A HIGH RESOLUTION INTEGRATED CIRCUIT BIOMEDICAL TEMPERATURE SENSING SYSTEM

by

Kenneth S. Szajda

Bachelor of Science
Massachusetts Institute of Technology
(1987)

Master of Science
Massachusetts Institute of Technology
(1989)

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1995

Signature of Author_____
Department of Electrical Engineering and Computer Science
January 13, 1995

Certified by_____
Professor Charles G. Sodini
Thesis Supervisor

Certified by_____
Dr. H. Frederick Bowman
Thesis Supervisor

Accepted by_____
Professor Frederic Morgenthaler
Chair, Department Committee on Graduate Students

A High Resolution Integrated Circuit Biomedical Temperature Sensing System

by
Kenneth S. Szajda

Submitted to the Department of Electrical Engineering and Computer Science on January 13, 1995 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering

ABSTRACT

The effectiveness of hyperthermic treatment of tumors is critically dependent on achieving therapeutic temperature and thus on the ability to precisely monitor temperature during treatment. The "active needle" integrated circuit multi-site temperature measurement system is demonstrated. The system consists of up to sixteen 580 micron wide, 7.5 millimeter long microelectronic "smart sensors" and a single 580 micron wide, 5 millimeter long digital controller that interfaces the instrument to a personal computer. The chips are mounted collinearly on a specially machined 22 gauge (710 micron) stainless steel needle. Electrical connections between chips are made using ultrasonic wedge bonding techniques; connection to the personal computer is made using a microribbon cable that is attached to the digital controller using ultrasonic ball bonding. The entire multi-chip system is passivated to prevent cross-contamination between the system and the surrounding tissue environment.

Each of the "smart sensor" chips contains a high resolution temperature sensor, preamplification circuitry, and an analog modulator for A/D conversion. The temperature transducer is a fully differential, p-n junction diode based circuit that uses feedback to reduce noise and improve linearity. Low noise preamplification is accomplished using switched capacitor techniques combined with correlated double sampling. The analog modulator, the front-end of an oversampled A/D converter, uses fourth-order noise shaping with one-bit quantization to generate a single-bit data stream that contains transducer measurement information. This data is sent to a custom microcontroller chip, also resident on the needle, that handles interfacing between the sensor chips and a personal computer. The personal computer issues instructions to and receives sensor data from the microcontroller. Once read, the digital data is processed on the personal computer to produce the final result. Measurements from test chips show a resolution of approximately 2 m°C over the 30-5C °C biomedical temperature range of interest. Linearity is approximately 0.012%. Data from a completed single-sensor needle system show a resolution of approximately 4 m°C over the same range. Because of the difficulty in testing at this level, it is thought that these figures may result from limits of the temperature testing system and not inherent limits of the sensing system.

Thesis Supervisor: Dr. Charles Sodini
Title: Professor of Electrical Engineering

Dedication

Since I have been so dedicated to this thesis, I dedicate this thesis to me.

No, just kidding. Not only would I be crucified for being so selfish as to dedicate the thesis to myself, but I'd also be doing several people a great injustice. In fact, I can't even take credit for that line since it comes from my old officemate Dr. Mike Curley. And when I say old I mean OLD--so old that he actually used troff on a MicroVax to write his entire doctoral dissertation.

Which brings me to the main point of all this rambling. I have indeed been at MIT for quite some time; long enough to remember what Kendall Square was like when the F & T diner was there and the T-station had temporary lighting because the electrical system didn't work. And through it all I have made many friends and learned quite a lot. But there have been certain people who have suffered along with me, and whose lives have been directly influenced by the ups and downs of this project.

First and foremost is my family: My parents **Richard and Cloe Szajda**, my brother **Douglas**, and my sisters **Nancy** and **Kimberly**. More than anyone else they have provided me with the strength to keep plodding ahead in search of my Ph.D. More importantly, they put up with my incredible frugality during my tenure in graduate school, where you get paid less than 25% of your market value for 10 times the effort of an industrial job. When I left for MIT many years ago they were all there to lend me support and encouragement, and they have been behind me 100% through 11 years and three degrees. Besides, I promised in my Master's thesis that I'd dedicate my Ph.D. to them.

Although she has not been with me through my entire tenure at MIT, my girlfriend **Pamela DiFiore** has perhaps suffered more than anyone, even my family, as a result of the ups and downs of my doctoral work. When I first met her, I told her I'd be done "in 2 months." That was nearly 2 years ago. Since then, she has put up with the highs, the lows, the delays, and my very long hours in the office, through it all giving me the support and encouragement I needed to keep from getting discouraged. Unlike my family which is 250 miles away, she had to deal with me on a daily basis, 7 days a week, 365 days a year. I will never forget the sacrifices she has made for me. I only hope I can return the favor now that she is in graduate school.

Finally, I must mention my extended family, namely, **John Cook** and **Michael Hummieney**. They came on the scene during my time at MIT and have put up with my whining every time I go home to New Jersey. More importantly, they have put up with my lame Christmas presents for many years because of my abject poverty. I can't thank them enough.

So I dedicate this thesis to all of you. All I can say is thank you.

Acknowledgments

This thesis is the culmination of a project that literally began with nothing. Although my name appears on the title page, there are many people who contributed time and effort to the successful completion of the project. I would like to take the opportunity here to try and make sure that after all their work they at least get their names in print. So here they are, in no particular order:

First, I'd like to thank my thesis committee: Dr. H. Frederick Bowman, Prof. Charles Sodini, and Prof. Alan Grodzinsky. They were faced with the very unusual task of supervising work that had significant components outside each of their areas of expertise. In some sense I think that working on this project has been just as much of a learning experience for them as it has been for me. I can honestly say that no matter what happens to me in the future I will never forget the last few years.

A special thank you also goes to Greg Martin and Dan Sidney, my partners in crime. Throughout the project, Greg served as my resident heat transfer guru, especially with the thermal modelling presented in chapter 2. Dan, perhaps because of his genetic links to the math community, was always there to check my math on the toughest problems. But technical assistance aside, Greg and Dan have been two of my closest personal friends throughout the entire project, and have been a great source of moral support when times were tough. The most pleasant memories of this project that I will carry with me will be the daily *dejeuner* at the coffeehouse, eating on the wet grass during the summers, and the great xpilot breaks. Thanks a lot guys.

Some of the most helpful technical discussions I had were with Peter Roman, a fellow engineer and a great friend. Although he is a digital designer (cough cough), he has always been willing to drop everything to help me *en caso emergencia*. He was instrumental in getting the parallel interface board working. He has always been one of my strongest supporters, and I have him and his mother to thank for getting me hooked on Europe as a vacation spot. One of these days we'll both be in Geneva at the same time, and I'll get to hear his French in action.

Several people at MIT deserve a thank you for all of the great technical assistance (and friendship) they have provided. Craig Keast was integral to the success of this project: When I first started working on it, Craig was the "mentor" who was always willing to answer my questions and look over my shoulder while I was learning to process. Later, he served as my liaison with Lincoln Laboratories, and was instrumental in getting some processing done there when I was in a pinch. Jeff Gross provided invaluable advice on the device noise issues. Jen Lloyd was nice enough to design and build a test board for her work that could easily be modified for some of my testing. She also put up with all of my terrible jokes, as well as my relentless mooing. Perhaps one day she'll learn how to use capital letters in her e-mail.

There were a multitude of other people that made this experience unique. Amongst them are Kathy Krisch, Joe Lutsky, Steve Decker, Fritz Herrmann, Jeff Gealow, and Chris Umminger, all members of Professor Sodini's research group who helped make

my work in building 39 that much more pleasant. One person who deserves special mention is Andy Karanicolas, with whom I began the quest for a Ph.D. back in 1983. Yes, 1983. We both started as clueless freshmen and ended as Ph.D.s. In the process of going from one to the other we managed to take many classes together (something like 30 I believe), and have remained friends throughout the last 11 years.

While on the topic of building 39, I must say thank you to the staff of the Integrated Circuits Laboratory, who have always done everything they could to help me get my work completed. I have always appreciated the effort they put in to help me get things done quickly. All I can say to them is that I never expected that there would be a Needle6, let alone a Needle10. The champagne is on its way.

And lest I forget from whence my paycheck (if you want to call it that) came, I must thank Keiko Oh, Christopher Newell, Sally Mokalled, Patricia Cunningham and all of the staff over in E25. Keiko singlehandedly insured that I would not be homeless even when Dr. Bowman didn't remember to renew my R.A. Christopher saved many many hours of waiting by making sure Dr. Bowman showed up only 15 minutes late to our meetings. He also introduced me to the wonderful world of scanning, the culmination of which is my photo on the last page of this thesis. Sally and Patty were my "insiders" in my dealings with the MIT bean-counters, always pointing me to the people at MIT who *really* do all the work.

Thanks too go to all of my colleagues and coworkers in building 20, who are too numerous to name individually. Although the building is officially unnamed, we refer to it affectionately as the Plywood Palace. It has certainly been an experience. Let's just hope they really did get all the asbestos out of this place.

On the personal side, I have to mention all of my friends who have had to put up with me over my 7 years in graduate school. They shared both the joy and frustrations I have experienced, all the while offering me tremendous support. So thank you to:

- Julie Hardy, who I promise not to sue for pain and suffering;
- Mary Kay Roman, who kept Peter and myself from turning into geeks in social settings, and who put up with our marathon Nintendo-nights;
- Peter and Hollie Schmidt, Erik and Pirjo Heels, and Matt and Lori Birkholz, who provided me with a semblance of a social life for many years;
- Dimitry Rtischev, Scott Lichtman, John Bulzacchelli, Raghu Krishniah, and Lynn Pekmezian, my Poisson-process e-mail correspondents;
- Lisa Magnano, to whom I owe lots and lots of lunches at the Mandarin (and a job when I hit it big in business);
- Kathleen Donahue, who I promise to visit any time I'm in the state of Wisconsin;
- Doug and Coleen Smith, who have kept in touch despite moving 2000 miles away;

- and Simone Pottenger, who always reassured me that someday, somehow, I'd finish this thesis.

My apologies to anyone I inadvertently left out. There have been so many of you who have helped me at one point or another that it's difficult for me to keep track of everyone.

"No man is a failure if he has friends."

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Advancement in medical technology is increasingly limited by problems which are multidisciplinary in nature. The physical sciences have played a role in medical ''innovations'' for centuries. Rapid advances in technology, however, have drawn engineering even deeper into the realm of medical research. Electrical engineers, for example, have become, in an increasing number of areas, as valuable to medical research as physicians and biologists. Instruments such as the magnetic resonance imager, the ultrasound scanner, and computer aided tomography systems all have at their core complex electrical systems. Recent advances in integrated circuit and integrated sensor technology are now breaking down the barrier between instrumentation and biology, making ideas such as implantable drug delivery systems and minimally invasive local physiological monitoring systems a reality.

One of the most promising areas in which electrical technology is finding application is in cancer treatment, where radiotherapy and chemotherapy continue to be the two primary therapeutic alternatives to surgery. Recently, the use of hyperthermia, a technique that uses heat to alter the biological state of tumors, as adjuvant therapy has been shown to increase the effectiveness of treatment. Through controlled elevation of tumor temperature, tumor physiology, including oxygenation and blood flow, can be altered. This control of the tumor microenvironment allows the clinician to better plan and execute the therapy. Clearly, the effectiveness of such a treatment depends

critically on the knowledge of the local tumor conditions.

This document describes a course of research that addresses the application of integrated circuit and sensor technology to the minimally invasive measurement of local tissue temperature, as it has been shown that the tumor microenvironment can be largely described by the temperature profile throughout the tumor and its surroundings [1]. A temperature sensor employing only silicon process compatible devices was developed; this sensor, as well as specially designed excitation circuitry and a high resolution data converter form a single sensing ''unit'' that is the basis of the ''active needle'' system. A series of these integrated circuit sensing units is mounted on a specially machined needle, bonded, and passivated to form the ''active needle'' minimally invasive measurement system for use in the hyperthermic treatment of tumors.

1.1 Hyperthermia

As the name implies, hyperthermia is the artificial elevation of tissue temperature for the purpose of modifying the tissue environment. In most cases, heating is applied in order to generate local conditions that are toxic to the cancerous cells being treated. Although the exact threshold varies, at temperatures above approximately 41°C cell mortality occurs. As the temperature is raised above this critical temperature, the rate of cell mortality increases. In short, the total thermal dose, defined as the weighted time integral of the tissue temperature, is the key parameter for determination of local treatment efficacy in this situation [2].

In other cases, heating is used to enhance the susceptibility of tumor tissue to other forms of treatment, usually by modifying tissue perfusion (volumetric blood flow) in the treated region. Since all tissue transport processes depend on capillary level blood flow, treatments which are transport controlled are dependent on perfusion. Radiotherapy depends on blood flow to deliver oxygen to the tumor which interacts with the ionizing radiation to produce free-radicals ($\dot{O}_2^-$ and $H\dot{O}^-$). Chemotherapy depends on perfusion

to carry the chemotherapeutic agent to and into the tumor. Immunotherapy relies on blood flow to carry monoclonal antibodies to the tumor site. Hyperthermia itself is more effective in achieving higher temperatures in regions of low flow: Elevated temperature will generally induce higher flows in the tumor, thus increasing the effectiveness of both radiation therapy and oxic chemotherapeutic agents. In summary, in all cases the fundamental basis for hyperthermia is the strong temperature dependence of local tissue properties.

1.1.1 Early Hyperthermia Techniques

Early hyperthermia systems relied on the difference in thermosensitivity between normal and cancerous tissue. This enhanced sensitivity to heat is believed to be a result of the unique biological nature of cancers: Studies have shown that tumors are nutritionally deprived, unusually acidic, and chronically hypoxic when compared to healthy tissue. Compromised tissue perfusion (volumetric blood flow) is believed to be largely responsible for this environment, as blood flow is the primary path for oxygen and other nutrient supply to the tissue. The hypoxic areas of the tumor would then be the least well perfused; conversely, one would expect the well oxygenated areas to be concentrated around regions of adequate perfusion. In addition, the primary thermoregulation or "energy removal" mechanism in tumors is blood flow. The absence of oxygen indicates a decrease in perfusion and a corresponding increase in the ability to heat the tumor.

Preferential killing of tumors (over healthy tissue) was then possible: If this differential in thermal sensitivity could be obtained with regularity, heat would become an invaluable tool in cancer therapy. Hyperthermia has been shown to be more effective in acidic tumors [3,4], and has at least the same success on hypoxic cells as it has on oxygenated cells [5,6,7,8]. It was demonstrated that tumor regeneration following radiotherapy is largely due to the high survival rate of hypoxic cells, which are quite radioresistant [8]. This indicated that the combination of hyperthermia with other

treatment modalities could significantly decrease tumor survival rates, and indeed this was demonstrated [9].

