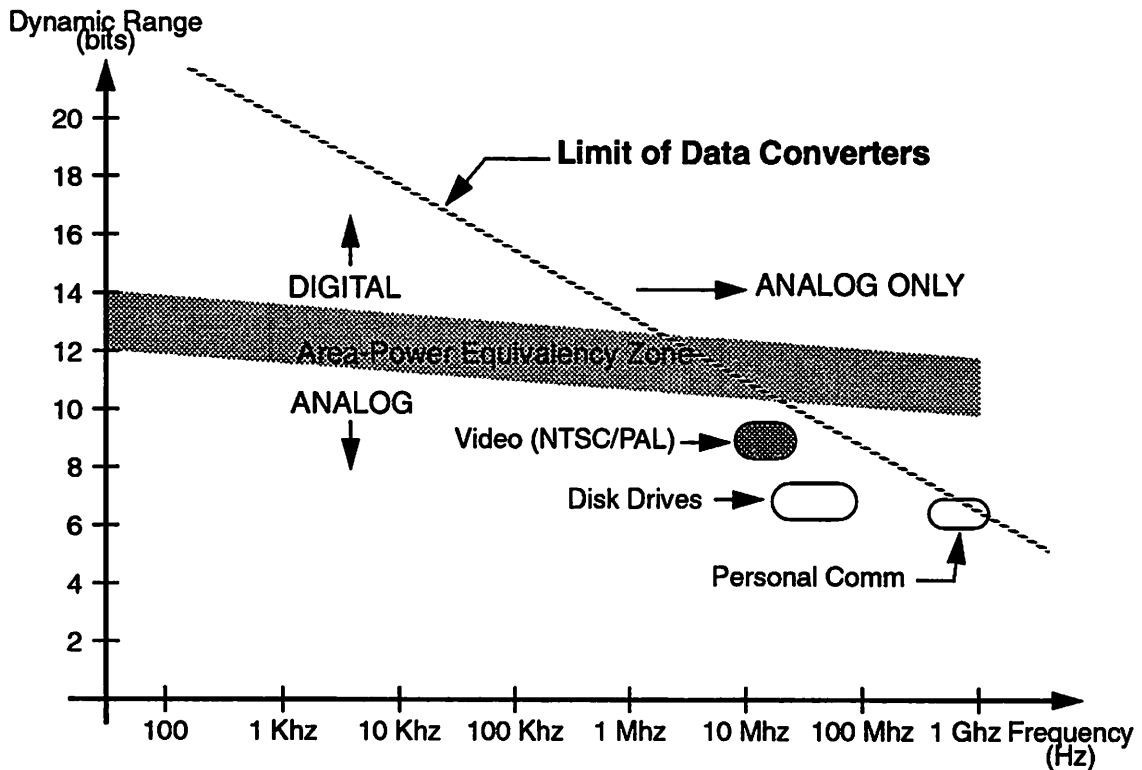


Figure 2.12 Preferred Areas of Operation for Analog and Digital Signal Processors

data converter with sufficient speed exists. Analog signal processing proves to have an area/power efficiency advantage with lower resolution systems, where kT/C and other noise is not an overriding factor. Finally, applications which lie to the right of the dashed line are analog only as data converters to translate between the analog and digital domains are not yet available.

While the above graphs were determined using a $1\mu\text{m}$ technology, recall that a rough estimate of the theoretically minimum power and area independent of technology showed that analog techniques were more efficient at lower resolutions, which is consistent with the results shown above. This is expected, because independent of technology, the trends which make analog processing costly at higher resolutions remain. To add one-bit of SNR to a system, the capacitors in the analog system must be quadrupled in size, while a digital system requires only an incremental addition to the circuit.

Figure 2.12 also shows some common signal processing tasks — it is interesting to note that they all lie within the area where analog processing is preferred. While the convenience and flex-

ibility of digital processing has taken over many of the tasks in the “analog” area, careful circuit design can result in analog processors which are more power and/or area efficient than existing digital signal processors performing the same task.

2.6 References for Figure 2.5

- [a] B. DelSignore, D. Kerth, N. Souch and E. Swanson, “A Monolithic 20b Delta-Sigma A/D Converter,” in *ISSCC Digest of Technical Papers*, Feb. 1990, pp. 170-171.
- [b] P. Ferguson, Jr., A. Ganesan, et. al., “An 18b 20 KHz Dual $\Sigma\Delta$ A/D Converter,” in *ISSCC Digest of Technical Papers*, Feb. 1991, pp. 68-69.
- [c] G. Miller, M. Timko, H.-S. Lee, et. al., “An 18b 10 μ s Self-Calibrating ADC,” in *ISSCC Digest of Technical Papers*, Feb. 1990, pp. 168-169.
- [d] G. Yin, F. Stubbe, and W. Sansen, “A 16-b 320-KHz CMOS A/D Converter Using Two-Stage Third-Order $\Sigma\Delta$ Noise Shaping,” in *IEEE Journal of Solid-State Circuits*, vol. 28, no. 6, pp. 640-647, June 1993.
- [e] A. Karanicolas, H.-S. Lee and K. Bacrania, “A 15b 1Ms/S Digitally Self-Calibrated Pipeline ADC,” in *ISSCC Digest of Technical Papers*, Feb. 1993, pp. 60-61.
- [f] Y.-M. Lin, B. Kim, and P. R. Gray, “A 13-b 2.5 MHz Self-Calibrated Pipelined A/D Converter in 3- μ m CMOS,” in *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, pp. 628-636, April 1991.
- [g] R. Jewett, J. Corcoran and G. Steinbach, “A 12b 20MS/s Ripple-through ADC,” in *ISSCC Digest of Technical Papers*, Feb 1992, pp. 34-35.
- [h] K. Sone, N. Nakadai, Y. Nishida, et. al., “A 10b 100Ms/s Pipelined Subranging BiCMOS ADC,” in *ISSCC Digest of Technical Papers*, Feb. 1993, pp. 66-67.
- [i] J. van Valburg and R. van de Plassche, “An 8b 650 MHz Folding ADC,” in *ISSCC Digest of Technical Papers*, Feb. 1992, pp. 30-31.
- [j] A. Matsuzawa, S. Nakashima, I. Hidaka, et. al., “A 6b 1 GHz Dual-Parallel A/D Converter,” in *ISSCC Digest of Technical Papers*, Feb. 1991, pp. 174-175.

2.7 References

- [1] P. R. Gray, Private Communication
- [2] H. Nyquist, “Certain Factors Affecting Telegraph Speed,” *Bell System Tech. Journal*, vol. 3, pp. 324-346, 1924.
- [3] S. H. Lewis, “Video-Rate Analog-to-Digital Conversion Using Pipelined Architectures,”

- University of California, Berkeley, ERL Memorandum M87/90, 1987.
- [4] C. C. Shih, *Precision Analog to Digital and Digital to Analog Conversion Using Reference Recirculating Algorithmic Architectures*, University of California at Berkeley, Ph.D. Thesis, July 25, 1985.
- [5] R. Gregorian and G. Temes, *Analog MOS Integrated Circuits for Signal Processing*, Wiley, New York, NY, 1986.
- [6] K. C. Hsieh, *Noise Limitations in Switched-Capacitor Filter*, University of California at Berkeley, Ph.D. Thesis, May 18, 1982.
- [7] B. DelSignore, D. Kerth, N. Sooch and E. Swanson, "A Monolithic 20b Delta-Sigma A/D Converter," in *ISSCC Digest of Technical Papers*, Feb. 1990, pp. 170-171.
- [8] B. J. Hosticka, "Performance Comparison of Analog and Digital Circuits," in *Proceedings of the IEEE*, vol. 73, no. 1, pp. 25-29, January 1985.
- [9] G. M. Jacobs, D. J. Allstot, et. al., "Design Techniques for MOS Switched Capacitor Ladder Filters," in *IEEE Trans. Circuits Syst.*, vol. CAS-25, no. 12, pp. 1014-1021, Dec. 1978.
- [10] R. Gregorian, K. Martin and G. Temes, "Switched-Capacitor Circuit Design," in *Proc. IEEE*, vol. 71, no. 8, pp. 941-966, Aug. 1983.
- [11] R. Castello, *Low-Voltage Low-Power MOS Switched-Capacitor Signal-Processing Techniques*, University of California at Berkeley, Ph.D. Thesis, Aug. 20, 1984.
- [12] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3 ed., Wiley, New York, NY, 1993.
- [13] S. M. Sze, *Physics of Semiconductor Devices*, 2 ed., Wiley, New York, NY, 1981.
- [14] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation," in *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 224-231, June 1977.
- [15] Y. Tsvividis, *Operation and Modeling of The MOS Transistor*, McGraw Hill, New York, NY, 1987.
- [16] C. Turchetti, G. Masetti, and Y. Tsvividis, "On the small-signal behavior of the MOS transistor in quasi-static operation," in *Solid-State Electronics*, vol. 26, pp. 941-949, 1983.
- [17] L. B. Jackson, *Digital Filters and Signal Processing*, Kluwer Academic Publishers, Boston, MA, 1986.
- [18] H. J. M Veedrick, "Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits," in *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 468-473, August, 1984.
- [19] H. De Man, *The Digital Filtering Alternative*, Proceedings of Summercourse 1979 — Sampled Analog Signal Processing, Katholieke Universiteit Leuven, pp. 10.1-10.38, June, 1979.
- [20] L. A. Schmidt, "Designing Programmable Digital Filters for LSI Implementation," in *Hewlett-Packard Journal*, vol. 29, no. 13, pp. 15-23, 1979.

The first step in the design process is to determine the requirements of the system. This involves defining the input and output signals, the operating frequency, and the desired performance characteristics. Once the requirements are established, the next step is to select the appropriate components and topology for the implementation.

In the case of an analog implementation, the designer must consider the bandwidth, gain, and noise characteristics of the components. The choice of components is often dictated by the operating frequency and the required accuracy. For example, at high frequencies, the parasitic capacitance and inductance of the components can significantly affect the performance of the system.

On the other hand, a digital implementation offers several advantages over an analog implementation. Digital systems are generally more accurate and less sensitive to component tolerances and environmental conditions. They also offer greater flexibility in terms of functionality and can be easily reprogrammed or modified. However, digital implementations may require more complex circuitry and higher power consumption, especially at high frequencies.

The decision between analog and digital implementations often depends on the specific requirements of the application. For example, in a low-frequency, high-precision application, an analog implementation may be preferred. In contrast, for a high-frequency, low-precision application, a digital implementation may be more suitable.

Ultimately, the designer must carefully evaluate the trade-offs between the two implementation methods and select the one that best meets the requirements of the system. This process often involves prototyping and testing both implementations to compare their performance and identify any potential issues.

CHAPTER 3

Overview of NTSC System

3.1 Introduction

Developed in 1953 as a method of transmitting color video signals compatible with existing monochrome transmissions, the NTSC [26],[27],[28] system represents a compact, relatively robust system which is used throughout North America and parts of the Far East. Although the intricate details of the system are beyond the scope of this dissertation, the first portion of this chapter will attempt to serve as an overview aid in understanding the theoretical basis for the prototype circuit described later. Emphasis will be placed on the electrical nature of the video signal and topics that are of direct interest to the experimental work. Those that seek a more detailed explanation are encouraged to seek out one of many excellent references on the NTSC system.

The latter part of this chapter will introduce basics of NTSC decoding — that is, how to recover the original RGB values from the encoded NTSC signal. In particular, advanced methods such as comb filters will be discussed to give a theoretical background for the prototype circuit described later in this dissertation.

3.2 Raster Scanning of an Image

A primary task of a video system is to transform an image, which is inherently two-dimensional (three, in the case of a motion picture) into a single-dimensional electrical signal. This transformation is accomplished using a process known as raster scanning. A form of sampling,

raster scanning traces a locus across the image which moves at a constant rate in both the horizontal and vertical directions, with the rate of motion in the horizontal axis much greater than that in the vertical axis. If a restriction is placed such that the horizontal rate is an integral multiple of the vertical rate, then a repetitive scanning pattern results. This action allows the two-dimensional image to be closely approximated by values of the image directly under the locus. In the case of a monochrome image, the value is simply the localized brightness of the image. For color images, the image needs to be described by a more complex system, which will be discussed later in this chapter. A pictorial representation of the scanning process is shown in Figure 3.1. The locus begins at point *A*, scanning across the image to point *B*. During the period

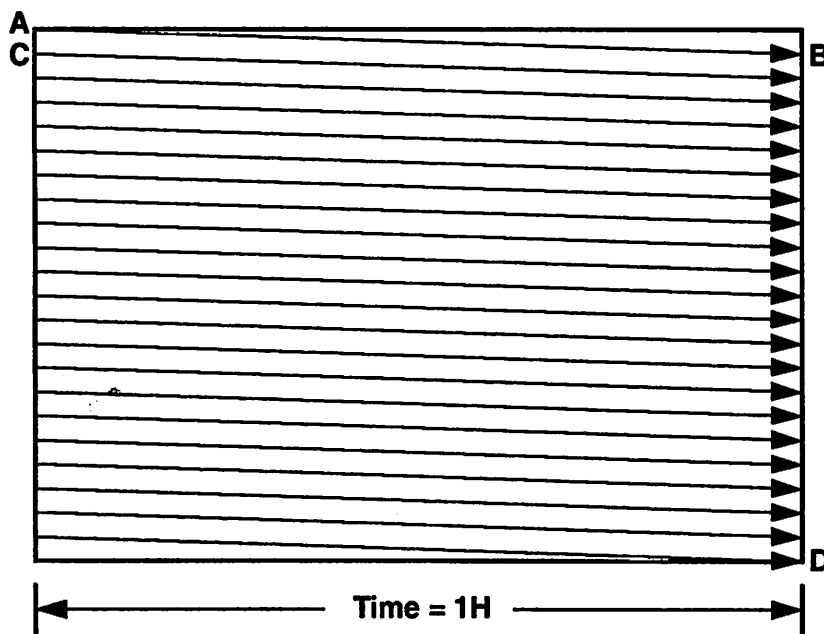


Figure 3.1 Raster Scanning of an Image

known as horizontal blanking, the locus quickly moves back to point *C* at the left edge. The above process is repeated until the lower right corner is reached at point *D*. The single scanned image is referred to as a frame, and motion pictures can be transmitted by repeating the scanning process for each frame of the motion picture. The period of time taken for the locus to move from point *D* back to point *A* in the next frame is known as the vertical retrace. Note that even with a still image, the resultant image is time-varying because of the 2-D to 1-D transformation. In most video related literature, and in this dissertation, the period of time required to traverse the

image horizontally will be called the line period or H . Likewise, the frame period, the time taken to scan a complete image, will be denoted V .

The NTSC system calls for 525 scan lines with a frame rate of 29.97 Hz, yielding a line rate of 15.734 KHz. In practice, however, the scanning is interlaced to reduce flicker effects visible at the 29.97 Hz rate. Interlacing divides the raster into odd and even lines, effectively doubling the vertical displacement during horizontal retrace. The area of the complete image is scanned in half the frame rate. This “half-frame” is known as a field. During the second scan (field), the scanning locus is shifted slightly to fill in the “gaps” left by the first scan. This process creates the illusion of doubling the frame rate as far as flicker is concerned, without increasing the overall amount of information in the signal. The vertical retrace interval consumes 41 of the 525 lines in the frame, yielding 484 visible lines. However, interlacing causes a visible reduction in vertical resolution known as the Kell Factor [26], yielding an approximate vertical resolution of 340 lines. The horizontal resolution is determined by the allowable bandwidth of the signal as shown in the next section.

At the receiver, the raster must be recreated to reconstruct the image. In order to synchronize the scanning processes at the source and receive points, synchronization signals are inserted at the beginning of each line and at the end of each field. These synchronization signals are negative voltage pulses that extend below the nominal DC reference voltage.

3.2.1 Spectral Analysis of the Raster Scanned Image

The resultant time-varying signal, representing the value of the image as the locus traces its path can be analyzed using Fourier transform techniques to obtain the frequency domain representation. For simplicity, consider a stationary, monochrome image — the principles developed can be easily extended to modern color images. Using H and V to represent the horizontal and frame rates respectively, 2-D Fourier analysis results in the following [29]:

$$x_{mn} = \frac{1}{HV} \int_0^H \int_0^V F(h, v) \exp \left[-j2\pi \left(\frac{mh}{H} + \frac{nv}{V} \right) \right] dh dv \quad (3.1)$$

$$F(H, V) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x_{mn} \exp \left[j2\pi \left(\frac{mh}{H} + \frac{nv}{V} \right) \right] \quad (3.2)$$

where $F(H, V)$ is the image quantity of interest (intensity) and x_{mn} is the Fourier component at spatial frequency (m, n) . Thus, the video signal can be represented by:

$$y(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x_{mn} e^{j2\pi(mf_h + nf_v)t} \quad (3.3)$$

where f_h and f_v are the horizontal and vertical scanning rates respectively. A key property of this signal is that it is doubly periodic in f_h and f_v . The quantity x_{mn} is usually a monotonically decreasing function of m and n because most images contain less energy corresponding to high spatial frequencies. This result is expected as raster scanning is akin to sampling in the vertical direction.

Because f_h is much higher than f_v (15.734 KHz vs. 59.94 Hz in the NTSC system), the maximum frequency component (total bandwidth) of the video signal $y(t)$ in Equation 3.3 is determined by the highest spatial frequency component in the horizontal dimension of interest. This directly corresponds to the horizontal resolution of the scanned image — a more detailed image will have larger values of m for which x_{mn} is a significant quantity. The NTSC system allows a total bandwidth of 4.2 MHz, which provides a horizontal resolution of approximately 340 lines, which agrees with the perceived vertical resolution, thereby satisfying the desire to equate resolution in both axes.

Motion pictures removes the periodicity between frames, resulting in the blending of distinct frequency components at intervals of f_v to form a continuous spectrum (Figure 3.2). The widths of each “clump” of energy — spaced at intervals of f_h — is dependent on the spatial frequency of the image. Images with high spatial frequencies in the vertical dimension (poor line to line correlation) will tend to spread out the clumps, while high frequencies in the horizontal dimension will tend to extend the series of clumps into higher frequencies, thereby increasing the overall signal bandwidth.

3.3 Principles of Color Video

3.3.1 Additive Color Theory

Before discussing specifics of the NTSC color system, a brief review of color theory is in order. Video images, unlike those printed on paper, consist of light generated by the screen rather

than reflected light. Therefore, additive color theory is applicable rather than the more familiar subtractive color theory. As a result, the primary colors are red, green and blue (RGB) instead of cyan, magenta, and yellow. Any color can be represented as a mixture of the three primary colors, and this is the system most often adopted in computer graphics. However, an equivalent method of representing color known as HSV (hue, saturation, brightness) is more applicable to video.

A color can be divided into its hue (tint), saturation (deepness of color), and value (brightness). Hue is the actual color or shade (e.g. red, green, blue), while saturation is the intensity or purity of that color. Red and pink consist of the same hue (red), but pink is not fully saturated, while red is — any color can be desaturated by the addition of white light. Finally, brightness is the overall intrinsic luminosity of the color. Colors such as yellow are brighter than others such as blue, and for a given hue, desaturated colors are in general brighter than their saturated counterparts. When dealing with a monochrome system, this intrinsic luminosity is all that is transmitted. For accurate color transmission, all three of these components need to be reproduced. Because the luminosity by itself carries a good portion of the image information, it is typically

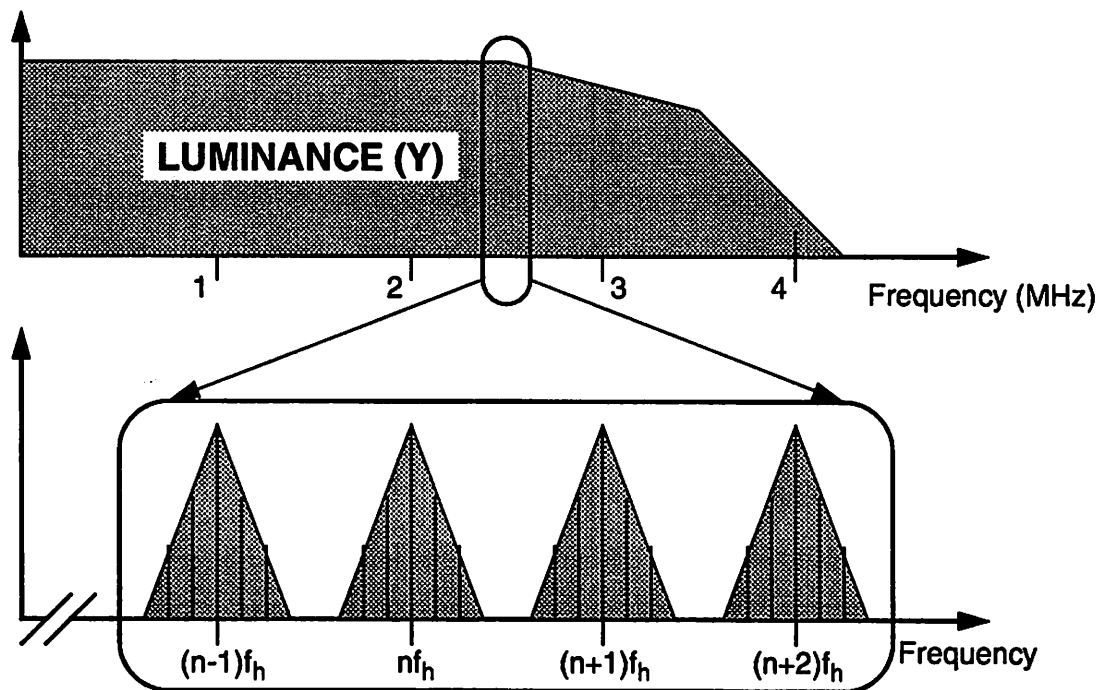


Figure 3.2 NTSC Luminance Spectrum (magnified inset)

separated out and treated as one component called luminance (Y), while the hue and saturation are coupled and treated as another component called chrominance (C).

3.3.2 Color Vision Characteristics of the Human Eye

Although the goal of the NTSC system is to provide as precise as possible reproduction of the original image at the receiver, the goal to reduce the overall amount of information transmitted makes it advantageous to match the characteristics of the system to the final receiver of the image, the human eye [30]. The human eye utilizes two separate light sensitive structures to render images. Rods, which are more plentiful, are sensitive to the presence or absence of light, and hence contribute to monochrome vision. Cones, on the other hand, are color sensitive but fewer in number. As a result, spatial resolution of a color pattern is lower than that for a monochrome pattern. The sharp edges in a color image are distinguished by the change in the luminosity of the color, rather than the change in hue. Thus, a color transition with a relatively long transition in hue will appear sharp if it is accompanied with a sharp edge in luminosity. This property makes separating a color image into the luminance and chrominance components even more advantageous since the resolution in the chrominance channel can be reduced without adversely affecting the overall image quality as long as the luminance channel retains the full resolution of the original image.

Moreover, the resolution of the human eye is not constant for all colors, with the highest resolution in the orange and cyan hues, and correspondingly less in the magenta and green. This fact is used in the NTSC system to allocate more bandwidth toward those colors which require more resolution. Thus making full use of the knowledge of the physiological limitations of the eye allows the NTSC system to reproduce a high quality image while occupying limited bandwidth.

3.4 The NTSC Color Video Signal

3.4.1 Methods of Transmitting Color Information

The most straightforward method of transmitting a color image would be to record the red, green, and blue components of each image point under the raster scan locus in the same way as brightness is recorded for a monochrome image. However, this method, known as RGB transmission requires three times the bandwidth of an equivalent monochrome system, and would be incompatible with the previously set monochrome transmission standard. The FCC mandate that

any new color transmission standard be compatible with the existing monochrome standard, and also result in no net signal bandwidth increase ruled out RGB as a method of color transmission.

The first requirement of monochrome compatibility dictates that the luminance information of the image must be transmitted in roughly the same method as in a monochrome system independent of the present of additional chrominance information. Thus, the RGB values from scanning the color image are decomposed into the luminance and chrominance components as described earlier. Specifically, the luminance component is a linear combination of the red, green and blue values as follows:

$$Y = 0.30R + 0.59G + 0.11B \quad (3.4)$$

Because color theory follows the rules of linear space, two other signals are required in addition to the Y signal to transmit the equivalent RGB value. Given Y , the two components $(R - Y)$ and $(B - Y)$ provide sufficient information to recover the original RGB values. (The value $(G - Y)$ could have been used in lieu of one of the other two, but as green is the largest component of Y , the $(G - Y)$ signal would be statistically smaller in value thus more susceptible to noise.) In literature, $(R - Y)$ is sometimes referred as Pr , and $(B - Y)$ as Pb — collectively, these Pr and Pb signals are also known as color difference signals. Transforming the RGB values into the YP_rP_b values can be viewed as a basis transformation as shown by the following matrix equation:

$$\begin{bmatrix} Y \\ Pr \\ Pb \end{bmatrix} = \begin{bmatrix} 0.30 & 0.59 & 0.11 \\ 0.70 & -0.59 & -0.11 \\ -0.30 & -0.59 & 0.89 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3.5)$$

In summary, the resultant Y signal is used to transmit the monochrome component to maintain compatibility with the existing monochrome system. The full 4.2 MHz bandwidth is allocated to the Y signal to preserve as much resolution as possible. The remaining two signals, Pr and Pb , can be viewed as the “color” portion of the images, and contain the remainder of the information necessary to recreate a full color image. However, because the limited resolution requirements of the color component of the image, the bandwidth allocated to the Pr and Pb signals can be substantially less than the 4.2 MHz allocated to the Y signal; experimentation has shown that 1 MHz is sufficient bandwidth for the chrominance (Pr and Pb) component.

3.4.2 The NTSC Color Encoding System

Transmitting the low bandwidth chrominance information concurrently within the original 4.2 MHz bandwidth of the luminance signal without upsetting the original luminance component is a key aspect of the NTSC system. The salient feature of the system is the frequency division multiplexing of a subcarrier which has been quadrature amplitude modulated by the two chrominance components. This multiplexing is possible due to the periodicity of the spectrum due to the scanning process. For most images, the energy of the luminance signal is concentrated in clumps centered at multiples of the line frequency, f_h . Thus, a method that inserts the color information in between these clumps will allow recovery of the original luminance signal, although considerable effort may be necessary to achieve this separation — the topic of the prototype circuit introduced in this dissertation.

The sampling process which results in the periodicity of the luminance signal also results in periodicity of the chrominance signal. Left alone, the energy peaks would also coincide with multiples of f_h . However, amplitude modulation of a subcarrier by the chrominance signal effectively shifts the spectrum of the signal by the frequency of the subcarrier. Thus, if the subcarrier is chosen such that its frequency, f_{sc} , is in between two adjacent multiples of f_h , ($f_{sc} = \frac{1}{2}(2n+1)H$), the energy peaks of the modulated subcarrier will lie in between the peaks of the luminance signal, as shown in Figure 3.3.

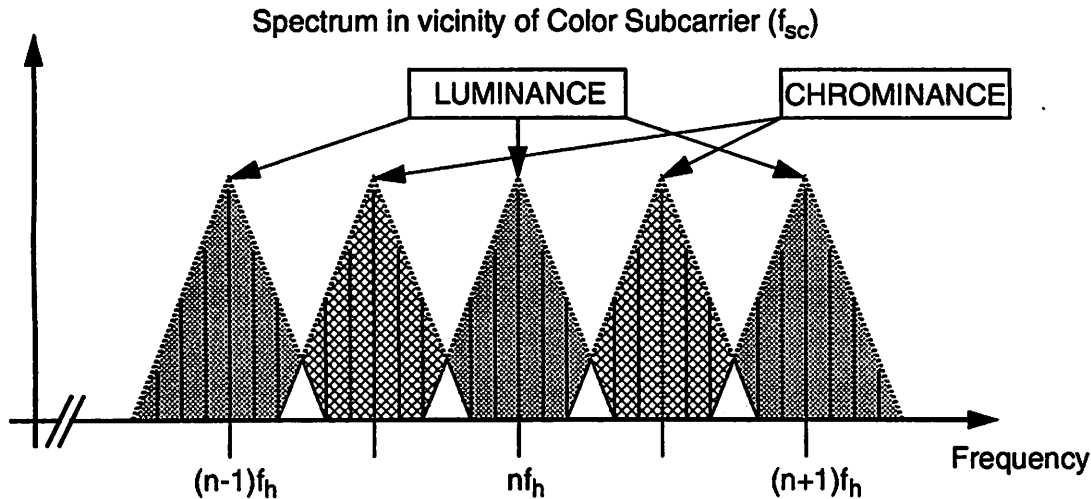


Figure 3.3 Frequency Interleaving of Luminance and Chrominance Signals

The use of quadrature amplitude modulation (QAM) allows encoding of both components of the chrominance signal (P_r and P_b) on a single subcarrier. The NTSC standard sets the subcarrier frequency, f_{sc} , at $455/2$ times f_h , or 3.57954525 MHz. A small problem with this choice of frequency is that the highest frequency of the modulated chrominance signal would extend beyond the 4.2MHz limit. Recalling that the eye is sensitive to some colors more than others, a small linear transformation is made from $YPrPb$ space to YIQ space. This change in color space basis vectors allows the I (in-phase) component to represent colors which have higher spatial resolution, while representing the less sensitive colors with the Q (quadrature) component. Prior to modulation, the I signal is bandlimited to 1.3 MHz and the Q signal to 600 KHz. After modulation, the composite ($Y + \text{modulated } IQ$) signal is bandlimited to 4.2 MHz. The I signal therefore has some of its upper sideband cut off, resulting in some crosstalk between the I and Q channels, requiring a more complex receive filter. The final NTSC spectrum is shown in Figure 3.4 — note the relative placement and energy density of the three components. In practice, it is common to

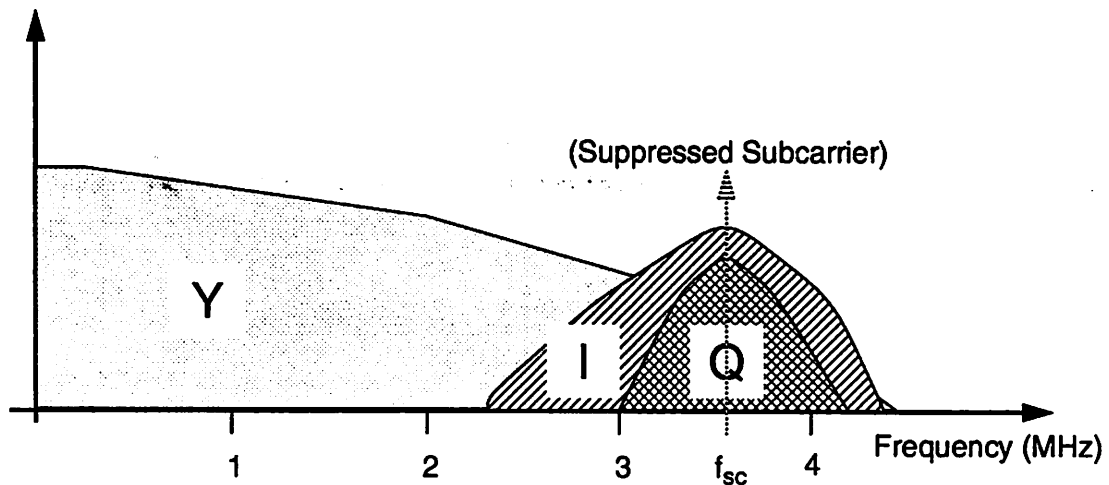


Figure 3.4 Overall NTSC Composite Frequency Spectrum

bandlimit both components to simplify the receiver. The modulated IQ signal is commonly referred to as the chrominance or C component. To avoid confusion with the unmodulated color components, the abbreviation Y and C will be used for the remainder of this dissertation to represent the luminance and modulated chrominance signals respectively.

Successful demodulation of the C signal at the receiver requires knowledge of the phase of the suppressed subcarrier used at the transmitter. Thus, a reference signal or colorburst signal is added immediately after each horizontal synchronization signal. This burst signal is simply a

short segment of the subcarrier of known phase. The receiver uses this information to lock its local demodulating oscillator to retrieve the I and Q components.

In summary, the NTSC signal is the summation of four components: (1) the luminance (Y) signal representing brightness, (2) the chrominance (C) signal representing the color information, (3) the synchronization pulses to recreate the raster and (4) the colorburst to provide phase reference information for the QAM demodulator. The NTSC standard calls for a 1V(p-p) signal with the reference level being 286 mV above the bottom of the sync signal tip. Peak white, representing the brightest area of an image, is 714 mV above the reference level, with black being roughly 5 mV above the reference level (Figure 3.5). A typical NTSC waveform is shown in

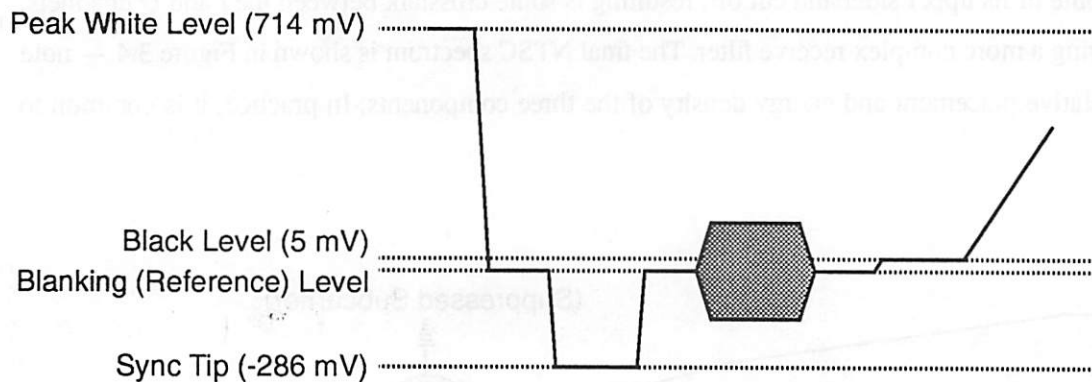


Figure 3.5 NTSC Signal Levels and Construct of Horizontal Blanking Signal

Figure 3.6.

3.5 PAL and SECAM Color Encoding Methods

In addition to the NTSC system, two other systems are in use within the world to achieve the same goal. PAL (Phase Alternation by Line) is prevalent in Western Europe and South America, while SECAM (Séquentiel Couleur à Mémoire) is used in France, Eastern Europe and the Middle East. Both these systems are similar to NTSC in that they separate the image into the luminance and color components. However, the specific method of encoding and adding the color information to the existing luminance information differs among the three systems.