1.1.2 Modern Hyperthermia Approaches

Initially, hyperthermia systems were forced to rely on the differential heat sensitivity because of the lack of applicators capable of locally heating small volumes of tissue. Under these conditions, the thermal energy deposited into a patient during treatment was not well controlled, and the only mechanism for generating cytotoxicity in tumor tissue while minimally harming healthy tissue was the differential response to heat between the two tissue types. This proved to be a major obstacle to effective use of hyperthermia since there is generally a large variation in local tissue properties. Large tumors, for example, usually are modelled as a central core of very lowly perfused, necrotic tissue, a middle region in which tissue properties are similar to normal tissue, and an outer growing margin in which perfusion is abnormally high due to the formation of new vasculature as the tumor expands.

Recent advances in technology have greatly improved the viability of hyperthermia as a tumor treatment modality and have eliminated the dependence of treatment efficacy on differential thermosensitivity. Many systems have been developed to more accurately apply local deep heating, including interstitial RF electrodes [10,11,12], implanted antennas [13,14,15], thermoseeds [16,17,18,19], and various ultrasound applicators [20,21,22,23,24]. Each of these has the advantage of selectively elevating tissue temperature over small volumes, allowing excellent local control of heating. As a result, hyperthermia can be successfully applied even in tumor regions that show thermosensitivity comparable to that of normal tissue.

Consequently, the establishment of local dose-response relationships is now critical for hyperthermia therapies, since the locally applied thermal dose can be modified to adequately heat tumor regions regardless of their thermosensitivity characteristics. This, in turn, requires the capability to locally quantify tissue temperature, thermal

properties, and perfusion. The importance of temperature measurement in cancer thermal therapy cannot be overemphasized, since tissue temperature history directly correlates with success or failure of a treatment. As the database of thermal effects on various types of tissue expands, clinicians can more effectively treat each tumor, and can optimize both individual treatment sessions and the multi-treatment therapies of which they are a part. Ultimately, with the development of real-time local thermal dose quantification capabilities, treatments could be modified "on the fly" to adjust for local tissue property and perfusion changes. The optimized heating patterns that result would ensure that the appropriate therapeutic thermal dose is delivered over the entire treatment area.

1.2 Clinical Hyperthermia Treatment

Although the previous section outlined the importance of temperature and perfusion information during hyperthermia, no mention was made of the methods for providing this information. This is because the requirements of a clinical temperature measurement system are more easily understood in the context of the clinical treatment. Furthermore, because accurate measurement of perfusion relies on accurate temperature measurement, as will be discussed below, the most stringent requirements for temperature measurement systems are actually dictated by the desired perfusion resolution.

When a patient is receiving hyperthermia therapy, there are three distinct phases to each treatment session. The first phase is the treatment planning or pre-treatment period; this is when the clinician examines the current state of the tumor and the tumor environment and derives from that information an appropriate hyperthermia treatment (applicator, treatment portal, thermal dose and the like) for the session. During this evaluation phase, the primary physiologic measurements of interest to the clinician are the tumor blood flow, oxygenation, and pH, as each of these parameters strongly affects not only the treatment plan but also the ultimate effectiveness of the therapy session.

The second phase is the actual hyperthermia treatment, when heat is applied to the

tumor using either microwave or ultrasonic energy deposition. At the hyperthermia treatment center at the Dana-Farber Cancer Institute, for example, one of several ultrasound heating devices can be used, including the 4 element Sonotherm 1000, the 16 element Sonotherm 2000, or the focussed segmented ultrasound machine (FSUM), which is an array of 56 ultrasound transducers that can heat deep seated tumors to a depth of approximately 15 centimeters. During this part of the treatment, knowledge of the physiologic state of the tumor is essential; as the treatment evolves, the clinician can adjust both the position and the magnitude of the thermal energy deposition (and thus thermal dose) to optimize the treatment effectiveness. The primary physiologic parameters of interest during heating are the tumor blood flow, which will dictate which areas require greater local power deposition, and the tissue temperature, since that is the direct measure of the residual effect of the energy deposition. When hyperthermia is used in conjunction with radiation therapy, the clinician would like to know the applied radiation dose also. The end result is a global feedback loop, in which heat and/or radiation is applied, the resulting tumor state monitored, and the doses adjusted to increase treatment effectiveness. This loop is illustrated in figure 1.1. Although currently the clinician is included in the loop, it is hoped that in the future the loop can be entirely automated, so that the dose can be electronically adjusted for optimum effectiveness without clinician intervention.

The final phase of a hyperthermia treatment session is the post-treatment or treatment evaluation period, where the 3-D thermal field achieved during therapy is reconstructed from discrete measured temperature sites and at the more sophisticated centers from 3-D thermal models as well. These local thermal histories permit local dose to be calculated and compared with observed tumor response. After heating is discontinued, tumor perfusion is monitored and compared with pre-therapy values. In most cases, there is an "immediate" response (generally an increase in perfusion) as well as a longer term response (which is observed during the pre-treatment perfusion monitoring of the next hyperthermia session). The volume of tumor achieving therapeutic thermal

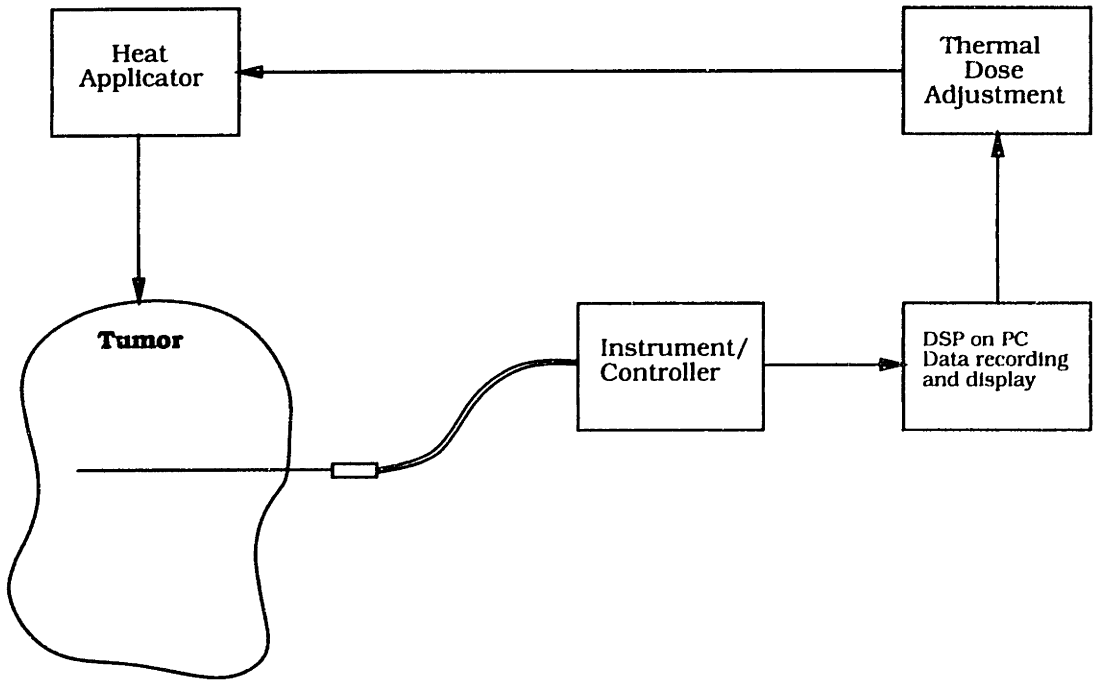


Figure 1.1: Hyperthermia therapy feedback loop

doses in equivalent minutes at 43°C is the best indicator of treatment outcome. Trends in blood flow over the course of several sessions also appear useful as an indicator of the success or failure of the therapy.

These trends are illustrated in figures 1.2 and 1.3 [25]. These diagrams are actual perfusion measurements taken over the course of treatment for a typical patient who has received hyperthermia therapy. The first graph shows tumor perfusion during a single session; as one can see, the perfusion is affected by several factors, including motion (and anesthesia) as well as the specific tumor heating. Note also that the variation in perfusion is approximately 100%. Clearly these perfusion-altering events will have a dramatic outcome on the local temperature achieved during the heating phase and ultimately on the treatment efficacy. It is therefore critical that this information be provided to the clinician. The second diagram shows the macroscopic trend in perfusion over several treatment sessions, with the average pre-heating and post-heating perfusion measurements for each treatment session.