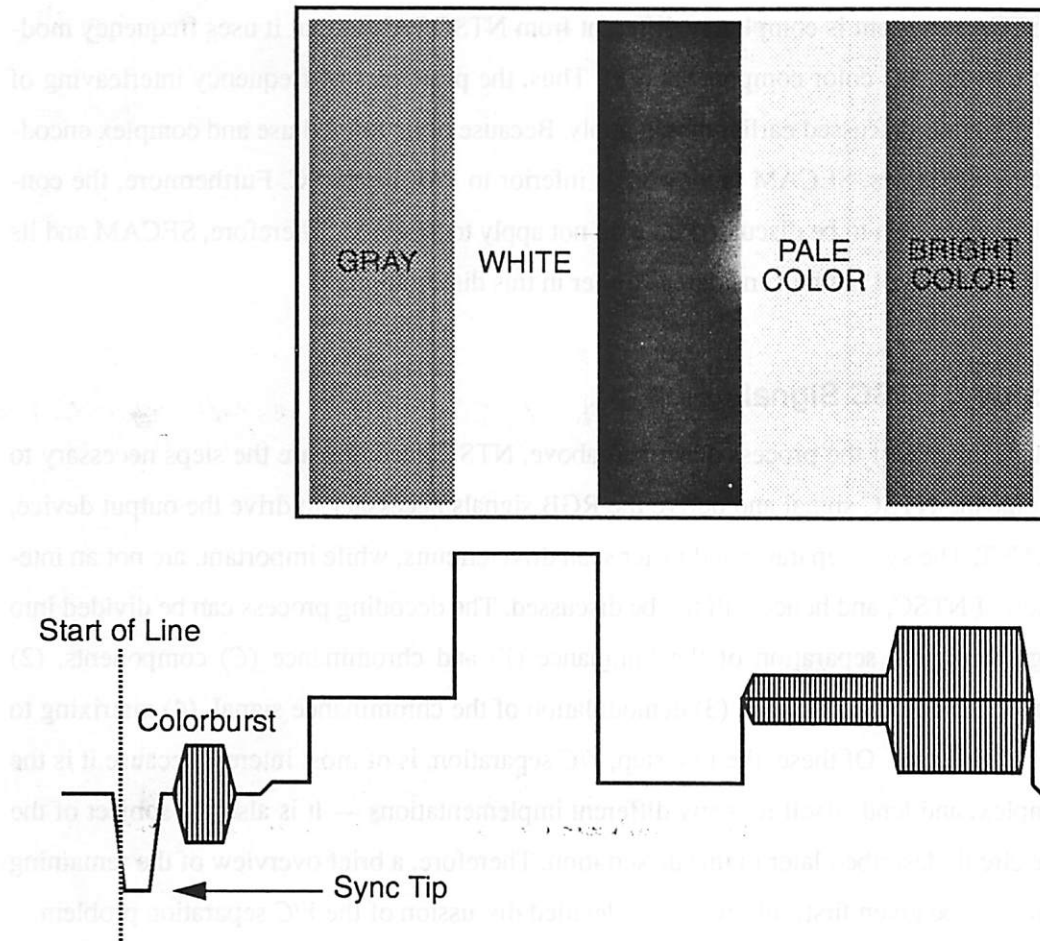


Figure 3.6 NTSC waveform resulting from image at top of figure

3.5.1 PAL

PAL is quite similar to NTSC in that it utilizes a QAM subcarrier. The main difference is that the reference phase of the subcarrier used to modulate the color information is inverted from line to line [31]. The advantage of this complexity is that the effect of phase distortion in the transmission path, which affects the QAM chrominance information, is cancelled. The resultant phase distortions will cause hue shifts in one direction on one line, and in the other direction for the remaining line. The eye will then integrate and cancel out the hue shift. This system is therefore more robust than NTSC, and affords color stability to the degree that hue and tint controls are not usually needed on PAL receivers. As the work discussed in this dissertation focuses on NTSC, the specific changes necessary to encode and decode PAL signals vs. NTSC signals will not be discussed.

3.5.2 SECAM

The SECAM system is completely different from NTSC/PAL in that it uses frequency modulation to transmit the color components [28]. Thus, the principles of frequency interleaving of the Y and C signals discussed earlier do not apply. Because of its limited use and complex encoding/decoding processes, SECAM is viewed as inferior to PAL or NTSC. Furthermore, the concepts of Y/C separation to be discussed later do not apply to SECAM. Therefore, SECAM and its derivative systems will not be considered further in this dissertation.

3.6 Decoding NTSC Signals

A rough inverse of the process described above, NTSC decoding are the steps necessary to take a composite NTSC signal and derive the RGB signals necessary to drive the output device, usually a CRT. The sync separator and raster scan drive circuits, while important, are not an integral portion of NTSC, and hence will not be discussed. The decoding process can be divided into four rough steps: (1) separation of the luminance (Y) and chrominance (C) components, (2) extraction of the colorburst signal, (3) demodulation of the chrominance signal, (4) matrixing to yield the RGB output. Of these, the first step, Y/C separation, is of most interest because it is the most complex, and lends itself to many different implementations — it is also the subject of the prototype circuit described later in this dissertation. Therefore, a brief overview of the remaining three steps will be given first, followed by a detailed discussion of the Y/C separation problem.

3.6.1 Colorburst Extraction

Because of the QAM encoding of the two color difference signals, the phase of the suppressed carrier must be known to properly demodulate the chrominance signal. To aid this process, the colorburst signal is added to the NTSC signal at the beginning of each scan line. Consisting of 9 cycles of subcarrier with a known phase, the receiver circuit must phaselock to this signal to provide a reference carrier for demodulation. Early receivers used injection locked oscillators which are basically high-Q resonant circuits tuned to the subcarrier frequency. As the time interval of burst is known relative to the sync pulse, a switch could be used to apply the subcarrier signal to the resonant circuit during the burst interval. The high-Q nature of the circuit provided a lasting oscillation during the remainder of the line period.

The injection oscillator, while robust, required a number of precision tuned components — the vast majority of current receivers use PLL techniques. Here, a quartz crystal at the subcarrier

frequency is used to run a VCO. A phase comparator is used to compare the burst phase with that of the VCO. Feedback is applied to the oscillator to lock the frequency and phase to provide the proper demodulating signal. Because the period of the burst signal is small compared to the line period, direct phase comparison can be difficult. A better method is to use the VCO output to demodulate the burst signal itself. As the burst signal has known phase, a properly tuned VCO should result in a known output. An error in the output of the demodulator can be used to correct the VCO phase.

3.6.2 Chrominance demodulation

Demodulation of the chrominance signal is fairly straightforward — the oscillator output from the VCO described in the last paragraph is applied to two signal paths with a 90° phase shift is applied to one path. These two signals are mixed with the chrominance signal to yield the I and Q color difference components along with some out of band components which are subsequently filtered out.

While two balanced mixers are required in the continuous time domain, a judicious choice of the sampling rate in discrete time systems can make the demodulation process trivial. A sample rate of $4f_{sc}$ allows the demodulation to be performed by polarity inversion. This is because a sine or cosine at f_{sc} sampled at $4f_{sc}$ is the stream $-1, 0, 1, 0, -1, \dots$. This greatly simplifies the demodulation, and in absence of overriding reasons, makes $4f_{sc}$ the choice of sampling rates for chrominance processing in the sampled data domain.

In situations where the sampled bandwidth is important, the NTSC signal can be sampled at $3f_{sc}$. In this case, the sampling points are no longer vertically aligned, and the number of samples per line is no longer constant. This makes construction of vertical filters such as comb filters difficult. The PALE (Phase Alternating Line Encoding) method [32] reverses the phase of the sampling instants from line to line, and produces vertically aligned samples. However, the added complexity of such a system is rarely justifies the extra circuitry involved.

The out of band components are removed by low pass filtering, which in the sampled data domain can be realized as a $1/8$ rate low pass filter. As with all other video filters, the group delay characteristics are important, and linear phase throughout the passband is very desirable, thereby making sampled data FIR filters very attractive.

3.6.3 De-matrixing of the chrominance signal

Subsequent to demodulation, the Pr , Pb and Y signals must be matrixed to yield the RGB signals necessary for image output. Because the Pr , Pb and Y signals were originally obtained from the RGB signals using a linear transformation, the exact opposite operation will return the RGB values from the Pr , Pb and Y signals. [33]

3.7 Y/C Separation

The task of separating the luminance and chrominance components of the NTSC signal is perhaps the most challenging and critical aspect of NTSC decoding. Techniques to perform the separation vary from the very simple to extremely complex, with varying degrees of performance. Moreover, degree of separation affects the number and nature of image artifacts due to mixing of the chrominance and luminance information [34],[35],[36]. While simple techniques such as bandpass separation produced results satisfactory for consumer use, the advent of large screen picture tubes and multimedia applications, where artifacts are much more visible, have created a need for more advanced and effective methods of Y/C separation.

3.7.1 Effects of Incomplete Separation

Incomplete separation results in two effects, cross-color and cross-luma. Cross-color occurs when high frequency luminance information is misinterpreted as chrominance information. Demodulation of these high frequency Y components, especially those close to the subcarrier frequency, result in spurious signals in the color difference signals. These unwanted signals manifest themselves as image artifacts — for example, the striped shirt of a sports referee contains high frequency luminance patterns which will cause a rainbow like color pattern to appear over the stripes of the shirt.

The second effect, cross-luma, is the opposite, and is the result of chrominance components being interpreted as luminance. Because of the choice of subcarrier frequency, leakage of the chrominance into the luminance channel is masked to a considerable degree. However, in severe cases, it manifests itself as a dot pattern in the image. This is often seen in large areas of bright color with little change in luminance information.

3.7.2 Bandpass separation

Referring to Figure 3.4, separation of the Y and C components can be performed by band-splitting techniques. Because the significant bandwidth of the C component is approximately 1 MHz, a 2.5 MHz low pass filter will produce a signal that is nearly completely Y . Similarly, a 3.58 MHz bandpass filter with a bandwidth of 1 MHz will yield a signal that contains most of the C signal. However, there are two serious drawbacks to this method. First, the resultant Y signal has substantially reduced bandwidth causing a serious degradation of the overall horizontal resolution of the image. Second, the output of the bandpass filter contains high frequency Y components.

The first limitation can be overcome by the use of a notch filter in lieu of a low pass filter. However, this increases the chance of cross-luma because sharp transitions in the chrominance signal will cause a considerable amount of the chrominance energy to lie outside of the notch. Because of this effect, receivers that use this technique will often suffer hanging dots at vertical color transitions — for example, between the bars of a colorbar pattern.

The second limitation of this method is harder to circumvent. Because a bandpass filter allows all the energy within a range of frequencies to pass, it cannot distinguish between the peaks of luminance energy and peaks of chrominance energy. Thus, images with large amounts of high frequency luminance information will cause substantial amounts of cross-chroma to occur, with its image artifacts.

3.7.3 Comb Filters for Y/C Separation

The choice of the subcarrier frequency in the NTSC system places the peaks of the chrominance signal in between peaks of the luminance signal. Therefore, a specialized filter whose frequency response is adjusted to pass the peaks of the luminance signal while rejecting peaks of the chrominance signal would act as an ideal Y/C separator. Such a filter would have multiple transmission zeros at multiples of the line frequency, f_h , while allowing signals between the zeros to pass unhindered [37]. Such a frequency response is known as a comb filter response due to the characteristic shape of its magnitude response (Figure 3.7).

Such a filter may appear to be difficult to construct at first, with multiple zeros spaced at relatively close intervals. However, a simple two tap FIR filter will result in such a response if the delay in between tap weights is the reciprocal of the zero spacing. Thus, for this application, the

delay element should be $1H$ in length, or $63.5 \mu s$ in the NTSC system. Figure 3.8 shows a simple

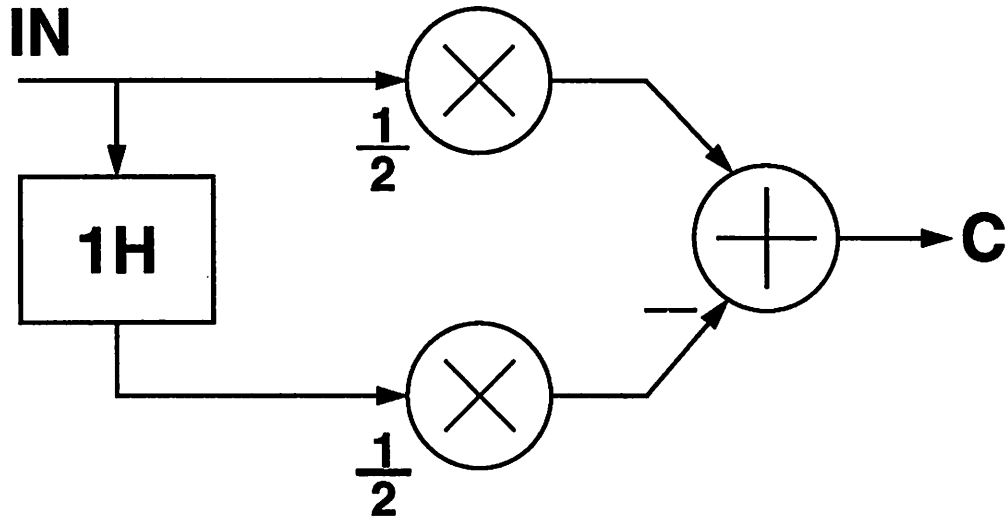


Figure 3.8 2-Tap (1H) Comb Filter Structure

schematic of such a filter. Note that due to the difference being taken at the output summer, this filter removes luminance peaks and leaves chrominance peaks. A true addition would result in the opposite output, with the filter removing the chrominance signal.

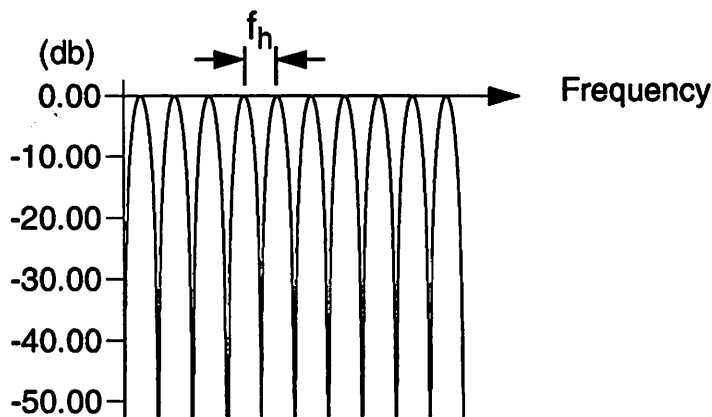


Figure 3.7 Magnitude Response of Video Comb Filter

A parallel approach to understanding comb filters can be undertaken in the time domain. Recall that the frequency of the subcarrier is $455/2 f_h$. This implies that there are an integral number plus one-half cycles of subcarrier per line period. As a result, the absolute phase of the unmodulated subcarrier inverts from line to line. Most real images have a high degree of correlation in image information between adjacent lines in both the luminance and chrominance domains (Figure 3.9). As a result, addition of two adjacent lines reinforces the luminance while

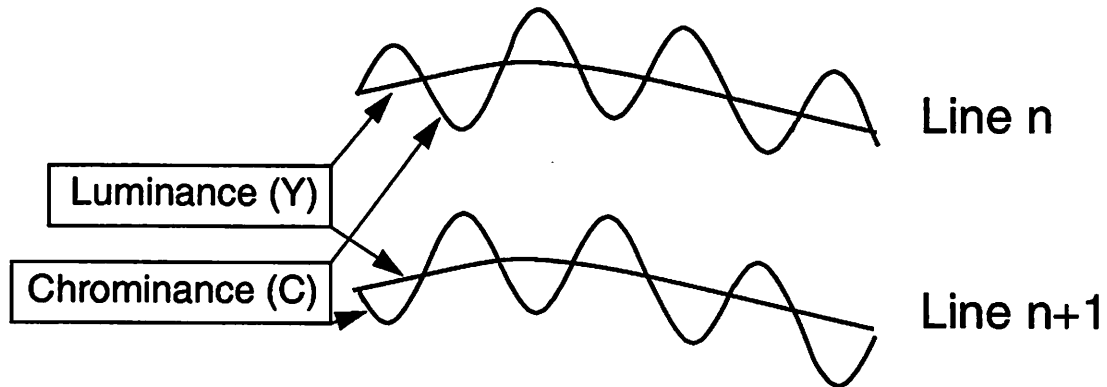


Figure 3.9 Diagram showing correlation of Y and C components between adjacent scan lines

cancelling the chrominance; subtraction results in chrominance with cancellation of luminance. As expected, these results are identical to those obtained by frequency domain analysis. In limited cases, the assumption of correlation fail, with incomplete cancellation or even enhancement of cross color/luma artifacts— these will be discussed later.

3.7.3.1 2H Comb Filters

The structure shown in Figure 3.8 provides a single zero in the frequency response at intervals of f_h . More delay elements can be cascaded to achieve a higher order filter. Of particular interest is the “2H” or 3-line filter shown in. This structure utilizes two delay elements to take information from three adjacent lines, resulting in a double zero spaced at intervals of f_h . This leads to better separation of the two components. However, a bigger advantage is that the group delay of the filter is one line, as opposed to half a line for the simple 1H filter. This makes it easy to compensate for the group delay — the structure shown in Figure 3.10 can be used to provide a chrominance output and a group delay compensated composite output. Finally, as will be discussed below, 2H filters allow use of adaptive algorithms to provide the best performance under all image conditions, including those that would give trouble with simple 1H filters.

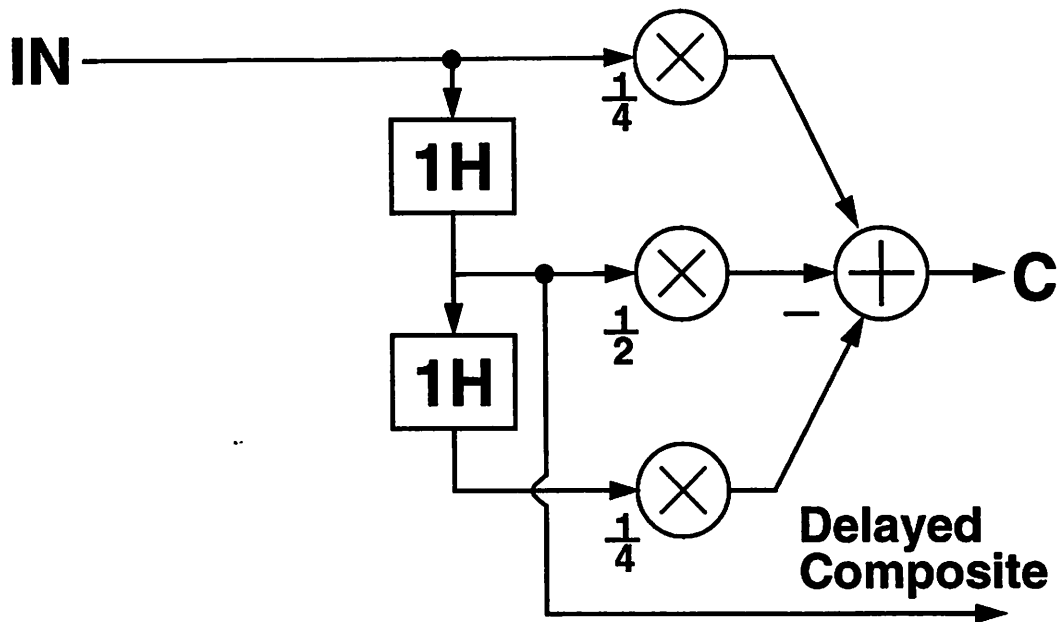


Figure 3.10 Diagram of 2H Video Comb Filter

In all these comb filters, a design decision is made to generate chrominance or luminance from the comb filter. The complementary signal is derived by subtracting the comb filter output from the composite signal. The integral line group delay of the $2H$ filter makes this operation easier — $1H$ comb filters require some type of half-line delay compensation before the subtraction, which adds complexity and cost.

3.7.3.2 Adaptive Comb Filters

A major problem with comb filters is that they depend on the line to line correlation present in most images. However, there will always be a class of images or portions of images where this assumption is false. The most common problem area is that of a vertical color transition — where the color changes dramatically from one scan line to next. With the advent of computer generated graphics, such transitions are becoming more common. To demonstrate the problem caused by such a transition, assume that one line is red and the next line is blue. Where there should have been a phase inversion (or very nearly so) in the chrominance signal, there now exists a phase coherence. Thus, a signal intended to be devoid of chrominance energy now contains a large amount of unwanted chrominance signals. This also upsets the luminance signal

because it is usually derived by subtraction of the now corrupt chrominance signal from the composite signal.

Adaptive comb filters [35],[38],[39] address this problem by looking at the chrominance signal across several scan lines to detect marked changes in chrominance information. At that point, they either switch to a simple bandpass Y/C separator, or switch to a simple IH filter where the two lines being summed do not have a chrominance transition. Switching to a bandpass algorithm for one or two scan lines causes little visible artifacts because the area of impairment is so small — switching to a IH algorithm is even less noticeable. Because of the requirement to “look-ahead,” adaptive comb filters usually utilize a 2H structure. Detection of a chrominance transition can be accomplished by amplitude detection or phase correlation detection. These circuits can range from the simple to very complex, and are beyond the scope of this dissertation.

3.7.4 Delay Elements for Comb Filters

The key component in all comb filters is the IH delay, which provides the necessary delay to implement the desired transfer characteristic. The relatively high cost of comb filter circuits is primarily due to the difficulty of obtaining high performance delay elements, which have a very high ratio of delay time to bandwidth⁻¹. Traditionally, there have been four major classes of devices used to implement these delays: bulk devices, CCD technology, digital memories and analog memories.

3.7.4.1 Bulk Delay Devices

These devices make use of the finite propagation time of an acoustic wave to pass through a block of material. Usually constructed from a sheet of glass about 2 cm long, the electrical signals are converted into ultrasonic waves by way of a piezoelectric device. The waves received at the other end of the sheet are reconverted into an electrical signal by another piezoelectric device. The delay of the device is governed by the bulk wave propagation speed of the glass sheet which varies with thickness and temperature. As a result, these devices require a tuned LC network at the output to trim the delay to the required value. Moreover, these devices have a serious insertion loss which requires a recovery amplifier. In addition, these devices are physically large and not integrable on a silicon circuit. Finally, the frequency response and group delay characteristics of bulk devices is optimized for only a very narrow range of frequencies [40]. Thus, comb filters utilizing these devices are usually chrominance output with bandpass filters to

remove components that would upset the bulk devices. Because of these drawbacks, bulk delay lines are slowly being replaced by more advanced and less costly alternatives.

3.7.4.2 CCD Delay Lines

Charge coupled devices (CCDs) configured as shift registers can be used to delay video signals for comb filters in the sampled analog domain [41]. By cascading an appropriate number of charge storage locations and clocking at a specified frequency, a delay of $1H$ can be achieved. Traditionally, the $4 f_{sc}$ frequency is chosen with 910 storage locations. CCDs enjoy a relatively mature technology, and lower cost than bulk devices [42]. Moreover, because the delay is strictly a function of the clock frequency and the number of sample locations, the delay period is fixed and will not drift independent of the clock frequency. Insertion loss is not an issue as the charge readout amplifier can be configured to give unity gain through the delay element. However, the technology needed to produce high quality CCDs for video delay lines is not compatible with modern analog CMOS processes. In addition, CCDs suffer from noise problems and require relatively high voltages to achieve proper operation. Thus, their future in scaled, low-voltage circuits is uncertain.

3.7.4.3 Digital Delay Lines

The advent of low-cost digital memories and video-rate analog to digital converters has made possible the use of digital memories to implement the delays required for comb filters. There are two main classes of digital delay lines, shift registers and circular buffers. Shift registers work much like CCDs, shifting each bit of data from one storage location to the next at some specified clock frequency until the delay period is reached. Circular buffers, on the other hand, store bits of information in successive locations and read them back out after the delay period. Much like CCD delay lines, the delay period is governed by the clock rate and number of sample locations. Digital delays themselves are transparent to the data — noise and other degradation to the signal can be attributed to the A/D and D/A conversions processes incumbent with digital signal processing.

The main drawbacks of digital line delays are the requirement of an A/D and D/A converter and associated power and silicon area and the power and silicon area of the delay memory itself. While it can be argued that the cost of the data converters should be amortized over the entire DSP task performed (comb filtering is usually only a fraction of the signal processing performed in a digital signal path), the power and area directly attributable to the delay lines can be consid-

erable. An analysis of these costs versus the last implementation, analog line memories, is investigated in detail within this dissertation, and culminates in the prototype circuit.

3.7.4.4 Analog Line Memories

The general principle of digital memories can be implemented in an analog circuit by replacing them with continuous valued storage locations. Within the realm of large scale integrated circuits, such circuits are best realized using switched capacitor techniques where the value is stored as a charge on a capacitor. Both shift register and circular buffer topologies can be duplicated, but the latter is superior for analog switched capacitor circuits owing to its inherently lower power and reduced complexity. Implementing a long delay using analog shift register techniques would involve moving packets of charge many hundreds of times. Not only would this consume large amounts of power to move stored charges once per clock period, but it also has a very high probability of corrupting the signal due to the large number of transactions. The circular buffer topology moves the stored value only twice, once to write and once to read, independent of the delay period, and there is no power consumption associated with a given stored value once it is written into its storage location.

The main challenges to implementing analog line-delays that are comparable to performance of digital memories are noise and power consumption. Because analog circuits have no "noise margin," great care is required to devise a topology that protects the signal from extraneous noise and circuit non-idealities while maintaining necessary circuit speed and meeting power constraints. The large number of storage locations required to implement analog video line memories accentuates these problems because of the parasitics associated with array structures. These parasitics, especially stray capacitances, greatly impact the speed of the circuit and increase the power consumption. Therefore, many of the circuit techniques used in the prototype circuit address problems that are the result of these parasitics. Although these disadvantages may appear to make analog implementations less desirable than digital implementations, the projected power and area savings of the analog implementation merits its use in cost and power critical areas.

3.8 References

- [26] D. G. Fink, *Color Television Standards: Selected Papers and Records of the National Television System Committee*, New York, NY, McGraw-Hill, 1955.
- [27] H. E. Ennes, *Television Broadcasting: Equipment, Systems, Operating Fundamentals*, Indi-

- anapolis, IN, H. W. Sams, 1979.
- [28] F. G. Stremler, *Introduction to Communication Systems*, Reading, PA, Addison-Wesley, 1982.
- [29] A. B. Carlson, *Communication Systems*, New York, NY, McGraw-Hill, 1986.
- [30] A. N. Netravali, B. G. Haskell, *Digital Pictures*, New York, NY, Plenum Press, 1988.
- [31] G. B. Townsend, *PAL Colour Television*, London, Cambridge University Press, 1970.
- [32] U. S. Patent No. 3,946,432.
- [33] W. N. Sproson, *Colour Science in Television and Display Systems*, Belfast, Universities Press, 1983.
- [34] W. F. Schreiber, "Improved Television Systems: NTSC and Beyond," in *SMPTE Journal*, vol 96, no. 8, pp. 734-744, August 1987.
- [35] Y. Faroudja and J. Roizen, "Improving NTSC to Achieve Near-RGB Performance," *SMPTE Journal*, vol. 96, no. 8, pp. 750-761, August 1987.
- [36] J. O. Drewery, "The Filtering of Luminance and Chrominance to Avoid Cross-Colour in a PAL Colour System," *BBC Engineering*, vol. 9, pp. 8-39, Sept. 1976.
- [37] L. B. Jackson, *Digital Filters and Signal Processing*, Boston, MA, Kluwer Academic Publishers, 1989.
- [38] Y. Faroudja, "Adaptive Comb Filtering," U. S. Patent No. 4,179,705, Mar. 13, 1978.
- [39] Y. Faroudja and J. Campbell, "Processing Methods Using Adaptive Threshold For Removal of Chroma/Luminance Cross-Talk in Quadrature-Modulated Subcarrier Color Television Systems," U. S, Patent No. 4,731,660, Mar. 15, 1988.
- [40] Asahi Glass Co., "Technical Data Sheet for Delay Lines for Video Applications."
- [41] S. M. Sze, *Physics of Semiconductor Devices*, 2 ed., New York, NY, John Wiley & Sons, 1981.
- [42] Y. Maki, T. Kondo, A. Izumi, et. al., "A CMOS-CCD Comb Filter with Dropout Compensation for a VCR," in *ISSCC Digest of Technical Papers*, pp. 46-47, 1988.

CHAPTER 4

Prototype Comb Filter

4.1 Introduction

The prototype comb filter was designed and fabricated as a proof-of-concept circuit designed to perform a 2H comb filtering function while obtaining at least 8-bit equivalent resolution and linearity. In addition, the focus was to minimize power and silicon area while maintaining full functionality. Because the ultimate application of this circuit would be as a part of a larger mixed-signal circuit, only components available in a standard double poly, double metal CMOS process were used. The resultant work, which exhibits full functionality and has performance exceeding the 8-bit level in both dynamic range and linearity, requires no adjustments, a single reference current and a single external clock. This chapter will discuss the design of this prototype chip at the system and circuit levels.

4.2 System-Level Design Considerations

System-level considerations center on the actual architecture of the comb filter used to perform the Y/C separation. As discussed in earlier chapters, the 3 line or 2H architecture is a good compromise between complexity and performance. Other variables include the system clock speed, the type of outputs provided, the choice of single-ended or differential architecture, and choice of clocking scheme. While NTSC processing can be performed at $3f_{sc}$ using the PALE concept as discussed in Chapter 2, the circuit complexities necessary to implement a $3f_{sc}$ comb filter do not justify the 25 percent reduction in circuit speed requirements. For these reasons, and

the fact that downstream processing of the chrominance signal can be greatly simplified, the sampling rate of this circuit is set at $4f_{sc}$ or 14.318 MHz for the NTSC system.

A fully differential architecture is utilized to minimize the possibility of power supply and clock signal feedthrough from corrupting the video signal. However, for this particular circuit, a differential topology consumes roughly 70% more silicon area than would a equivalent single-ended design. Similarly, because all the OTAs use double the bias current, and the metal line parasitics are larger due to the greater area of the circuit, the power is increased by roughly 75% over a single-ended alternative. Because power and area are critical aspects of this prototype, a single-ended design was considered. However, because of the large ratio of substrate to signal line capacitance, excessive coupling of supply noise to the signal is likely to occur. This is especially critical at higher frequencies where the PSRR of active circuits begins to fall off due to the reduction in loop gain. This finding corresponds well with past experience in analog switched capacitor circuits which indicate that the advantages of differential topology outweigh the costs [43]. However, this conjecture has not been rigorously proven for this particular application, and in areas where power and area are of absolute concern, an investigation into a single-ended alternative may be warranted.

Finally, to simplify, and hence to first order reduce power in the clock generator, a single phase master clock scheme is used. This has the added advantage that all clock transitions on the chip occur at roughly the same instant, greatly reducing the chance of clock noise being injected into a signal line during a critical interval. Only one master clock is needed — all clocks are derived internally from this signal.

4.3 System Overview

The architecture of the prototype circuit is that of a chrominance output 2H comb filter whose block diagram is repeated below as Figure 4.1. The actual implementation will be through a sampled data analog system, utilizing switched capacitor like techniques to realize the line delays. Figure 4.2 shows a block diagram of the prototype circuit which implements the function outlined in Figure 4.1. The comb filter is constructed from two sample/hold stages, two line delays and a scaler/summer output stage.

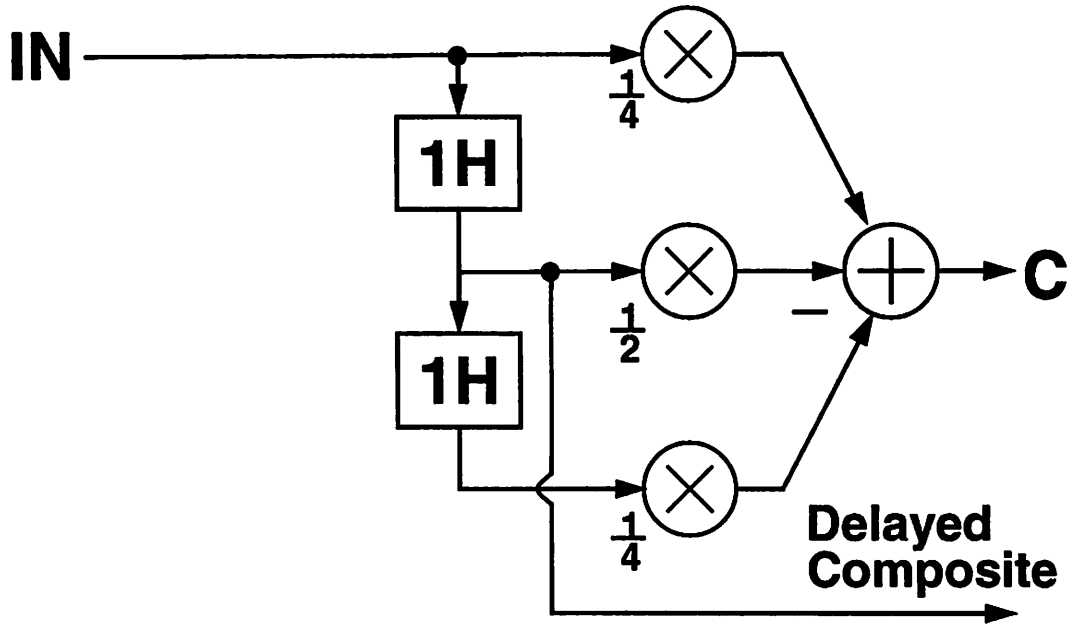


Figure 4.1 Block Diagram of 2H Comb Filter

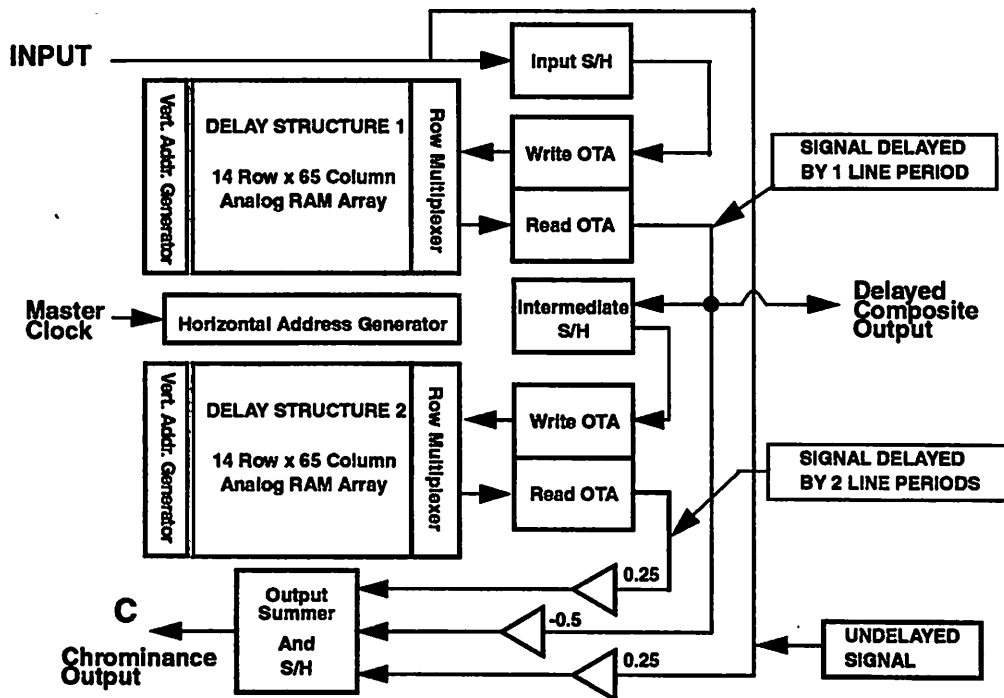


Figure 4.2 Block Diagram of Prototype Chip

4.3.1 Operational Overview

The data path is kept serial as much as possible to minimize the potential for fixed pattern noise to be introduced into the signal. The use of parallelism to achieve the necessary throughput results in a repetitive modulation of the signal due to offset and gain mismatches between the parallel channels. This modulation is in effect a fixed pattern noise, and in video, is especially noticeable. While the use of sampled data analog for video signal processing is not new [44],[45], the proposed architecture [46] is unique in that critical areas of the signal path are serial. Therefore, any non-idealities affect each pixel of information in the same way, dramatically reducing the chance of fixed pattern noise effects.

All operations of the chip are done in one of two phases, ϕ_1 and ϕ_2 , of a single clock. During ϕ_1 , the input is sampled by the input S/H and also by the “zero delay” input of the output stage. At the same time, the signals stored in both delay lines 910 clock cycles ago are read out; the output of the first delay line is sampled by the intermediate S/H stage and the “single delay” input of the output stage, while the output of the second delay line is sampled by the “double delay” input of the output stage. During ϕ_2 , the values stored in the input S/H and intermediate S/H are written into the first and second delay lines respectively. The signals applied to the output stage are scaled appropriately and summed producing the desired output. A signal from the intermediate S/H is also provided as a group delay equalized output signal for external use.

Design and operation of each component is detailed in the following sections. Contained within the delay structure itself are the storage cells arranged in an array and the row multiplexers. Amplifiers (OTAs) are used to interface to the storage cells; access to individual storage locations is accomplished through horizontal and vertical address generators. Not shown in Figure 4.2 are the bias generator, clock generator, and the output buffers.

4.4 Description of Silicon Technology

Although the analysis and architectural design of this prototype should be first order independent of the technology, it is useful to have in mind the particular silicon process used to fabricate the prototype. The circuit was designed with the Orbit Foresight™ 1.2 μm double poly, double metal CMOS process. Although the process utilizes 1.2 μm gates, many of the design rules mimic that of a 1.5 μm technology with its larger parasitics and area consumption. To facilitate the most robust design possible, actual test devices were tested on a HP4145B to extract an accurate level 3 SPICE model. In addition, g_m - g_{ds} curves were extracted to provide operating

point information for critical devices. Some key parameters of the technology are summarized below — in the subsequent description of the design of key elements of the prototype, the parameters below are assumed.

TABLE 4-1 Parameters of 1.2 μm Double Poly Double Metal Orbit Foresight™ Technology

Parameter	Value	Unit
Minimum Gate Length	1.2	μm
NMOS V_{t0}	0.8717	V
PMOS V_{t0}	-0.9637	V
NMOS μ_0 (SPICE LEVEL 3)	524.5	cm^2/Vs
PMOS μ_0 (SPICE LEVEL 3)	160.5	cm^2/Vs
t_{ox}	22.5	nm
NMOS γ	0.3121	
PMOS γ	0.1670	
Minimum MOS Device Width	2.4	μm
Minimum Diffusion Spacing	1.2	μm
Minimum Poly Spacing	1.8	μm
Effective Minimum PMOS to NMOS Distance	6.0	μm
Minimum Poly to Active Spacing	0.4	μm
Minimum Gate Overhang	1.0	μm
Metal 1 Minimum Width	2.2	μm
Metal 2 Minimum Width	2.0	μm
Metal 1 Minimum Spacing	1.2	μm
Metal 2 Minimum Spacing	1.6	μm
Poly1 - Poly 2 Capacitance	0.63/0.435*	fF/ μm^2

* (Expected/Actual) — actual interpoly oxide thickness much greater than expected.

4.5 Architecture of the Analog-RAM

The most critical section of the prototype circuit is the structure used to delay the video signal by the period of one scan line. As discussed in previous chapters, an implementation based on an analog equivalent of a digital circular buffer, whose high level diagram is shown in Figure 4.3, will be used to implement the delay line. The delay line is constructed from an array of storage locations, with two commutators, controlled by non-overlapping clocks, ϕ_r and ϕ_w , to permit storage and retrieval of information. The clock phases are arranged such that for a given storage location, the stored value is read first, then a new value is stored. It is important to note

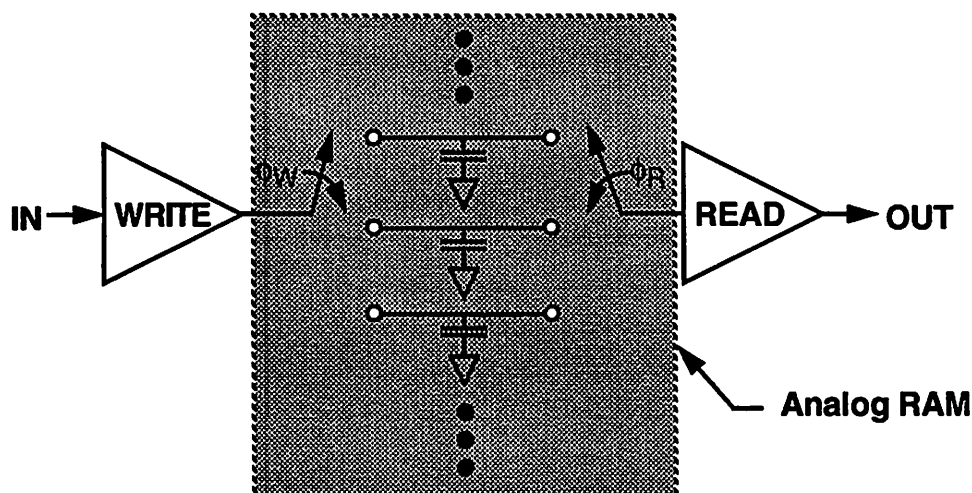


Figure 4.3 Generalized Concept of Analog Circular Buffer (Analog RAM)

that there is no restriction that storage locations be accessed in order, hence the name “Analog RAM.” However, for this particular application, a comb filter, the requirement is for a $63.5 \mu\text{s}$ delay line. Thus, the storage locations are accessed sequentially, akin to a circular buffer, to form a delay line. Note that the delay period is the product of the clock period and the number of cells. For reasons given earlier, NTSC processing is best carried out at $4 f_{sc}$ — this implies that 910 storage locations will be required to provide the proper delay.

4.5.1 Storing Information onto the Analog RAM

The actual process of storing a value involves the transfer of the analog quantity into the selected storage location, which is simply a small capacitor. For reasons that will become evident later, the quantity of interest is the charge on the capacitor plates rather than the voltage across the device. Referring to Figure 4.6, an operational transconductance amplifier (OTA) is used as a transfer device to move charge from a sampling capacitor to the selected storage capacitor. The action of the commutator is realized by using a select switch in conjunction with each storage capacitor. This capacitor/switch combination will be referred to as a storage cell as it represents the unit structure needed to hold one sample of information.

During the first phase, ϕ_1 , the input is sampled on C_S , a sampling capacitor whose value is nominally equal to those of the storage cells, using a parasitic insensitive, “bottom-plate” sampling method to minimize the effect of clock feedthrough. The charge stored on C_S is then transferred using the OTA during ϕ_2 onto a storage cell (C_A , C_B , C_C in the above figure) within the

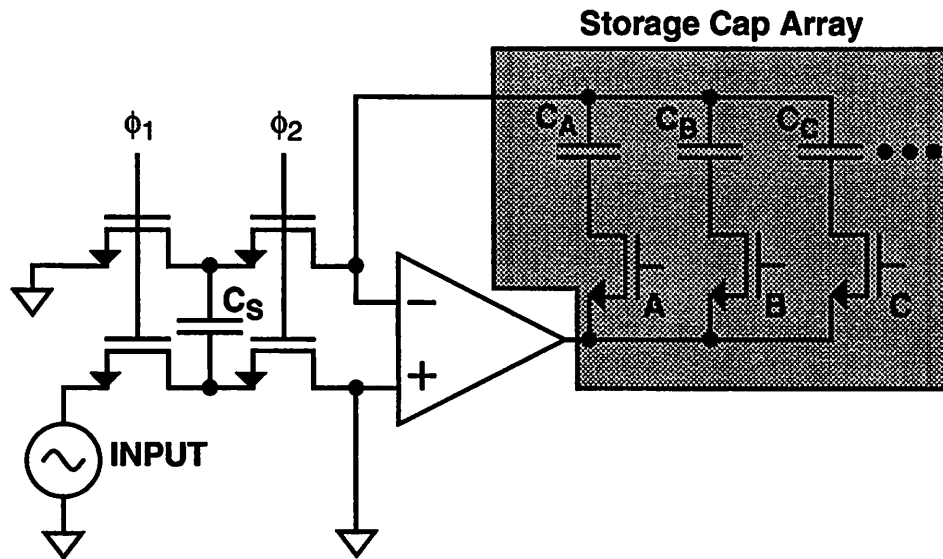


Figure 4.4 Simplified Schematic of Analog-RAM During Write

array through a storage switch (A, B, C). This process can be repeated to sample new values and store them in selected storage cells as needed.

4.5.2 Reading Stored Values from the Analog-RAM

The inverse to the above operation, reading a stored value, is accomplished using a similar structure, in which an OTA is used to take charge from a selected storage cell and integrate it onto a capacitor to produce an output voltage (Figure 4.5). During the first clock phase, a common grounding switch and a switch which selects a particular storage cell within the array are closed. The OTA then takes the charge and integrates it on C_I , which is nominally the same value as the C_S in Figure 4.4 and the storage cell capacitors. This results in a voltage at the output of the OTA which is proportional to the charge read out of the selected capacitor. The other clock phase is used to reset C_I so that the circuit is ready to perform another read operation.

4.5.3 Overall Read/Write Architecture of Analog-RAM

The storage arrays shown in Figure 4.4 and Figure 4.5 are the same, and hence, the circuit schematics can be combined to demonstrate the overall operation of the Analog-RAM as shown in Figure 4.6. Two switches have been added to isolate the read and write circuits to prevent data corruption. All switches are driven from one of the two non-overlapping phases, which are arranged such that the read and write operations are time interleaved. In this particular applica-

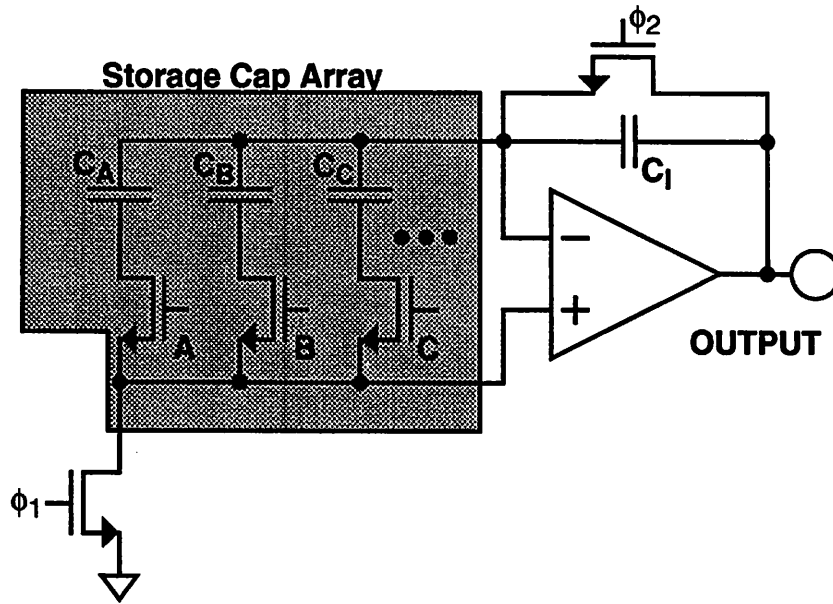


Figure 4.5 Simplified Schematic of Analog-RAM During Read

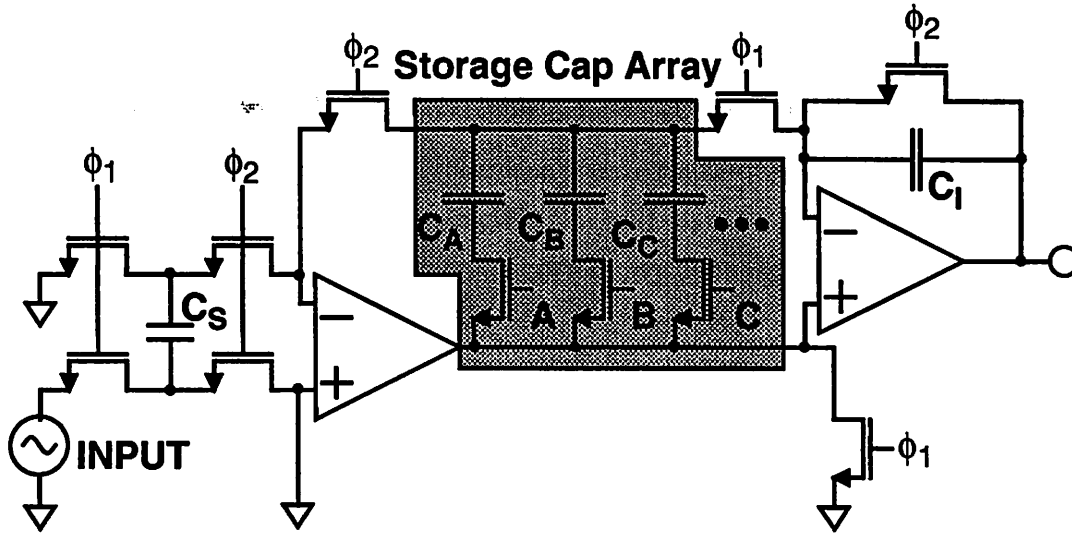


Figure 4.6 Simplified Analog RAM Schematic

tion, each clock is asserted for half of the 70 ns clock cycle, with the read clock, ϕ_1 , asserted before the write clock, ϕ_2 , for a given selection of a storage cell as shown in Figure 4.7.

A key concept which has been alluded to is the fact that the stored value is represented as a charge, rather than a system where the values are stored as capacitor voltages. This has two major advantages: (1) first order insensitivity to storage capacitor nonlinearity, and (2) first order

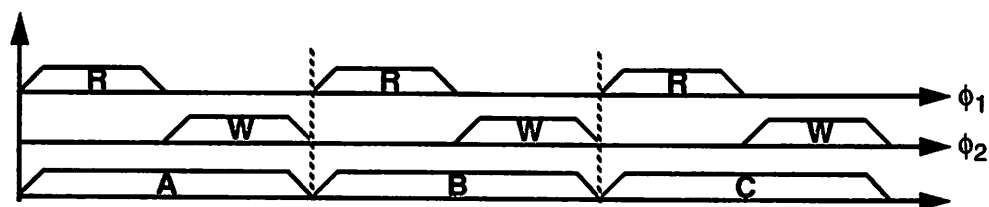


Figure 4.7 Timing Diagram for Circuit of Figure 4.6

insensitivity to storage capacitor mismatch. Given that there are nearly 1000 storage cells per delay line, these two factors allow a great deal of flexibility to minimize the cell area. Although not implemented in this prototype, the allowance of nonlinearity suggests that MOS capacitors could be used to substantially reduce the storage capacitor area. The lack of critical matching requirements mitigates the need to maintain minimum capacitor sizes or use specialized geometries due to lithographic requirements.

The transfer function of the Analog RAM is governed by the ratio of C_I to C_S . The sample and hold circuit that feeds the write OTA should be viewed as a voltage to charge converter — C_S must be linear, but as there are very few of these capacitors within the circuit, this requirement is not burdensome. Similarly, the read OTA and integrating capacitor C_I , which also must be linear, should be viewed as a charge to voltage converter. Thus, as long as the capacitors within the storage array do not leak charge, the output voltage will be a linear function of the input voltage.

Second-order effects that would contribute to a deviation from a true unity gain transfer function include incomplete settling of the amplifier, clock feedthrough, random noise, incomplete charge transfer and amplifier offsets. The causes, effects and circuit solutions to these problems will be discussed in detail in subsequent sections in this chapter.

4.6 Topology of the Storage Array

An examination of the overall circuit showed that unlike most switched capacitor analog circuits, this circuit would be parasitic dominated. That is, the dynamics, frequency response and transfer characteristics of the circuits would be dictated more by the parasitics present than the driven circuit element. This finding is not surprising given that the majority of the circuit area is consumed by two large arrays of storage elements. As with DRAMs the parasitic capacitances

associated with long metal lines and multiple source-drain diffusions associated with arrayed transistors can total into the tens of picofarads. Therefore, unlike conventional design processes which begin with the active circuitry and later compensate for the parasitics, the design of this circuit must begin with the storage cell to determine the approximate parasitics that will be seen by the remainder of the circuit.

4.6.1 Contents of the Cell

The first item to be determined is the contents of the cell. At a minimum, each storage cell must contain a storage capacitor and a method of selecting the cell. The most straightforward, two switches and a capacitor immediately showed the difficulties of parasitic dominated circuits. A simplified, single-ended equivalent circuit is shown in Figure 4.8. Although this circuit will

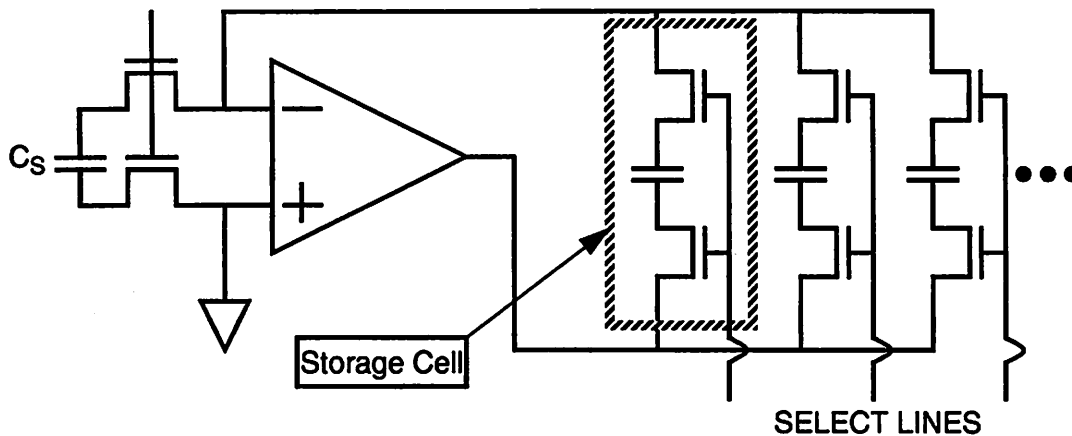


Figure 4.8 Prototype Storage Cell Configuration

operate with a single cell, adding a large number of cells as needed to realize a large delay line results in an unworkable solution. First, the switches on either side of the storage capacitor contribute a significant amount of parasitic capacitance on both the output of the OTA and the summing node. Second, the select switches which allow access to the storage cell possess a considerable amount of channel resistance. This results in a feedback zero due to the series resistance and the large parasitic capacitances which result in loop instability. Enlarging the switches to reduce the resistance increases the parasitic capacitance due to the source-drain areas. As a result, the frequency of the zero is a weak function of the switch size, making it difficult to remove the instability. In addition, the total load on the OTA increases, thereby slowing the

entire loop down. As a result, it was determined that this circuit with two series switches is impractical, at least in the given technology.

4.6.2 Two-Dimensional Array

One other fact discovered in early experimentation is that a linear array is impractical for a 910 tap delay line. Even with minimum size switches, an array of 910 storage cells would result in 910 parallel source-drain parasitics totalling into the tens of picofarads. Although it may be possible to make the circuit work, driving such large parasitic capacitances would result in a large amount of CV^2f power to charge and discharge the parasitics. Such a penalty is inappropriate for a low-power design. This finding led to the concept of a 2-D array to reduce the parasitics.

The total parasitic capacitance seen by the OTA at any given time can be reduced if all 910 elements are not in the circuit at the same time. Thus, if the 910 storage cells could be broken into sub-arrays, and only the active sub-array is connected to the OTA at any one time, then the parasitics seen by the OTA can be substantially reduced. This is the concept of the 2-D array. By breaking the 910 cells into a rectangular array consisting of M rows of N cells each, then the parasitics could be reduced by approximately a factor of M . Unfortunately, 910 is not an easily factored number, its factors being 2, 5, 7, and 13. Reasonable choices for M are therefore 10, 13, 14, and 26. The corresponding choices of N are 91, 70, 65, and 35. Determination of the optimum combination of M and N depends not only on the actual values of the parasitics, but also the costs associated with switching the rows in and out.

Row (de)-multiplexers are used to route the signal from the OTAs to the proper row. The switches within these "muxes" have channel resistance and their source-drain areas contribute to parasitic capacitance. Therefore, increasing M carries with it the penalty of adding mux parasitics while reducing the parasitics due to the cells themselves. Because it is impossible to optimize the loop dynamics without knowing the relative values of the parasitics, an overall circuit architecture for the 2-D array and the specific cell schematic and layout are necessary prior to computing the final values of M and N .

4.6.3 Single Switch Storage Cell

An obvious extension of the analysis of Section 4.6.1 on page 78 implies that reducing the number of switches in the storage cell would result in an immediate win with respect to loop sta-

bility, area and power. The principle advantage of the two switch design is that each storage capacitor is uniquely isolated. That is, once the switches are opened, the charge on the capacitor is fixed and cannot be altered inadvertently by any another part of the circuit. A closer view of Figure 4.8 shows however, that both switches are not necessary to uniquely address each cell. Note that the common line connected to the summing node is always at ground (or the common mode voltage in a differential circuit). During the read operation as depicted in Figure 4.5 and Figure 4.6, the common line remains at ground. Thus, connecting the plates of each of the storage capacitors to that line as shown in Figure 4.6 should, in principle, not affect operation of the circuit.

Computer simulations of the circuit of Figure 4.6 prove that the lack of switches on both sides of the capacitor does not affect the proper operation of the circuit. However, the circuit does become very sensitive to capacitive loops — that is, a loop of two or more storage cells where the charge in the capacitors can be transferred between them. This occurs when there is either a conductive or a parasitic capacitive shunt across the select switch. Finally, although not unique to the single-switch topology, care must be taken to insure that any charge injection into the storage cells is charge independent. The single-switch of this topology shown in Figure 4.6 has a V_{GS} that is signal dependent. As a result, unless other precautions are taken in the circuit, a charge that is a function of the signal is impressed upon the circuit at that instant will be injected into the storage capacitor leading to harmonic distortion. This limitation is overcome by moving the single switch to the opposite side of the storage capacitor (Figure 4.9), thereby accomplishing “bottom-plate” sampling. As a result, provided the OTA is settled, the switch has a V_{GS} that is independent of the signal, thus preventing signal dependent charge injection. Isolation of the cells is accomplished even after moving the switch as one plate of the storage capacitor is disconnected from the remainder of the circuit. Thus, the charge representing the stored value is trapped on the plate and remains to be read out at a later time independent of the other plate of the capacitor, assuming no leakage or breakdown.

4.6.4 Cell Addressing

Given the 2-D array, a method must be provided to access each individual cell. A row-column accessing method, similar to those used for DRAMs, has been shown to be efficient and robust method of cell access. However, in a DRAM, the AND function of the row and column signals is performed implicitly by the pass transistor. This method is not usable for the analog

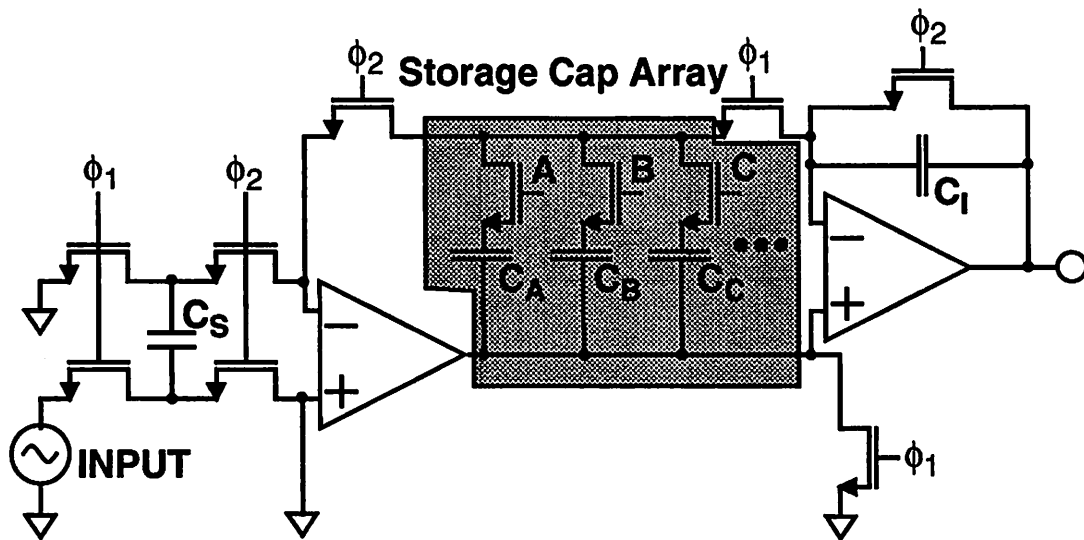


Figure 4.9 Simplified Analog-RAM Circuit with "bottom-plate" sampling

RAM as described here because activating a column line would enable all the cells in that column. While appropriate for the selected row, for all the unselected rows, a capacitive loop between the row parasitics and the selected column would exist as shown in Figure 4.10 which would corrupt the signal.

DRAMs avoid this problem by reading the entire contents of the row when the row select is asserted. However, this requires N (number of columns) read channels, which implies N sense amplifiers, or in the case of an analog RAM, N read OTAs. Clearly, this would result in a large static power dissipation, and is not appropriate for a low power design. Thus, a method of using row-column addressing and activating only the actual selected cell is needed.

Activation of an unique cell is accomplished by embedding a simple decoder within each cell which logically ANDs the row and column signals. The output of this AND is used to drive the select switch. As a result, only the actual addressed cell has its storage capacitor connected to the circuit and its parasitics. The main drawback of this approach, the overhead of embedding a logic gate in each cell, is minimized by using a two transistor transmission gate logic AND circuit as shown in Figure 4.11. Although this 2T AND circuit does not produce full logic swing given a certain sequence of inputs, for use within this analog RAM circuit, proper outputs to drive the select switch are obtained.

Prototype Comb Filter

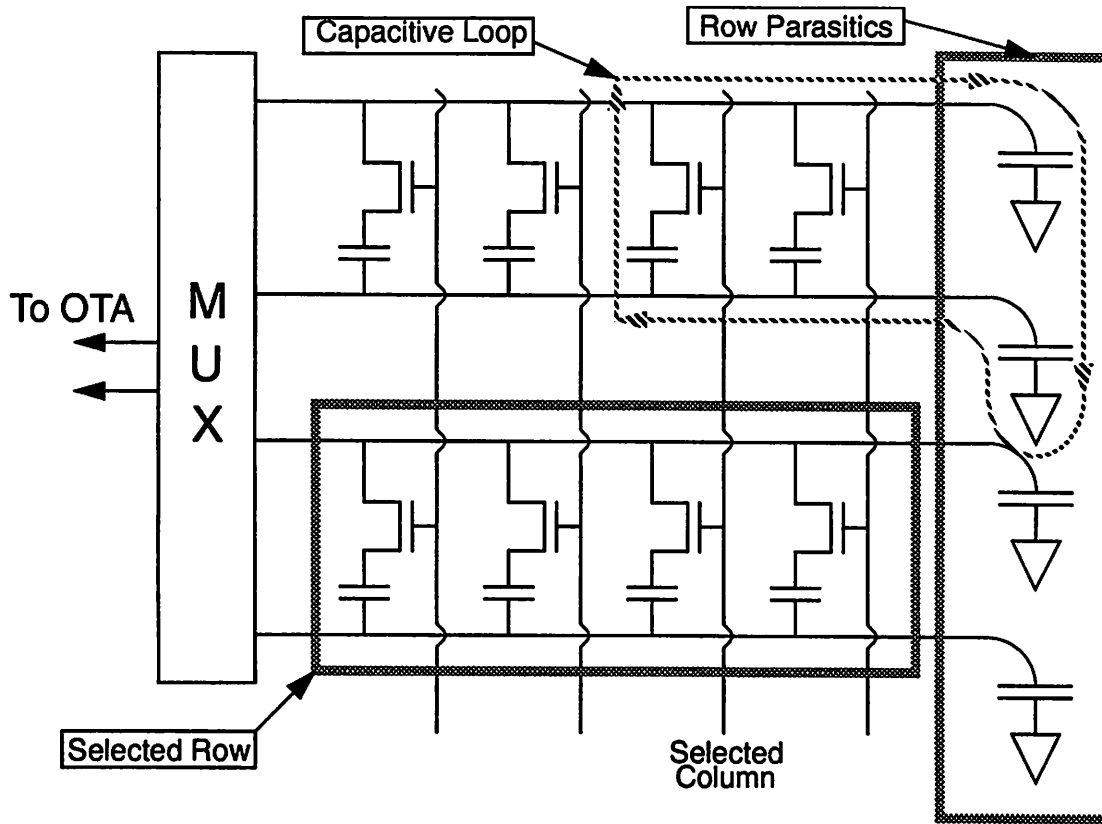


Figure 4.10 Circuit Showing Capacitive Loop Caused by Direct Row-Column Addressing

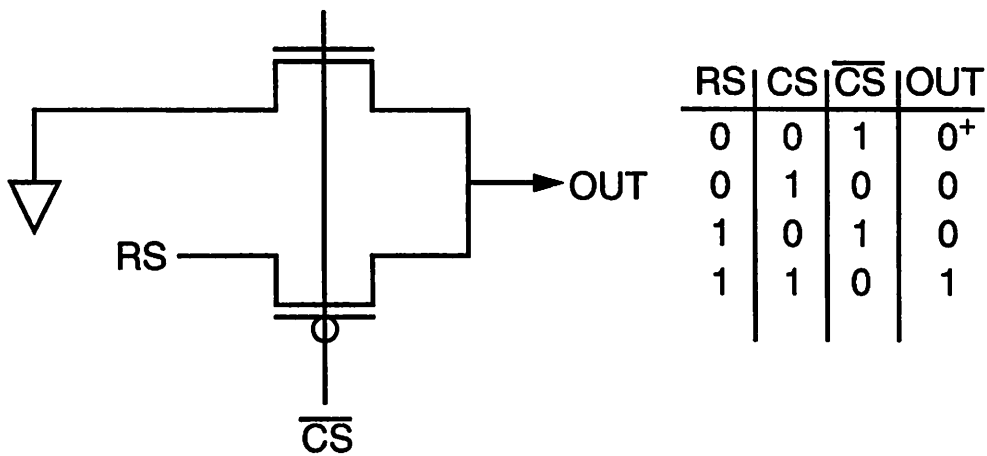


Figure 4.11 Two Transistor TG AND Gate for Cell Select

Thus, the circuit elements and topology for the storage cell has been determined. A fully differential schematic of the storage cell is shown in Figure 4.12. The two remaining variables are

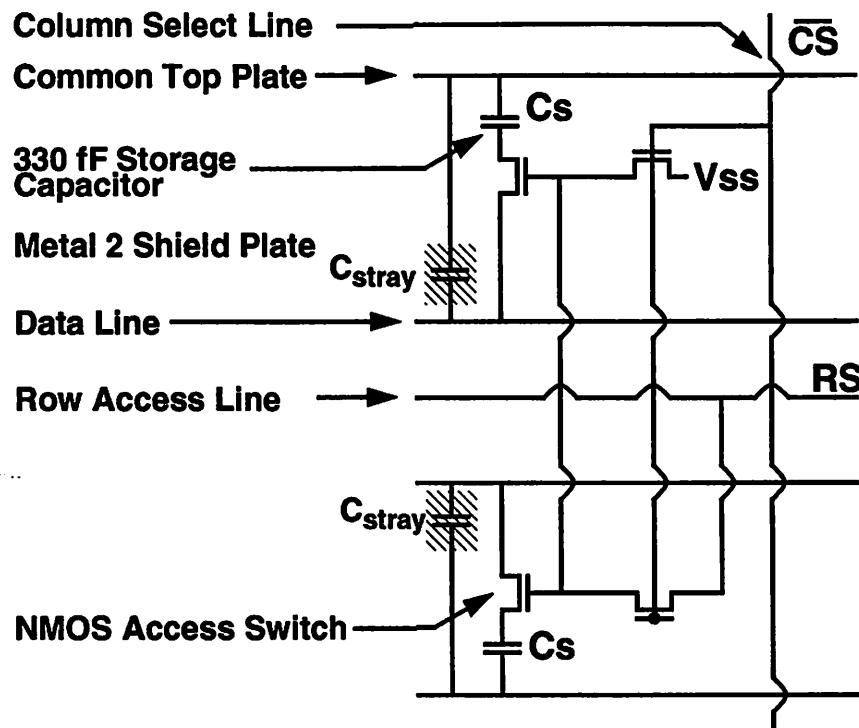


Figure 4.12 Complete Schematic of Storage Cell

the storage capacitor size and the size of the select transistor. Once these have been fixed, it is possible to perform a layout and determine the magnitudes of the parasitics so that the remainder of the circuit can be designed. Of particular note are the stray capacitances marked C_{stray} in Figure 4.12. These are the capacitances that will lead to a capacitive loop within the array which will corrupt the charge stored within the cells. As a result, it is crucial that the value of these stray capacitances be kept to a minimum with judicious use of shielding.

4.6.5 Sizing of Storage Capacitor

The actual size of the storage capacitor is driven by a desire to make it as small as possible while remaining within the limits imposed by thermal noise, matching requirements, immunity to clock feedthrough and loop dynamics. Minimum size is desirable because of the large number of storage cells required — a small increase in capacitor value, hence its area will lead to a substantial increase in the overall area of the delay line. Thus, careful sizing of the capacitor is an important design step.