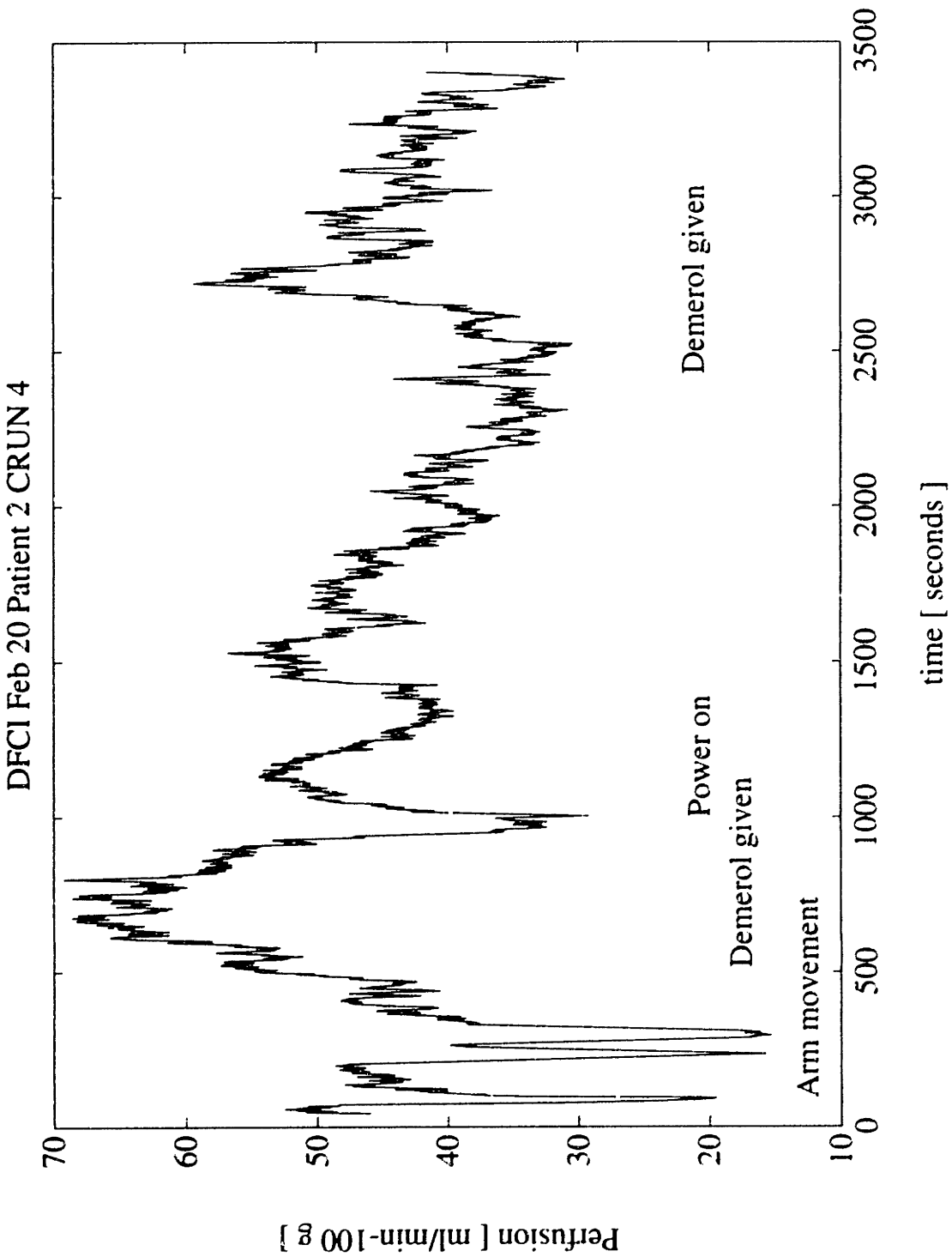


Figure 1.2: Perfusion trends during a single treatment session

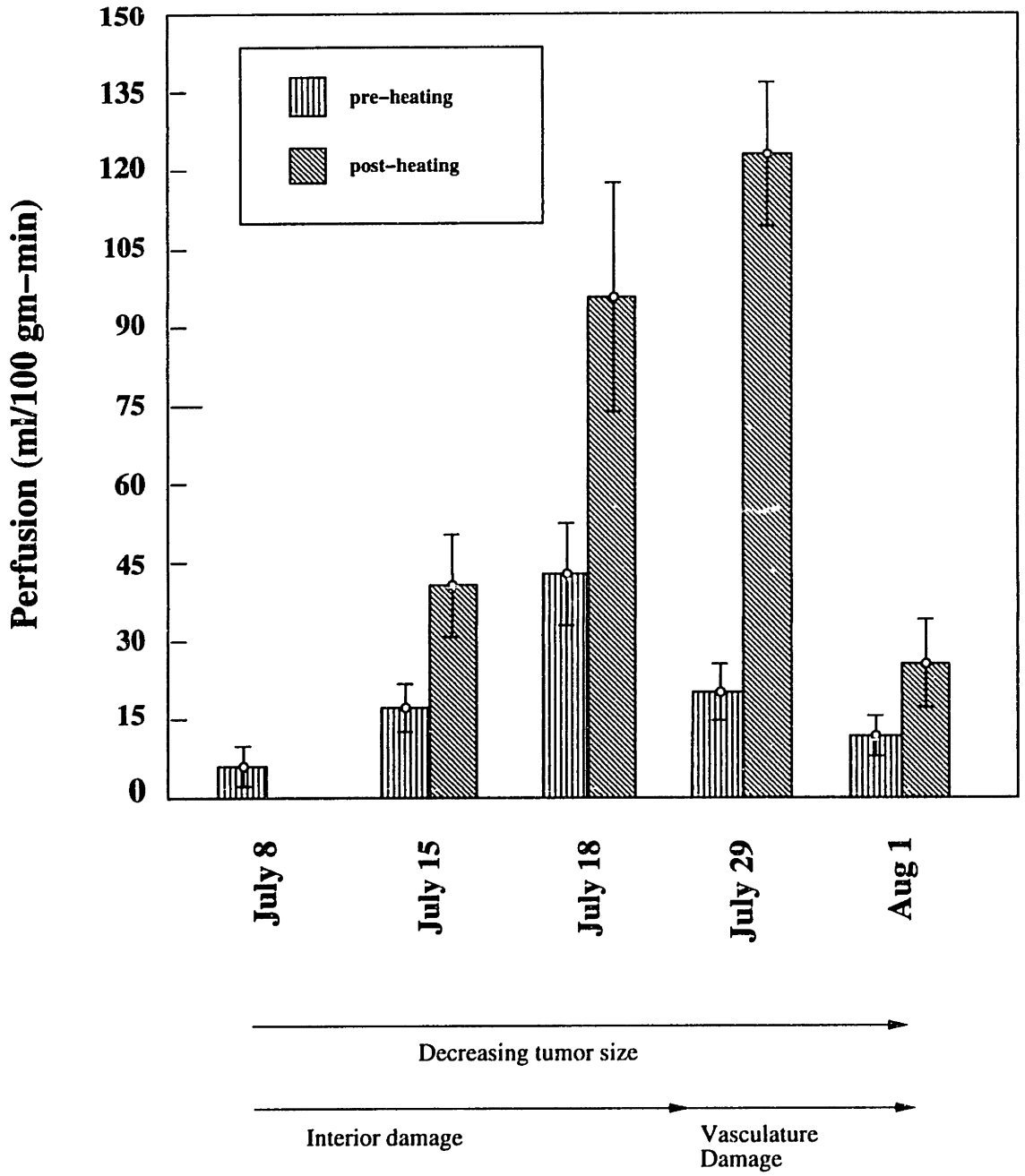


Figure 1.3: Perfusion trends over five treatment sessions

1.2.1 The Relationship between Temperature and Perfusion

Whether it be chemotherapy, radiotherapy, hyperthermia, monoclonal antibodies designed to target tumor cells or efforts at new modalities of therapy such as anti-oncogenes or cytokines, these therapeutic techniques all depend on local perfusion distributions in the target tissue. Therefore, knowledge of the perfusion level and distribution characteristics to deliver optimal dosage, regardless of the anti-cancer agent, would be desirable. To the extent that tissue environmental factors such as pO_2 and pH are important determinants of the effectiveness of a given anti-cancer agent, perfusion is a primary controlling influence. Thus, there is a clear need for local blood perfusion and temperature measurement in the context of both individual hyperthermia therapy sessions and long-term hyperthermia regimens, especially when hyperthermia is used as part of a multimodal treatment. With this foundation, it is now possible to examine the particular requirements of a clinically useful thermometry instrument.

As mentioned above, measurement of perfusion relies critically on the ability to measure temperature accurately. Since perfusion is so strongly linked to treatment planning and effectiveness, any temperature measurement sensor is evaluated in terms of the perfusion resolution that would result if it were used as part of a perfusion measurement system. Thus it is necessary to explore perfusion measurement techniques and the relationship between local temperature, thermal properties, and blood flow so that the design goals of a clinical temperature measurement system can be better understood.

1.2.2 Methods of Perfusion Measurement

There are two major methods used to measure blood perfusion; both can be described as ''indicator-dilution'' techniques in which an indicator is added to the medium and the resulting dilution of the indicator is used to extract the blood flow in the tissue surrounding the measurement probe. In both techniques, the indicator is heat, and the dilution is detected by monitoring the power supplied to the heat source. The

relationship between the heat supplied and the resulting dilution depends on which of the two methods is used, and will be examined further below.

The first technique is the constant temperature method, in which the heat source is used to create a small local temperature increment above the baseline tissue temperature [26]. This increment, (ΔT), is applied and maintained for the duration of the measurement; the power required to maintain the increment is perfusion dependent. The exact course of the measurement is as follows: First, the unperturbed (baseline) temperature is measured. The temperature step ΔT is then applied; initially, during the transient part of the measurement, the power supplied to the heating element is large. As the thermal field surrounding the heat source spreads out into the tissue, the power requirement decreases. When steady state is reached, the power required becomes constant. Blood perfusion around the heat source dramatically affects heat transfer in the vicinity of the probe, and therefore both the transient power required to create the increment and the steady state power are related to perfusion.

To demonstrate the effect of perfusion mathematically, the heat source (usually a thermistor or other resistive heater) is modelled as a sphere in a continuous, infinitely perfused medium. In the absence of perfusion, the volume average temperature increment of the heat source is [27]:

$$\Delta T = \frac{P_{ss}}{4\pi a k_b}\left[\frac{k_b}{k_m} + 0.2\right] \tag{1.1}$$

where P_{ss} is the steady state power required to maintain the temperature increment, a is the radius of the sphere, k_b is the thermal conductivity of the sphere, and k_m is the thermal conductivity of the medium (tissue). The first term ($P_{ss}/4\pi a k_m$) represents the temperature rise that results from the finite conductivity of the medium; the second term accounts for the heating that occurs due to the finite conductivity of the bead. The value of 0.2 falls out from the geometry of the problem and the determination of the volume average bead temperature from the actual thermal profile. In the presence of

perfusion, the volume average temperature increment is [28]:

$$\Delta T = \frac{P_{ss}}{4\pi a k_b}\left[\frac{k_b}{k_m}\left(\frac{1}{\sqrt{\frac{wc_{bl}a^2}{k_m}}+1}\right)+0.2\right] \tag{1.2}$$

where w is the local blood flow and c_{bl} is the specific heat of the blood. Comparing the two equations, it is evident that the result with perfusion is equivalent to the unperfused solution if an "effective" thermal conductivity is defined:

$$k_{eff} = k_m\left[\sqrt{\frac{wc_{bl}a^2}{k_m}}+1\right] \tag{1.3}$$

Which in turn gives:

$$\Delta T = \frac{P_{ss}}{4\pi a k_b}\left[\frac{k_b}{k_{eff}}+0.2\right] \tag{1.4}$$

Solving equation 1.3 for the blood perfusion w gives:

$$w = \left(\frac{k_{eff}}{k_m}-1\right)^2\left(\frac{k_m}{c_{bl}a^2}\right) \tag{1.5}$$

Thus, if the probe radius a, tissue thermal conductivity k_m and blood heat capacity c_{bl} are known, the blood perfusion can be found by first determining the effective thermal conductivity from equation 1.4 (since ΔT and P_{ss} are known), and plugging the resulting value into equation 1.5.

The relationship between the uncertainty in the applied temperature step ΔT and the corresponding measured perfusion w can be found by differentiating equations 1.4 and 1.5 with respect to k_{eff} to give:

$$\frac{\partial \Delta T}{\partial k_{eff}} = \frac{-P_{ss}}{4\pi a k_{eff}^2} \tag{1.6}$$

$$\frac{\partial w}{\partial k_{eff}} = \frac{2}{c_{bl}a^2}\left(\frac{k_{eff}}{k_m}-1\right) \tag{1.7}$$

After combining these equations (by solving for ∂k_{eff}) and rewriting the resulting expression in terms of ΔT and w, the desired relationship is found:

$$\frac{\partial w}{w} = \frac{2k_{eff}(k_b+.2k_{eff})}{k_b(k_{eff}-k_m)}\cdot\frac{\partial \Delta T}{\Delta T}$$

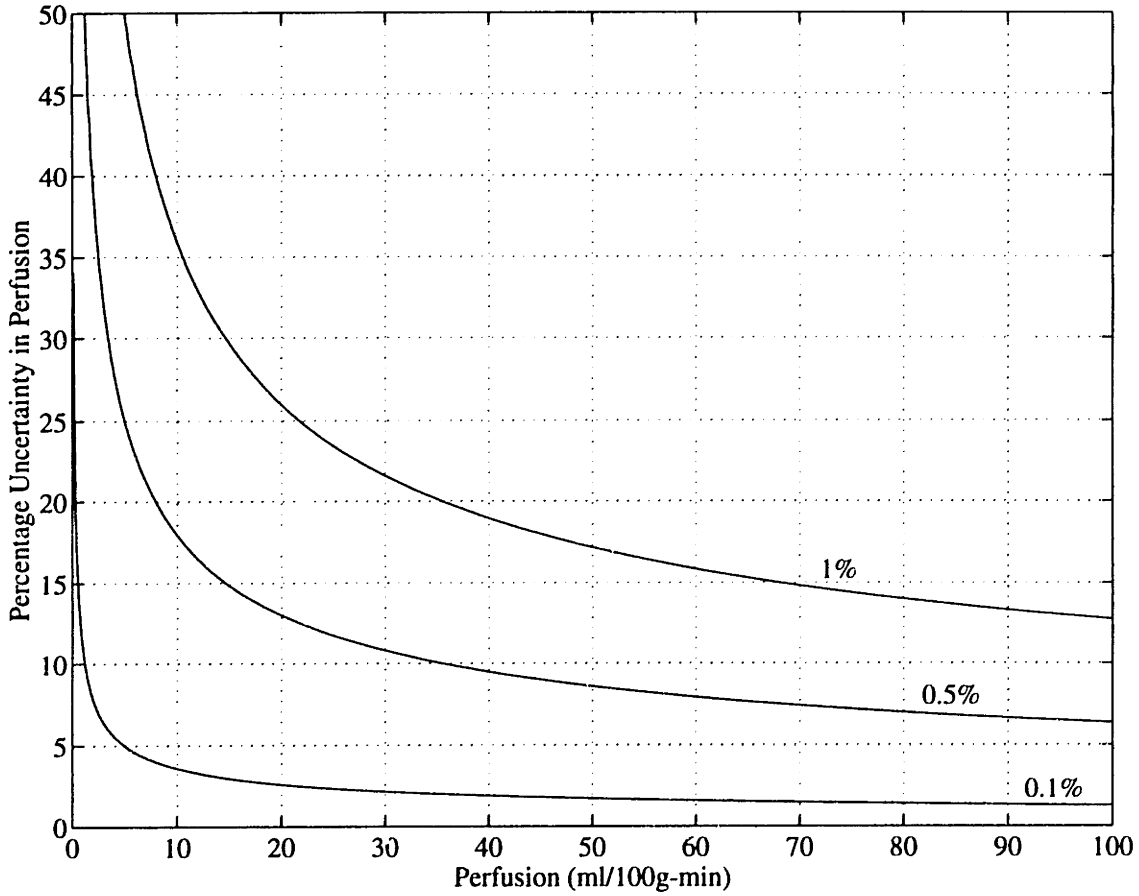


Figure 1.4: Relationship between percentage uncertainty in temperature and percentage uncertainty in perfusion as a function of perfusion

$$= 2\left[1 + \frac{1}{\sqrt{\frac{wc_{bl}a^2}{k_m}}}\right] \cdot \frac{\partial \Delta T}{\Delta T} \tag{1.8}$$

Thus for a given resolution in temperature $\partial \Delta T$ the corresponding uncertainty in measured perfusion can be found. Figure 1.4 shows this function plotted as a function of perfusion for various temperature uncertainties. For this graph, $k_b = .2k_m$, which is typical of currently available systems. It is evident from this figure that large uncertainties in perfusion result from small inaccuracies in temperature: For blood perfusion on the order of 5 ml/100g-min (a typical value for resting muscle tissue), the required uncertainty in temperature must be less than 0.1% in order to resolve perfusion to within 5%. For an applied temperature step of 5°C (again typical of current systems)

this translates into a temperature resolution of 5 m°C. As a result, one of the design goals of any perfusion measurement system must be to maximize the resolution in temperature. Since one of the long term objectives of the active needle project is to measure perfusion, the temperature system must be designed for very high resolution.[1] Thus, to maintain a level of uncertainty in perfusion of less than 5% over the entire physiologic range of interest, the temperature measurement system must resolve less than 5 m°C. In order to make the performance of this system comparable to the best discrete systems, the design goal is to resolve 1 m°C.

The second technique used to measure perfusion is the constant power heating method, in which the power supplied to the heat source is held constant. The constant energy deposition will result in a thermal field in the heat source and the surrounding tissue. This ΔT generated by the supplied power is clearly dependent on the thermal parameters of the tissue surrounding the heat source, since the only mechanism for heat loss is through the tissue. This coupling provides the means for extracting perfusion information. The exact course of an experiment is as follows: First, the baseline temperature is measured. The power step (Γ) is then applied; initially, this power step does not alter the thermal field in the tissue. As the heat source temperature rises in response to the step, the thermal region of influence expands into the tissue. Steady state is reached when the heat lost to the tissue equals the heat being generated by the heat source. At this point, the heat source temperature is measured; the blood perfusion is extracted from this measurement.

Mathematically, the relationship between perfusion uncertainty and temperature uncertainty is found using the same method as for the constant temperature experiment. The heat source is modeled as a sphere embedded in the tissue. For constant power

[1] It is important to note that it is the *resolution* of the system that is important and not the *accuracy* of the measurement, since the measurements are always made with respect to the baseline (unperturbed) temperature of the tissue.

heating, the steady state temperature profile in the sphere is given by [29]:

$$\Delta T_b(r) = \frac{P_{ss}}{\frac{4}{3}\pi a k_b}\left[\frac{k_b}{3k_m\left(1+\sqrt{\frac{wc_{bl}a^2}{k_m}}\right)} + \frac{1}{6}\left(1-\frac{r^2}{a^2}\right)\right] \tag{1.9}$$

The volume average temperature of the sphere (which is actually the measured parameter) can be found by integrating the above expression over the volume of the sphere:

$$\begin{aligned}\Delta T_b &= \frac{1}{\frac{4}{3}\pi a^3}\int_0^a 4\pi r^2 \Delta T_b(r)dr \\ &= \frac{P_{ss}}{4\pi a k_m\left(1+\sqrt{\frac{wc_{bl}a^2}{k_m}}\right)} + \frac{P_{ss}}{20\pi a k_b}\end{aligned} \tag{1.10}$$

By substituting equation 1.3 into this equation and factoring the result, equation 1.4 is obtained. Consequently, the same sensitivity analysis applies. The result is that the uncertainty in perfusion is related to the uncertainty in temperature resolution in the same way as before; the only difference now is that the steady state power is being controlled and not the temperature increment. For reasonable experimental parameters, however, the power used to generate the thermal field will be such that the resulting temperature step is on the order of 5°C; in other words, the thermal fields created by the constant temperature and constant power techniques will be approximately the same. It is therefore apparent that high resolution temperature data is also required for constant power perfusion measurements, and that the design goals stated earlier will satisfy the requirements for a constant power perfusion measurement system also.

1.3 Biomedical Temperature Measurement Systems

1.3.1 Thermistor-based Schemes

The highest resolution systems to date are those based on the thermistor temperature transducer, which in its various forms has exhibited temperature coefficients of resistance ranging between +70 and -6.5%/°C [30], with nominal room temperature

resistances on the order of 1 to $10K\Omega$.[2] In addition, although thermistor sensors are nonlinear, it has been shown empirically that these transducers can be calibrated to within 1 m°C over the narrow physiologic temperature range of interest (typically 30-50°C) [31]. Because at present there is no way of integrating thermistors into a conventional silicon process, these systems are discrete. Since there is therefore no thermal coupling between the sensor and the electronics, the instrumentation can be optimized independently from the sensor; the resulting circuitry can be constructed so that it fully exploits the resolution capabilities of the thermistor.

The sensor biasing and signal processing circuitry is usually based on constant current source excitation or temperature to frequency conversion. In the former case, the current source is made constant and with negligible temperature drift, so that the voltage developed across the thermistor is directly proportional to the temperature via ohms law $V = IR$.[3] The general form of this configuration is shown in figure 1.5. Coupled with this excitation, however, is the self heating of the sensor resulting from the dissipation of power in the bulk of the thermistor. This limits the maximum excitation current, which in turn results in low signal levels (on the order of millivolts). In addition, the circuit is sensitive to noise from the power supply, the current source, and the amplifier. Constant current source approaches, therefore, require first a very well controlled, low noise excitation current as well as low noise, low drift voltage amplifiers. In spite of these limitations, however, these simple constant current thermistor sensors have recorded the highest temperature resolution to date (3.5 m°C) in a biomedical environment [32].

Thermistor based temperature-to-frequency converters attempt to bypass both of the above requirements by transforming the problem into one of frequency measurement. These circuits are typically relaxation oscillator circuits, with the thermistor configured

[2]The nominal resistance is clearly material and geometry dependent. Most small thermistors used in biomedical temperature measurement fall within these limits.

[3]Clearly a drift-free current source is not essential. Inaccuracies in the current source can also be tolerated if they are well quantified.

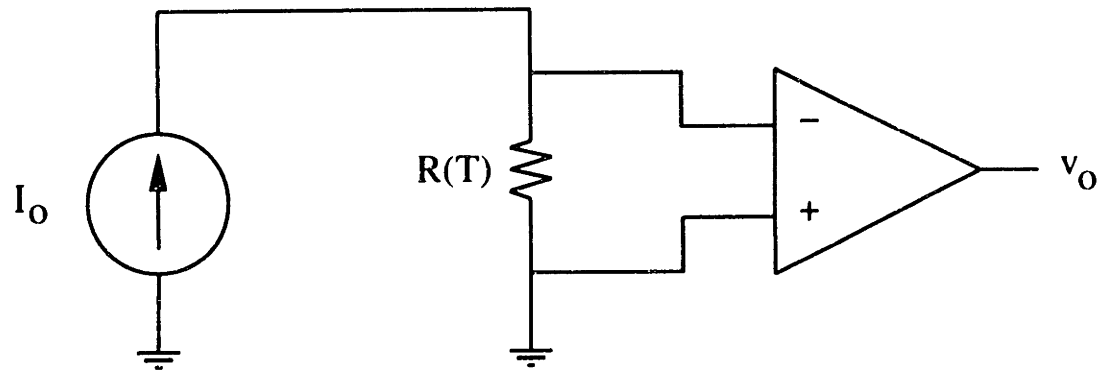


Figure 1.5: Thermistor excited by current source

as one of the timing control elements; a typical example of this is shown in figure 1.6. The variation in the R_TC time constant with temperature generates a temperature-dependent variation in the oscillator output frequency. Numerous examples of this type of design and many variations of it can be found in the literature [33,34,35,36,37]. The major difficulty with this approach, however, is that the current through the thermistor varies widely during the oscillation cycle, making compensation for self-heating a very difficult if not impossible task. There are also several practical problems associated with the post-processing of the frequency output, especially when making high accuracy systems. As a result, attempts to use this technique for biomedical measurements have met with little success; the best results have been obtained with wide dynamic range sensing rather than high resolution measurement.

1.3.2 Integrated Measurement Systems

By far the most active research in the temperature measurement field is concerned with integrated silicon temperature sensors, primarily because in theory an integrated sensor can be manufactured on the same substrate as the signal processing circuitry. The resulting chip is advantageous not only because of its size but also because the circuitry can perform elaborate on-chip preprocessing of the signal prior to transmission of data off-chip. However, because the "boundary" between the sensor and the signal processing circuitry is eliminated, a degree of freedom is lost in the design process;

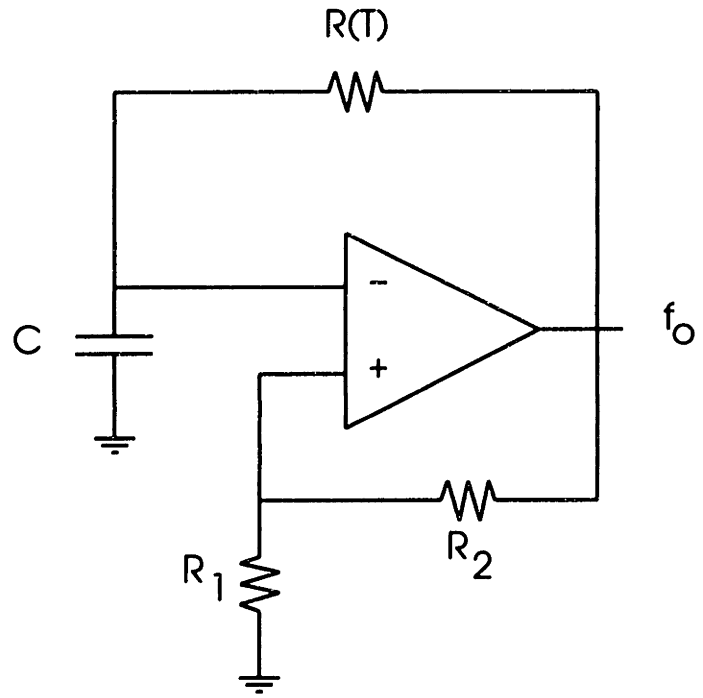


Figure 1.6: Typical temperature-frequency converter

namely, it is no longer possible to keep the processing circuitry in an isothermal state during measurement. The on-chip instrumentation must therefore be temperature insensitive or else be compensated in some way for the temperature fluctuations that will invariably occur due to the proximity of the sensor.

From a biomedical standpoint, however, the possibility of an integrated sensor/instrumentation system is quite desirable. The vast reduction in size would give the clinician the ability to make measurements in areas of the body where larger sensors would prove too invasive. Although some of the discrete systems with small remote thermistor sensors can be relatively non-invasive, the integrated sensors eliminate the need for additional instrumentation alongside the patient, creating a more mobile and less intimidating environment during therapy.

Perhaps the greatest advantage, however, is the relative ease with which an integrated system can be expanded or modified. A multi-parameter, multiple-site measurement with a discrete system would at a minimum require insertion of another

sensor and its attachment wiring into the patient for each added site; each type of sensor would also require a separate signal processing box, a rather unwieldy prospect at best. With an integrated sensor system, a single chip could be manufactured that contained multiple sensor/circuitry cells, with control circuitry added so that only one set of leads into the patient would be required. The ability to make measurements of different parameters could also be included with the addition of an appropriate sensor/circuit system on the chip. It is just this sort of flexibility afforded by an integrated system that is exploited by the active needle system, as will be described below.

Recent work has focused more on exploiting the temperature sensitivity of semiconductor circuit elements rather than on the development of a highly sensitive silicon temperature transducer such as a semiconductor thermistor. For this reason most of the integrated systems currently under development use the p-n junction diode as the temperature sensor, since such diodes are easily manufactured in a standard bipolar or CMOS process. Furthermore, these diodes exhibit a temperature sensitivity larger than most circuit elements ($\approx$ -2 mV/°C change in the forward voltage) that has been well characterized [38,39]. Also contributing to the wide use of junction diodes in integrated temperature measurement systems is the inability to manufacture good thermistor-type materials and circuitry on the same wafer using standard processing techniques.

Just as in the case of the discrete systems, one can find circuit designs employing both the constant current excitation [40,41] and temperature to frequency conversion [42,43] techniques. These designs face the same limitations as discussed for the discrete systems, but can exploit the advantages of integrated processing to improve performance. In the case of constant current source excitation, for example, the improved matching of transistors results in better control and characterization of the excitation current since superior temperature compensated circuits can be manufactured. The *decreased* sensitivity of the sensor also reduces the effect of self-heating. Of course, this decrease in sensitivity also makes the measurement more difficult to make, and in fact the best resolution to date (about 0.01°C) obtained with integrated semiconductor-

based sensors is nearly three times worse than the resolution quoted above for discrete systems.

Other attempts have been made to integrate multiple sensor arrays for biological temperature measurement, but none to date have met with great success. One of the more recent efforts was work performed at Stanford University on a thin linear thermometer array for use in hyperthermia therapy [44]. The array was a series of silicon p-n junction diodes connected into a flexible sensor array using fine stainless steel wire. The resulting multi-site thermometry system was connected to an external signal processing system that recorded the data from the sensors. Among the difficulties encountered with this approach was the inability to make accurate measurements due to thermal conduction down the stainless steel wires as well as the failure to properly passivate the system, leading to air bubbles in the sensor coating [45] and contamination of the sensors over a relatively short time period. In addition, the sensor configuration was inferior: Diode "arrays" were constructed in such a way that leakage through the inactive diodes in the array limited the measurement from an active diode. These problems are avoided here by modularizing the system, eliminating sensor crosstalk; by integrating the sensing circuitry as well, eliminating the need for long, relatively large wiring; and by incorporating newer and better packaging techniques than were available at the time of the Stanford work.

Ongoing research into the development of the bulk-barrier diode (BBD) as a biomedical temperature sensor is also encouraging. Physically, bulk-barrier diodes are bipolar transistors with a very thin base region and no base contact. In fact, the base region is made so thin that punch through occurs under zero bias conditions; i.e., there is no neutral base region in equilibrium. The charge distribution and energy band diagram for a BBD are shown in figures 1.7 and 1.8 respectively; the depletion approximation has been assumed in making these figures. Because there is no neutral base region, the energy band diagram does not show any "flat" middle region; instead, there is a potential "well" in the middle of the device. The device therefore behaves

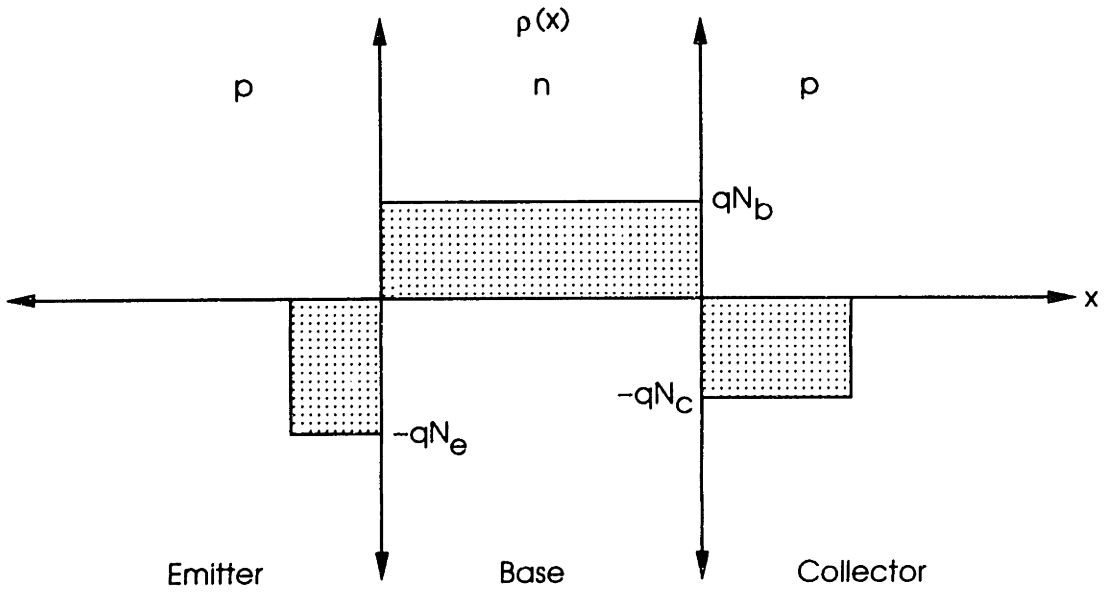


Figure 1.7: Charge distribution in a bulk-barrier diode

more like a resistor, with the depth of the potential well controlling the height of the conduction barrier and hence the conductance between the "collector" and "emitter" of the transistor. The depth of the well is a strong function of temperature: Although the theory explaining the temperature sensitivity has not been fully developed, it has been hypothesized that the space charge of thermally generated carriers in the well partially compensates the background depletion charge, resulting in temperature induced barrier reduction [46].