4.6.5.1 Thermal Noise

The combined channel resistances of the switches in the feedback circuit of the write OTA contribute a broad-band thermal noise which is sampled onto the storage capacitor along with the input signal. As the switch begins to open, the channel resistance increases thereby increasing the noise power. However, the increased resistance combined with the storage capacitance form a low pass filter network which effectively reduces the noise bandwidth. As a result, the standard deviation of the expected noise power sampled onto the storage capacitor is independent of the switch size, and is given by $\sqrt{\frac{kT}{C}}$, where k is the Boltzmann's constant and T is the absolute temperature [43]. As a point of reference, a 1pF capacitor would have 64 μ V of noise impressed upon it. For this prototype, the desired dynamic range is on the order of 50 dB, and a conservative value for the signal swing is $1.5V_{p-p}$, which implies that the noise must be less than $1.67 mV_{rms}$. Thus, the minimum capacitance dictated by thermal noise is 1.47 fF. Although the actual value of the noise due to " kT/C " effects will be several times this value due to multiple sampling operations, it is obvious in this technology, thermal noise is not the limiting constraint.

4.6.5.2 Matching Requirements

Although the circuit is first-order insensitive to mismatch of capacitors within the array, size constraints due to matching are still an issue. Because the sampling capacitors and the integrating capacitors in the R/W circuit are nominally the same value as the storage capacitors, mismatches in the capacitor sizes can result in a non-unity gain through the delay line. Although it would not result in fixed pattern noise, it will be a random value from chip to chip and will affect the final transfer function resulting in incomplete separation of Y and C components. As determined in a previous section, mismatches as small as 1 percent can cause serious degradation of the transfer function notch depth. Studies have shown that 1 percent capacitor matching requires a minimum capacitor area of $36 \mu m^2$ in a $1.2 \mu m$ technology [47],[48],[49]. For the technology used in the prototype, this translates to a minimum capacitor value of 20 to 30 fF depending on the degree of fringing fields considered. However, achieving this degree of matching assumes that the fringing fields are subjected to the same conditions on both capacitors. Hence, either dummy devices or shield plates are required. As a result, the actual area of the storage capacitor will be considerably larger than that implied by the capacitance. Thus, the areal advantage of reducing the capacitance is not as strong as might have first been imagined.

4.6.5.3 Clock Feedthrough Rejection

The “bottom-plate” sampling method used in the storage array mentioned in Section 4.6.3 on page 79 does not eliminate clock feedthrough, it merely attempts to assure that the injection of charge is independent of charge. In an ideal differential circuit, such a injection of charge would result in a common mode shift which would be eliminated provided the allowable common mode range is not exceeded. In practical differential circuits, there is a finite CMRR, and a non-ideal matching between the two symmetric devices in the circuit. As a result, there will always be a small differential signal that results from clock feedthrough. The critical parameters that determine what the magnitude of this error will be are the size of the switch in question, the switch gate swing, and the capacitive load seen at both terminals of the switch. Prior work regarding clock feedthrough has shown that the channel charge will tend to flow into the side of the switch with lower impedance. Therefore, to obtain a repeatable splitting of the charge, the impedances that the switch sees must be kept consistent. An examination of the circuit in question

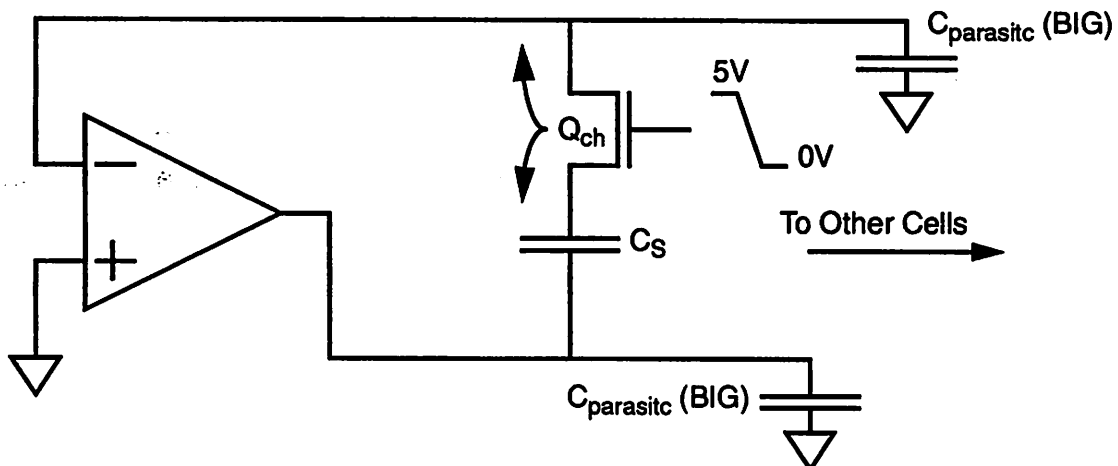


Figure 4.13 Circuit to analyze clock feedthrough

(Figure 4.13) shows that there is a large parasitic capacitance on one side of the switch, with the storage capacitor, which is much smaller in relative size, on the other side. Therefore, a majority of the charge is expected to flow into the parasitic. However, a portion will flow into the storage capacitor — care must be taken that the portion that flows into the capacitor is consistent from cell to cell and that the positive and negative halves of the circuit inject charge equally [50].

The first criteria, constant division of the charge is helped by the large parasitics. Because the parasitics originate from the accumulated row capacitance, and remains constant, the impedance on that side of the select switch will remain constant. Moreover, it is larger than the storage cell capacitance by a significant factor thereby sinking a large portion of the charge. What fraction does flow into the cell capacitor is dependent on matching of the value of the capacitor from cell to cell. More important though, is insuring that the two sides of the differential circuit inject an equal amount of charge to the cell capacitor so that a differential error voltage does not occur. Strictly speaking, for this circuit, the variance in the differential error is more important, as that will contribute to noise. A constant differential error will cause a DC offset to arise, but is unimportant in this circuit so long as the magnitude is sufficiently small to prevent saturation of the subsequent circuit stages.

The switch size will be stated here *a priori* as $(20/1.2)$, as the process to obtain this number is described later. As a result, there will be a channel capacitance of approximately 40 fF leading to a channel charge of 200 fC assuming a 5 volt switch gate swing. A conservative 1 percent mismatch of this charge was assumed giving rise to a 2 fC error charge, half of which or 1 fC would flow into the storage capacitor. Assuming a 1 volt signal on the capacitor, and "8 bits" of desired SNR, this sets the minimum storage capacitance at 250 fF. In reality, the error should be significantly less than this due to the non-equal splitting of the channel charge and better than 1 percent matching of such a large transistor. However, for this design, 250 fF was determined to be the minimum size of the capacitor for these purposes.

4.6.5.4 Loop Stability and Dynamics

The criteria imposed on storage cell size by loop stability and dynamics is the most difficult to compute of those considered because it is inherently an iterative process. The limitation arises from the fact that there is in reality a complex network in the feedback loop of the OTA due to the non-zero resistance of the switches. As a result, an improper mix of parasitics, switch sizes and storage cell capacitor (which acts as a feedback element), will lead to poor settling or even oscillations. At the same time, a solution that preserves stability may result in an exceedingly long settling time which would prevent proper operation of the circuit.

The major problem with an analytical solution to this problem is that the parasitics are a complex function of storage cell capacitor size, cell layout, switch size, and array topology, as each of those affect the parasitics, which are the dominant elements in the feedback loop. The

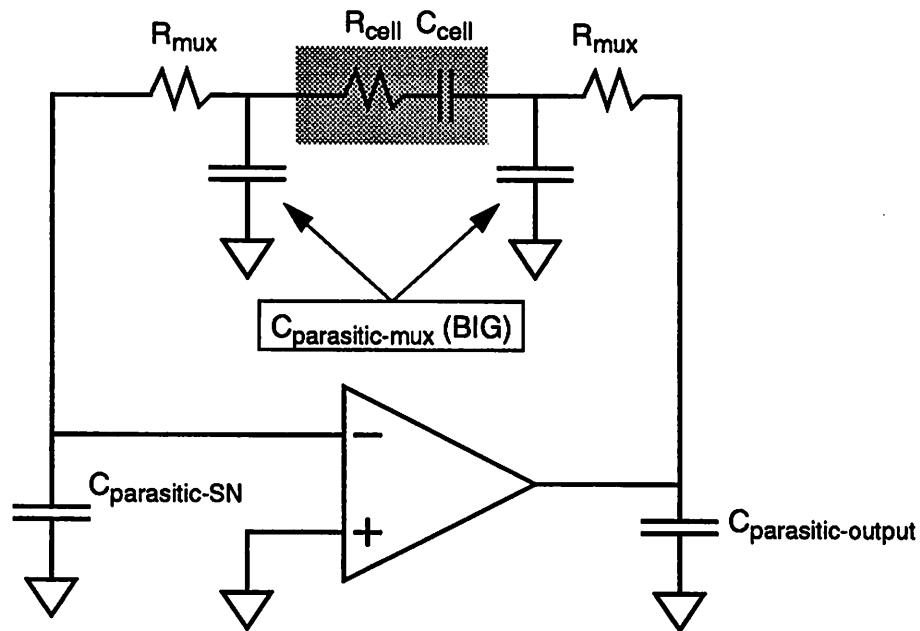


Figure 4.14 Simplified circuit showing parasitic capacitances and resistances

only practical method found to solve this problem was through iterative computer simulation based on extracted parasitics from actual layout.

Early simulations revealed that there is an upper limit on the total series resistance in the feedback loop for reasons of stability. This is expected as the resistances combined with the parasitic capacitances form a lag network which contributes to excess phase. Unfortunately, this implies large devices as the $(V_{GS}-V_T)$ of these devices is limited by the difference of the supply voltage and the common mode voltage. Some improvement can be obtained by using complementary devices at the cost of significantly increased source-drain parasitics. As there is a strong desire to minimize circuit area, a decision was made to place most of the allowable resistance in the cell, while making the multiplexer switches large, thereby minimizing cell area. Complementary switches in the cell were considered, but rejected in favor of a slightly larger (20/1.2) NMOS device due to the added parasitics, added area, and complexity in driving both PMOS and NMOS switches.

A lower limit on the cell capacitance was determined using simulation based on settling time. Hand analysis of the circuit of Figure 4.14 is difficult due to the multiple poles in the feedback loop. However, if the simplification is made that R_{mux} is sufficiently small, then the circuit reduces to a second order loop which is easily tractable. Consider the small signal simplified cir-

circuit of Figure 4.15. If a test current is injected at the output, and the closed loop transfer function

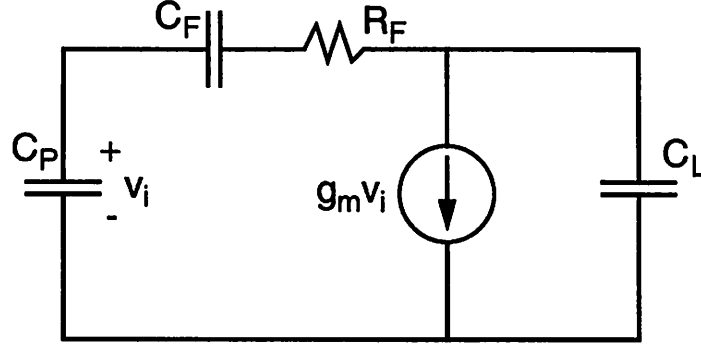


Figure 4.15 Simplified circuit for computation of loop dynamics

computed, the result is that of a typical second order system with a $\frac{N(s)}{D(s)}$ form. In this case, the loop dynamics are given by $D(s)$ which is of the form:

$$1 + s \left[\frac{1}{g_m} \left[C_P + C_L + \left(\frac{C_L C_P}{C_F} \right) \right] \right] + s^2 \left(\frac{C_L C_P R_F}{g_m} \right) \quad (4.1)$$

which can be expressed in the standard form of $1 + \frac{2\xi}{\omega_n} s + \frac{s^2}{\omega_n^2}$ where ξ is the damping factor. Expressing (4.1) in this form results in:

$$\omega_n = \sqrt{\frac{g_m}{C_L C_P R_F}} \quad (4.2)$$

$$\xi = \frac{C_P + C_L + \frac{C_L C_P}{C_F}}{2\sqrt{C_L C_P R_F g_m}} \quad (4.3)$$

A good value of ξ as a compromise between stability and settling is 0.707, where the system has poles 45 degrees off the real axis [51]. An examination of (4.3) shows that under the conditions of large parasitics ($C_L, C_P \gg C_F$) the value of ξ is affected most dramatically by the ratio of C_F to the parasitics. Too small a value of C_F will result in an overdamped and hence slow response. The value of ξ can also be affected by changing the forward G_m of the OTA, but note that there is only a square root dependence on g_m , making this approach not as attractive as altering the capacitive ratio.

The settling behavior of a two-pole system with ξ equal to 0.707 is closely approximated by a single-pole system whose pole frequency is equal to ω_n in the two-pole case. Thus, an addi-

tional constraint must be included so that an equivalent “8-bit” settling behavior is obtained. Approximately 5.5 time constants are required to obtain adequate settling — assuming a settling period of 25 ns, this means that ω_n must be at least 240 Mrad/s. Combining this with (4.2) and (4.3), with the estimated value of C_L and C_P of 4 pF, the required G_m is only 1 mS. However, if the previously computed value of 250 fF for C_F (based on matching requirements) is used with this value of G_m , the resultant value of ξ is nearly 8.0, which leads to a very slow settling tail. In fact, the desired value of 0.707 is not achievable with this value of G_m . Thus, either C_F or G_m or both must be increased to bring the value of ξ to the desired value.

An examination of (4.3) shows that given a set of fixed capacitances, ξ can be decreased by increasing ω_n , and hence G_m :

$$\xi = \frac{C_P + C_L + \frac{C_L C_P}{C_F}}{2C_L C_P R_F \omega_n} \quad (4.4)$$

It is obvious from (4.4) that increasing the value of C_F will also aid greatly in decreasing the value of ξ . However, it is in fact the ratio of C_F to the parasitics which affects the value of ξ . In this particular circuit, making C_F larger implies a larger storage cell area and hence larger parasitics. Because the majority of C_P and C_L are due to the accumulated parasitics associated with the array, simply increasing the value of C_F does not necessarily improve the ratio of the parasitics to C_F . However, increasing the value of ω_n through increased G_m does not, to first order, affect the value of the parasitics. Thus, a combination of these two approaches will be needed to obtain the desired value of ξ .

The value of C_L and C_P are a complex function of the cell layout and the choice of dimensions for the 2-D array, and a closed form expression for optimization through analytical methods is practically impossible. After considerable computer simulation, a choice of $M = 14$, $N = 65$, and C_F of 330 fF was determined, based on a desire to make the cell size as small as possible yet meet all the constraints on the size of C_F outlined above. These values of M , N , and C_F roughly give a balanced parasitic of 3 to 4 pF each. Because the square law relationship between settling time and C_F is much stronger than the relationship between G_m and settling time, an emphasis has been placed on increasing C_F as much as possible given size and parasitic constraints. Increasing the value of C_F beyond 330 fF was investigated to reduce the requirements on the OTA, but was deemed counterproductive due to the escalating cell size and parasitics

which would increase both chip size and power consumption. Thus, the resultant value of G_m required to maintain ξ in the neighborhood of 0.707 is approximately 100 mS assuming a value of $1\text{K}\Omega$ for R_F .

Obtaining a value of 100 mS for the forward G_m may appear a bit unreasonable if the design of the OTA is restricted to a single state topology. However, as will be discussed later, the prototype circuit will use a modified OTA to achieve a higher forward G_m than would be achievable with a standard design.

4.6.6 Select Switch Sizing

It was stated earlier *a priori* that the select switch size is (20/1.2). The sizing of this switch was performed interactively using computer simulations to achieve loop stability and speed. The expression for ξ found in the previous section, Equation 4.4, shows that the value of R_F — determined by the switch size — combines with the forward G_m of the OTA to improve loop dynamics. Because of the relative difficulty in obtaining a large forward G_m compared to lowering R_F , the approach of using a larger than minimum switch was taken. Unfortunately, the effective “on” resistance of the switch in this circuit is quite high due to the low ($V_{GS} - V_T$) available. The common mode voltage present at the switch is the OTA input common mode voltage, which is fortunately less than half-supply — allowing for body effect, the ($V_{GS} - V_T$) of the switch can be expected to be about 1V. Thus, to obtain the value of $1\text{K}\Omega$ used in the above computations, a switch size of (20/1.2) was selected. Cell layout considerations discussed in the next section set the limitations on the size of the select switch. This was considered in obtaining the final size of the switch, as increasing the switch further would increase the parasitics, C_P and C_L , thus reducing the effectiveness of a decreased switch resistance.

4.6.7 Cell Layout

Given all the values of the elements in the storage cell depicted in Figure 4.12, the task now turns to producing a cell layout that is compact and minimizes parasitics yet provides isolation and shielding of stray capacitances. In addition, as the array is configured as a 2-D array, a method of bussing signals in a row-column manner is required. The cell must be fully differential, and strict adherence to symmetry is required to achieve the necessary common-mode rejection.

Each cell requires V_{DD} , V_{SS} , the row and column address lines, two pairs of data lines, two storage cells, two select switches and the logic necessary to decode the access signals. The aspect ratio of the cell needs to be chosen such that an array of 14 rows and 65 columns retains a reasonable aspect ratio. To achieve this, the cell is laid out such that it is roughly twice as tall as it is wide, with the width of the cell dictated by the size of the storage capacitor. The capacitor aspect ratio is kept below 2:1 to minimize fringing capacitance as a proportion of the plate capacitance. The axis of symmetry is a horizontal line through the center of the cell, with mirror symmetry for all analog signal parts. The logic is placed at the center of the cell, and is shielded from the analog sections by the supplies which act as an AC ground. Thus, the inherent asymmetry of the logic is shielded from the analog sections. A n-well shield is placed underneath the storage capacitor and is connected to a clean V_{DD} supply. The select switch is placed immediately adjacent to the storage cell — a Metal 2 shield plate is placed over the switch and a portion of the storage capacitor to minimize stray capacitance across the switch. This stray capacitance acts as a shunt capacitance across the entire array, and if it is large enough, will contribute to an IIR like response, degrading frequency response. The area to be shielded is a trade-off between shielding effectiveness and added parasitics. The actual shielded area was computed to reduce the value of the stray parasitics to less than 0.05% of the storage capacitance.

The supplies, data lines, and the row address lines are bussed horizontally across the cell in Metal 2 with the column select running vertically in Metal 1 along one side of the cell. Once again a design trade-off exists, this time between metal width and parasitics. The distributed capacitance and resistance of the metal lines must not contribute to excessive delay, especially on the row select line. Fortunately, there is no static power consumption in any of the cell components thereby alleviating the need for heavy supply and ground lines, nor is there any section of the cell that requires an accurate voltage reference. The resultant cell layout can be seen in Figure 4.16. Bussing of signals in both vertical and horizontal dimensions allows construction of an array by butting the cells next to each other, minimizing the overhead of array construction. Ample substrate and well contacts are present to minimize the possibility of stray noise coupling from underneath the active circuit. As discussed earlier, a Metal 2 shield plate extends over the access switch and part of the capacitor. The overall cell size is $40\ \mu\text{m} \times 77\ \mu\text{m}$, with the size being limited by both active area and metal pitch. Using these cells, the array of 910 storage cells occupies an area of $2.8\ \text{mm}^2$.

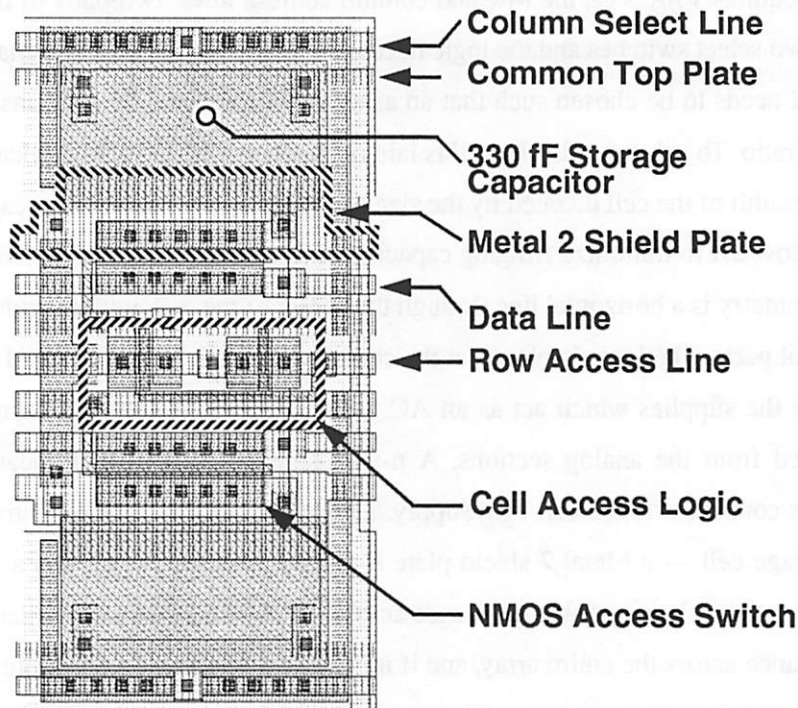


Figure 4.16 Storage Cell Layout

4.6.8 Row Multiplexers

Incumbent with the 2-D array outlined in the previous sections are the row (de)multiplexers required to route the signals to and from the read/write OTAs to the proper row with minimum insertion resistance while isolating the unwanted parasitic capacitances. The strategy involved to optimize the design of the muxes is similar to that of the cells, but not as critical. Here, the goal is to minimize the series pass resistance while keeping the added parasitics to a reasonable level. Although each row will only see the parasitics associated with each mux, the OTA will see the source/drain diffusions of all the rows combined. This is also another reason that $M = 14$ was chosen over an alternative layout with more rows. Finally, these switches are used to isolate the write and read OTA circuits from each other to prevent run-through of the data.

4.6.8.1 Mux Switch Design

Each row contains one write mux and one read de-mux, each of which contains four switches — differential sets of the summing node and OTA output lines. Because these switches serve to isolate the write and read OTAs from each other, they must switch at the clock rate. Thus, one limitation of the switch size is the amount of gate drive available to toggle the

switches and the power consumed by the operation. However, the limiting factor is the parasitics associated with large switches — multiplied by the number of rows, these appear at the terminals of the OTAs affecting the circuit by adding more load capacitance and by reducing the overall loop gain. Complementary pass gates are used in certain locations unlike the cell to provide the lowest series resistance. Switches connected to the summing node of the OTAs use a single NMOS switch to minimize parasitics, as they will be at the input common mode voltage of the OTA, which is fairly low. However, the switches at the output of the OTA, which swing about the output common mode voltage (approximately half-supply), use complementary switches because under certain conditions, the NMOS switch will either cut off or have a very high pass resistance (Figure 4.17).

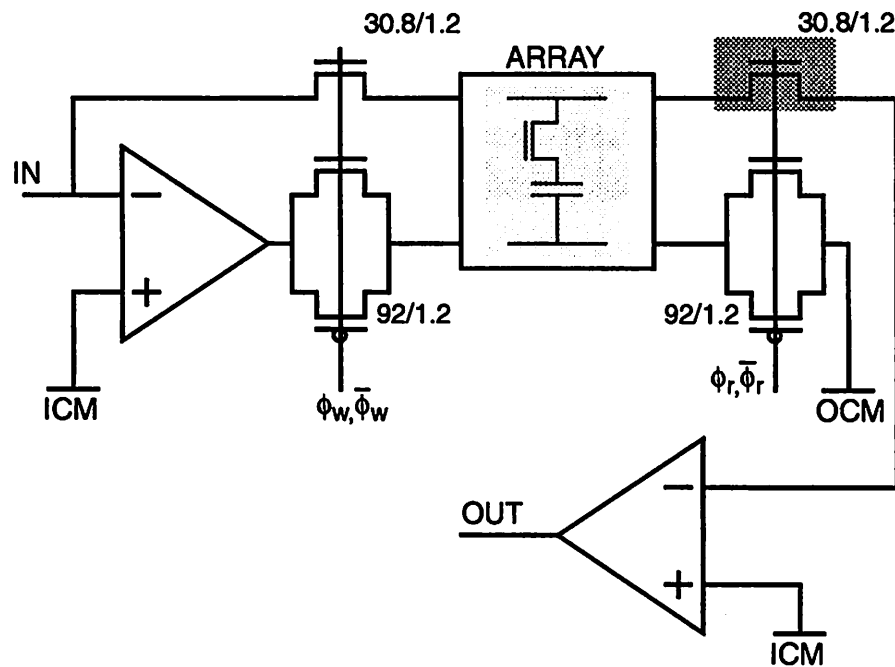


Figure 4.17 Arrangement of Write/Read Muxes (SE equivalent)

The actual sizes of the switches was determined through a combination of analysis and computer simulation. Most critical is the sizing of the PMOS gate, as this contributes the most resistance under high swing conditions as well as the largest parasitic capacitance per unit on resistance. The final values (30.8/1.2) for the NMOS switches and (92/1.2) for the PMOS are based on an asymmetrical folded layout, in which the majority of the source-drain diffusion is placed on the array side of the switch — this minimizes the load seen by the OTAs.

4.6.8.2 Multiplexer Logic

The muxes contain simple logic to perform a logical AND between the row select line and the write or read strobe. The output of this logic provides the proper signal to enable the switches to perform the write and read operation. The circuit consists of a static CMOS NAND gate followed by an inverter to generate true and complement outputs as there are both NMOS and PMOS switches present. Devices were sized to provide a 1ns (nominal) rise/fall time for the switch gates.

4.6.8.3 Auxiliary Output for Shunt Capacitor

A technique used to speed the circuit dynamics which was implemented in the prototype will be discussed in Section 4.9 on page 112 which uses an auxiliary shunt capacitor. However, it should be noted here that in addition to providing the gate drive signals for the main switches, a portion of the signal is used to drive a SPDT NMOS switch used in this capacitor. This increases the load on the logic gate within the mux resulting in a slightly larger than expected gate size.

4.6.8.4 Read Switch Delay Circuit

Included in the demux circuit is a one shot circuit used to delay the turn on of the read enable switch (shaded in Figure 4.17). This is a safety measure to insure that the remainder of the circuit is switched completely to the read mode before the switch is closed to prevent loss of charge and data corruption. Two critical events must occur before the read enable switch is closed. First, the write enable switch which connects the array to the write OTA must be open. Second, the read OTA output shunt switch, (required because the OTAs are used only during half the clock period — during the other half, they are disabled by shunting the differential outputs), must open before the read enable switch is closed. Although the entire chip is clocked with two non-overlapping clocks, this safety measure is included to allow for process variations which may cause unmatched clock skew, and because of the fatal nature of the potential mistiming (e.g. if the read enable switch closes too early, the charge in the storage cell will flow either into the write OTA or be shunted to ground — in either case, the data is irrecoverably lost).

The circuit makes use of a delay stage and logic (Figure 4.18) to produce the desired waveform. The gate delay of the two inverter chain determines the one-shot delay. In the prototype circuit, inverters with longer than minimum channel length were used to increase the delay available to provide the required safety margin.

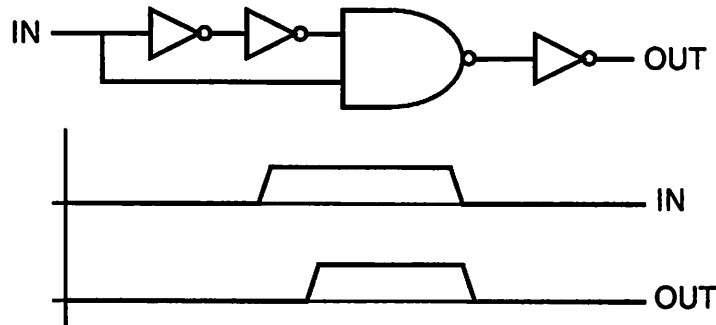


Figure 4.18 One-Shot Delay Circuit

4.6.8.5 Layout Considerations for the Multiplexers

An area of substantial concern is the mixed signal nature of the multiplexers. One part is involved with the actual switching of the analog signals from the OTAs, while adjacent to the analog switches are digital circuits switching fairly large loads with fast edges. Although the circuit is synchronously clocked, steps were taken to minimize the coupling possible both capacitively over the active areas and through the substrate.

To achieve this isolation, the analog and digital areas were separated by at least one metal line at AC ground potential. In addition, a heavily doped grounded substrate stripe lies between the analog and digital sections. Finally, the well/substrate connections for the digital devices are tied to separate V_{DD}/V_{SS} lines to prevent a noisy digital supply line from coupling noise into the substrate.

Although the write mux and read de-mux are essentially independent circuits, in order to minimize the routing of clocks and to segregate as much as possible the digital sections from the analog sections, both mux and de-mux circuits are laid out in one block. The layout allows tiling in the vertical direction with pitch equal to that of the vertical dimension of the storage cell. Signals to and from the OTAs as well as supply and clock signals are tapped at either the bottom or top of the tiled stack. The resultant cell size is $70\mu\text{m} \times 520\mu\text{m}$ (HxW).

4.6.9 Assembly of the Storage Array

The 910 storage cells and the 14 row multiplexers are combined together to form the storage array used by the circuit to provide a 1H delay. Due to the bussing of the metal lines across the

cells, all components of the array are simply tiled together — there are no extra metal lines required. Figure 4.19 shows the relative floorplan of the array. In the completed circuit, there are two of these storage arrays as two lines of delay are required to realize the desired transfer function.

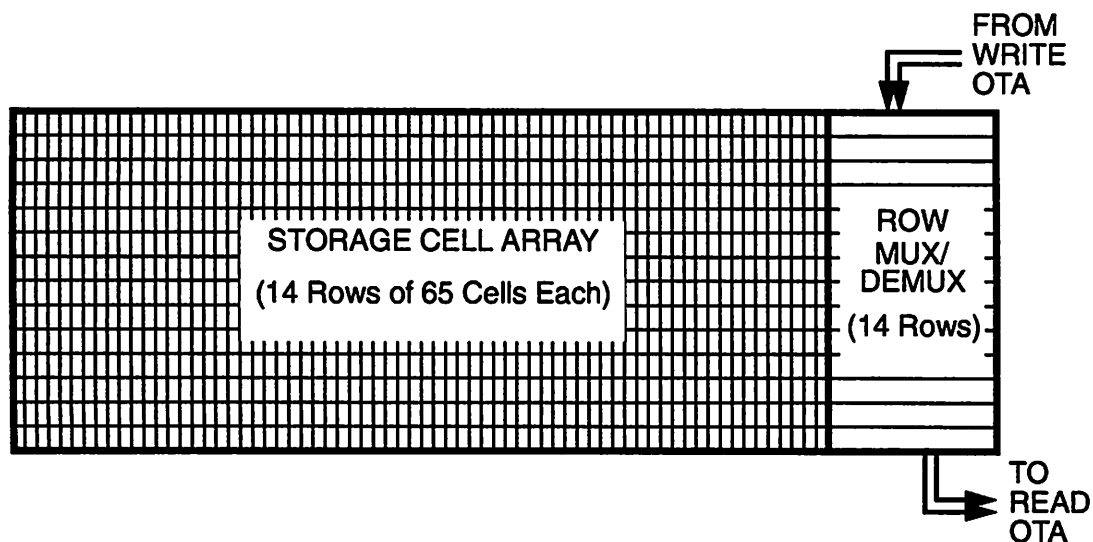


Figure 4.19 General Floorplan of Storage Array

4.7 Design of the OTA

4.7.1 Introduction

As was discussed in earlier sections, especially Section 4.6.5 on page 83, the OTA must, in addition to usual requirements of speed and stability, have a very high forward G_m , in the neighborhood of 100 mS. Such values of G_m are rarely found in traditional switched capacitor circuits, as these values of G_m are not usually needed and imply large devices and/or high bias currents. Moreover, loop stability problems may arise when the forward G_m is increased to too high a level. However, in this prototype circuit, the overwhelming dominance of parasitic capacitances mitigates the stability problem to a large degree, at the same time increasing the requirement on the amplifier G_m to achieve adequate speed. This section will discuss the various topologies considered and an analysis of the chosen design.

4.7.2 Amplifier Specifications

4.7.2.1 Effect of Finite Amplifier Gain

While the analysis and design of switched capacitor circuits is most easily undertaken using the ideal op-amp approximation, at some point, the effects of finite amplifier gain must be considered [52]. For the prototype circuit, the OTA will be operated as a SC integrator performing charge transfer as shown in Figure 4.20. For a perfect ideal OTA, the transfer function at DC is

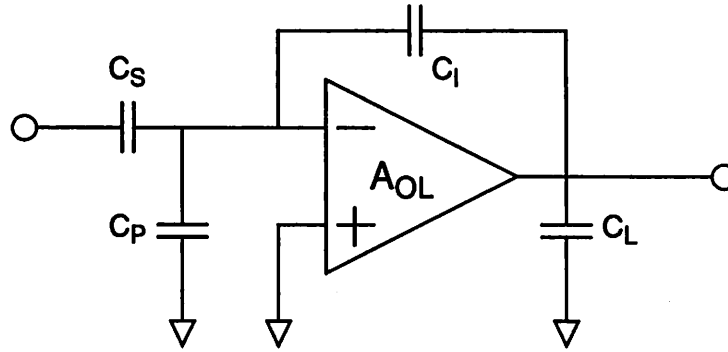


Figure 4.20 Circuit to demonstrate effect of finite amplifier gain

given by $(-C_S/C_I)$ as expected, and can be determined to the limit of capacitor matching, which can be on the order of 100 ppm. However, if the conditions are relaxed to allow a non-infinite forward gain, A_{OL} , in the amplifier, the DC transfer function is given by:

$$-\left(\frac{C_S}{C_I + \frac{C_S + C_I + C_P}{A_{OL}}} \right) \quad (4.5)$$

which shows another error term due to the non-infinite gain of the amplifier. Note that the error term is adversely affected by large parasitics at the summing node. For conditions such as those found in the prototype circuit, where $C_P \gg C_I$, the additional error factor can be approximated as the ratio of C_P/A_{OL} to C_I . Thus,

$$A_{OL_{min}} \approx \frac{C_S}{\epsilon C_I} \quad (4.6)$$

where ϵ is the allowable error. From earlier chapters, C_P is approximately 4 pF, while C_I has been determined to be 330 fF. Thus, to maintain a 1 percent gain error, an open loop gain of 1200 is required. Although this effect of finite amplifier gain manifests itself as a gain error, which can

be compensated for later in the circuit, because it is dependent on the value of parasitics rather than actual circuit elements, the degree to which the gain is affected is somewhat uncertain. As a result, it is desirable to minimize this effect.

4.7.2.2 Secondary Requirements

In addition to the G_m and gain requirements, the OTA must, minimize power consumption and maintain reasonable output swing. Characteristics that are not as important in this circuit are input offset voltage and noise. Because the underlying reasoning for this work is to demonstrate the lower power requirements of analog circuits, the OTA needs to be as efficient as possible, while at the same time be as simple to boost reliability and speed and provide adequate signal swing to improve the systems SNR. Due to the fully serial nature of the signal path, and the fact that video signals are DC corrected at each subsequent signal processing block, the absolute offset voltage of the amplifiers is not critical, so long as they do not have short term drift, and do not cause the subsequent stages of the circuit to saturate. Finally, as the overall SNR goal of the circuit is in the low 50dB range, the noise performance of the OTA is not a primary concern.