One of the major drawbacks of this sensor, however, is the lack of a well-developed theory explaining the physical operation of the device. This makes it very difficult to calibrate, as no clear understanding of the mechanisms governing the temperature sensitivity exists; the theory espoused above is merely published theory with only a small amount of laboratory work to support it. Furthermore, observed device characteristics lead one to believe that the device is operating in a breakdown mode; if this is the case, the noise level in the device will be significantly larger than that of an ordinary forward biased p-n junction diode, ultimately limiting the temperature resolution. Also, bulk-barrier diodes are not easily integrated into CMOS processes:

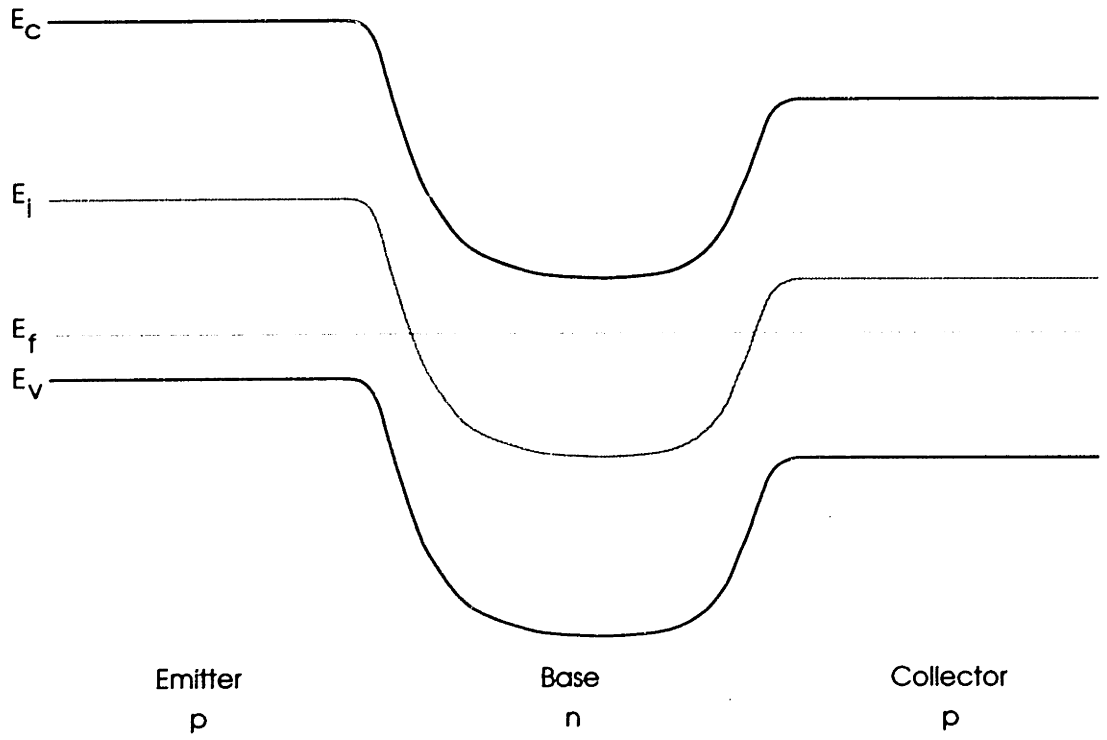


Figure 1.8: BBD energy band diagram

Although the device is fabricated like a bipolar transistor, the zero-bias punchthrough requirement necessitates additional processing even if a BiCMOS or bipolar process is used. Because most digital circuitry is fabricated in straight CMOS, this complicates the manufacturing process significantly. In short, although these sensors certainly look promising, practical application of the BBD requires a much more extensive body of knowledge about the device and its operation.

1.4 The Active Needle System

The purpose of this project is not only to measure temperature but also to demonstrate the feasibility of the "active needle" techniques. The system approach used for this project can easily be extended to other types of sensors, including (but not limited to) oxygen, blood perfusion, radiation, and pH sensors. Furthermore, any sensors that can be manufactured within the framework of conventional silicon processing can be mixed

on a single sensor array, allowing multiparameter measurements from a single needle. This is a major long term benefit of this project, as a single needle could potentially replace a large number of individual measurement instruments while providing a more complete characterization of the tissue microenvironment.

More specifically, the goal of this research is to develop and demonstrate an "active needle" integrated circuit system for the measurement of temperature at up to 16 sites in tissue. A sketch of the needle system, before bonding and coating, is shown in figure 1.9. Each of the chips is approximately 600 μm wide and 8 mm long. The distal integrated circuits are "smart sensor" chips, with a sensor, a detector, and an analog modulator (the front end of an oversampled analog-to-digital converter) on each one. The chip farthest from the needle point is the interface driver, which controls communication between the sensor chips and the external environment. The needle/circuit system is passivated to prevent contamination by the operating environment. The connection from the needle to the external environment is made with a small microribbon cable, which allows the system to communicate with a computer or other data recording device.

1.4.1 System Overview

Many different circuit functional blocks are required for this project. Each chip on the needle can be classified as one of two types: a sensor/detector/converter (S/D/C) chip or an interface driver. The S/D/C chips are the core of the system; they perform the actual data measurement. Each of these chips contains a single sensing unit, an amplifier for buffering and gain, and a dedicated analog modulator for generating the digital result. In addition, this chip also contains all of the sensor excitation circuitry and a small amount of digital control circuitry. Each sensor/detector/converter subcell produces a single bit digital data stream output that is fed back to the PC through the interface chip. This bit stream is then digitally processed by the PC to form the final result. The technical details of all of the circuitry are presented in succeeding chapters.

The important point is that there is no separation between the sensor, the detector, and the converter; they are designed to work together as a sensing "unit." There are two major benefits to this approach. First, the S/D/C chips act as "black boxes" that produce a single bit stream at their output. The rest of the circuitry need not know how this bit stream was generated; all that is known is that this bit stream corresponds to the output of the data converter. The personal computer processes the bit stream without regard to the source that generated the analog voltage that in turn caused the output bit stream. Partitioning the system in this way allows several different types of sensors to be part of a single needle system. Each different sensor/detector/converter chip has its own specialized excitation and detection circuitry; as long as a chip produces the bit stream output the personal computer can process the measurement.

The second advantage of this approach is the elimination of signal corruption as a result of long analog lines travelling from the measurement site to the control and processing instrument. Because each sensor/detector cell has a dedicated data converter built alongside it, data is never transmitted off the chip in analog form. Because of the much higher noise tolerances of digital signals, noise introduced by analog coupling of signals to the data lines (through crosstalk or spurious environmental signals) is practically eliminated. This indirectly increases the resolution of the measurement, since the data conversion is done at the sensing site and the usual signal degradation that occurs between the sensor and the processing circuitry does not occur.

The digital controller chip is the "brain" of the system. This chip accepts the output data stream from the S/D/C chips, tags the data with the appropriate sensor identification information, and transmits the results off of the needle. The chip also controls each of the sensor chips--turning sensor chips on and off as desired during the measurement sequence. In the future, this chip will also handle bidirectional communications between the S/D/C chips and the host computer, so that sampled-data active parameter measurements (such as the perfusion measurement described earlier) can be made.

In summary, there are many benefits realized by the active needle. First, the needle provides a broad platform for medical instrumentation because of the modularity of design and the development of a standard interface between chips. The elimination of the barrier between the sensor and the processing circuitry permits optimization of the electronics for each type of sensor. As a result, the active needle system vastly improves instrumentation by taking advantage of available microelectronic technology and improved circuit design. The following chapters describe the entire system in detail. Chapter 2 presents an analysis of the thermal perturbation created by the system. Chapter 3 discusses the temperature sensing circuitry. Chapter 4 describes the analog modulator and the analog-to-digital conversion technique and implementation. Chapter 5 explains the digital control system. Chapter 6 details the silicon processing, manufacturing and packaging of the system. Finally, chapter 7 presents results from manufactured sensor systems, and some suggestions for further work.

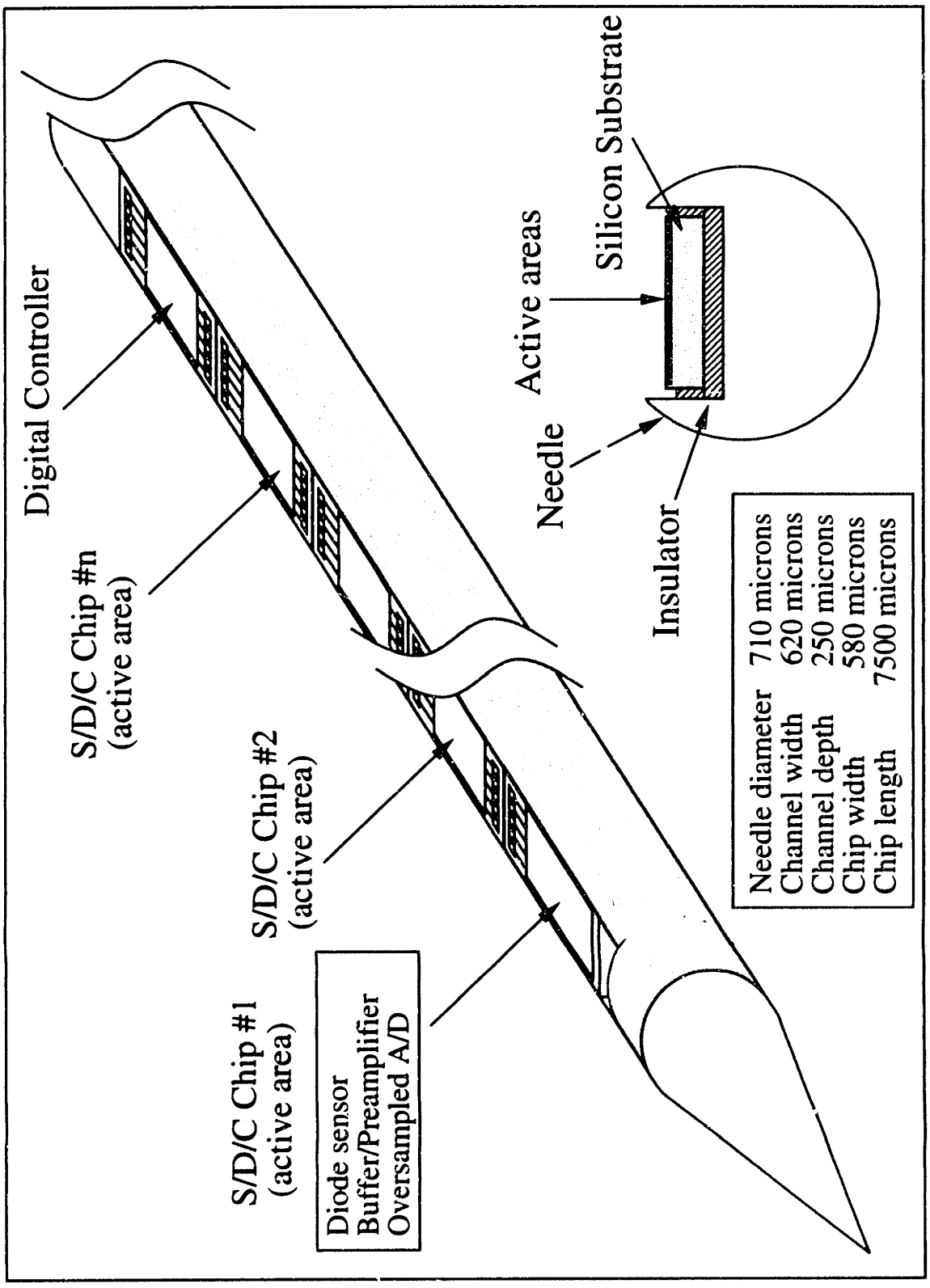


Figure 1.9: Needle/chip system

Chapter 2

Thermal Modelling

One of the fundamentally difficult issues with local on-site temperature measurement is that the sensor and detection circuits consume power. This power, dissipated through joule heating of the sensing system, couples into the sensed temperature and results in an artifact in the measurement. A first approach to solving this problem would be to minimize the consumed power; while this does indeed lower the artifact, it does not and cannot completely eliminate it. Instead, what is required is an understanding of the nature and behavior of this error, so that the data provided by the sensor can be correctly interpreted. The active needle geometry is quite complex, and involves heat transfer between macroscopic bulk materials and microscopic thin films. Consequently, any attempt to analytically examine the nature of the temperature artifact generated by the on-chip power dissipation quickly becomes an exercise in futility. This chapter examines the thermal behavior of the active needle system in greater detail using a large finite element computer model. Correlation of the simulation results is performed using comparison with an approximate analytical thermodynamic model. The results of these finite element studies are then examined and interpreted in the context of the temperature artifacts generated by the active needle system.

2.1 Finite Element Modelling

2.1.1 Model Setup

The most important aspect of the finite element analysis is the accurate representation of the relevant geometry and dominant thermal processes. In the case of the active needle system, the geometry was shown in figure 1.9; although this is the truest representation of the construction of the system, it is far more complex a model than necessary since most of the thermal processes of interest during the time of a single measurement will occur on smaller length scales. In addition, the computational complexity of fully modelling the entire needle would increase the solution time to unreasonable levels.

The major simplification that can be made is that the behavior of each of the sensors is approximately the same regardless of the location of the sensor on the needle. With this assumption, the model need only consider a section of the needle consisting of a single sensor and the material near it. Clearly the amount of material around the sensor that must be considered is determined by the expected region of thermal influence of the sensor; there is no need to finely model areas outside of this region since there will be, by definition, no significant perturbation of the thermal field in this area. The resulting model geometry is shown in figure 2.1. The needle is parallel to the z axis and is bounded by the planes $z = 0$ and $z = 4$ (all length units in the model are in cm). In order to model abutting sensors, silicon extends along this entire 4 cm distance in the groove of the needle; the excited sensor extends from $z = 1.6$ to $z = 2.4$; the sensor itself is at the $z = 1.6$ end.

Another simplification could be made depending on the nature of the problem being studied and the complexity/computation time tradeoff. If, for instance, the heat generation produced by the chip is considered uniform over the entire volume of the chip, the problem becomes symmetric in two dimensions and only a quarter "wedge" of the structure shown in figure 2.1 need be considered. Behavior in the other three quarters of the volume is determined from the assumed symmetry. This reduces the

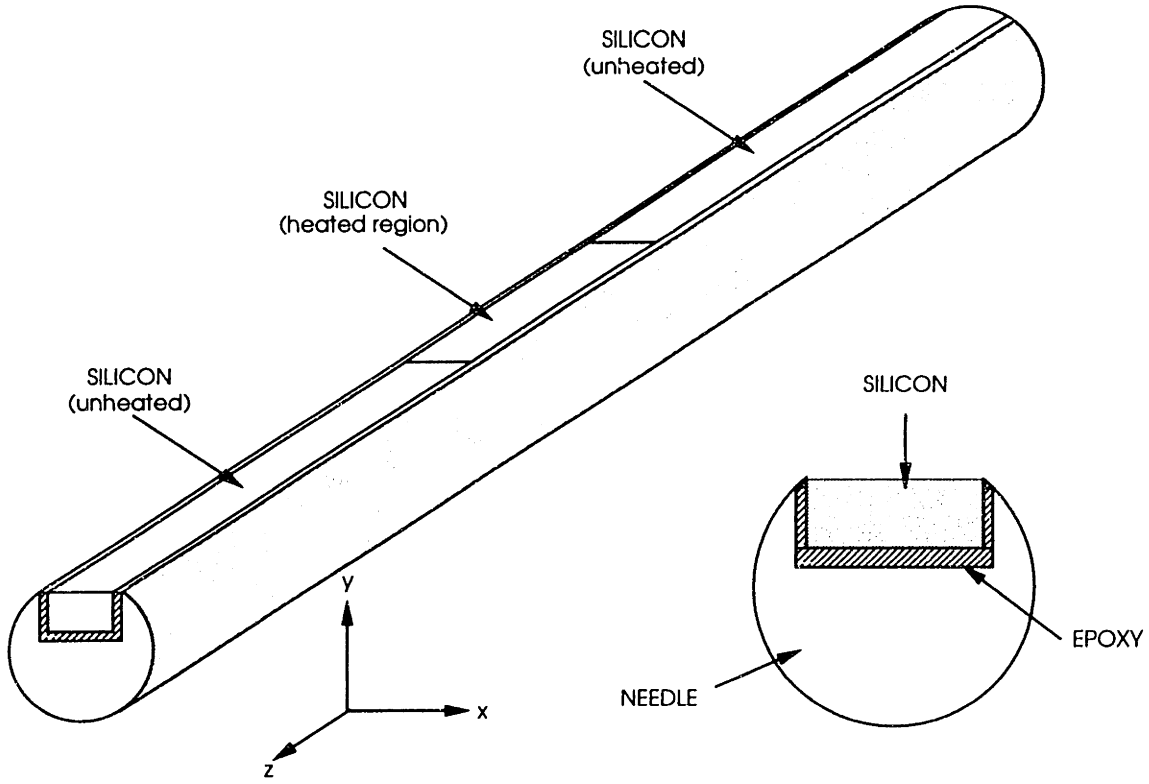


Figure 2.1: Thermal model geometry

computation time, but limits the heating functions that can be considered because of the symmetry requirement. In reality, the power dissipation (and hence the heating) will be nonuniform across the volume of the chip; this must be correctly modelled in order to accurately extract the short-time behavior of the system. Hence, these symmetry approximations are not made. As a result, the simulation time is longer, but the results of the simulation are more reliable, and the mesh that is developed can be applied to a wider range of problems.

Once the geometry has been selected, each of the components of the volume of interest must be characterized. The thermal problem consists of four significant materials. First is the tissue itself. This is modelled as a homogeneous material, with isotropic thermal conductivity, density, and heat capacity. The parameters for resting muscle tissue are used for this study: k_m = 5 mW/cm°C and $(\rho c_p)_m$ = 4.175 J/cm^3°C. Although the thermal properties of tissue vary depending on the exact tissue type, this

variation will not significantly alter the results since it is relatively small. In addition, since all of the heat generated by the chip is deposited in the tissue, the problem will scale with the thermal conductivity. The results obtained for one conductivity can therefore be easily adjusted if the thermal profile for a different tissue conductivity is desired.

The second material is the chip itself, which, for the most part, is monocrystalline silicon. Clearly, the thermal properties of silicon (k_s = 1.5 W/cm°C, $(\rho c_p)_s$ = 1.51 J/cm^3°C) are used to model this material. Technically, however, the chip is composed of several different materials grown or deposited during the fabrication process. These materials include silicon dioxide, polycrystalline silicon, boro-phospho-silicate glass (doped silicon dioxide), silicon nitride, and different types of doped silicon. Each of these films has different thermal properties, and should therefore be represented as separate regions in the thermal model. Doing so, however, drastically increases the complexity of the model, and does not offer a significant benefit, since the variation in thermal properties is minimal. In addition, the thickness of these films is so small, on the order of a few hundred atomic layers, that the thermal behavior of these each material will be significantly different than the behavior of the corresponding bulk material. Since the substrate silicon is much thicker, much larger in volume, and more conductive than the other thin films, and since the devices are in direct physical contact with the bulk silicon, using the thermal properties of silicon to model the entire chip is appropriate.

The third material is the needle, which for the purposes of this study is modelled as solid stainless steel. The relevant thermal properties of stainless steel are k_n = 250 mW/cm°C and $(\rho c_p)_n$ = 3.15 J/cm^3°C. Notice that the thermal conductivity of the stainless steel needle is significantly lower than that of the silicon itself, although the ρc_p products are on the same order. Since the diffusivity $\alpha = k/\rho c_p$, this implies that heat spreading will occur preferentially in the silicon and not the needle--over short time scales, the heat will diffuse significantly faster in the silicon than in any of the

Table 2.1: Thermal model parameters

Material	*Thermal Conductivity (mW/cm°C)*	ρc_p *(J/cm^3 °C)*	*Subscript*
Tissue	5	4.175	m
Silicon	1500	1.51	s
Needle	250	3.15	n
Epoxy	6	4.175	i

other materials. Although this is not important for the present work, this has interesting implications for a perfusion sensor based on the active needle concept: With proper heater geometry, the entire silicon chip can be considered isothermal.

The final material represented in the thermal model is the electrically insulating layer that isolates the individual chips from the carrier needle. This material can be either thermally insulating or thermally conducting, depending on the thermal characteristics desired. Using a thermally conducting layer will reduce the chip-induced temperature artifact, but will do this by "smearing" the thermal error over the entire needle. This will affect the measured thermal profile. A thermally insulating layer avoids this smearing, but increases the temperature artifact by making the chip look more like a "stand-alone" heater. Since the goal of this project is to develop a temperature sensing system that can be used with a blood perfusion sensor (in which the chip is indeed used as a heater), a thermally insulating layer is assumed. For the purposes of this model, the properties are k_i = 6 mW/cm°C and $(\rho c_p)_i$ = 4.175 J/cm^3°C. These values are approximately those of water and are very similar to the thermal properties of the tissue, and were selected because they are typical of many insulating epoxies. Table 2.1 summarizes the various thermal parameters for the model.

2.1.2 Boundary Conditions/Heating Functions

The placement of the boundaries and the boundary conditions are based on the assumption that the temperature perturbation of the tissue approaches zero as the

distance from the heat source (the chip) increases. The boundaries are therefore modelled as constant temperature surfaces; since the problem examines the relative thermal behavior of the system, this temperature can be arbitrarily chosen. For simplicity, it is fixed at $T_o = 0$, so that the temperature field produced by the simulation represents the temperature perturbation above the baseline tissue temperature. The problem is to determine how far away the boundary can be reasonably placed: If the simulation volume is very large, the computation becomes very difficult because of the widely varying density of nodes. If the boundaries are placed too close to the heat source (the chip), the solution error will be large because the boundaries will artificially influence the solution: They will be artificially forcing the temperature error to zero within the region of significant thermal perturbation.

The thermal region of influence of the problem can be determined by realizing that the heat transfer is due entirely to conduction from the chip to the surrounding materials. Thus, the characteristic length is the diffusion length, and can be found by examining an analogous conduction problem, namely, the heat transfer associated with a point source in an infinite medium of conductivity k, diffusivity α and thermal mass ρc_p. The general solution for the temperature $T(r, t)$ is [47]:

$$T = \frac{1}{8(\pi\alpha)^{\frac{3}{2}}} \int_0^t \frac{Q(t')e^{-\frac{r^2}{4\alpha(t-t')}}}{\rho c_p} \frac{dt'}{(t-t')^{\frac{3}{2}}} \tag{2.1}$$

In the case where $Q(t) = Q_o\delta(t)$, the impulse response of the system is:

$$T = \frac{Q_o}{8\rho c_p(\pi\alpha)^{\frac{3}{2}}} \cdot \frac{e^{-\frac{r^2}{4\alpha t}}}{t^{\frac{3}{2}}} \tag{2.2}$$

It is clear that the exponential term governs the spreading of the thermal field over time; this term can be expressed as $e^{-\frac{r^2}{L^2}}$, where $L = 2\sqrt{\alpha t}$ is the characteristic length of the problem. Note that the characteristic length is a function of time, and hence there will always be a time t for which the boundary placement will "interfere" with the solution. Using the fact that

$$\alpha = \frac{k}{\rho c_p} \tag{2.3}$$

it is found that the solutions obtained with this placement will be valid for times $t < \frac{L^2 \rho c_p}{4k}$. Transient solutions will be valid for times that fall in this interval. Since the thermal field theoretically becomes infinite in extent at infinite time, steady state solutions are valid only when the boundaries are placed far enough away so that the field is near steady state when the thermal field reaches the boundary. This is equivalent to saying that the temperature perturbation created by the presence of the boundary is negligible.

In the problem considered here, four of the six boundary surfaces surrounding the needle are only in contact with the tissue, and were placed 1 cm away from the needle; the mesh generation software was unable to produce a usable mesh for distances significantly larger than 1 cm. With this value for L, and the values of k and ρc_p for tissue given above, the interval of validity of transient solutions obtained with this mesh is $0 \leq t < 208$ sec. The two remaining surfaces (the front and back surfaces in figure 2.1) intersect the tissue, the needle, and the unheated silicon. Since the thermal conductivities of the needle and the silicon are so much larger than the tissue, these boundaries were placed 1.6 cm away from the heat source to allow greater room for the spread of heat. If the parameters for the stainless steel are used, it is found that the solutions are valid for times $t < 6.1$ sec. Clearly this is much shorter than the time associated with the transient response of the tissue, as would be expected. However, since the thermal field is nearly steady state at this time, the results of the simulations of interest, namely, those associated with taking a measurement, will be valid with these boundaries.

The most important element of the simulation is the modelling of the power dissipation on the chip. This joule heating is modelled as a source of internal heat generation in the silicon. In order to more accurately model the behavior without significantly increasing the computation time, a piecewise constant heating function is used to model the varying power dissipation. The volume of the chip is divided into three sections, each of which has an internal heat generation rate equal to the power

dissipated in that section divided by the volume of the section. The power dissipation is not so significantly nonuniform over the volume of the chip to warrant a larger number of sections; separate sections are used primarily to differentiate the power dissipation in the sensor itself from the power dissipation in the signal conditioning and data conversion circuitry.

The first section models the power dissipation of only the temperature sensor circuitry. This accounts for only a small volume of the chip, approximately $V = 2.4 \times 10^{-4}\,\text{cm}^3$. In this region, the power dissipation is approximately P=1.24 mW; this number is based on the design specifications for the circuit. The internal heat generation in this region is therefore $P/V = 5.16\,\text{W/cm}^3$. The second subdivision of the chip models the power dissipation in the analog modulator (the front end of the A/D converter). This is where most of the power (approximately 2.5 mW) is dissipated, but it also occupies the largest volume of the chip, $6 \times 10^{-4}\,\text{cm}^3$. Thus, the internal heat generation of this section is $4.17\,\text{W/cm}^3$. The final section models the power dissipation in the biasing and control circuitry. This region consumes a power of .86 mW and occupies a volume of $0.6 \times 10^{-4}\,\text{cm}^3$, resulting in an internal heat generation rate of $14.33\,\text{W/cm}^3$. Note that all three of the values for the volumetric heat generation are on the same order of magnitude. This is a direct consequence of the design and layout, as will be shown in Chapter 3; since the circuits used in the modulator and the sensor share many common pieces, the power dissipation per unit volume in those areas is very similar. The large power density in the bias and control circuitry is due largely to a much higher density of circuitry in this area.

2.1.3 Perfusion

An additional factor influencing the proper behavior of the model is the tissue perfusion, which will effectively enhance the removal of heat from the chip. Without accounting for perfusion, the temperature perturbation predicted by the model would always be larger than the actual temperature error. Although perfusion will always be present to

some extent in the interrogated tissue, the perfusion can vary widely depending on the exact tissue type. In order to maintain the conservative characteristics of the unperfused model while at the same time increasing the accuracy of the solutions, a perfusion of 5 ml/100g-min, typical of resting muscle tissue, is used in the model.

The perfusion itself is modelled as a temperature dependent heat sink function in the tissue [48]. This formulation can clearly be seen by examining the heat conduction equations in each of the four materials. In the needle and the insulating layer, there is no internal heat generation, and the temperature satisfies the conduction equation:

$$k_m \nabla^2 T = (\rho c_p)_m \frac{\partial T}{\partial t} \tag{2.4}$$

where k and ρc are the thermal conductivity and thermal mass as given above. The chip itself is affected by conduction as well as internal heat generation due to the power dissipation:

$$k_s \nabla^2 T + Q(r, T, t) = (\rho c_p)_s \frac{\partial T}{\partial t} \tag{2.5}$$

where the term $Q(r, t)$ is the internal heat generation function. In the most general case, this heat generation could be a function of space, temperature, and time, although in the model examined here it is a function of space only. The thermal field in the tissue is governed by the bioheat transfer equation [49]:

$$k_m \nabla^2 T + q_{met} - \omega \rho_{bl} c_{bl} (T - T_o) = (\rho c_p)_m \frac{\partial T}{\partial t} \tag{2.6}$$

where the subscript m refers to the properties of the tissue and bl refers to the properties of blood, ω is the blood perfusion, and q_{met} is the metabolic heat generation. In this analysis, this term is neglected since it is assumed small compared to the perfusion term. Since $T_o = 0$, the last term on the left hand side can be written as $\omega \rho_{bl} c_{bl} T$. Comparing this equation with equation 2.5, it is clear that perfusion behaves exactly like a heat generation Q_{eff} in the tissue, where:

$$Q_{eff} = -\omega \rho_{bl} c_{bl} T \tag{2.7}$$

The negative sign in Q_{eff} reflects the fact that heat is lost through perfusion. Unlike the power dissipation in the chip, the heat generation term is a function of temperature only and not space. This is accounted for in the model by updating Q_{eff} each time step, using the most recent nodal temperature to compute Q_{eff} at that node. This capability is built in to the finite element simulator so no modifications to the software are required.

2.1.4 Simulation Results

The model as described was used to perform four simulations: The first two simulations calculated the steady state temperature distribution in both the perfused and unperfused case. These simulations were used to correlate the scale of the model by comparing the finite element solution with an approximate analytical model for which the solution was known. The analytical model used and the results of these simulations are discussed further in section 2.2 below.

The second two simulations examined the temperature artifact due to the chip power dissipation over the course of a single measurement experiment. Since the measurement time is short relative to the thermal time constants of the system, steady state is not reached; transient simulations are required. The simulations provide both the temporal behavior of the artifact and, more importantly, the maximum artifact that can be expected. As before, one simulation was performed assuming zero perfusion and another was performed assuming minimal expected perfusion.

The results of the transient simulations are shown in figures 2.2 through 2.5. Figures 2.2 and 2.3 show the results when perfusion is not included. Figures 2.4 and 2.5 show the results with perfusion. Figure 2.2 shows the temperature distribution across a plane containing the surface of the chip, where maximum heating will occur. The temperature is represented in greyscale, with white corresponding to the highest temperature elevation. The front edge corresponds to the plane $z = 4$; the back edge corresponds to the plane $z = 0$. The sensor at $z = 1.6$ is located near the back end. Figure 2.3 shows the temperature profile in this plane along a line parallel to the

needle that cuts the top surface of the chip in half. Notice that, as would be expected, maximum heating occurs near one side of the long axis of the chip, where the internal heat generation is largest. At the sensor end, the artifact is significantly lower but is still on the same order of magnitude as the peak error. The qualitative results are the same for the perfused case, which looks almost identical to the nonperfused case. The numbers show that the peak temperature error is slightly lower as would be expected from the perfusion-enhanced heat transfer, but at such a low perfusion the effects are minimal and no significant reduction is seen. Figure 2.4 shows the same planar surface as figure 2.2 with perfusion; figure 2.5 shows the temperature error along the same line as figure 2.3, but again with the effects of perfusion included.