4.7.3 Basic Single-Stage Amplifiers

Because of the relatively fast settling times required by the circuit (~25 ns), simple single-stage amplifiers are of interest because they have the minimum number of devices — hence poles in the transfer function — while providing the amplification desired. Within this class of amplifiers, there are two main divisions, unfolded (“telescopic”) cascode amplifiers and folded cascode amplifiers. Each has its own strengths and weaknesses as outlined below.

4.7.3.1 Unfolded “Telescopic” Cascode OTA

Unfolded cascode OTAs, also known as “telescopic” due to the shape of the schematic, are single stage amplifiers with active loads and cascoding to achieve high gain with minimum number of active devices [53]. The transfer characteristics of such an OTA are easily defined, since the circuit can be assumed to have a single-pole response when placed in a capacitive feedback loop. The open loop gain of this circuit is simply the product of the g_m of the input devices and the output impedance at the output nodes. The overall G_m of the circuit is simply the g_m of the input devices, and the unity gain frequency is just the capacitive load at the output divided by the g_m of the input devices.

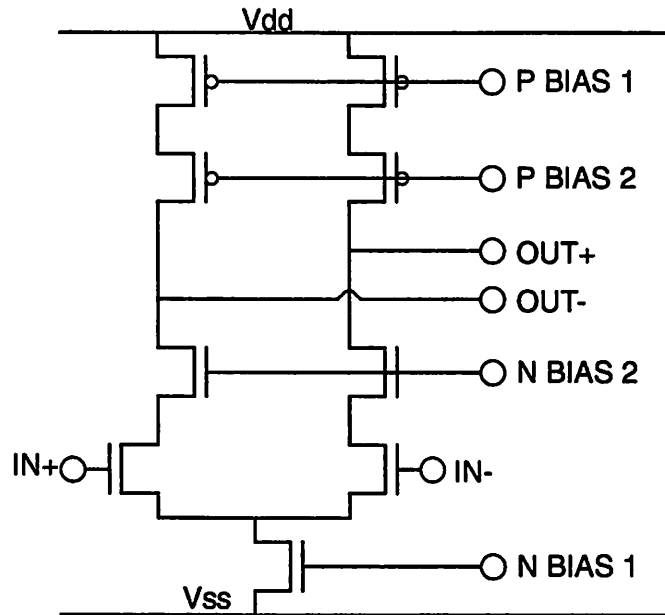


Figure 4.21 Schematic of "Telescopic" OTA

The simplicity of this design, however, also restricts the ability to change the specifications of the OTA. To achieve the desired 100 mS of Gm with this OTA in this technology, the $\frac{W}{L}I_D$ product would have to be approximately 80, implying that the input devices would have to be quite large with a correspondingly large bias current (e.g. 10000/1.2 at 10 mA). Such a large device would present an input capacitance of almost 20 pF, which would adversely affect the feedback network thereby negating the efforts of the previous section to optimize the value of the storage capacitor. In addition, the large bias currents necessary would result in excessive power dissipation, in direct conflict with the desire to produce a low-power analog solution to the task at hand. Moreover, as the gain of this type of OTA is inversely proportional to the square root of the bias current, operating the OTA at these bias levels would result in a very low open loop gain, which would result in substantial errors in the analog RAM transfer function. Finally, "telescopic" OTAs have a fairly limited output swing with a correspondingly small input common mode range due to the danger of placing the input or input cascode devices into the triode region. The effect of this particular limitation is dependent on the design of the overall system. In circuit where large dynamic range is of paramount importance, the loss of available signal swing is highly undesirable. However, in this application with its moderate dynamic range requirements (~50 dB), the effect of this limitation may not be of substantial concern.

The restriction on input common mode range is a bit more troublesome, as it implies that measures to insure that the input common mode voltage does not drift out of specification is required. Such drifts usually come from clock feedthrough and offsets in the system “AC ground” reference voltages. Such measures can usually be taken through the use of replica bias circuits to generate the precise voltages necessary for proper circuit operation while tracking drifts due to temperature and supply.

Coupled with this restriction is the fact that the input and output common mode voltages are by definition different for this architecture. This presents a problem for circuits that have cascades of switched capacitor circuits. A common mode level shift circuit is required in between each state — while this can be easily accomplished using the standard four switch S/H circuit, it introduces another potential source of offsets to the signal. This presents a limitation for circuits that are DC sensitive; however, the application of this prototype circuit, video, is first order insensitive to DC offsets as they are removed after processing. As a result, as long as the offsets are small enough to prevent saturation of the subsequent stages, the added penalty of common mode shifts are not that severe.

4.7.3.2 Folded Cascode OTA

The restrictions on input common mode range and output swing of the “telescopic” OTA can be overcome through a slight increase in circuit topology known as the folded cascode OTA [54]. In this circuit, the differential output current from the input devices is steered into a separate high impedance node which can swing a much larger range without the danger of placing critical devices into the triode region. At the same time, the input common mode range of the devices is increased as the V_{DS} of the input devices becomes first order independent of the output voltage. Finally, the input and output common mode voltages can be made equal, thereby simplifying the interstage circuits.

The price for this added flexibility is reduced bandwidth. As shown in the diagram of Figure 4.22, the signal path includes a PMOS device operating as a common gate device. Although the bandwidth of a common gate device is near the f_T of the device, the pole resulting from this can affect the achievable settling time of circuits that use this type of OTA. Some designs “flip” the circuit upside down, with PMOS input devices and a NMOS common gate device to take advantage of the higher f_T of the NMOS device. However, such a strategy reduces the available gain and G_m of the OTA due to the inherently lower performance of PMOS devices.

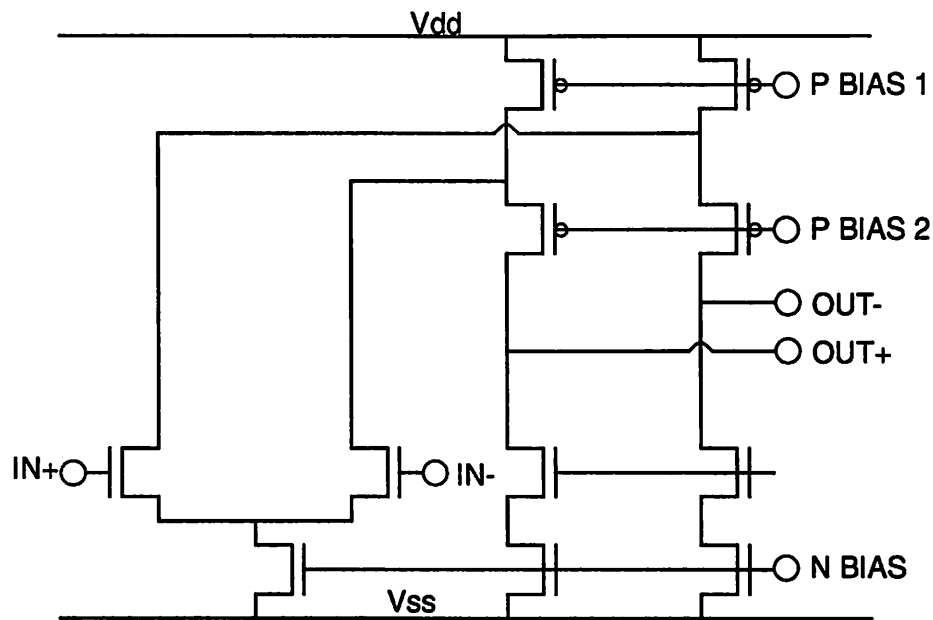


Figure 4.22 Schematic of Folded Cascode OTA

Due to the prospects of reduced settling performance, and low gain and G_m from PMOS input devices, and the fact that the difference in the input and output common mode voltages is not a critical issue in this system, the advantages of the folded cascode architecture did not appear to outweigh the loss in speed and added complexity.

4.7.3.3 Class AB OTAs and Dynamic Biasing

The OTA architectures detailed in the previous two sections operate in the class A region — that is, the devices in the circuit are biased in their active region over the full input and output range. However, class A amplifiers are inherently inefficient circuits and dissipate the majority of the power consumed in current sources. Circuits that need to drive a large amount of parasitic capacitances require large bias currents to achieve adequate slew rate as the maximum differential output current is limited to twice the quiescent bias current. Class AB amplifiers, on the other hand, allow the output current to exceed the quiescent bias current by a large factor to charge large capacitive loads, while keeping standby current at a very low level, thereby increasing overall power efficiency. This makes class AB amplifiers such as those proposed by Black, Castello, Lewis and Gray [55],[56],[57] attractive when power consumption is a key parameter.

Along with the class AB structure, many power efficient amplifier schemes use a form of dynamic biasing. In its simplest form, dynamic biasing adjusts the device bias currents to match

the instantaneous requirements of the amplifier. For example, the bias current can be adjusted to be large at the beginning of an integration cycle, then slowly taper off as the amplifier settles. During the period of time the amplifier is quiescent, the overall bias current can be kept very small. In addition to power efficiency, this method also provides improved open loop gain. An analysis of the forward open loop gain of a single stage OTA shows that the gain is a function of the $g_m r_o$ product of the devices. MOS devices exhibit increasing g_m and decreasing r_o with an increase in drain current. However, the increase in g_m follows a square root dependence while the output resistance is a linear function. Thus, the overall gain of actively loaded MOS amplifiers tends to decrease with increasing bias currents. As a result, an inherent trade-off exists between fast slewing and accuracy in switched capacitor circuits. By using dynamic biasing, however, the bias currents can be made to taper off during the last portion of the clock period allowing the amplifier to exhibit high gain while providing the large charging currents necessary at the start of the clock period.

Although the class AB architecture and dynamic biasing schemes present attractive methods of providing low power solutions to providing high gain and output current capability, as with the folded cascode amplifier, the added number of devices necessary to implement these circuits results in reduced frequency response and correspondingly lengthened settling times. In addition, it is unclear as to whether the additional high frequency poles would interact adversely with the complex feedback network presented by the prototype chip. While it is not clear that a viable solution to this particular circuit design problem could or could not be addressed using class AB amplifiers, for purposes of simplicity and robustness, it was decided not to pursue these styles of amplifiers for the prototype circuit.

4.7.4 Two Stage OTAs

Two stage amplifiers as shown in Figure 4.23 are an alternative to the cascode amplifiers of the previous section [54]. A short analysis of the open loop gain, and the frequency response of single stage cascode amplifiers versus this style of amplifier, however, shows that this simple two stage amplifier is inferior to the basic single stage amplifier. Because this style of amplifier has two poles in the forward path, stability in a closed loop configuration cannot be guaranteed. As a result, a compensation capacitor must be included in the circuit. Miller compensation using C_C provides the dominant pole, while the output load C_L provides the non-dominant pole. In the single stage amplifier, the output load forms the dominant pole, with the common-gate (cascode)

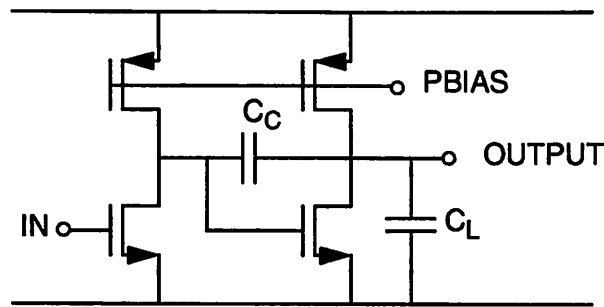


Figure 4.23 Schematic of Two Stage Amplifier (single-ended equivalent)

device contributing a non-dominant pole near the device f_T . Any two pole system with sufficient closed loop gain will result in pole splitting — the time constant of the loop is approximately proportional to the reciprocal of half the non-dominant pole. Thus, comparing the non-dominant pole of the two stage amplifier, g_m/C_L , to the non-dominant pole of the cascode amplifier, g_m/C_{gs} , it is obvious that for circuits with large C_L , the single stage amplifier provides a faster settling solution.

Another concern with the two stage amplifier in this prototype circuit is the additional pole introduced by the feedback network unique to the circuit. As a result, a three pole system results, which exhibits stability problems even with moderate amounts of loop gain. Because of this and the fact that two stage amplifiers are likely to be slower than their single stage counterparts, the two stage architecture was not considered to be viable for the prototype and will not be discussed further.

4.7.5 Single Stage Amplifier with Preamp Stage

Unfortunately, the available single stage architectures do not provide a simple solution to the requirements of high gain and large G_m that are required in this circuit. As a reminder, the desired G_m is on the order of 100 mS, while the desired minimum A_{OL} is 1200. Because the factors which increase forward G_m tend to lower the forward gain, a combination of device sizes and bias currents which meet the requirements while maintaining a reasonable power and area is not feasible in this technology. Therefore, a modification to the standard single stage amplifier is proposed.

4.7.5.1 Low-Gain Wide-Bandwidth Preamp

Addition of a low-gain, wide-bandwidth voltage preamplifier stage to the circuit has the effect of increasing both the forward G_m and A_{OL} while affecting the stability of the system to a minimal degree. By using the simplest structure possible, the gain bandwidth product of the preamplifier can approach the inherent device f_T . Preamplifiers with a gain of 5 V/V can have bandwidths exceeding 500 MHz in this technology, while reducing the required G_m and A_{OL} requirements for the main amplifier to 20 mS and 250 V/V, both of which are achievable in this technology.

The structure shown in Figure 4.24 uses an all NMOS g_m/g_m stage to provide the gain

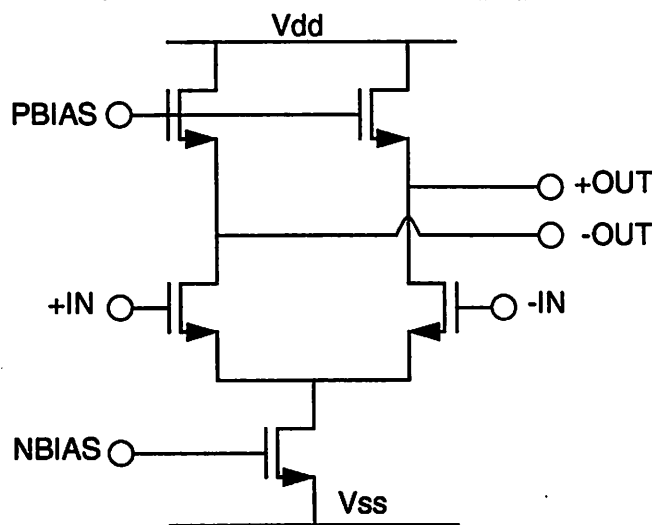


Figure 4.24 Schematic of Wide-Bandwidth Preamp Stage

required in the preamplifier. The voltage gain is adjusted by sizing the load devices appropriately. In the prototype circuit, a drawn β ratio of 18.3 was used to give a voltage gain of approximately 4.3 V/V. However, the effective β ratio is closer to 20 giving a voltage gain of 4.5 V/V. Attempts to raise the voltage gain are hindered by the increasing $PBIAS$ voltage necessary for operation. Making the load devices smaller would require that the $PBIAS$ voltage be above 4.25V, which would substantially affect the supply margins of the design. The input devices are sized at 470/1.2 with a tail current of 800 μ A.

The common mode output is determined by the value of $PBIAS$, and the load device size. Because the output common mode voltage is $(PBIAS - V_{GS-load})$, this voltage is not uniquely

defined across process variations, necessitating the use of a replica bias circuit to generate the *PBIAS* voltage. The input common mode range is quite narrow for this circuit and is bounded on the lower side by the minimum V_{DS} of the tail current source, and on the upper side by the required output common mode voltage — for the combined amplifier, it is quite low as will be seen shortly.

4.7.5.2 Complete OTA Circuit

The preamp of the previous section is cascaded in front of a standard “telescopic” OTA discussed in Section 4.7.3.1 on page 98. The resulting amplifier is shown in Figure 4.25. Because

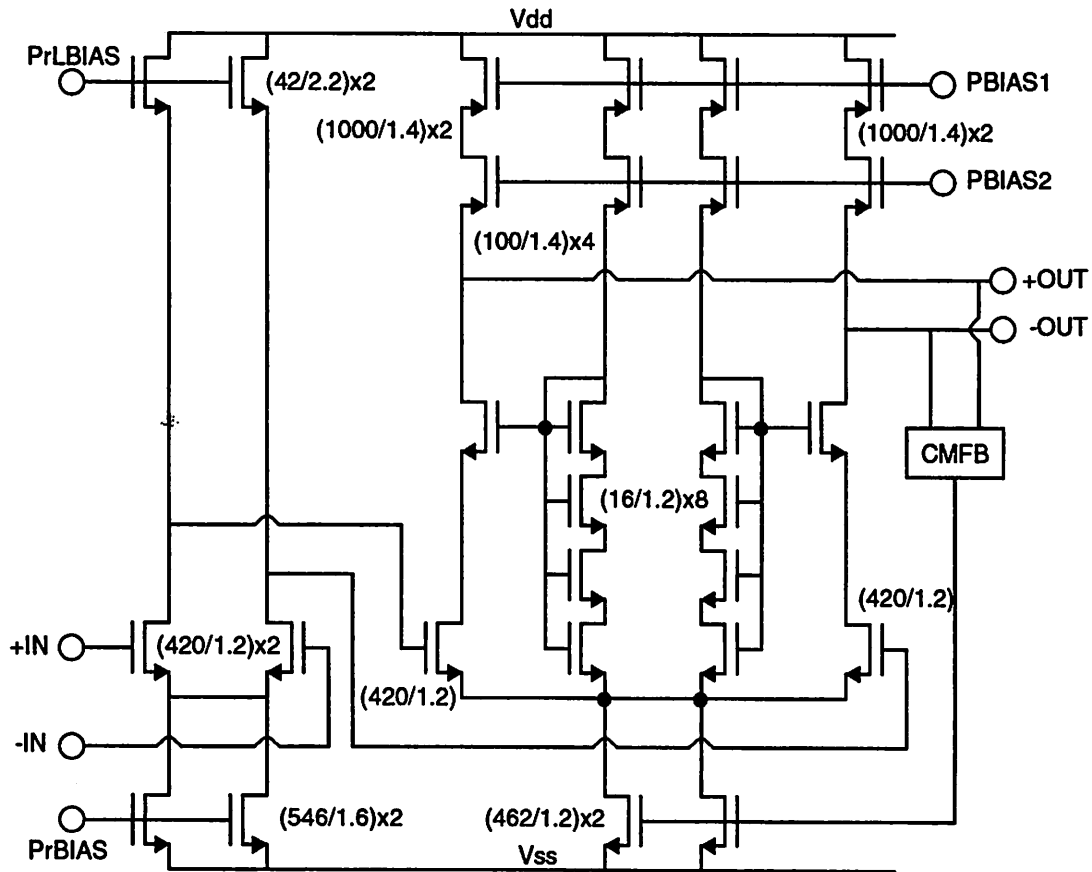


Figure 4.25 Schematic of OTA with Preamp

the output common mode is undefined for this type of configuration, an auxiliary circuit is necessary to fix the common mode output voltage. This circuit, denoted CMFB in Figure 4.25, is detailed in Figure 4.26. It uses a capacitive divider to determine the common mode voltage with a switched capacitor circuit to replenish the charge on the capacitors. The size of the capacitors

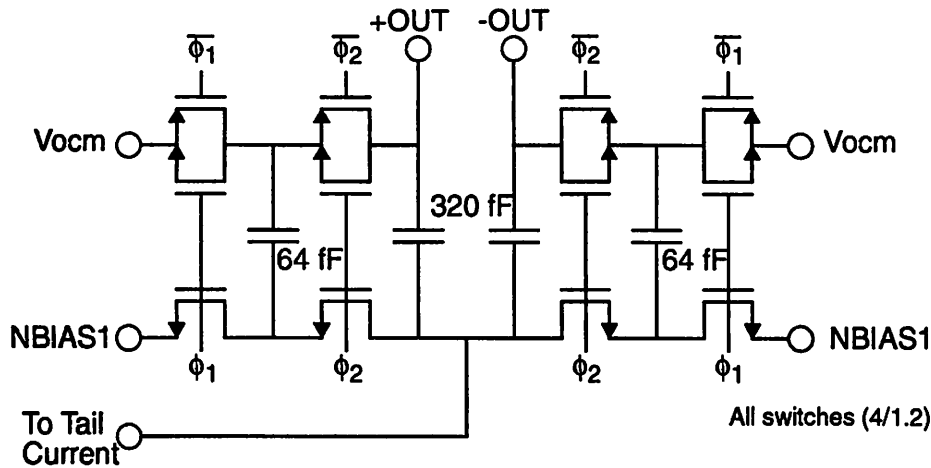


Figure 4.26 Detail of Common Mode Feedback Circuit

used in the CMFB circuit are a function of the allowable common mode pedestal, and the loop dynamics of the common mode circuit. V_{OCM} is the desired output common mode voltage of the OTA, while $NBIAS1$ is the voltage determined by the replica bias circuit necessary to balance the circuit. The switches replenish the charge on the center capacitors each clock cycle to compensate for leakage. Because the $NBIAS1$ and tail current bias voltages are quite close to the negative rail, single NMOS circuits are acceptable, while the switches for V_{OCM} use complementary devices to minimize the switch on resistance.

The sizing of the devices in the main amplifier is a trade-off between area and speed, keeping in mind the performance of the devices as a function of current density. Normally, the size of the input device of a single-stage amplifier is dictated by the total G_m needed in the amplifier. However, in this topology, the total G_m is the product of the A_{OL} of the preamplifier and the g_m of the input device. Thus, by increasing the gain of the preamplifier, the main amplifier input device can be made to provide a smaller amount of g_m . The proper partitioning of these two parameters is dictated by the relationship between total G_m and the bandwidth of the amplifier.

The g_m of a MOSFET in saturation is given by:

$$g_m = \sqrt{2\mu C_{ox} \frac{W}{L} I_D} \quad (4.7)$$

while the input capacitance of the same device is simply $\frac{2}{3}WLC_{ox}$. The pole formed by the preamplifier stage is due to the input capacitance of the main amplifier combined with the output resistance of the preamplifier, which is, for all intensive purposes, the resistance of the diode

load, and is simply the inverse of (4.7), adjusted for the size of the load device. Moreover, the gain of the preamplifier, for a fixed preamplifier input device, is proportional to the diode load in the preamplifier. These give several relationships between the device sizes, bandwidth and overall G_m which are important to the design of this style of amplifier. Because of the square-root relationship between main amplifier input device size and total G_m , and the linear relationship between the preamplifier diode load resistance and the total G_m , coupled with the linear relationship between the main amplifier device size and the pole frequency, there is a net G_m -bandwidth product gain by making the preamplifier gain as high as possible by raising the diode load resistance while using a small device for the main amplifier. As mentioned previously, the limitation to this approach is the maximum gain available in the preamplifier due to headroom problems. The final input device size was fixed at (420/1.2) keeping these considerations in mind. Unfortunately, a device of this size would require about 8mA of drain current to provide the required G_m of 20mS. This creates a serious problem, for the resultant $(V_{GS} - V_T)$ of the devices would seriously impact the available output swing.

Clearly, some method of reducing the G_m requirements of the amplifier are required, as the above bias current is unreasonable for a low power design. Fortunately, a technique described in Section 4.9 on page 112 uses an auxiliary capacitor to effectively double the value of C_F present in the circuit. Recalling the square-law dependence of ξ as a function of C_F , it becomes clear that doubling the effective C_F to 660 fF reduces the minimum required G_m of the cascode amplifier to 6.6 mS. This in turn reduces the bias requirement to 800 μ A which is very reasonable for this application.

The active load is a PMOS cascode current source formed with two (1000/1.4) devices. Actual test devices characterized demonstrated a large difference in available r_o between 1.2 μ m and 1.4 μ m length devices. While increasing the channel length further would improve the r_o of the current source, there would have to be corresponding increase in the device width adding area to the circuit. In addition, the required gain in the main amplifier is only 250 V/V; thus, maximizing the output resistance is not critical to this circuit. The gates of the devices are biased by voltages generated in the bias circuit described later.

The NMOS cascode devices are sized identically to the input devices, and are biased internally to prevent potential OTA to OTA crosstalk problems that could arise by sharing a common bias line. The PMOS current sources are extended to provide a bias current of 1/10 the main cur-

rent. This current is fed into a stacked diode connected device to generate a V_{GS} plus a ΔV to keep the input devices fully saturated. The choice of bias current in this section of the amplifier is a trade-off between power dissipation and chip area versus the impedance at the gates of the NMOS cascode devices. The value of ΔV is set at approximately 300 mV based on the characteristics of measured NMOS devices in this technology. The stacked devices are composed of four devices in series — although only two devices are required to form the bias circuit, four devices were used to allow the use of 1.2 μm devices to improve matching between the bias leg and the main circuit leg.

The circuit is biased at the negative rail using a pair of non-cascoded tail current sources sized at (462/1.2) which are biased by the common mode feedback circuit described above. Minimum length devices were used as the CMRR at this point is not as critical, with much of the common mode being removed by the preamplifier. Finally, the bias for the preamplifier is provided by a pair of (546/1.6) devices whose gate is biased by the bias generator. Here, longer channel lengths were used to provide a higher output resistance giving rise to better CMRR.

4.7.5.3 Layout Considerations

In addition to the usual precautions of symmetry for differential circuits, the circuit Figure 4.25 was laid out with care in several other respects. First, most all devices are compound devices composed of many smaller devices in parallel. This is not only necessary due to the large aspect ratio of these devices, but also facilitates matching. Thus, the input devices and tail current sources are composed of (42/1.2) segments, while the PMOS current sources are composed of (100/1.4) segments and so forth. Thus, the current source for the cascode bias could be made very close to 1/10th the main current by using one (100/1.4) segment. To balance this extra current, the tail currents were made 10 percent larger by adding an extra (42/1.2) segment.

A factor unique to the overall layout of the circuit dictated a Metal 2 shield be placed over a portion of the main amplifier input and cascode devices. This is because the amplifier output lines were routed fairly close to these devices, and any stray capacitance would be Miller multiplied and result in degraded performance.

Finally, each section of the circuit is isolated using metal lines at AC ground potential and ample substrate contacts and substrate diffusions are provided to minimize noise and potential differences. On chip bypass capacitors are provided for the *PBIAS1*, *PBIAS2* and preamplifier

tail current bias lines to minimize on chip noise pickup. These are formed using poly-poly structures and are situated underneath the supply rails — thus, they consume no extra area on the die.

4.8 Bias Generator

4.8.1 Introduction

To simplify the circuit and maximize the probability of first silicon working, a simplified bias circuit was used for the prototype. This circuit takes an external reference current and generates the necessary voltages to bias the OTAs in addition to providing the input and output common mode voltages for the OTA which are required in various locations throughout the circuit.

4.8.2 Reference Voltage Generation

The bias circuit is shown in Figure 4.27 takes a single external reference current which is then mirrored as necessary to provide the required output voltages. Cascode devices are used on the input diode so that the input device has a V_{DS} comparable to the tail current sources in the amplifier and the other current setting devices in the bias circuit. A small bleeder resistor formed by a (3/160) PMOS device insures start-up of the circuit. The input bias current is mirrored twice to form three replica currents from PMOS devices. The first output is used to determine the proper input bias voltage. Using a four device structure similar to that in the OTA, the voltage developed at *Node A* represents the proper V_{GS} for the preamplifier input device. This voltage is then raised a diode drop and to provide the gate drive for a source follower to provide low output impedance. The second PMOS current source is used to generate the preamplifier load bias. Note that this bias point indirectly determines the input common mode voltage present at the main amplifier inputs. As discussed earlier, the “telescopic” OTA has a fairly limited input common mode range, thus control of this voltage is important to proper operation of the circuit. Once again, a four device structure is used to generate the “ideal” input common mode voltage for the main amplifier. This is then increased by the V_{GS} of the preamplifier load device to give the proper bias voltage for the preamplifier. Although the impedance at the *PrLBIAS* node is quite high due to the fairly small device (42/2.2), no source follower scheme similar to the input common mode bias voltage generation was used. This is for two reasons — first, while the input common mode bias voltage sees glitches in the load due to switching of sample/hold stages, the preamplifier load bias line sees a fairly constant load. Second, including a source follower would limit the maximum *PrLBIAS* voltage to V_{DD} minus the sum of the V_{GS} of the source follower

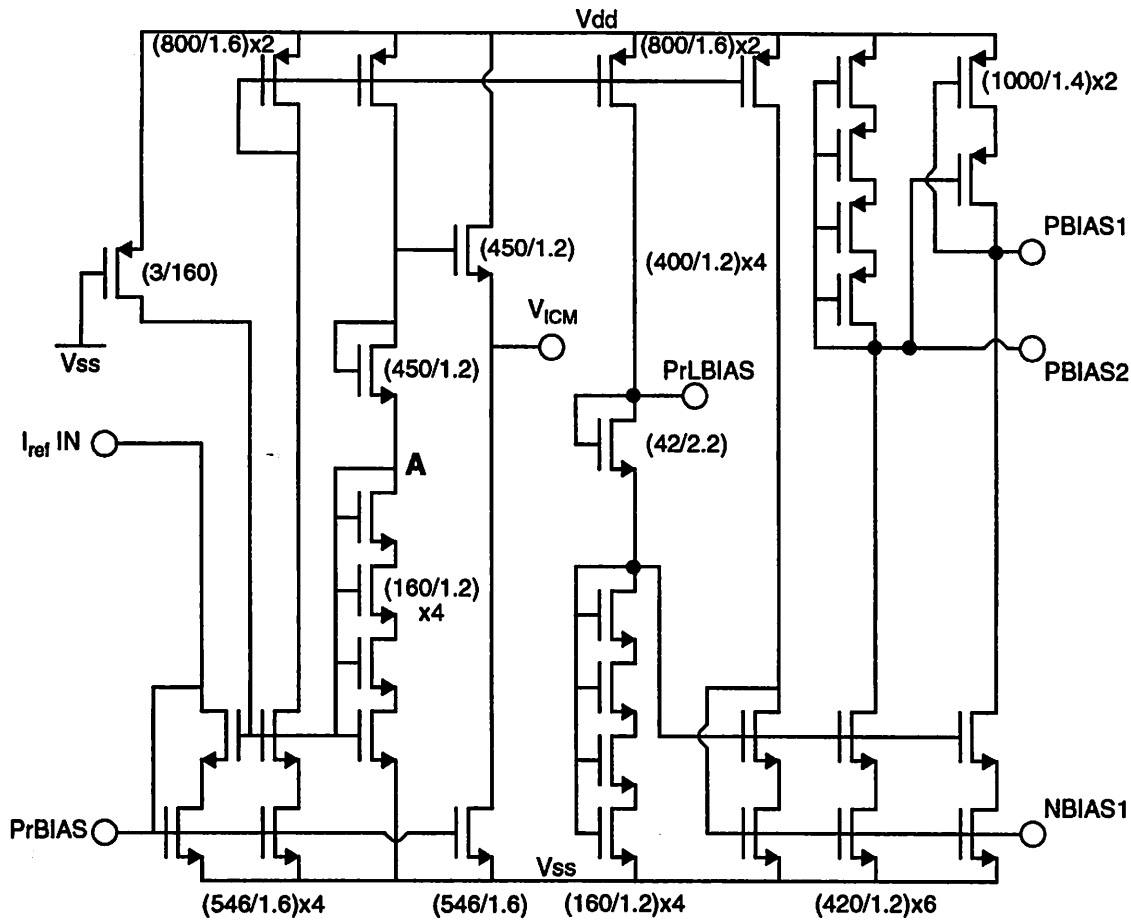


Figure 4.27 Schematic of OTA Bias Generator

plus the V_{Dsat} of the PMOS current source. Because of the fairly high voltages at these points, the V_{GS} of a source follower could be well above 1V due to body effect. This would limit the maximum voltage available for $PrLBIAS$, and thus limit the gain available in the preamplifier. As determined earlier, maximizing gain in the preamplifier is advantageous; thus, omission of the source follower stage produces a net overall win for the system. The final PMOS current source is used to bias up the NMOS and PMOS reference voltages for the main amplifier. The NMOS bias voltage ($NBIAS1$) which is used indirectly by the OTA through the common mode feedback circuit, is generated by a diode which is cascoded to emulate the V_{DS} seen in the actual OTA to provide best matching. The PMOS reference voltages ($PBIAS1$ and $PBIAS2$) are generated in the same way, with another four stacked device structure to generate the ΔV to insure saturation of the cascode devices.

4.8.3 Output Common Mode Voltage Generator

The common mode output voltage of the OTA is not defined by any circuit. The CMFB circuit within each OTA will force the common mode output voltage of each OTA to be the OCM voltage. Thus, it is advantageous to make the OCM voltage a value such that the output swing of the OTA is maximized. The device sizes and other characteristics of the OTA have been designed for a 5 volt supply — under these conditions, maximum swing occurs with a OCM of approximately 2.7 volts.

The internal OCM generator uses a polysilicon resistive divider to generate the 2.7 volt signal, and buffers the voltage to provide a low impedance output. Because the ideal OCM voltage can drift due to process and temperature variations, an option is made to allow an external voltage to be applied to the buffer amplifier so that the OCM voltage can be altered from outside.

The buffer circuit (Figure 4.28) is open-loop to avoid stability problems, and drives a source follower circuit for low output impedance. Minimum channel length is used in the mirror as the error due to output resistance is minimized by the low V_{DS} across both devices. The circuit is biased with the *NBIAS1* line from the bias generator.

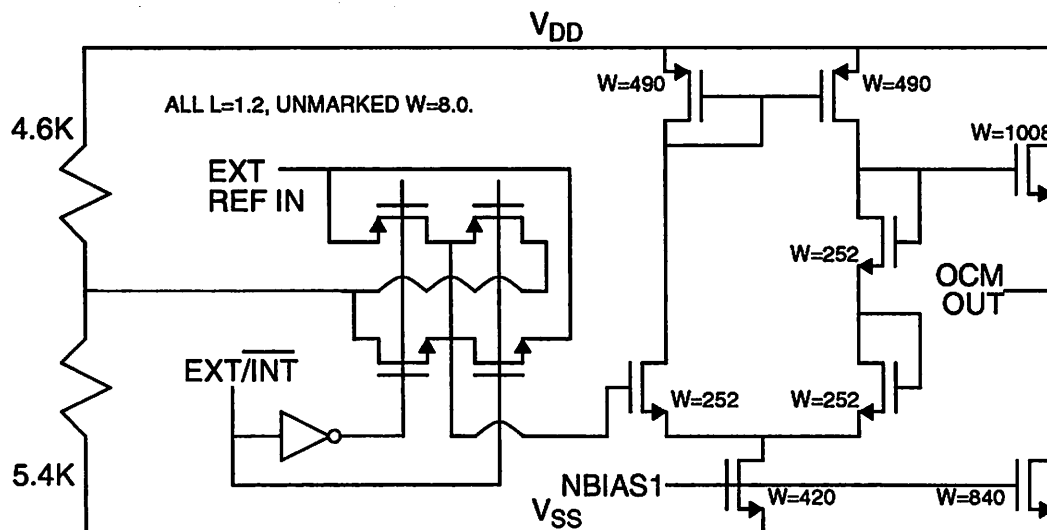


Figure 4.28 Output Common Mode Generator Schematic

4.9 Auxiliary Feedback Capacitor Stage

4.9.1 Introduction

In the previous sections, it was found shown that if some method of increasing the effective value of CF, the feedback capacitor, could be found while minimally impacting the magnitude of the parasitics, a huge gain in circuit speed could be achieved. This section will describe the use of an auxiliary feedback capacitor placed in the circuit which accomplishes the above. Recall that the majority of the parasitics are due to the accumulated parasitics from each cell. Thus, any addition to the cell itself, including adjustments to the component sizes/values, would probably increase the parasitics seen by the circuit and hence negate any benefits of the modification. As a result, an auxiliary capacitor is added outside of the storage cell, but within the row multiplexer to increase the effective feedback factor.

4.9.2 Write Cycle Auxiliary Capacitor

The auxiliary capacitor is required only during storage cell write, for only during the write phase does the adverse feedback ratio exist. Recall the simplified circuit discussed earlier, repeated here as Figure 4.29, except that an additional capacitor has been added. Note that this

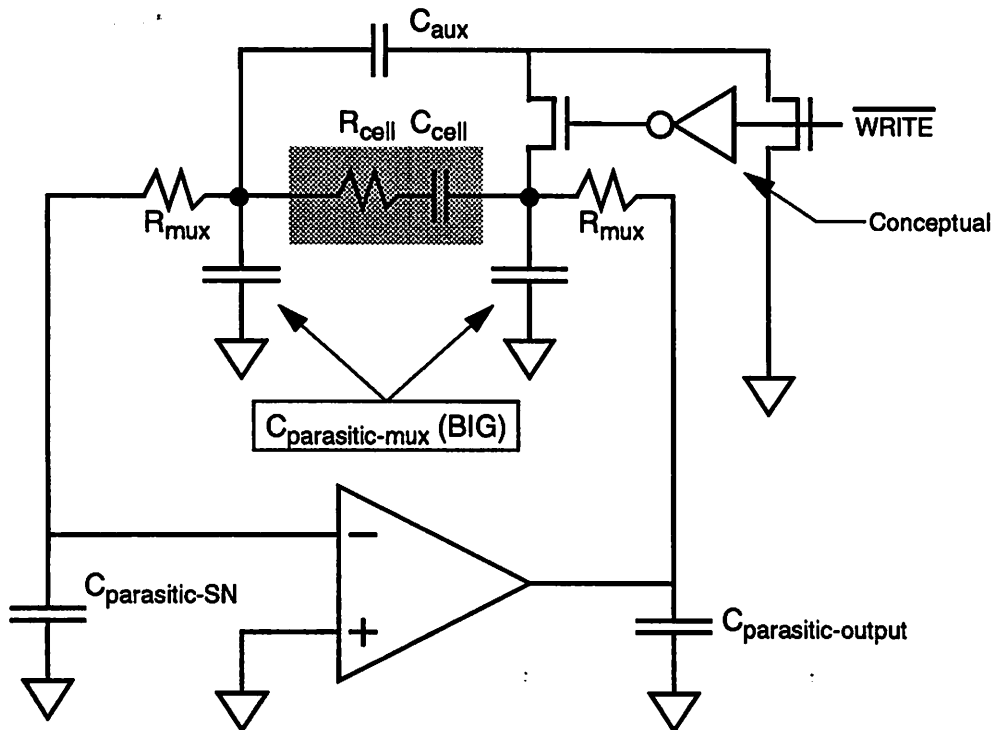


Figure 4.29 Simplified Write Circuit with Auxiliary Feedback Capacitor

capacitor is not part of the storage array, and hence adds minimal parasitics of its own. This capacitor, C_{aux} , is sized nominally to be equal in size with C_{cell} , such that the parallel combination of C_{cell} and C_{aux} is twice C_{cell} , or 660 fF. Note that this results in half the charge flowing into C_{aux} , thereby reducing the magnitude of the stored signal. This loss is overcome by using a double size capacitor in the input sample and hold to the write amplifier.

The capacitor is connected to the circuit through a double throw NMOS switch. One plate of the capacitor is permanently connected to the summing node of the write OTA. Because this node, in theory, does not move in voltage, this connection should not affect the circuit operation except that it adds a small amount of parasitic capacitance. The other plate of C_{aux} is connected to the “wiper” of the switch. One pole of the switch connects to the output of the OTA, while the other pole connects to AC ground. It is important that the charge stored in C_{aux} is completely discharged before the next write cycle. Otherwise, it would contribute to charge sharing between two cells, and affect the transfer function of the Analog-RAM. In particular, when used as a line delay, the frequency response will suffer, as an IIR-like response will be added. Incidentally, this is precisely the effect avoided by shielding against C_{stray} in Figure 4.12. Therefore, an adequately sized device must be used for the SPDT MOS switch. Based on “on-resistance” studies similar to those used to size the mux switches, a device size of (18.6/1.2) was chosen based on layout considerations.

The gates of the switch are controlled by the auxiliary capacitor output of the mux described in Section 4.6.8.3 on page 94, and are driven with opposite non-overlapping clocks to prevent charge from being lost from the OTA output to ground. These clocks are derived from the gate drives to the main mux switches, thus insuring that the auxiliary capacitor operates in synchronism with the remainder of the circuit. Because of the relatively small size of the switch compared with those used in the mux, the additional clock load is not significant, and is easily compensated by slightly increasing the size of the mux clock drivers.

The auxiliary capacitor and SPDT switch are laid out as part of the multiplexer — although only one of these circuits is necessary for the entire array, for purposes of layout and clock simplicity, it was decided to include an auxiliary capacitor stage for each row of the array. Because all the required clock and signal lines are present in the multiplexer, the impact on the layout is minimal, and no lengthy, parasitic filled routing is necessary.

4.9.3 Effect of Auxiliary Capacitor

The primary goal and effect of the auxiliary capacitor is to increase the effective feedback around the write OTA. As far as this is concerned, the auxiliary capacitor accomplishes this goal well, and as predicted earlier, the loop damping factor of the circuit is reduced while at the same time reducing the G_m requirements on the OTA. However, like all things, there is a cost paid with this approach. In particular, because C_{aux} in parallel with the selected cell capacitor appears to be a charge divider, a fraction of the signal is diverted and subsequently thrown away. While the loss of the charge can be made up as described earlier in the input stage, the exact fraction of the input charge that flows into C_{aux} cannot be governed well. The relative fractions of charge that flow into C_{cell} and C_{aux} are a function of the relative capacitances. The danger is that due to random mismatches in the capacitors, there will be a random variation in the ratio of C_{cell} to C_{aux} contributing to a random variation in charge sharing.

Because each row has a single C_{aux} , while the C_{cell} is unique to each storage cell, variations in the value of C_{cell} from cell to cell will result in gain variations as each cell is selected. This is highly undesirable, for if the cells are used in a video line delay, it can easily contribute to fixed pattern noise. Fortunately, the size of the capacitors is such that in this technology, relatively good matching is expected from cell to cell, on the order of 0.1%. Moreover, because any fixed pattern noise will be on a pixel to pixel basis, and hence relatively high in frequency where the eye is less sensitive to noise, the degradation to image quality should be negligible. Thus, the original advantage of capacitor mismatch insensitivity described for the storage array in general is curtailed with the use of this technique. However, given the specifics of this implementation, access to high-quality capacitors, and the speed limitations of this technology, the trade-off of capacitor accuracy for speed is justifiable.

One other issue is that of capacitor linearity, also discussed earlier. Unlike the matching issue, as long as the non-linear characteristics of the capacitors are matched between C_{cell} and C_{aux} and the two capacitors are roughly equal in size, the effects of the non-linearity will be negligible. Therefore, high-density capacitors such as MOS devices can still be used with this approach, although matching requirements specified above may prove to be difficult to meet with lower quality capacitors.

The importance of this auxiliary capacitor technique will probably fade in the future, as it is a method of extracting a bit more speed out of a technology, which is bound to improve. In this

application, a clock speed of $4f_{sc}$ is required — falling short of that, even by 10 percent would make the circuit unusable. This inflexible requirement, combined with the fact that this particular technology is marginal for this application justify the use of this additional technique. With faster technologies, however, it is anticipated that the auxiliary capacitor will be removed and attempts made to utilize only the cell capacitance as feedback.

4.10 S/H and Summer Stages

4.10.1 Introduction

At the input to each write OTA is a sample and hold stage, which not only stores the incoming value, but performs the necessary voltage to charge conversion. There are two such stages in the entire circuit, one at the input to the first delay line, and the second located in between delay lines to feed the second delay. In addition, there is a summer stage which takes outputs from both delay lines in addition to the current sample, scales and sums them with a final output OTA to form the output signal. Because these are basically similar structures, their design will be discussed as a group.

4.10.2 Input S/H Stage

The input sample and hold stage shown in Figure 4.30 is a fully differential S/H stage which

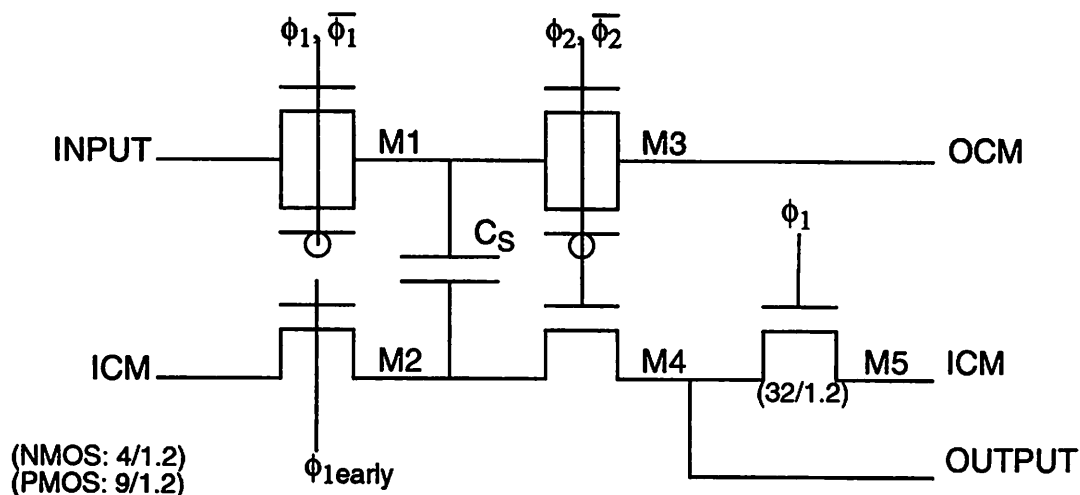


Figure 4.30 Input S/H Stage (1/2 Differential Circuit)

employs “bottom-plate” sampling to avoid signal dependent charge injection. During ϕ_1 , the

sampling switches are closed, and the input signal is applied across the sampling capacitor, C_S . The sampling switch, M2, is switched by an “early” ϕ_1 , or a clock that is asserted along with ϕ_1 , but is de-asserted slightly earlier. Note that opening M2 samples the signal, for one plate of the capacitor is not totally isolated. Because M2 always has the same source and drain voltages, ICM or input common mode, the charge injected from the channel due to M2 being opened is same, regardless of the input signal. Thus, a signal independent sampling operation is performed. Opening M1 then completes the sampling operation. During ϕ_2 , M3 and M4 are closed, thereby allowing charge stored in C_S to flow into the summing node of the write OTA. Because the OTA is configured as a charge transfer circuit, the charge in C_S is pulled out and transferred onto the storage cell as described earlier in this chapter.

Note that the top plate of the capacitor is switched to OCM (output common mode) in this case. This difference is necessary because the OTA used has unequal input and output common mode voltages. It can be seen that if a “zero output” from the OTA is applied to the input of this circuit (i.e. the absolute voltage equals OCM), then the output of this circuit is ICM , or the equivalent “zero” output as far as the input of the OTA is concerned. Thus, this circuit also acts as a level shifter to compensate for the level shift introduced by the OTA. Cascading of OTA and S/H stages is possible without adverse effects from common mode level shifts using this technique. However, care must be taken that the same reference voltage is used throughout the circuit. Therefore, a central bias circuit is critical to successful operation of the circuit; mismatches between the ICM voltages throughout the chip contribute to a fixed DC offset. Fortunately, the nature of video processing allows for a DC offset to be removed easily, and for this application, the sensitivity of this technique to offsets is not an issue.

In addition to providing sampling, an extra transistor, M5 is introduced to clamp the OTA inputs. Referring to Figure 4.9, it is evident that the OTA is unused half the time, namely during ϕ_1 . Moreover, there is no feedback applied to the amplifier, and due to random offsets, the amplifier will saturate when left disconnected. Because the amplifier could possibly enter a funny state of operation, and take a long time to recover, allowing the amplifier to saturate is undesirable. Thus, M5 serves to clamp the input to ICM during ϕ_1 . Similar switches are used to short the differential outputs of the OTA together at the same time.

Transistors M1 and M3 are complementary switches, while M2, M4 and M5 are single NMOS switches. The ICM voltage is fairly close to the negative supply rail, and hence, enough

gate overdrive is supplied with the gates at V_{DD} such that there is sufficient conduction through the channel if the source of the transistor is at ICM , as is the case with M2, M4 and M5. However, the source voltage of M1 and M3 are not defined, as it represents the signal voltage. It is possible that with high swings, the source voltage of M1 and M3 could be high enough to prevent a single NMOS switch from functioning. As a result, a complementary switch was used to insure proper closure of the switch for all conditions. Near minimum size switches are used to minimize channel charge injection. However, M5 must be sized fairly large due to the large amount of parasitic capacitance located at the input of the write OTA. A (32/1.2) switch is used to insure that the inputs are clamped securely to the ICM voltage.

Layout of the circuit is critical to insure matching of between the two differential halves and with other components on the chip. Recall that the auxiliary capacitor scheme requires that the input capacitor be twice as big as the storage capacitor. Two "unit" capacitors are arranged in parallel so that not only the areas but the perimeters are doubled to match capacitance due to fringing effects. Dummy Poly 1 – Poly 2 structures are placed on the outside edges of the capacitors to minimize the effect of a "bias" in processing. The entire capacitor array is covered with a Metal 2 shield to prevent stray fields from affecting the input signal. Finally, a grounded well with a ohmic substrate ring is employed to reduce the effect of substrate noise coupling.

4.10.3 Intermediate S/H Stage

This stage is essentially identical to the input S/H stage described above, and is used between the two delay lines. As with the input S/H, the storage capacitor is double size to provide the extra charge that will be thrown away in the auxiliary capacitor. Clocking, layout and interface to this circuit is identical to that of the input S/H stage.

4.10.4 Output Scaling and Summing Stage

This section serves to take the signals that represent vertically aligned samples of a video signal, scale them and sum them to yield the desired signal. Recall that the idealized transfer function of a 2H comb filter is:

$$H(z) = 0.25z^{-2} - 0.5z^{-1} + 0.25 \quad (4.8)$$

Because the gain through the delay lines is nominally unity, the output stage must first scale each signal (zero delay, 1H delay and 2H delay) appropriately and sum them together. The circuit is

designed in a fashion very similar to that of the input S/H, but with three sets of inputs, sampling switches and sampling capacitors. Scaling of the signals is accomplished by altering the size of the capacitors according to the desired coefficient. Again, a set of unit size capacitors are used to insure that a high degree of matching is achieved. During the output clock phase (ϕ_2), the charge in all three sampling capacitors are transferred using an output OTA to form the output. A signal inversion for the 1H delayed signal is accomplished by cross coupling the differential signals.

4.10.4.1 Effect of Mismatch in Scaling and Non-Unity Gain in Delay Line

It is important to discuss the effects of a mismatch in the scaling capacitors, as they would affect the overall transfer function obtained. The effects caused by non-unity gain through the delay line as a result of finite OTA gain can also be lumped in with the capacitor mismatch as a source of “coefficient error.” In order to analyze the effect of such errors, it is necessary to investigate the transfer function given in (4.8) more closely. Specifically, non-ideal coefficients will affect the stopband performance of the filter — in this type of filter, zeros are formed by cancellation of the weighted signals. Any deviation in the coefficients will result in incomplete cancellation with resultant loss of attenuation. Thus, the overall comb filter notch depth is most severely affected by the matching of the gain through the delay stages and the weighting capacitors.

Consider a generalized three tap FIR filter with tap coefficients A, B, and C. The resulting z-transform transfer function is given by:

$$H(z) = A + Bz^{-1} + Cz^{-2} \quad (4.9)$$

The magnitude squared response of such a filter is given by:

$$|H(e^{j\omega T})|^2 = [A + B \cos \omega T + C \cos 2\omega T]^2 + [B \sin \omega T + C \sin 2\omega T]^2 \quad (4.10)$$

The zeros of such a filter will be located at:

$$\frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad (4.11)$$

The locations of the zeros determine the frequency response, especially that of the stopband — there are three distinct cases to consider:

(1) The argument inside the square root is greater than zero. Then, there are two real zeros, on either side of the point $(-1,0)$ in the z -plane at a distance determined by the argument inside the square root (Figure 4.31 (b)). If the goal is to place a double zero at the frequency corresponding to half the sampling rate, then minimizing the argument, $B^2 - 4AC$ is critical. the tap weights for the prototype comb filter nominally have a ratio of $1 : 2 : 1$ ($A : B : C$); thus, if the product AC is less than one, real axis zeros will occur. This is especially undesirable as the frequency response will contain no zeros. A reduction as small as one percent in the values of A and C with respect to B causes a reduction in the notch depth from infinity to -46 dB, with a corresponding worsening of the response at other frequencies.

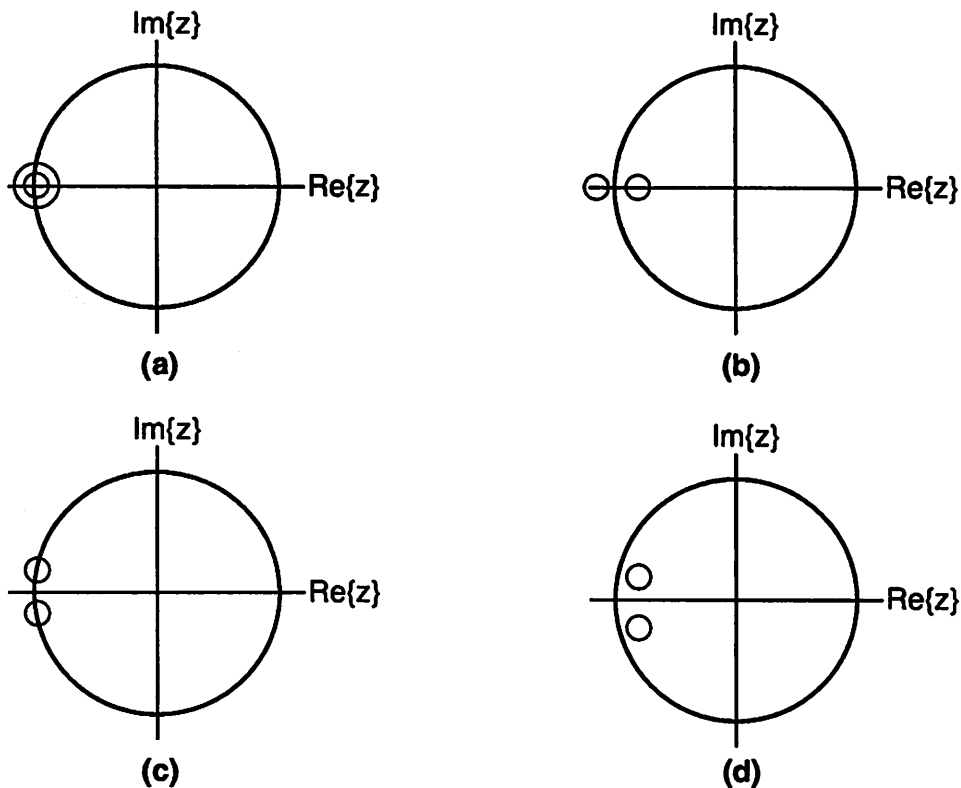


Figure 4.31 (a) Ideal Double Zero at $(-1,0)$; (b) Real Axis Zeros; (c) Complex Zeros on Unit Circle; (d) Complex Zeros off Unit Circle

(2) The argument inside the square root is zero. This is the ideal situation, with a double zero, giving the classical second order “cosine” filter (Figure 4.31 (a)).

(3) The argument inside the square root is less than zero. Complex zeros will result, given by:

$$\operatorname{Re}\{z\} = \frac{-B}{2A} \quad (4.12)$$

$$\operatorname{Im}\{z\} = \pm \frac{\sqrt{4AC - B^2}}{2A} \quad (4.13)$$

Two sub-conditions then exist. If A and C are essentially equal, then the zeros will lie on the unit circle, as the sum of the squares of the real and imaginary parts of the zeros equals one (Figure 4.31 (c)). Otherwise, the complex zeros will lie off the unit circle at a distance proportional to the ratio of C to A (Figure 4.31 (d)). Complex zeros lying on the unit circle are actually advantageous, as it gives two nulls in the response of the filter, spread slightly apart in frequency, with a zone of high attenuation in between. This results in a higher degree of separation of the Y and C components than in case (2) above. The same holds true to some extent with complex zeros off the unit circle, with the caveat that the nulls will no longer be true zeros. However, the effect of mismatches is lessened, as the error is distributed between the real and imaginary components of the zeros. A one percent error results in nulls of -63 dB, compared with -46 dB as was the case with real zeros (Figure 4.32).

The conclusion of this analysis is that complex zeros are much preferred to real zeros. Thus, an error such that coefficients A and C become slightly greater than ideal as opposed to slightly less will provide the greatest immunity against coefficient mismatches.

4.10.4.2 Compensation of Mismatch in Coefficients

Given the results from the previous section, it is highly desirable to skew the coefficients such that the ratio will nominally become (1.01 : 2.00 : 1.01). Considering the signal path of the circuit, the final output summer has three inputs, that of zero delay, one of $1H$ delay and one of $2H$ delay. The two sources of coefficient error are non-unity gain through the delay stage, and scaling capacitor mismatch. The first effect is to a degree deterministic — negating capacitor mismatches between C_S and C_I in the delay stage, the non-infinite voltage gain of the OTA will cause an error which is given by (4.5), which shows that in this case, there will be a reduction in gain of approximately the ratio of C_P/A_{OL} to C_I . The OTA can be expected to have a minimum

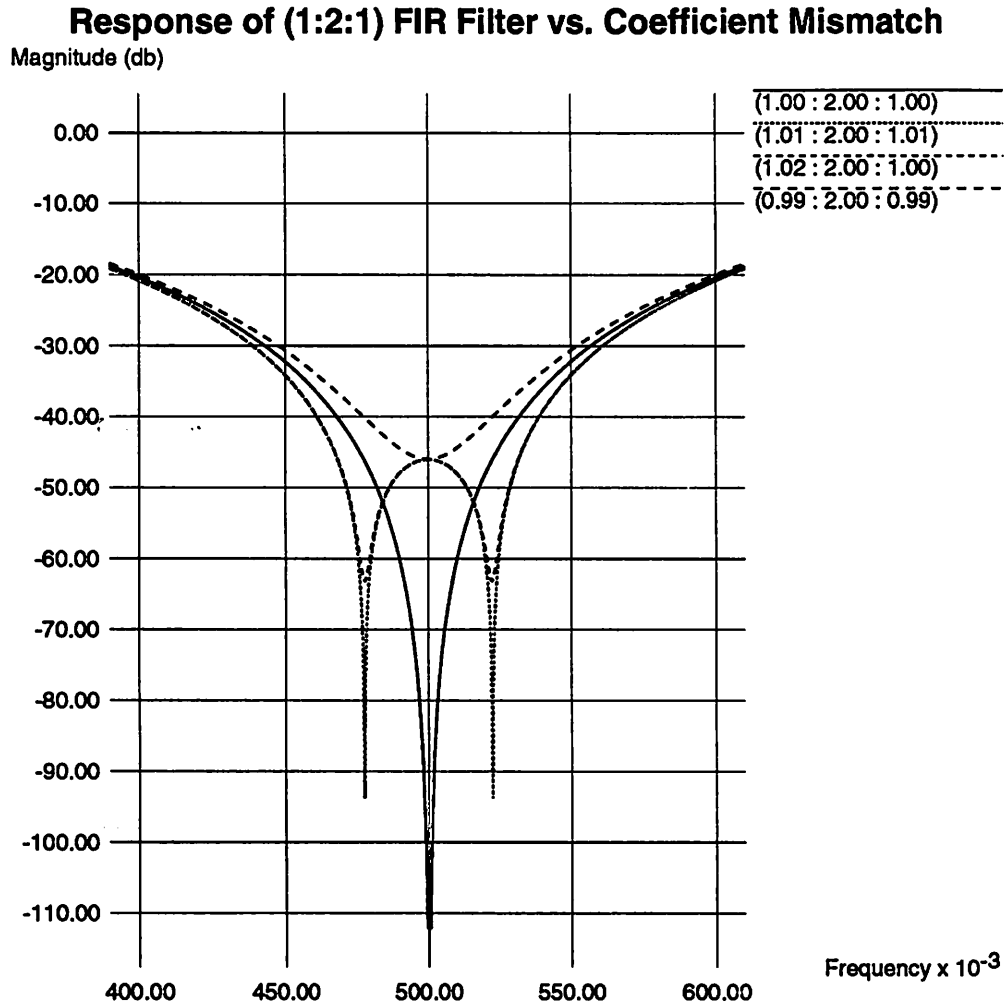


Figure 4.32 Frequency Response of a (1:2:1) FIR Filter as a Function of Coefficient Mismatch

A_{OL} of approximately 1500, while implies a slightly less than 1 percent gain loss through the delay line.

This loss in the signal can be easily made up by adjusting the size of C_j ; however, in the prototype, was not done. Therefore, based on just the gain loss, the effective gain at the input to the output summer is (1.00 : 0.99 : 0.98), which implies final tap coefficients of (0.25 : 0.495 : 0.245). Although these are not ideal tap coefficients, analysis of the previous section shows that it will still result in a double zero — however, the filter will not be linear phase, and will exhibit

a narrower stopband compared with the ideal case. Since the effect is minimal, modifications to the circuitry to compensate for the lost gain in the delay stages were not made.

The second source of errors is random and is the result of mismatches in the capacitors, both in the delay lines and in the output summer. While they are independent of each other, the effect is the same, in that they both affect the integrity of the coefficients in the final filter function. While care has been taken in the layout of the output scaling capacitors, a random error of 0.3% can be expected with the unit capacitor size used in the output summer ($22\mu\text{m} \times 22\mu\text{m}$). Similar sized capacitors are used for C_S and C_I in the delays. Under worst case conditions, a 0.3% error will occur in both the line delays and in the summer. Various combinations of errors are possible, but the set which most accentuates potential problems is the one which causes the zeros of the FIR filter to split and lie on the real axis. Based on the 0.3% error band for each set of capacitors, the worst case set of final coefficients is (0.250 : 0.503 : 0.250). This will result in a notch depth of 50 dB, rather than the theoretical infinite notch. As a figure in excess of 30 dB is considered adequate, steps to diminish these non-idealities were not pursued.

4.11 Output Buffers

One of the great advantages of implementing this filter function in a standard analog CMOS technology is that the circuit can become a part of a much larger VLSI signal processing circuit. Such was the intent with this circuit; the circuit concept and design made no efforts to drive signals to the outside world. It is intended in a real application that the output of this circuit will be processed by a on-chip chroma demodulator and matrix. However, in the interests of testing this chip with minimal downstream processing, a method of observing the output of the filter from the outside is necessary.

Because this the output of the circuit is sampled data analog, an accurate buffer circuit capable of driving tens of picofarads of stray capacitance to video rates is necessary. In the interest of simplicity, a simple source follower based buffer is used. Because of the inherent non-linearity of source followers due to the body effect, a PMOS device was used as the output device to allow the source of the device to be tied to the bulk through an isolated n-well. While this presents a fairly large capacitive load to the device, compared with the capacitance of external loads, the added penalty is minimal compared with the gain in linearity.

To further improve the linearity, a feedback loop with a simple $g_m r_o$ amplifier is used to linearize the circuit and stabilize the gain. This simple amplifier with a gain of approximately 30 V/V is expected to reduce the nonlinearities of the buffer to well under 1 percent. The circuit diagram of the buffer is shown in Figure 4.33.

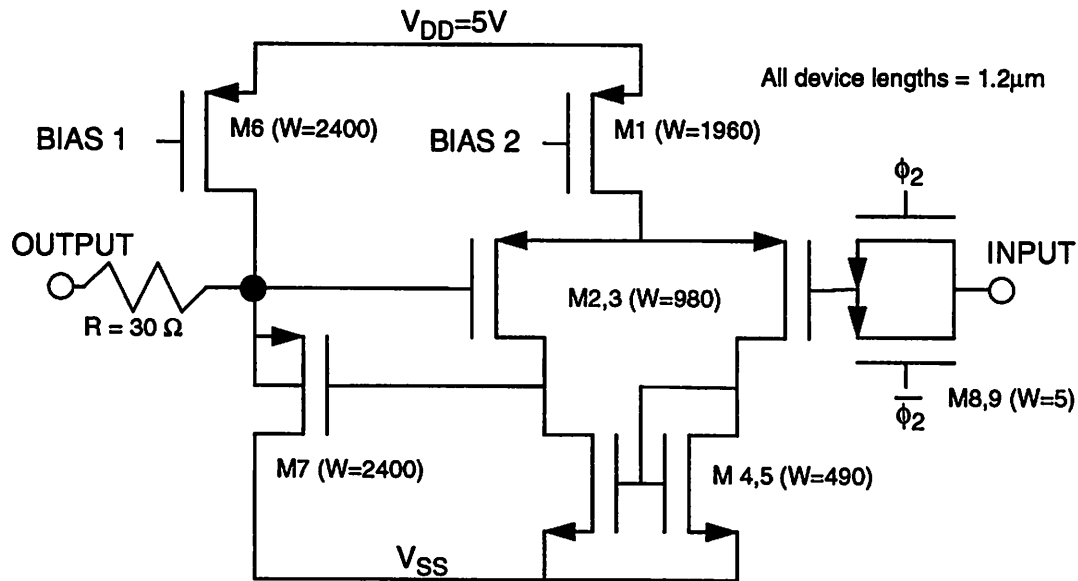


Figure 4.33 Circuit of Output Buffer Circuit

The buffer is a single-ended circuit; two identical stages are used to buffer each half of the differential output. While this means that the outputs are not truly differential, due to the linear nature of the output buffer, common mode signals are transmitted through the buffers. This will allow external reconstruction of a single-ended equivalent of the differential signal with full rejection of common mode noise. The output driver, M7, is biased with sufficient current (set externally) to drive the capacitance at the output. An output resistance of 30 ohms is used to decouple the load capacitance, preserving stability. The two bias voltages (*BIAS 1* and *BIAS 2*) are generated by forcing external currents into diode-connected devices of equal size. This allows tailoring the characteristics of the amplifier to the load, which is a function of the external circuitry connected to the chip.

M8 and M9 form a complementary switch which takes acts as an output sample and hold. This is included as the signals present within the chip are of the return-to-zero (RTZ) form. If an output sampling process were not included, the RTZ waveform would have to be replicated at the output resulting in large signal swings regardless of the frequency of the output waveform.

By sampling the signal, the RTZ waveform can be converted into a zero-order hold waveform which reduces, on average, the magnitude of the signal swings. In addition to reducing the power dissipation in the output drivers, this also reduces the amount of noise coupled into the substrate from the output drivers back into the analog circuitry.

The sampling clock is timed to close a short period after the settling cycle begins. At the start of ϕ_2 , the OTA which produces the output signal is beginning to settle. As the signals within the circuit are RTZ, the first portion of the settling period is spent bringing the output from zero to near the final voltage. Delaying the closing of the switch prevents the output from attempting to track the first part of the settling curve. The switch is timed to close roughly two time constants after the beginning of the settling cycle, and opens at the end of the settling cycle, so that the output of the buffer holds the final value for that clock cycle. This delayed clock is produced by a circuit very similar to that shown in Figure 4.18.

4.12 Clock and Address Generation

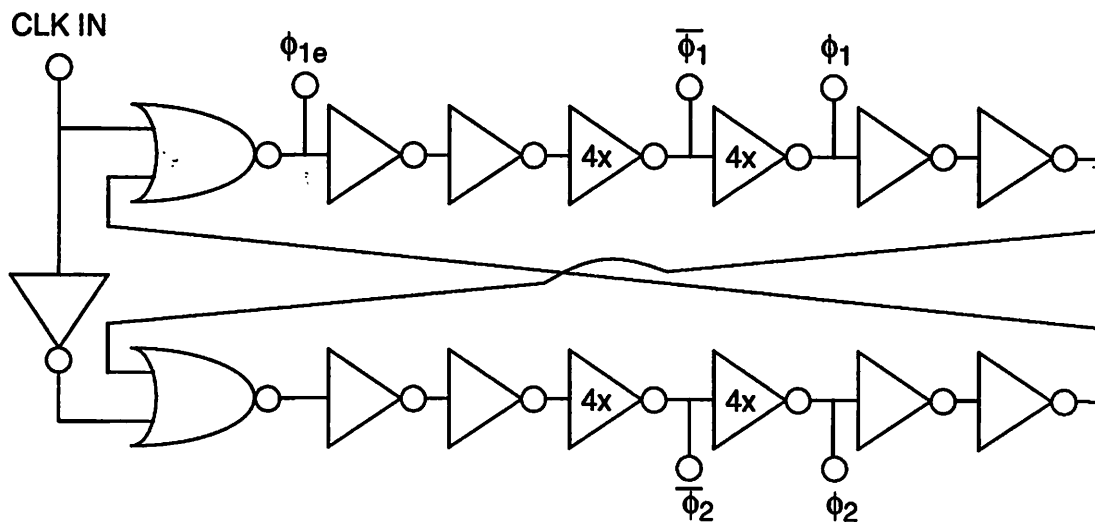
4.12.1 Introduction

The clocking philosophy of this circuit is to use a single two phase non-overlapping clock throughout the chip. The reasons for this are twofold: first, no complex clock generator is necessary to produce all the clock edges and keep the edges in order, and second, during the sensitive settling period, there are no clock transitions to inject noise into the circuit. This simple clock scheme is possible because all events that need to occur in the circuit are synchronous. All reads from the outside (input sampling), and reads from both delay lines occur during ϕ_1 , while all writes to the delay lines and the output buffers occur during ϕ_2 . Cell selection, which occurs after ϕ_2 and before ϕ_1 , adds the only degree of complexity to the clocking scheme. Cell selection should be complete prior to a read attempt and deselction should not occur prior to completion of a write. Clocks on the chip are separated into two groups, the analog clocks which control the OTAs and sampling switches, and the address generation clocks, which govern cell selection. The intent is to use a single external clock to drive both groups — however, to allow experimental adjustments, the two groups are driven separately and an adjustable delay is incorporated on chip to allow the relative phase between the two clock to be adjustable in roughly 1 ns steps.