2.2 Analytical Correlation of the Model

In order to correlate the scale of the finite element model, an approximate steady state analytical model was developed. Because of the complex geometry of the problem, finding an analytical model that closely approximates the actual heat transfer in both the transient and steady state and for which a simple closed form solution exists is quite difficult. Consequently, the approach is to examine the steady state solution produced by the finite element simulator and compare it with a much simpler steady state model for which a closed form solution is known. The results from the analytical model can then be used to coarsely predict changes in the temperature error that would occur with changes in the various model parameters (geometry, power distribution, etc). This section discusses the approximate model used and the correlation of the model with the finite-element solution.

In the steady state, it is assumed that all heat generated by the chip is deposited in the tissue, either through direct heat transfer with the tissue or through indirect heat transfer through the needle. Since the conductivity of the needle is significantly larger than that of the tissue, it is assumed that the heat is deposited indirectly into the tissue through the needle. The steady state problem is therefore modelled as a semi-infinite cylindrical

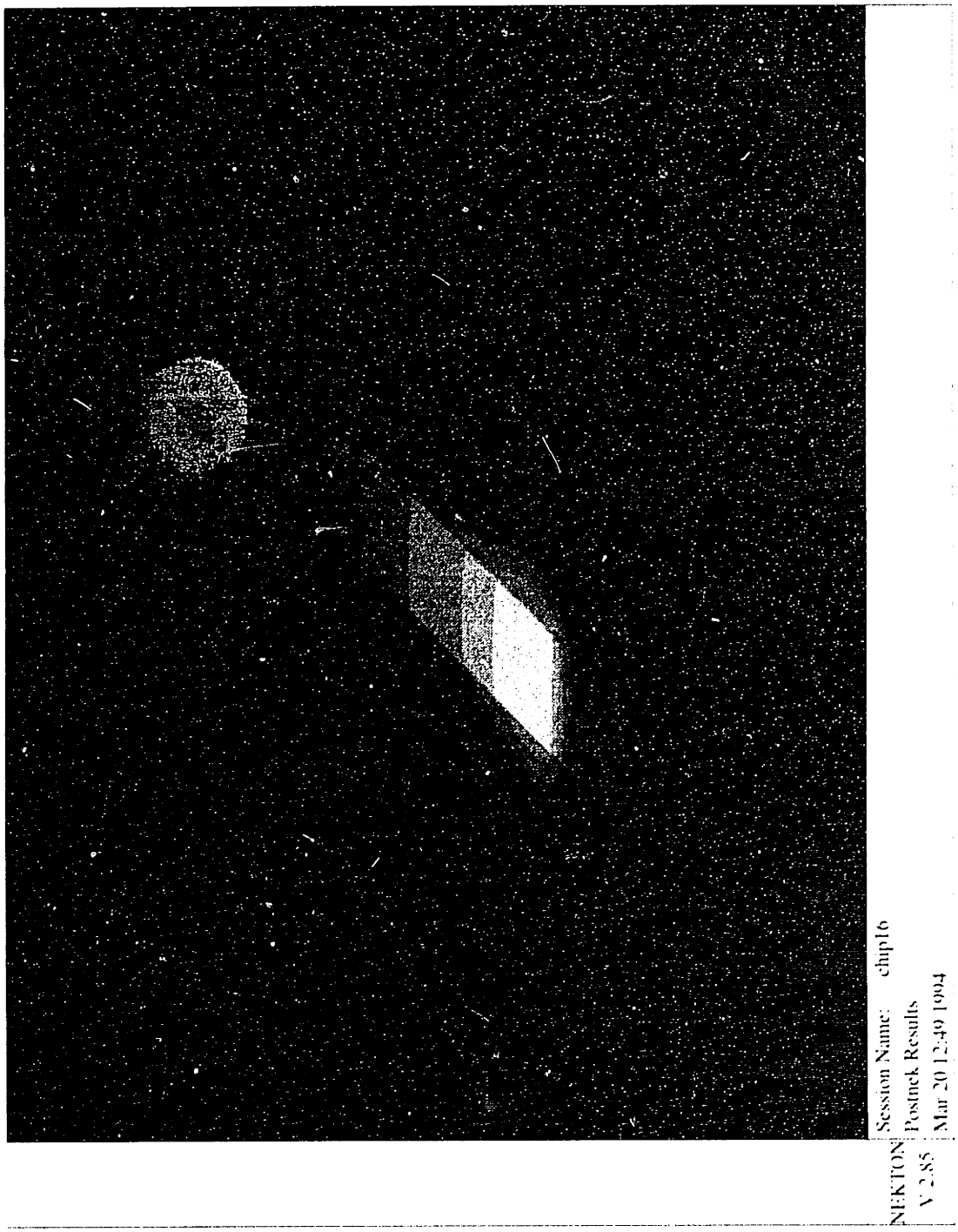


Figure 2.2: Temperature error in the plane containing the chip surface, without perfusion

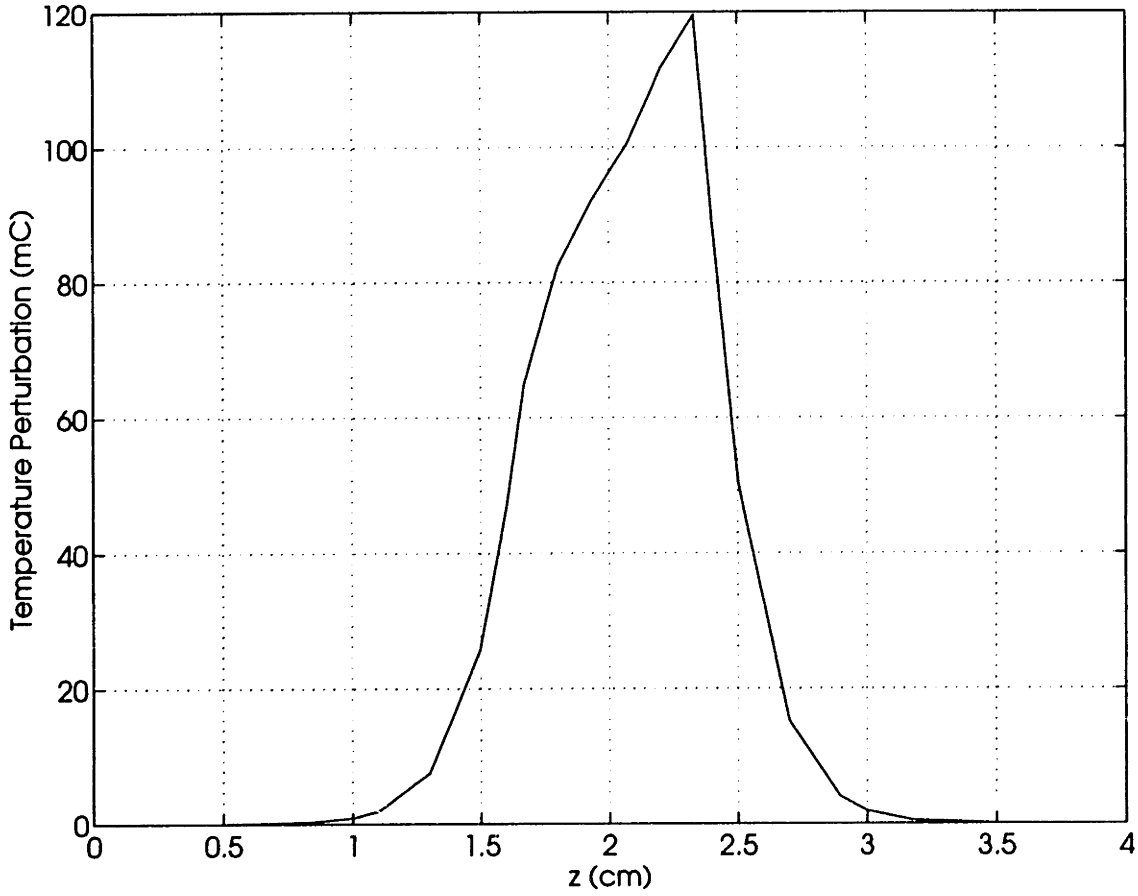


Figure 2.3: Temperature error along a line parallel to the needle, cutting through the middle of the chip surface, without perfusion

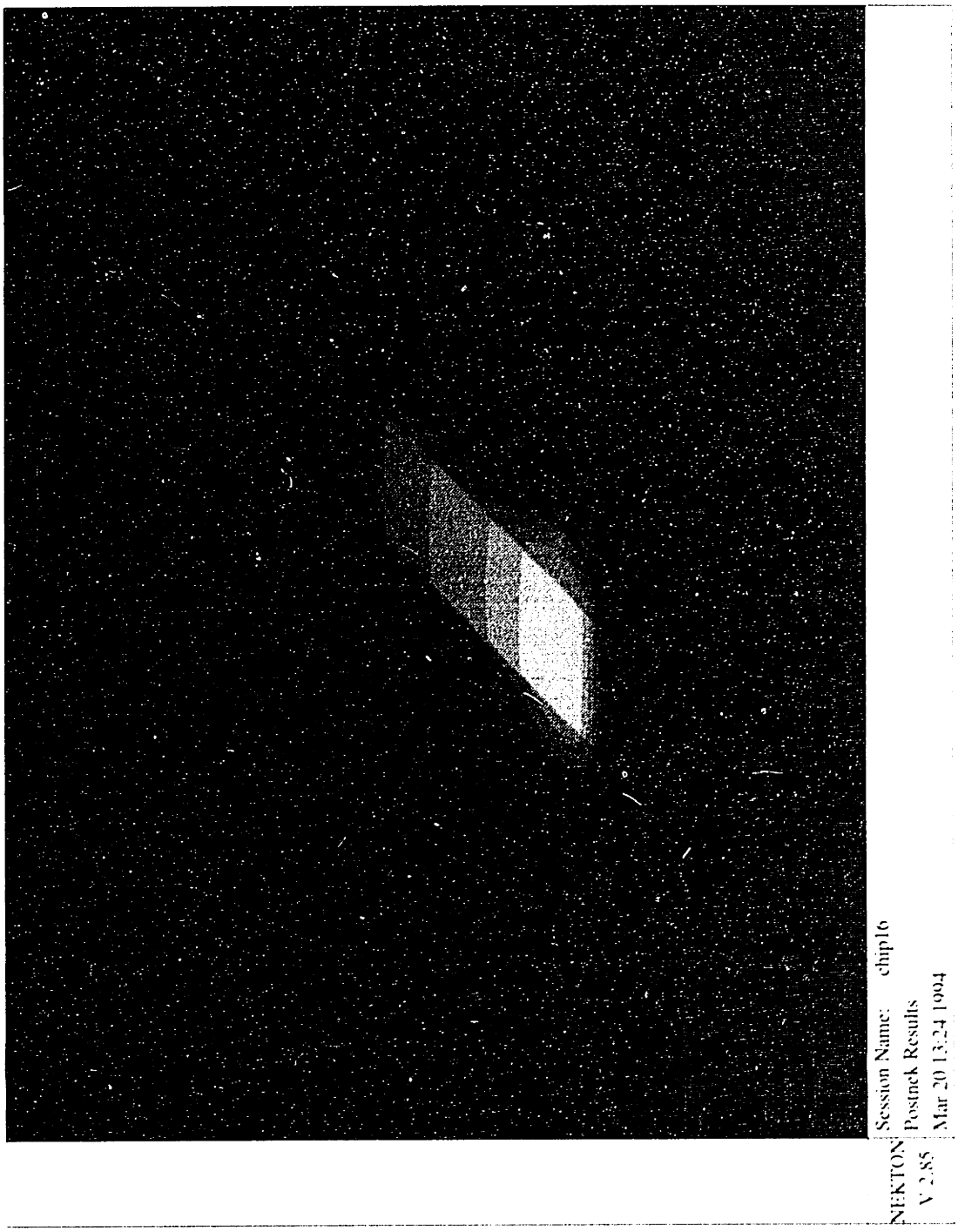


Figure 2.4: Temperature error in the plane containing the chip surface, with perfusion

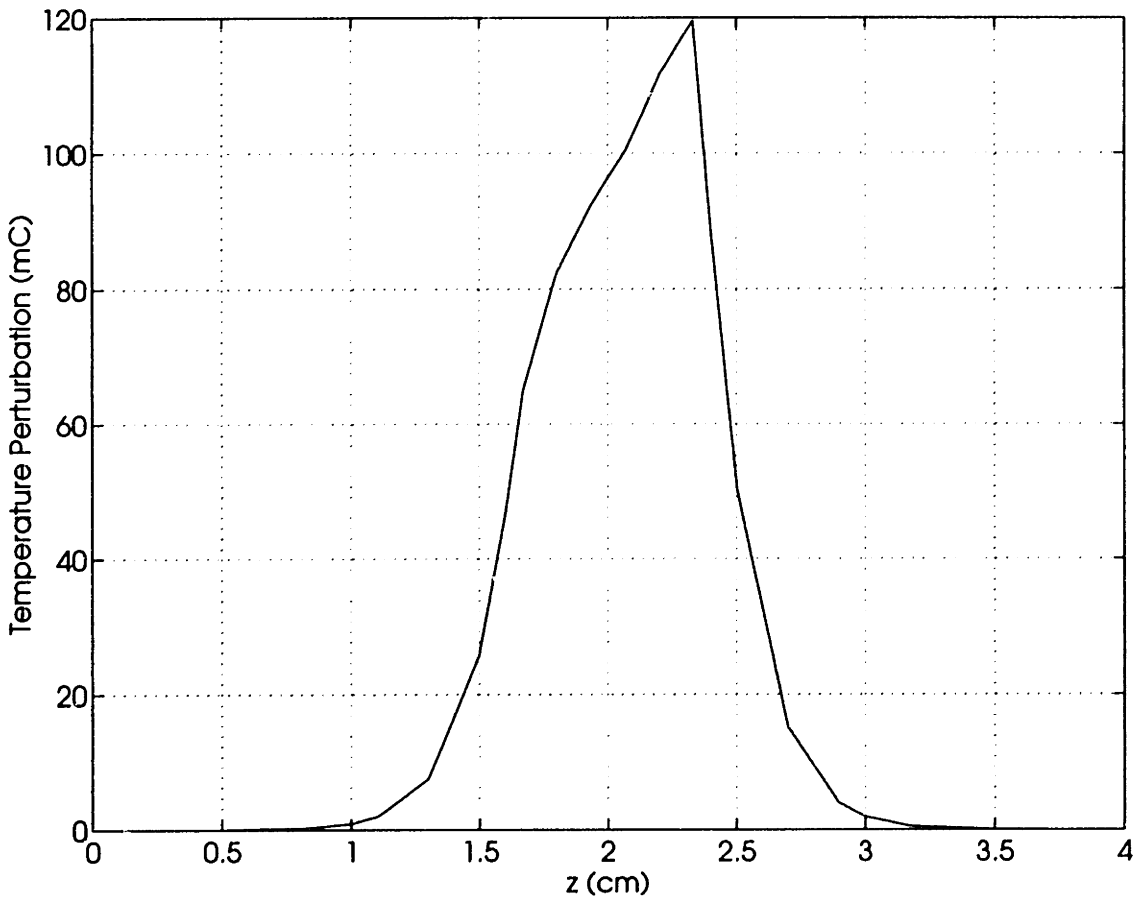


Figure 2.5: Temperature error along a line parallel to the needle, cutting through the middle of the chip surface, with perfusion