4.12.2 Analog Clock Generation

The analog read/write section of the circuit requires a two phase non-overlapping clock of both polarities plus some early versions of these clocks. Early clocks are asserted at the same general time as the equivalent “regular” clock, and share the same asserting edge. However, they are deasserted slightly earlier than the regular counterpart such that the early clock is completely deasserted prior to the deassertion of the regular clock. The purpose of this is to facilitate a signal independent charge injection sampling scheme, commonly known as bottom plate sampling and discussed earlier in this chapter.

Assurance that clock edges are non-overlapping is essential to prevent leakage of charge which would corrupt the signal. While gate delays can be used to implement the proper order of clock edges, the result is heavily dependent on the load presented to each clock line. Thus, a closed loop approach to generating a non-overlapping clock is used. Two cross coupled NOR gates in conjunction with an inverter string are used to produce the requisite clock edges. The



All device lengths = 1.2 μm . Unit inverter PMOS width = 114 μm , NMOS width = 62 μm . NOR gate PMOS width = 500 μm , NMOS width = 240 μm .

Figure 4.34 Non-Overlapping Clock Generator

circuit of Figure 4.34 takes as in input a single 50 percent duty cycle clock at $4f_{sc}$, and generates all the clocks required by the analog section of the chip. Use of the cross coupled NOR insures that the two clock phases are non-overlapping — two additional inverters are used in the feed-

back loop to add additional delays to insure non-overlapping edges. Two additional inverters are used at the outputs of the NOR to provide the delay for the early clock.

In addition to providing the early clock, the two inverters shift the time-axis such that the internal clocks are delayed with respect to the external clock input. This is important as it gives time for the cell selection circuitry setup time to complete addressing of the cells. The fact that the early clock is asserted slightly earlier is not an issue, as the early clock is used only for the sample-hold stages which are disjoint from the storage cells.

Simulations prior to fabrication of the circuit showed that adequate margins were obtained with the circuit as shown. As clock loads are relatively constant within the circuit, no further steps were taken to insure clock integrity.

4.12.3 Cell Selection and Address Generation

The ability to access each cell individually within a 2-D storage array requires that some method of row-column decoding be performed within each cell. As stated earlier in this chapter, each cell contains circuitry to connect the cell to the analog signal busses when both its row and column access lines are asserted. The task of producing the signals to properly access the desired cell to perform the delay line function is that of the horizontal and vertical address generators.

The general scheme of address is to choose a row, and access each cell within the row in order, incrementing the row when the last cell is reached. This is repeated until all rows have been accessed, when the cycle repeats. This implies that a relatively fast horizontal address generator is needed, while the vertical generator need only switch at a relatively low frequency. However, because the time available for row changes is the same as that for a horizontal cell change, the transition must be fast (~ 1 to 2 ns), though the frequency of the transitions is quite low (~ 220 KHz).

The relatively large capacitive load on both the horizontal and vertical access lines implies that large clock drivers are required to achieve sharp transitions. Moreover, the row access line is also used by the row multiplexers to determine which row is active, thus the load due to that logic is added to the load presented by the array itself. From layout and extraction, the estimated loads on the horizontal address generator (column select) is about 1pF, while the load on the vertical address generator (row select) is about 2.5 pF.

Both horizontal and vertical address generators are implemented as shift registers, connected in a circular fashion. A single “1” bit is injected at the start with a reset circuit, the remainder being “0”s, with the digital clock moving the “1” from column to column. In a similar fashion, the vertical shift registers operate in the same way, but are clocked with an increment signal generated when the horizontal shift register reaches the end of the row. The design of the two shift registers are slightly different, mostly due to a change in design philosophy made partway through the chip layout. The horizontal shift register uses the classical two phase non-overlapping clock design, while the vertical shift register utilizes a true single clock phase (TSCP) scheme [Yuen, Svensson 289]

4.12.3.1 Horizontal Shift Register

The horizontal shift register consists of 65 stages, of which 64 are body cells. The circuit is fed with a two phase non-overlapping clock of both polarities (4 clock lines), and was chosen over competing designs due to its compactness and robustness. A slight modification is made to allow for resetting the shift register bank. Because it is critical to have identically one active column at a time, a reset circuit is used to force all but one output to zero upon power up. Each body cell forces a zero when the reset line is asserted. The remaining “head” cell forces a one upon assertion of the reset signal. Thus, a single “1” bit is propagated down the shift register.

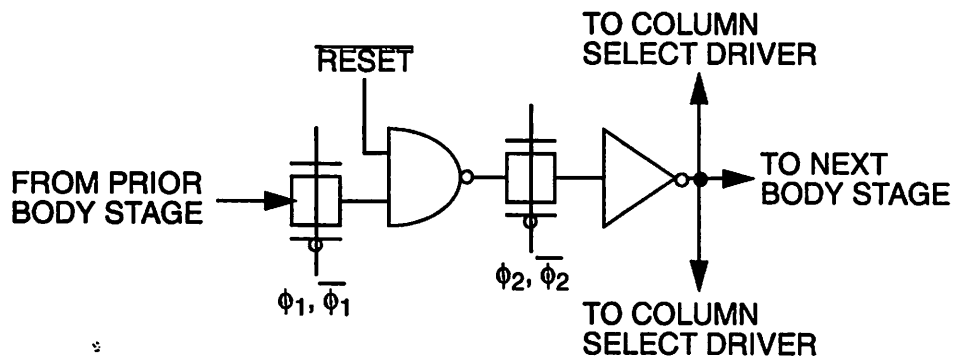


Figure 4.35 Schematic of Horizontal Shift Register Body Cell

Figure 4.35 shows a diagram of the “body” cell, and shows the modification made to the standard two phase clocked shift register for reset purposes. The inverter in the first stage is replaced with a NAND gate with one input tied to the $\overline{\text{RESET}}$ line. Normally high, this line, when low forces the output of the body cell to be a zero, regardless of the prior input clocked in. Thus, it allows setting all outputs of the shift register to a known value for reset purposes.

The head of the shift register is slightly different, in that the NAND gate is replaced with a NOR gate with an inverting input connected to the $\overline{\text{RESET}}$ line (Figure 4.36). Inspection shows

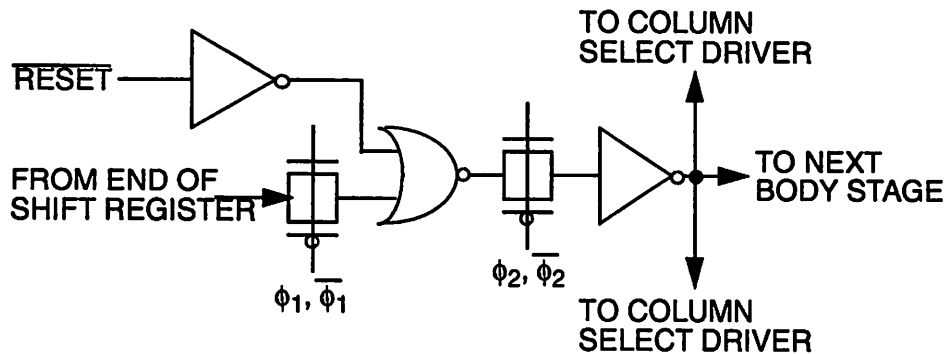


Figure 4.36 Schematic of Horizontal Shift Register Head Cell

that when the reset line is asserted, the output of the cell is forced to be a “1”. Thus, in conjunction with the body cells, this insures that a “1” is placed at the first column and zeros are placed in all other columns.

The outputs of the shift register are used to drive the column select drivers which are used to drive the bus lines in the array. Because the two arrays are clocked synchronously, a single horizontal shift register is used to drive two drivers, one for each storage array. The timing of the driver output is critical to insure that two cells are not activated at once. Thus, a one-shot delay is incorporated in the driver to allow the previous cell to completely deselect.

Earlier in the design process, a method which took the input from the previous driver to detect that the prior selection was deselected was used. While this method is in a sense more robust in that it looks at the actual logic level of the previous column select line instead of depending on gate delays, it was not used for two reasons. First, and most importantly, the circuit added delays to the driver which upset the timing required at the end of the write cycle, and second, it restricted cell accesses to a sequential pattern. Finally, it added complexity at the beginning and end of the shift register.

The final driver circuit (Figure 4.37) utilizes scaled inverters to drive the large parasitic load and uses two long channel inverters to provide the delay. Because the sense of the cell decode logic requires an inverted column select, the additional inversion is provided as part of the driver.

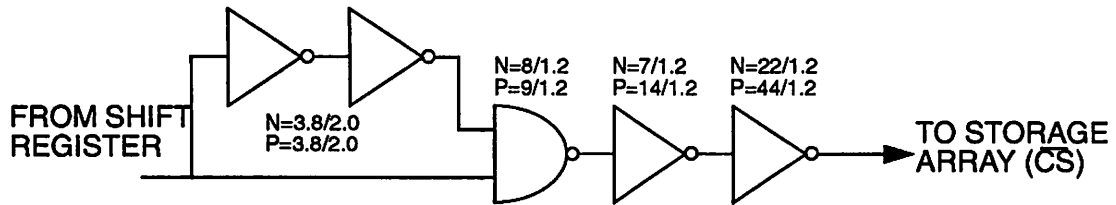


Figure 4.37 Column Select Driver Schematic

The overall latency of the driver provides enough time between the external clock edge and the cell selection waveform to give the analog circuit ample setup time.

4.12.3.2 Vertical Shift Register

The vertical shift register makes use of the TSCP logic style. Because the vertical shift register is clocked each time the “1” bit in the horizontal shift register returns to the head cell, a simple clocking mechanism is preferable to the two phase, non-overlapping scheme used in the horizontal shift register. In return for increased complexity and area in the logic, instead of four waveforms, only a single waveform needs to be generated by the row increment logic.

The body cell of the vertical shift register makes use of the basic TSCP latch, with the first inverter replaced by a TSCP NAND gate as with the horizontal shift register as shown in Figure 4.38. The head cell differs, just as in the horizontal case, in that the NAND gate is replaced by an inverter and a TSCP NOR gate. Thus, upon assertion of the reset line, a “1” bit is placed at the first row, while zeros are forced on all other rows.

As with the horizontal shift registers, the output of each body cell provides input to a buffer which drives the large load presented by the array. Built into the row driver is a lockout mechanism to prevent selection of two rows at the same time. While not used in the horizontal case, it was included here to provide an extra margin of safety against process variation (both device speed and parasitics). Using simple combinatorial logic, it looks at the final output of the previous row, and does not allow the current row to be asserted until the previous row is fully deasserted (Figure 4.39). Although the two storage arrays are clocked synchronously, each array has its own shift register due to layout considerations.

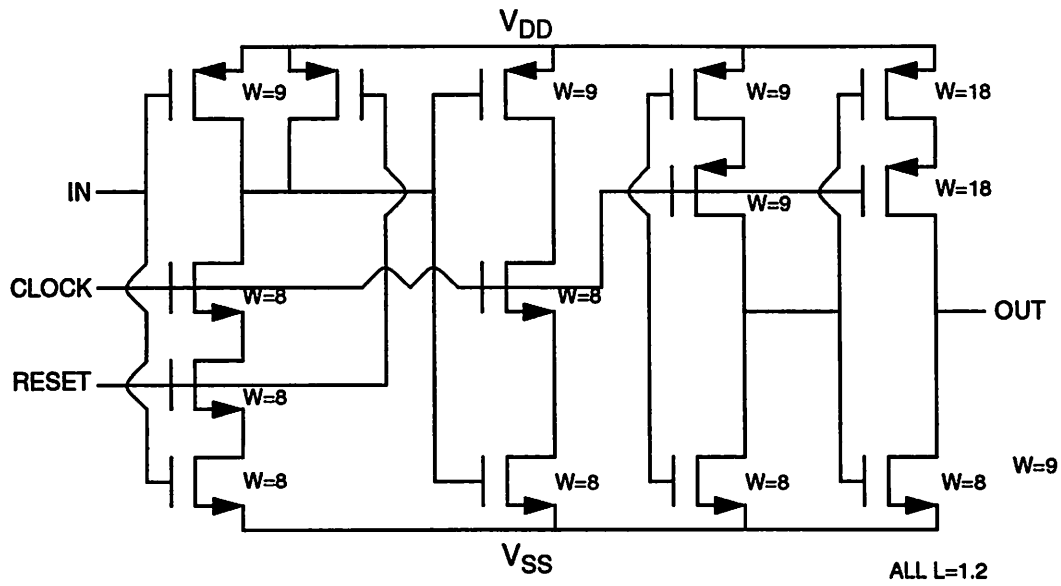


Figure 4.38 Vertical Shift Register Body Cell Schematic

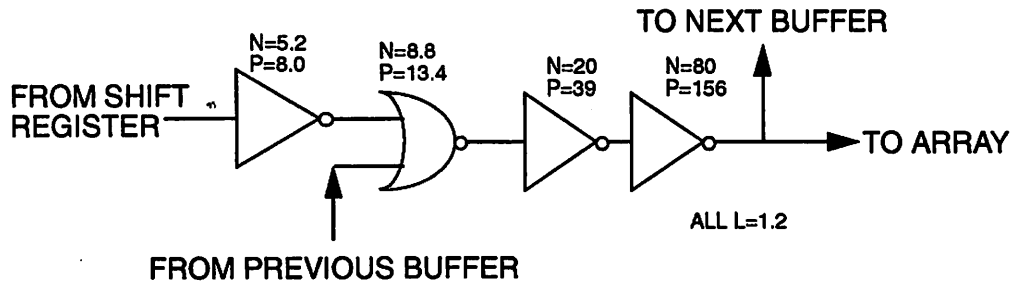


Figure 4.39 Row Buffer Schematic

4.12.3.3 Row Increment Logic

When the logical “1” bit is propagated to the end of the horizontal shift register, it is returned to the head cell to repeat the pattern. When the “1” bit is returned to the head cell, a logic signal is applied to the row increment circuitry to generate the clock for the vertical shift register. This signal then increments the position of the “1” bits in the vertical shift registers. The signal which is returned is logically ANDed with the ϕ_2 clock, to provide a clock signal which will result in a row change at the end of the current write cycle and before the next read cycle. Buffers are provided as the clock loads are quite large due to the large metal line area. The circuitry itself is very simple due to the use of TSCP logic in the vertical shift register. Earlier designs using the two

phase non-overlapping clock scheme required complex logic to generate the four clock signals necessary to effect a bit shift. The row increment logic is shown in Figure 4.40.

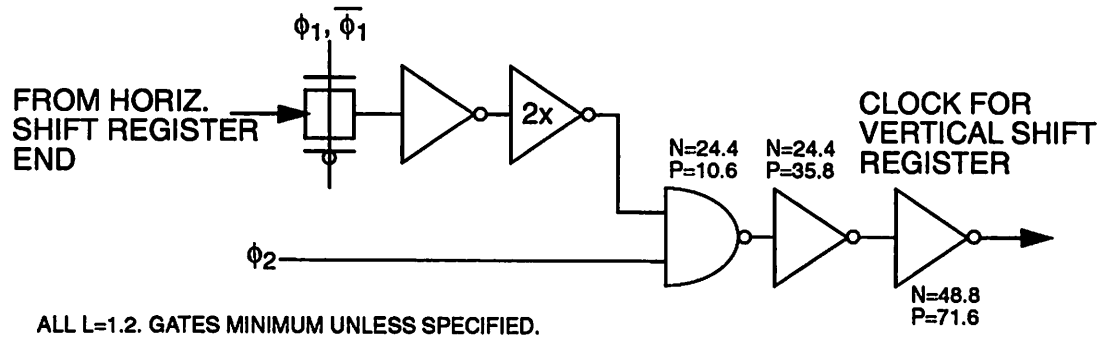


Figure 4.40 Row Increment Logic Schematic

4.12.3.4 Row Reset Lockout Logic

Special care is required to prevent two rows from being selected at once, or more specifically, an access being made to the storage array during the row change. Accessing a cell when more than one row is asserted will result in charge sharing and data loss. As the delay incurred in changing the rows is a function of parasitics and device speed, and is process dependent, a positive lockout mechanism to prevent the above is implemented. While most column select drivers are connected directly to the horizontal shift register outputs as outlined above, the driver to the first column is connected through the lockout logic.

The lockout logic will allow the first column select to be asserted only after the row increment logic has successfully propagated a row change instruction to the vertical shift registers. The "1" bit is propagated in the vertical shift register on the falling edge of the clock. Thus, a chain of inverters provides a delayed version of the vertical shift register clock. This is then ANDed with the column 1 output of the horizontal shift register to provide the drive to the buffer stage. By doing this, the chance of a false access during row transitions is minimized.

A side effect of this system is that the time available for cell reads during the first column is slightly less than that for the other columns. This is a potential problem of fixed pattern noise. While this is highly objectionable, a dual row access is a fatal error and the risk of injecting a small degree of fixed pattern noise was judged to be acceptable to insure proper operation of the circuit.

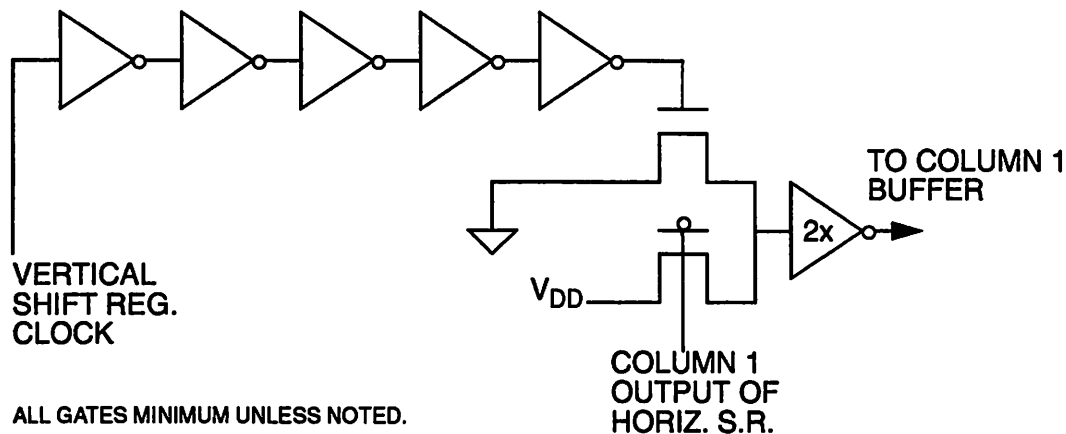
Prototype Comb Filter

Figure 4.41 Row Reset Lockout Logic Schematic

4.12.3.5 Power On Reset

Circuitry to insure that the address generators are in a known state after power up is required to insure that only one cell per storage array is selected at any one time. This is accomplished as described above by asserting the $\overline{\text{RESET}}$ line upon application of power to the circuit. This line must be deasserted at the proper time in the clock cycle to insure proper operation. Circuitry to accomplish this is provided by the power on reset logic.

The upper portion of the circuit (Figure 4.42) is used to produce the reset signal for the horizontal shift register. The circuit produces a reset pulse which is exactly one clock period long timed to coincide with the horizontal shift register independent of the length of the external reset signal. This is required so that the "1" bit which is placed at the head of the shift register is properly propagated on the next clock cycle.

Similarly, a similar mechanism is provided by the lower half of the circuit for the vertical shift register. This takes input from the horizontal reset circuit and the clock for the vertical shift register to produce a reset signal for the vertical shift register which is exactly one row period long. Both circuits reset on their own and do not depend on any gate delays to provide the needed signals.

4.12.3.6 Digital Clock Generation

The purpose of the digital clock generator is to provide the two phase non-overlapping clock for the horizontal shift register. All other clock edges necessary for address generation are

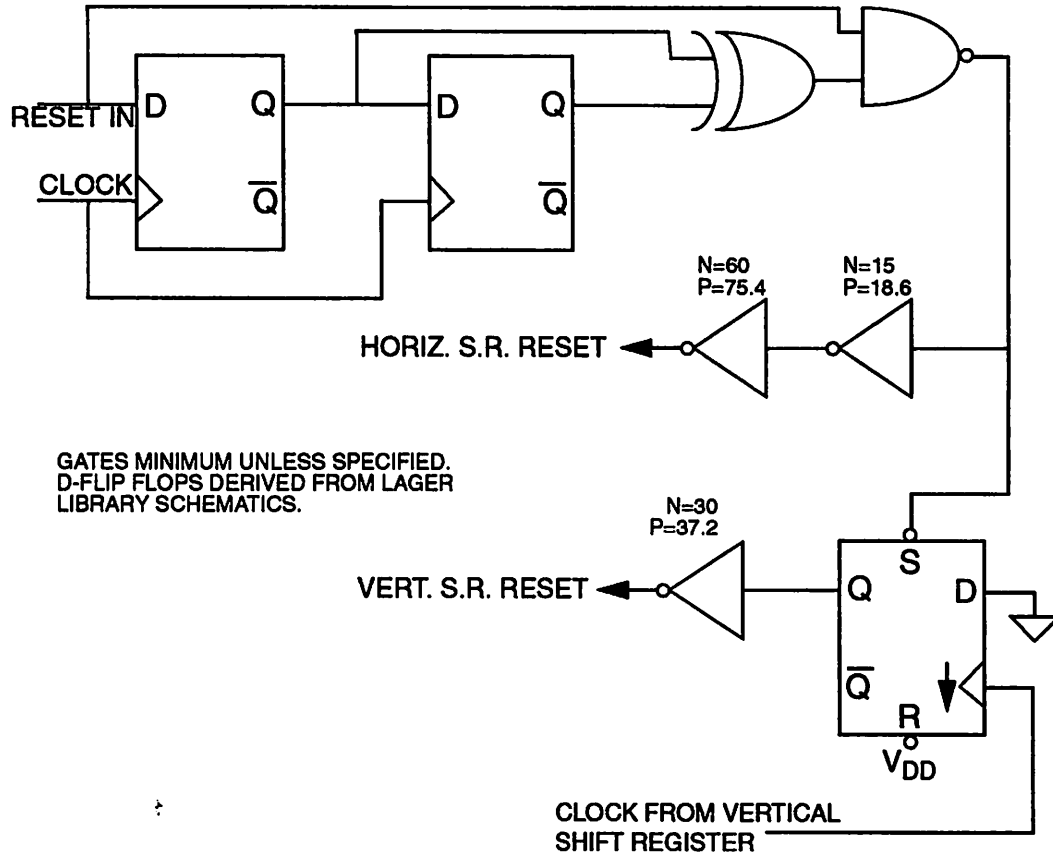


Figure 4.42 Shift Register Reset Circuit Schematic

derived from these two signals as described above. The strategy for accomplishing this is very similar to that used for analog clock generation. Cross coupled NOR gates (Figure 4.43) are used to generate the non-overlapped clocks; however, the overlap time is minimized by eliminating unnecessary inverter delays, as a slight leakage in the shift register does not cause irreparable harm to the circuit. Large clock buffers are used to drive the large loads presented by the horizontal shift register. The large load is due to both the long metal lines and the large number of gates driven by the clock line.

4.13 Circuit Layout

The layout philosophy of the circuit was to minimize the possibility of digital clock noise from upsetting critical analog signals. Bottom plates of capacitors and long analog signal lines are shielded with grounded substrates. The substrate connections of the digital gates are con-

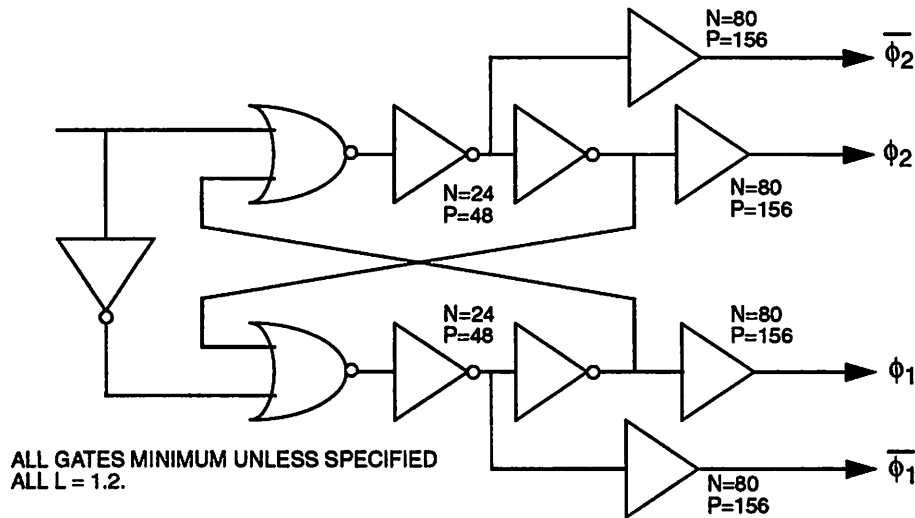


Figure 4.43 Digital Clock Generator Schematic

connected to their own supply lines in an effort to prevent the noisy digital power supplies from injecting noise into the substrate. The floorplan of the circuit follows the pictorial block diagram given in Figure 4.2 very closely, with the digital clocks at the left of the circuit, the storage cells in the middle, and the analog circuitry to the right of the chip. Ample supply connections are provided at all points of the circuit. Separate supplies are provided for digital, digital substrate, analog, output buffer and pad ring. Figure 4.44 shows a high-level layout of the prototype circuit.

ESD protection is afforded through the use of back biased diodes enclosed within a double guard ring. Series resistance was not used to prevent any degradation of frequency response. Pads connected directly to a large area source/drain diffusion do not contain ESD diodes. Digital input pads contain a buffer stage to insure that consistent edge waveforms are applied to all points in the circuit, and to compensate for any signal level mismatch.

Probe pads are attached to critical lines such as clocks, bias voltages and OTA inputs and outputs to allow troubleshooting of the circuit if necessary. In addition, it provided a means to monitor the bias voltages and internal circuit waveforms.

To avoid problems stemming from voltage drops in the power supply lines, very wide metal lines are used for the analog power supply. The computed supply drop is less than 2.5 mV. Because of the large area consumed by the bus lines, large bypass capacitors are fabricated

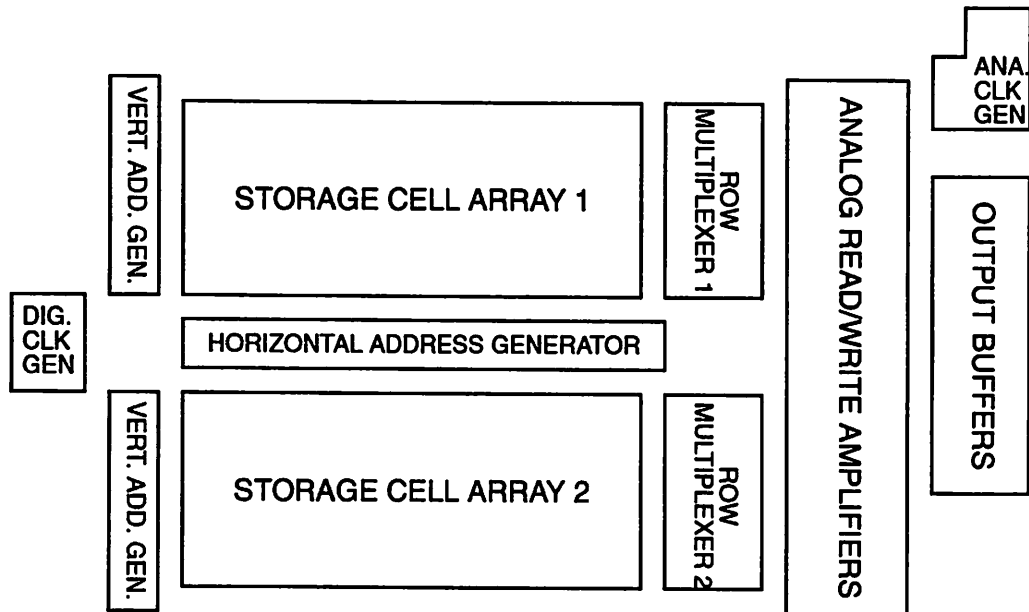
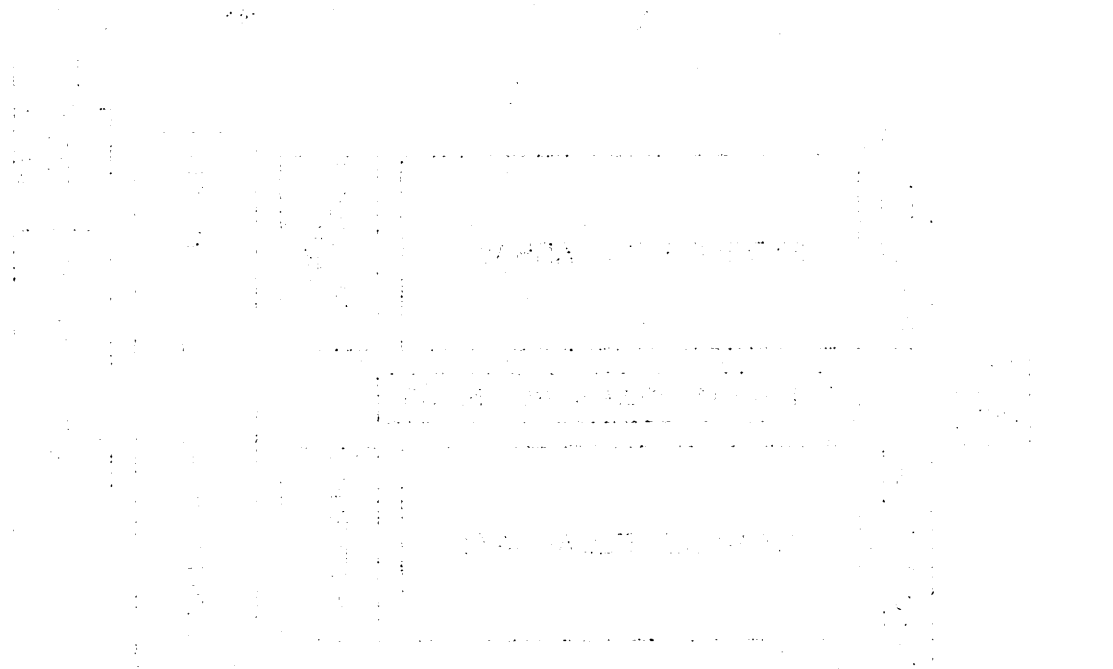


Figure 4.44 Prototype Chip Floorplan

underneath the metal lines. These capacitors are used as supply bypass for the OTAs, and are also connected to most bias lines to provide an extra degree of damping.

4.14 References

- [43] R. Gregorian and G. Temes, *Analog MOS Integrated Circuits for Signal Processing*, Wiley, New York, NY, 1986.
- [44] K. Matsui, et. al., "CMOS Video Filters Using Switched Capacitor 14-MHz Circuits," in *IEEE Journal of Solid-State Circuits*, vol. SC-20, pp. 1096-1101, Dec. 1985.
- [45] F. Volmari and T. Suwald, "Switch Capacitor Circuits for Analog Video Signal Processing," in *ICCE Digest of Technical Papers*, June 1984.
- [46] K. Nishimura and P. R. Gray, "A Monolithic Analog Video Comb Filter in 1.2- μ m CMOS," in *ISSCC Digest of Technical Papers*, Feb. 1993, pp. 30-31.
- [47] D. J. Allstot and W. C. Black, Jr., "Technological Design Consideration for Monolithic MOS Switched-Capacitor Filtering Systems," in *Proc. IEEE*, vol. 71, no. 8, pp. 967-986, Aug. 1983.
- [48] J. B. Shyu, G. C. Temes and F. Krummenacher, "Random Error Effects in Matched MOS Capacitors and Current Sources," in *IEEE J. Solid-State Circuits*, vol. SC-19, no. 6, pp. 948-955, Dec. 1984.
- [49] M. J. M. Pelgrom, et. al., "Matching Properties of MOS Transistors," in *IEEE J. Solid-State*



The diagram illustrates the internal structure of the prototype comb filter. It consists of a main rectangular frame containing several smaller rectangular components. These components are arranged in a regular grid, suggesting a multi-stage or multi-tap filter architecture. The labels within the diagram are extremely faint and illegible, but they likely represent various filter parameters or stages. The overall layout is symmetrical and organized, typical of a technical schematic for a digital filter.

CHAPTER 5

Experimental Results

5.1 Introduction

The prototype circuit described in the previous chapter has been fabricated using the 1.2mm n-well Orbit Foresight™ [58] process. Fifteen devices were fabricated, of which six were tested to produce the results reported here. A small layout error in the Metal 2 layer prevented complete operation of the circuit as originally designed. Of the 15 devices fabricated, three failed DC test, with excessive supply current. The layout error prevented operation of the read OTAs for both storage arrays. Of the remaining 12 chips, six were chosen on the basis of good experimental results found by monitoring the waveform at the output of the first OTA via probe pads. These six were then “repaired” by use of a focused ion beam to correct the error. Subsequent to this step, all six circuits functioned fully with no further adjustments.

This chapter will describe the test circuit used, and will report the measured results obtained by testing these six circuits both electrically and by visual inspection of the resultant video image.

5.2 Test Fixture

The prototype circuit requires a fairly simple interface. Required input signals are a differential video input signal, a $4f_{sc}$ clock, a reference bias current, an optional external OCM reference voltage, and a 5 volt supply. Two 3-bit digital control signals are used to adjust the time offset between the digital and analog clocks for experimentation. The circuit produces two dif-

ferential analog voltages, one representing the comb filter output, and one representing the output of the first line delay (delayed composite).

The test fixture, in addition to supplying the above signals, must perform the single-ended to differential conversion at the input to the chip, the inverse operation at the output, and output waveform reconstruction. Moreover, a means to derive the luminance signal from the delayed composite and chrominance signal is implemented. Finally, the test fixture provides several independent, regulated and filtered power supplies for use within the test fixture and to power the chip itself.

The test fixture assumes that a 75Ω single-ended $1V_{p-p}$ composite video will be applied to the input. The input is then scaled, and a DC restoration performed to clamp the sync tip level to a known value. A single-ended to differential conversion is then performed using a discrete differential amplifier, which also sets the common mode voltage to that appropriate for the chip under test.

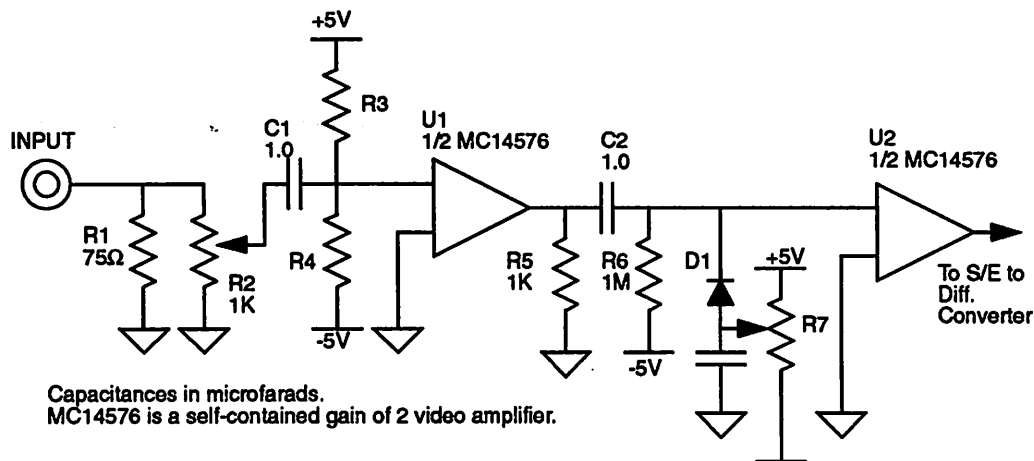


Figure 5.1 Test Fixture Input Circuit

The circuit for the input side of the test fixture is shown in Figure 5.1. A 75Ω resistor is used for termination. R2 sets the amplitude of the signal to be used. U1 acts as a buffer to drive the DC clamp formed by C2, R6, D1 and R7. The clamp operates by forcing the most negative voltage of the input waveform to be a diode drop from the voltage formed by R7. Because the sync tip is the most negative voltage of a video signal, and is consistent, this circuit performs the function well and has a wide bandwidth.

After the signal is DC clamped, it is then buffered by U2 to drive the single-ended (S/E) to differential converter (Figure 5.2). The circuit is comprised of a set of matched transistors forming a current source and a differential amplifier. An adjustment for offset (R13) and output common mode (R8) is available for trimming. The current source is an integrated PTAT device (LM334). The outputs are taken from the collectors of the differential pair and are fed directly to

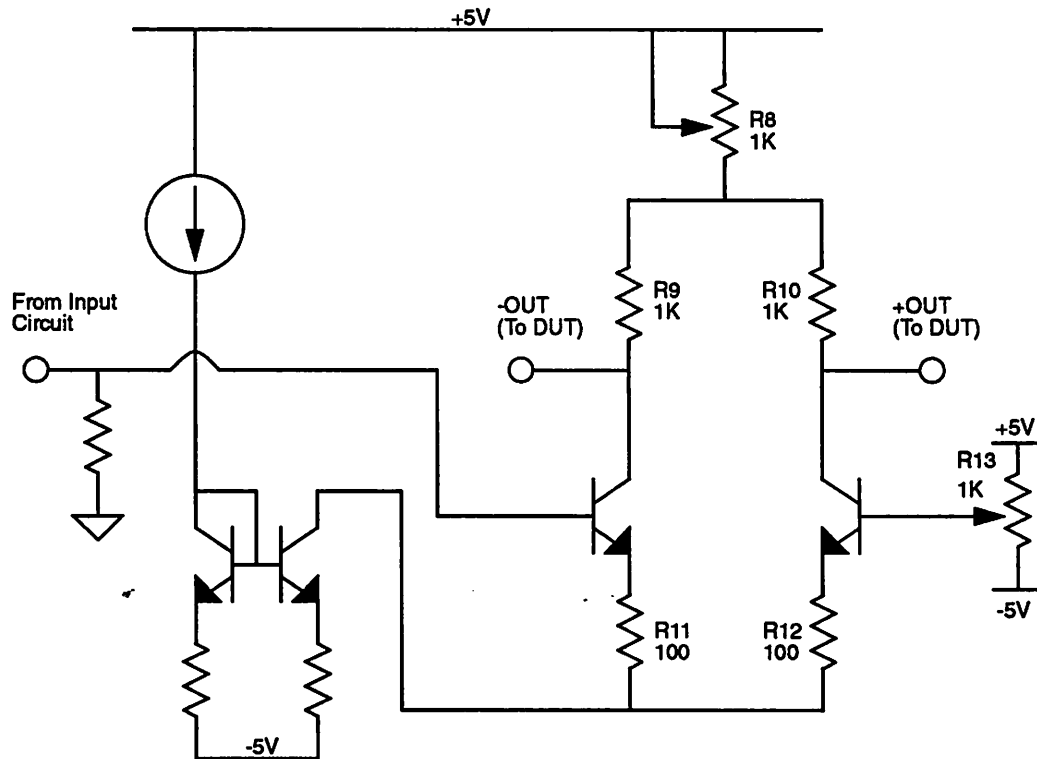


Figure 5.2 Test Fixture Single-Ended to Differential Converter

the chip under test.

The two outputs of the chip are differential, and need to be converted back to a single-ended signal for display on a video monitor. In addition, due to the sampled data nature of the signal, a reconstruction filter is required to eliminate higher-frequency components. The output then needs to be buffered to drive a 75Ω load.

Filtering of the output signal is accomplished using a pre-manufactured filter unit (TOKO 2722) which has a nominal $1K\Omega$ impedance. Because the source follower amplifier in the chip is large, run with large bias currents, and enclosed in a negative feedback loop, source termination of the filter is performed with a single $2K\Omega$ resistor across the differential outputs. The ground

terminals of the filter unit are tied together to form a virtual ground, and output termination is accomplished with another $2K\Omega$ resistor.

Conversion of a differential signal back to a S/E signal is accomplished using a current summing technique using current mirrors (Figure 5.3). Here, each signal is fed into a base of an emitter

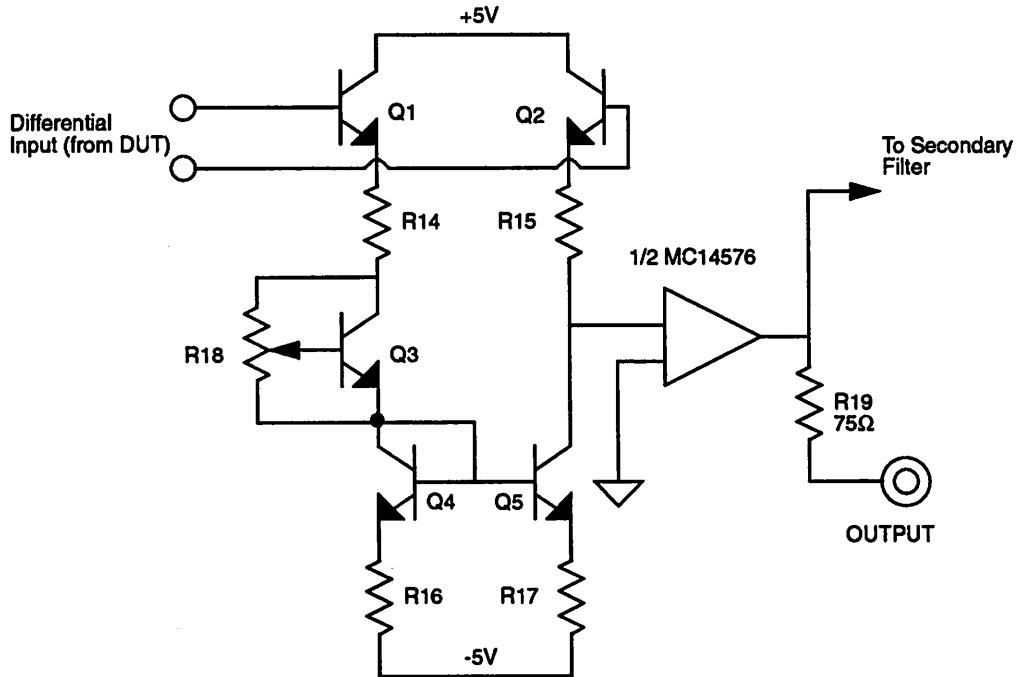


Figure 5.3 Test Fixture Differential to Single-Ended Converter Circuit

follower (Q1, Q2). A current proportional to the voltage drop across R14 and R15 flows in each leg of the circuit. Q4 and Q5 form a current mirror; the resultant two currents are summed at the node connected to the input of the final buffer amplifier. Q3 serves as a V_{BE} multiplier and adjusts the quiescent voltage of at the bottom of R14, which must match the voltage at the input of the amplifier (nominally ground).

Two of these differential to S/E circuits are contained within the test fixture, one for the delayed composite output, one for the comb filter (chrominance) output. The secondary filters are present to experimentally determine the effect of bandpass filtering the chrominance before subtracting from the composite signal. Because the bandpass filter (TOKO 6276) adds group delay, an equivalent delay is added as the secondary filter in the composite channel.

The single-ended outputs are then differenced to yield the luminance signal. Two difference circuits are included, one to operate on the direct (non-secondary filtered) outputs, while the other works with the bandpassed chroma. In most cases, the difference in performance between the two is negligible — some variation was noted in the comb notch depth measurement.

The remainder of the test fixture consists of power supply regulators, bias current generators and digital clock circuits which are of little interest. The entire test fixture is fabricated on a 4 layer printed circuit board with integral power and ground planes for low-noise.

5.3 Measured Performance

Evaluation of the chip is divided into two parts: (1) electrical measurements with quantitative results, and (2) subjective visual measurements based on the resulting image on a high quality video monitor.

The electrical measurements of interest are:

- Comb Notch Depth @ 3.58 MHz
- Composite Channel Frequency Response
- Power Consumption
- Dynamic Range (Random Noise)
- Fixed Pattern Noise (220.3 KHz)
- Full Scale Input Voltage
- Linearity
- Differential Gain and Phase

Measured results for the above parameters are summarized in Table I. These results are representative of chips taken from three separate wafers from the same processing run. The tests were conducted at room temperature and nominal supply voltage (5.0 VDC). No attempt was made to regulate the actual junction temperature of the chip. Based on self heating, it is estimated that the actual operating temperature of the circuit is in the vicinity of 305K. The measurements in the above table will be discussed individually in detail:

Table I: Summary of Measured Performance ($T=25^{\circ}\text{C}$, $V_{DD}=5\text{V}$)

Parameter	Value
Comb Notch Depth @ 3.58 Mhz	> 28 dB
Composite Channel Freq. Resp.	± 1.1 dB @ 4.2 MHz
Power Consumption	170 mW
Active Area	11.7 mm²
Dynamic Range (random noise)	> 51 dB
Fixed Pattern Noise (220.3 Khz)	0.2% FS (-55 dB)
Full Scale Input Voltage	2.6V_(p-p)
Linearity	better than 0.3%
Differential Gain	less than 1%
Technology	1.2-μm Double Poly CMOS

5.3.1 Comb Notch Depth

This measurement reflects the effectiveness of the filter itself. The measurement is taken in two ways. The first looks at the absolute depth of the comb notch by clocking the circuit at exactly $4f_{sc}$ and using a sine wave near the subcarrier frequency as input. The input frequency is then varied to determine the maximum attenuation. This provides a very good value, approximately 35.5 to 36.0 dB. The second method uses the delayed composite channel and the chrominance output to form the luminance output. The maximum attenuation of the chrominance signal in the resultant luminance signal is then measured. This value, which more closely represents a typical application of the circuit results in a minimum notch depth of 28 dB.

Both values are slightly lower than expected, and can be attributed to several causes. First, and foremost, the technique discussed in the previous chapter to skew the zero locations of the filter was not done, mostly due to oversight. As a result, the non-unity gain through the delay lines caused the filter zeros to be in non-optimal locations resulting in reduced filter efficiency. The second cause is related, and is due to the mismatch in the capacitors used to weight the three signals being summed inside the chip. As discussed in Chapter 4, the result of mismatch is

reduced filtering action. Finally, the figure derived using the second method described above is very sensitive to path matching between the luminance and chrominance signals. Because the circuit depends on cancellation of two signals to remove the chrominance component, slight variations in amplitude or delay can adversely affect this measurement. This measurement was taken with both the bandpass signal path and the direct signal path. Due to the group delay variation between the bandpass filter and the compensating delay, the measurement was 2 dB worse with the bandpass method, registering 26.3 dB. Using the direct path, the measurement was the reported 28.2 dB.

5.3.2 Composite Channel Frequency Response

This measurement reflects any degradation in the signal as a result of passing through the delay line. Because the delayed composite channel should ideally mimic the input except for a delay of exactly 1H, a comparison of the frequency response between this output and the input to the chip is useful. The channel exhibited a typical high frequency droop, probably as a result of parasitic capacitances forming a unwanted low pass filter within the delay line. The measured value of droop was manually compensated for $\sin x/x$, as the output of the chip represented a zero order hold. After compensating for the droop induced by the staircase output, the measured result showed that the maximum attenuation at the NTSC bandedge (4.2 MHz) is 1.1 dB for all chips tested.

5.3.3 Power Consumption

The DC power consumption of the chip was measured by summing the current drawn by the analog circuits and the internal digital circuits and multiplying the by the nominal supply voltage of 5.0V. Power dissipated by the pad drivers and the output buffers were not included in the computation, as they do not reflect the power dissipated by the circuit itself. The measured value, 170 mW, also includes the power dissipated in the analog bias circuit — if this circuit is incorporated within a larger circuit, this bias could be shared, and hence, the power attributable to the circuit would be lower, in the vicinity of 160 mW.

A significant portion of the power is due to the digital portion, drawn mainly as charging currents for the relatively large capacitive loads on the column select lines. The digital drivers used in the circuit were designed very conservatively to obtain very fast transitions. Optimization of these circuits could result in some savings of power while maintaining proper operation.

5.3.4 Dynamic Range (Random Noise)

This measurement reflects the noise present in the signal exclusive of any periodic signals. The measurement is conducted by applying a flat field of 50IRE (half scale) as input and measuring the RMS value of the noise present at the output of the signal. The measurement was conducted using both a spectrum analyzer and an oscilloscope, and yielded a value of greater than 51 dB below full scale using both methods. This measurement is consistent with the 8 to 9 “equivalent” bit resolution expected for this circuit.

5.3.5 Fixed Pattern Noise (220.3 KHz)

This measurement, made on a spectrum analyzer, determines the magnitude of spurious noise with a periodic nature. Although a serial data path was maintained as much as possible, the use of a two dimensional array and the accompanying switching of rows unavoidably injects some noise into the signal. Noise due to the changing of rows will appear at 1/14th the line rate, or 220.3 KHz.

The output spectra with a 50 IRE flat field input showed components at multiples of 220.3 KHz as expected. The square root of the sum of the squares of the first through seventh harmonic was down 55 dB (less than 0.2%) of full scale. No other significant periodic noise was observed.

Fixed pattern noise was also measured in a slightly different manner, using an oscilloscope. The signal was AC coupled and a flat field applied. The oscilloscope was adjusted to magnify the section of the trace representing the glitch due to fixed pattern noise. The amplitude of this noise is then compared to the full scale output voltage. The magnitude of the glitch is observed to be approximately 4 mV, which is 56 dB down from full scale. Thus, this measurement agrees well with that obtained using the spectrum analyzer.

5.3.6 Full Scale Input and Linearity

These measurements are complementary, as the linearity of the circuit is a function of the maximum signal applied. The circuit linearity was measured using both a 100 KHz and 1 MHz sine wave at the input with the circuit clocked at $4 f_{sc}$. A high-Q bandpass filter was applied to the output of the signal generator to remove spurious components which would affect the measurements. The output was observed on a spectrum analyzer, and the square root of the sum of the squares of the first through fifth harmonics noted. In order to minimize any contribution of

the external amplifiers or filters, measurements were taken at the input and the output of the chip using a differential probe. The final reported value represents the distortion added by the chip itself. Measurements at 1 MHz are slightly skewed as $\sin x/x$ and other factors affected the higher harmonics. More meaningful results were obtained using a 100 KHz signal. Incidentally, this represents a more realistic measurement, as most video signals do not contain large amplitude high frequency signals. The distortion measurements were a sensitive function of the input voltage when increased beyond $2.6 V_{p-p}$. A distortion measurement of 50 dB below fundamental was measured for a input voltage of $2.6 V_{p-p}$, increasing to 42 dB at $2.7 V_{p-p}$.

This measurement also exhibited sensitivity to the OCM voltage within the circuit. This is expected as the OCM voltage indirectly determines the output swing available at the OTAs within the circuit. These measurements were obtained using an external reference voltage trimmed to maximize the input range. The internally generated voltage results in an input range about 200 mV less than the values above.

5.3.7 Differential Gain and Phase

This measurement is a type of nonlinearity test [59]. Differential gain is the change in the gain of the circuit at 3.58 MHz as a function of the average DC level, while differential phase is the change in the delay of the circuit at 3.58 MHz as a function of average DC level reported in degrees of the 3.58 MHz signal. This type of distortion is usually a result of a non-linear capacitance such as a junction capacitance interacting with the circuit impedance to form a low-pass network. As the value of the capacitance is a function of the DC value, the pole of the network changes, and hence the gain and phase of the 3.58 MHz signal is altered.

Differential gain is measured using a signal known as a modulated ramp, which consists of 40 IRE (286 mV) of a fixed phase chrominance signal superimposed on a slow triangle wave whose value rises from zero to full scale in one line period. Differential gain is the difference in the chrominance amplitude when observed at a zero and full scale DC voltage.

Measurements taken on the prototype chip show that the differential gain is less than 1 percent, limited by measurement techniques. Similarly, the differential phase is known to be less than 1.5 degrees, but again is limited by the measurement technique. More precise measurements using a vectorscope were not available at the time of this writing.

Both the linearity and the differential gain/phase measurements are slightly suspect due to the presence of the output buffer amplifiers on chip. Although enclosed within a negative feedback loop, the gain of the forward amplifier is low, around 30 V/V. Thus, the amount of distortion cancellation available is limited. The source follower in the amplifier has the source tied to the well in an effort to reduce distortion due to body effect. Unfortunately, this introduces the well-substrate capacitance into the circuit. This capacitance is a fairly large valued, non-linear capacitance which will interact with the output resistance of the amplifier to form a DC voltage dependent pole, which is the exact cause of differential gain and phase. The magnitude of error contribution attributable solely to the buffer amplifiers is non measurable without destroying the test device, as access to the inputs of the buffer stage is limited.

5.3.8 Active Area

Although not an electrical parameter, it is useful to note the active area of the circuit and analyze the relative area consumption of the various parts of the circuit. The total active area, excluding pads and output buffers is 11.7 mm^2 . This area is divided among four main areas: (1) the storage arrays, (2) address generators, (3) analog read/write amplifiers and bias circuitry, and (4) clock generation and power-on reset circuitry.

5.3.8.1 Storage Array Area

The storage array consists of 910 storage cells each $40 \mu\text{m} \times 77 \mu\text{m}$ in area. Row multiplexers are added to each row at the right edge of the array. Tiling of each storage cell allows packing of the cells without any extra area for routing. The row multiplexers include all the circuitry and routing necessary for interface to the amplifiers, including the auxiliary capacitor circuit. The resultant area of each storage array is 3.425 mm^2 .

5.3.8.2 Address Generators

The horizontal and vertical address generators occupy the area surrounding the storage arrays and feed cell access signals from two sides of the array. As with the storage array, each cell of the address generator is tiled to match the pitch of the storage array. The combined areas of the address generators is 2.441 mm^2 .

5.3.8.3 Analog Read/Write Amplifiers

Immediately to the right of the storage array are the row of amplifiers which comprise the read/write circuitry to interface to the storage array. The five OTAs, bias generator, OCM generator and S/H stages are all arranged in one long block surrounded by the two wide supply buses. The area consumed by this portion of the circuit is 2.074 mm^2 .

5.3.8.4 Clock Generator and Reset Circuitry

The two clock generators and reset circuitry comprise the remainder of the active area. Each clock generator has a 3-bit adjustable delay for experimental purposes whose area is not included in this computation. Certain other miscellaneous circuitry essential to the operation of the circuit, but too small to be categorized individually are included in this area figure of 0.359 mm^2 .

Figure Figure 5.4 shows a photomicrograph of the circuit.

5.4 Functional Performance

Testing to insure that the circuit performs the *function* intended, in addition to meeting the electrical specifications, is carried out by applying an actual video signal to the circuit and noting the output. The signal used for this test is known as EIA colorbars, and consists of vertical bars of the primary colors plus all combinations of the three primary colors. The bottom portion of the signal contains white, black and the subcarrier signal. This signal is appropriate for testing the effectiveness of the Y/C circuit as it contains a staircase luminance value, sharp transitions and fully saturated colors.

Application of this signal to the circuit should yield a delayed composite signal at the one output, and a pure chrominance signal at the other. Figure 5.5 shows an oscilloscope photograph of the input signal on the top trace, and the output chrominance signal on the lower trace (vertical scale arbitrary). As can be seen, the luminance component represented by the staircase DC level is completely removed from the lower trace. Figure 5.6 shows a similar photograph, with the derived luminance in the middle trace. Note the cancellation of the chrominance signal in the luminance trace.

Further verification of the performance of the prototype circuit is carried out using real images rather than laboratory test signals, and are discussed in the next section.

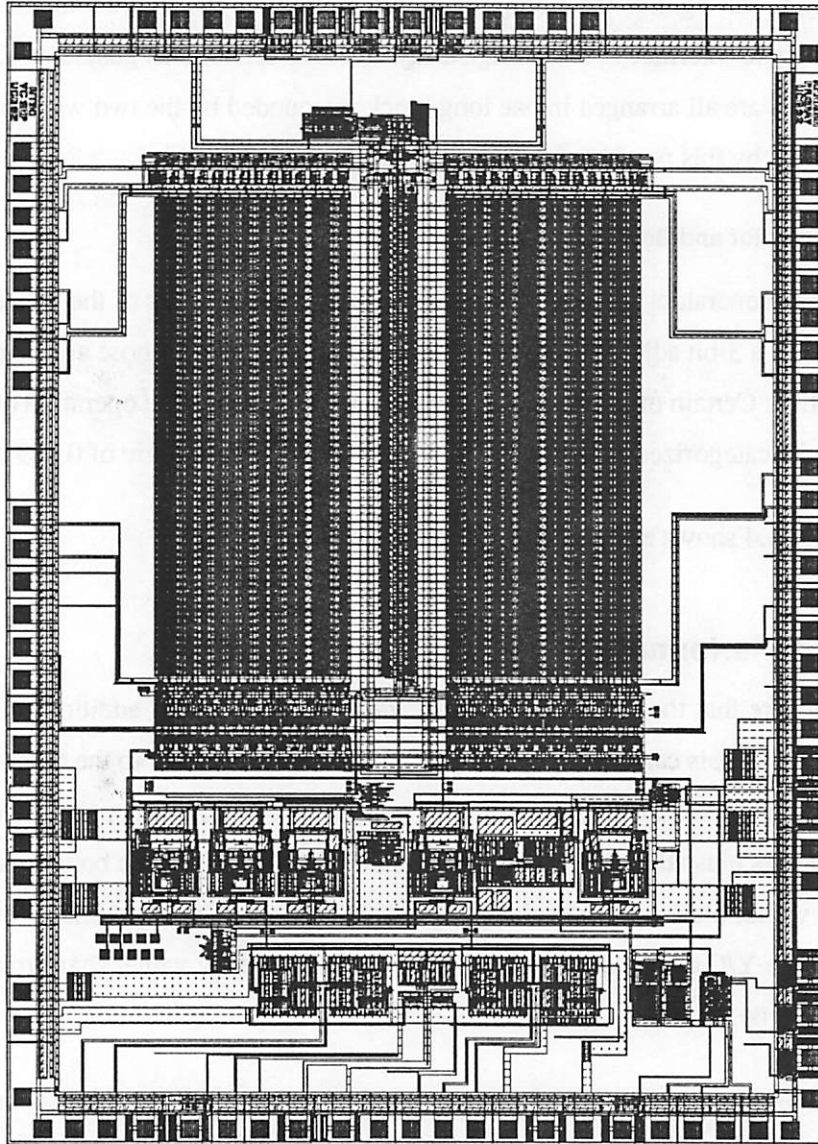


Figure 5.4 Layout Plot of Prototype Chip

5.5 Visual Inspection

Selection of appropriate images to determine the effectiveness of the circuit falls into two categories: (1) images that show gross impairment without comb filtering, and (2) images that will accentuate faults in the circuit, especially fixed pattern noise. The first category is fulfilled by images that contain a large amount of high frequency luminance. Most ordinary images contain only small areas of high frequency luminance. The impairments are visible, but are not of

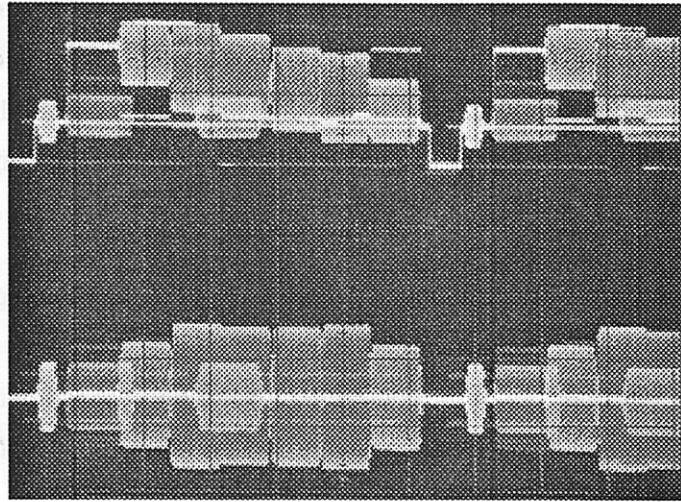


Figure 5.5 Input (EIA Colorbars) (top) and Chrominance Output (bottom)

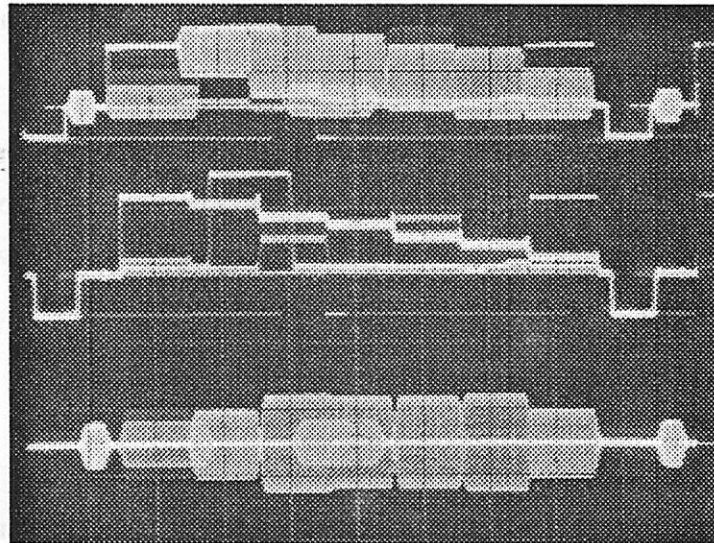


Figure 5.6 Input (top), Derived Luminance (middle), Output Chrominance (bottom)

enough area to accurately judge the performance of the comb filter. As such, the multiburst test signal is used for this test. Multiburst consists of an all luminance signal with vertical bars consisting of various frequencies, one of which is subcarrier frequency. It is important to point out that a luminance signal at subcarrier frequency does NOT make it a chrominance signal, for although it has the right instantaneous frequency, it lacks the proper line to line phase inversion between lines required of chrominance signals. NTSC decoders without comb filters however,

make this exact error, and process the high frequency luminance signal as chrominance. This results in a large area of rainbow-like colors, which are a totally fictitious artifact due to the failure of proper Y/C separation. Application of a properly functioning comb filter, on the other hand, should treat the multiburst signal as pure luminance and route nothing to the chroma demodulator.

Application of a standard NTSC multiburst pattern as input to the prototype chip results in complete removal of cross color components when viewed on a monitor. This provides subjective verification of the effectiveness of the comb filter. [Note: Because reproduction of the test images in black and white would provide no meaning, they are omitted in this dissertation.]

The second category of images, those to find visual impairments, are best performed with a real image, as opposed to an instrument signal. As the artifact most likely to contaminate the video output of this circuit is fixed pattern noise, an image with large areas of medium luminance colors with little background pattern is best. Fixed pattern artifacts are easily seen in such images as there is nothing to mask the impairment.

A test image meeting the above criteria was processed by the prototype chip. The results, when viewed on a monitor, showed no impairments due to fixed pattern noise, nor was there any evidence of distortion due to differential gain or phase. The overall noise floor was undisturbed, indicating that any additive noise is below the noise in the originating source.

5.6 Additional Tests

Two additional tests were performed, mainly to determine the operating margins of the devices. First, the circuit was operated at reduced supply voltages to determine the supply margin. Second, the chip was operated with increasing clock rates until proper operation ceased.

5.6.1 Operation with Reduced Supply Voltage

The device under test was subjected to reduced supply voltages (both analog and digital) and performance measured. As expected, the maximum input signal, and hence the magnitude of signals within the circuit, was reduced as the supply was reduced due to saturation of the OTAs. The circuit operated with reduced signal swing at room temperature with a supply voltage as low as 3.9 volts. Operation at lower supply voltages is probably possible with optimization of the OCM

voltage. However, if low voltage operation is the design goal, a redesign of the amplifier is probably in order. Nevertheless, successful operation at reduced supply voltages implies that a sufficient margin exists in the current design.

5.6.2 Operation with Faster Clocks

This test complements the above test; the frequency of the reference clock is increased while operating at the rated supply voltage of 5.0 VDC. Of particular interest is operation at 17.734 MHz, which is four times the subcarrier for PAL video signals. At this frequency, the circuit operated as designed, although visual tests were not possible due to lack of PAL test equipment. However, a proper transfer function was verified via oscilloscope tests. Operation at increased clock frequencies beyond 19 MHz results in reduced gain through the delay lines, presumably due to incomplete settling by the OTAs. No attempts were made to determine the ultimate limit of operation as non-unity gain through the delay lines implies that a proper comb filter cannot be constructed.

5.7 Future Modifications to Prototype Circuit

As only one run of silicon was fabricated of the prototype circuit, no opportunities existed to make changes to the design based on measured results. Some areas warranting a closer look for improved performance are:

- Use of Class AB/B amplifiers with faster technologies to improve power efficiency of the analog sections.
- Investigate the clocking scheme of the analog section so that the same amplifier can be used for both writes and reads. This would reduce the amplifier count by 40 percent.
- Implement changes to the output summer/scaler section to compensate for non-unity gain in the delay lines.
- Implement a ring of dummy cells around the perimeter of the storage arrays to insure better matching of cells.
- Change horizontal address generator to a single clock phase circuit, thereby simplifying the digital clock generator.
- Redesign output buffers (if needed) to improve distortion and linearity characteristics.
- Investigate low-voltage (3.3 V) operation of the circuit.
- Determine the effects of using single-ended circuitry.

In addition to the above, the obvious extensions to higher resolution and faster access times are of interest. However, as shown in this dissertation, it is unlikely that resolutions in excess of 11 to 12 bits will be practical with analog techniques. Continued pressure from digital designs using aggressively scaled technologies will continue to pressure the use of analog signal processing techniques in the near future.

5.8 Conclusion

The measured electrical parameters, function tests and visual inspection of the resultant video image all conclude that the prototype circuit performs as expected. The chrominance component of the signal is successfully separated from the incoming composite signal to avoid cross color and cross luma effects that plague traditional NTSC decoders. Power and area figures support the hypothesis that a properly designed analog circuit can perform this task in less power and area than an equivalent digital design [60]. The problems of fixed pattern noise which have prevented analog based solutions from working properly have been mitigated with the serial data path approach employed in this circuit. Noise performance of this circuit shows that video rate switched capacitor circuits are viable at the 8 to 9 bit range, and that this resolution meets the requirements for NTSC processing [61]. Finally, operation of switched-capacitor circuits with large parasitics has been demonstrated.

5.9 References

- [58] Orbit Semiconductor, *Foresight Users Manual*, 1992.
- [59] M. O. Felix, "Differential Gain and Phase," in *J. Society of Motion Picture and Television Engineers*, pp. 76-79, Feb. 1976.
- [60] Motorola MC141620 Data Sheet, *CMOS Application-Specific Standard ICs Databook*, 1991.
- [61] A. A. Goldberg, "PCM-Encoded NTSC Color Television Subjective Tests," in *J. Society of Motion Picture and Television Engineers*, pp. 649-654, Aug. 1973.

CHAPTER 6

Conclusions

6.1 Summary of Research Results

This research shows that analog signal processing is both viable and potentially more efficient (using power and area metrics) than equivalent digital methods for a class of signal processing tasks. In particular, there are four major results from this work:

- 1) An analysis of the relative performance of analog and digital signal processing technologies has been undertaken. Based on this, it has been concluded that considering only the power and area of the actual processing circuitry, signal processing tasks with modest (< 60 db) dynamic range requirements are more efficiently undertaken with analog processing compared to equivalent digital processing techniques.
- 2) A differential "Analog-RAM" circuit topology suitable for analog signal processing tasks involving transversal filters with long time delays has been developed. This circuit demonstrates a dynamic range in excess of 50 dB and operates with an access time of 25 ns.
- 3) A prototype circuit which utilizes the Analog-RAM has been conceived, designed and fabricated. Operating as a comb filter for NTSC video signals, the resultant circuit consumes less power and area than competing all-digital designs using a 1.2- μm CMOS technology with a 5 volt supply. This circuit also demonstrates circuit techniques useful

for the design of switched-capacitor circuits operating in the presence of large parasitic capacitances.

- 4) The performance of the Analog-RAM based NTSC comb filter has been experimentally evaluated. The resulting performance exceeded the requirements of consumer television receivers in all respects.

6.2 Projected Performance in Scaled Technologies

Analog circuitry is often at a disadvantage in scaled technologies due to limitations from thermal noise in the sampling process. Thus, it is important to project the performance of the prototype circuit described in this thesis with scaled technologies.

First and foremost is the effect of reducing the size of the sampling and storage capacitors. Although the circuit uses capacitors in the 300 fF range, such sizes are not required to maintain an adequate noise margin from thermal noise. If a projection to a low-voltage technology is made, the signal swing within the circuit may be reduced significantly. However, even with a 0.5 volt (RMS) swing, to maintain a "9-bit" dynamic range, a capacitance of less than 10 fF is required. This implies operation with a 0.2- μm gate length technology.

The limitation imposed on capacitor size within the circuit is that of the ratio of storage capacitance to parasitic capacitance. Thus, as the technology is scaled, the parasitics affecting the circuit should also scale. Thus, it is anticipated the performance limitation of this particular circuit with scaled technologies is that of parasitics, not thermal noise.

The effects of severe short-channel effects in the MOS transistors is less clearly defined. It is known that the gain available in the amplifiers falls off as the channel length is shortened; thus, additional measures to maintain the minimum required amplifier gain may need to be taken. The reduced supply voltages available with scaled technologies will affect the analog switches in the circuit — gate drive boosting circuits such as charge pumps may be needed to insure adequate conductance in the switches. However, it is anticipated that this circuit could be redesigned in a 0.8- μm technology with no major changes — use of this faster technology may mitigate the need for the auxiliary capacitor circuit, thereby restoring the immunity of the circuit to storage capacitor mismatches.