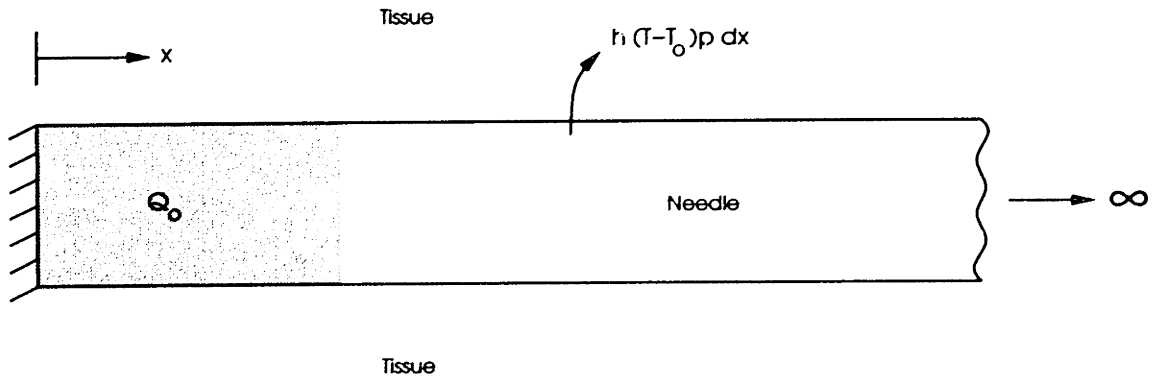


Figure 2.6: Approximate steady-state analytical model

fin problem, in which one end of the fin of cross-sectional area A and perimeter p is perfectly insulating and the other extends to infinity. The chip power dissipation is modelled as internal heat generation in the fin Q_o occurring from the insulating boundary out to a finite distance d. Heat transfer from the fin to the surrounding tissue is described by the heat transfer coefficient h. The thermal conductivity of the fin is k_n from above. This approximate model is shown in figure 2.6.

The internal heat generation rate Q_o for the analytical model is calculated using the actual chip power dissipation and dividing it by the volume of the needle segment over which heat generation occurs. This volume is clearly Ad. A is the cross sectional area of a 22 gauge needle, $4 \times 10^{-3}\,\mathrm{cm}^2$. Because of the insulating boundary, the problem models the heat transfer of half of the chip (the problem is symmetric with respect to the insulating boundary), and the distance d is half the chip length, approximately 3.5 mm. The total power dissipated by an individual sensor chip is approximately 4.6 mW; the power dissipation in the half-needle model is therefore 2.3 mW. With these numbers, $Q_o \approx 1.6\,\mathrm{W/cm}^3$.

The heat transfer coefficient h is the most important parameter and must be treated carefully. By definition, the heat transfer from the needle to the tissue over a differential length of the fin is given by

$$dq = h(T - T_t)p\,dx \tag{2.8}$$

Clearly, the temperature error will be highly correlated with the value of h since this heat transfer is the dominant heat loss mechanism. Qualitatively, there must be two components to h, a fixed (constant) term and a perfusion dependent term. This is clear from the physical problem: Even in the absence of perfusion, heat will be lost to the tissue. Therefore h is nonzero at zero perfusion. In the presence of perfusion, the heat transfer is enhanced, and therefore h must increase. The values of h used for this analysis are $h = .143\,\mathrm{W/cm^2{}^\circ C}$ in the absence of perfusion and $h = .172\,\mathrm{W/cm^2{}^\circ C}$ in the minimally perfused case. These values are calculated from measurements taken from existing biomedical temperature sensors.

The general solution to the problem is given by

$$\Delta T(x) = \frac{Q_o A}{Ph}\left(1 + \frac{\cosh(mx)}{\cosh(md) - \sinh(md)}\right) \tag{2.9}$$

where

$$m^2 = \frac{hp}{kA} \tag{2.10}$$

The peak temperature will occur at $x = 0$, the plane of symmetry. For the numbers used in this analytical model, this gives a peak temperature rise of .55°C. Comparing this value with the peak chip temperature obtained in the steady state with the finite-element model, it is clear that although the values are not equal, they are of the same order of magnitude, as expected. Although this does not provide verification of the simulation results to high accuracy, it does indeed verify that the solutions generated by the computer are indeed "reasonable" for this problem.

2.3 Interpretation of Results

The results of the simulation reveal that the temperature artifact that is caused by the on-chip power dissipation is approximately 46 m°C, which is quite large compared to the desired temperature resolution of 1 m°C. This result must be interpreted in terms of the use of this temperature sensor, however. In a strictly temperature monitoring mode, resolution at 46 m°C is certainly adequate; the temperature sensor design specifications

were selected so that the chip could be used as part of a future perfusion measurement system, and that is where the high resolution is important. In the perfusion measurement application, the sensor will be used to monitor an applied temperature step, as explained in Chapter 1. In this scenario, a heating element of some sort will be present on the chip, and will be used to generate the temperature increment. Thus, the chip will be purposely heated; this applied temperature step will be much larger than the small temperature artifact created by the power dissipated in the sensor. In short, the artifact will have no significance as long as the power supplied to the sensor is considered part of the total heater power. Presumably, in the case of a perfusion sensor, all of the power supplied to the chip will be measured, since all power supplied to the chip will be dissipated as joule heating, whether or not it is supplied to the heating element. Therefore, the temperature artifact does not affect the resolution of the sensor in the application for which it was designed.

In addition, the temperature artifact is clearly a function of time, and higher resolution measurements can be made by reducing the measurement interval; the one second interval assumed here is in fact larger than the anticipated measurement time, which is variable but will typically be less than half a second. At half a second, the artifact drops to 29 m°C; at one-quarter of a second the error is only 18 m°C. There is a fundamental tradeoff involved, however, in that the shorter the measurement interval, the less bits the modulator produces, which may compromise the A/D resolution. The key points are that the artifact does not represent a hard limit to the measurement resolution in the application for which it is intended, and that the artifactual effects can be altered by varying the system parameters.

What, then, is the real effect of the temperature artifact? It represents the limit of the resolution that can be obtained in a strictly passive transient temperature sensing mode. Since the peak error at the sensor over a typical measurement sampling interval is 29 m°C, the sensor will not be able to accurately resolve temperature changes below this error when used in a transient mode. In a steady-state situation in which the chip

is allowed to equilibrate with the medium under interrogation, the temperature artifact appears as an offset in the temperature measurement and can be subtracted out; the full resolution of the sensor chip can therefore be used. Since the temperature resolution requirement in a passive sensing mode is much less stringent (on the order of .1°C), this artifact will not prevent this chip from being used in a passive (non-heating) sensing mode. Reducing the power dissipation on the chip would permit greater resolution of temperature in this mode, but would not alter the resolution of the sensor in an "active" heated mode, where the temperature of the chip is purposely elevated.

Chapter 3

The Temperature Sensor

Clearly the most critical component of the temperature measurement system is the temperature sensor itself, since the sensor defines the maximum possible resolution achievable. For an integrated system, the choice of sensor is limited by the manufacturing process. As a result, discrete systems usually outperform integrated systems. With recent technological developments and advanced circuit design techniques, however, the resolution gap between discrete and integrated systems is closing. This chapter first discusses the temperature behavior of p-n junction diodes, the core element in most integrated temperature sensing systems. Several sensing circuits and their limitations are discussed, including a novel technique, the "chopped PTAT," developed as part of this project. Finally, the actual circuit implementation of the sensing scheme used on the active needle is presented, along with a detailed analysis of the sensor performance in the biomedical temperature range of interest.

3.1 Diode Temperature Sensors

The temperature sensors used in the active needle system are p-n junction diodes, as they are one of the most temperature sensitive devices that can be manufactured easily within the framework of a standard CMOS process. The devices are easy to operate and use little power, minimizing self heating of the sensor. Finally, the noise characteristics

of p-n diodes are significantly better than any of the other field-effect devices available in a standard CMOS process.

3.1.1 Basic Theory

The temperature behavior of the forward voltage of a p-n junction diode can be derived from the diode current equation:

$$I_D = I_S \left(e^{\frac{V_D}{V_{TH}}} - 1 \right) \tag{3.1}$$

where I_D is the diode current, I_S is the reverse saturation current, V_D is the diode voltage, and V_{TH} is the thermal voltage ($\frac{kT}{q}$). This can be rewritten in terms of the forward voltage V_D:

$$V_D = V_{TH} \ln \left(\frac{I_D}{I_S} + 1 \right) \tag{3.2}$$

The temperature behavior of the saturation current I_S is [50]:

$$I_S = AT^{\beta} e^{-\frac{V_{go}}{V_{TH}}} \tag{3.3}$$

where A and β are (temperature independent) material dependent parameters, and V_{go} is the bandgap voltage of silicon at absolute zero, approximately 1.205 V. This value as defined is temperature independent; the temperature dependence of the bandgap is lumped into β. If we assume that the temperature dependence of the excitation current is of the form:

$$I_D = BT^{\alpha} \tag{3.4}$$

where B is again a temperature independent process parameter, then the explicit functional form of $V_D(T)$ is:

$$V_D(T) = V_{TH} \ln \left(\frac{B}{A} T^{\alpha - \beta} e^{\frac{V_{go}}{V_{TH}}} + 1 \right) \tag{3.5}$$

If we assume that the diode current is much larger than the reverse saturation current ($I_D \gg I_S$), then the expression simplifies further to

$$V_D = V_{TH} \ln \left(\frac{B}{A} \right) + V_{TH}(\alpha - \beta) \ln(T) + V_{go} \tag{3.6}$$

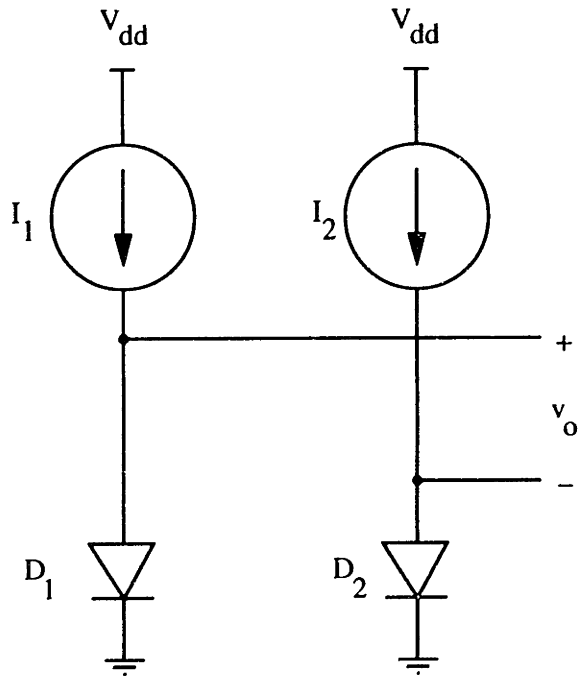


Figure 3.1: Difference-of-v_D circuit

which gives the desired temperature dependence in terms of the temperature independent parameters A and B and the bandgap voltage V_{go}.

3.1.2 Temperature Sensing Circuits

The derivation above expresses the relationship between the diode voltage and temperature when the excitation is a current source whose own temperature dependence is known. From a practical standpoint, there are two major problems with using this direct approach. First, there is the difficulty in deriving the inverse relationship, namely, the temperature as a function of the measured diode voltage. This problem stems from the $V_{TH} \ln\left(\frac{B}{A}T\right)$ term in the governing equation; the output voltage is very nonlinear, which complicates the signal processing as well as the initial sensor calibration. Second, there is the large turn-on voltage of the diode. Changes of -2 mV/°C are much more difficult to detect when they are superimposed on a nominal 0.7 V bias voltage.

Both of these problems can be avoided by using two diodes in a difference-of-V_D

configuration as shown in simplified form in figure 3.1. In this case, the output voltage is given by:

$$\begin{aligned} V_o &= V_{D1} - V_{D2} \\ &= V_{TH} \ln\left(\frac{B_1}{A_1}\right) - V_{TH} \ln\left(\frac{B_2}{A_2}\right) \\ v_o &= V_{TH} \ln\left(\frac{A_2 B_1}{A_1 B_2}\right) \end{aligned} \tag{3.7}$$

where it is assumed that the temperature behavior of the driving sources are the same, as is the temperature behavior of the reverse saturation current. This relationship between the output voltage and temperature is linear; rearranging the equation gives explicitly:

$$T = \frac{V_o}{\ln\left(\frac{A_2 B_1}{A_1 B_2}\right)} \cdot \left(\frac{q}{k}\right) \tag{3.8}$$

$$T = \left(\frac{q}{k}\right) \cdot \frac{V_o}{\ln\left(\frac{I_1}{I_2}\right)} \tag{3.9}$$

The corresponding temperature sensitivity is therefore:

$$\frac{\partial v_o}{\partial T} = \left(\frac{k}{q}\right) \ln\left(\frac{I_1}{I_2}\right) \tag{3.10}$$

It is important to note also that the turn-on voltages of the two diodes effectively cancel each other when the difference is taken. The end result is that both of the problems outlined above have been avoided by using the difference in the diode voltages to sense the temperature.

There are now two primary sources of error. The first error is caused by mismatch in A for the two diodes; this problem is unavoidable due to random process variations. This causes both an offset and a gain error, but the linear nature of the measurement is preserved. As a result, this error can be quantified and eliminated using a simple two point calibration. Since these process variations are typically small, the error usually does not adversely affect the signal processing circuit design. The second error is caused by mismatch in the current sources and in their temperature coefficients

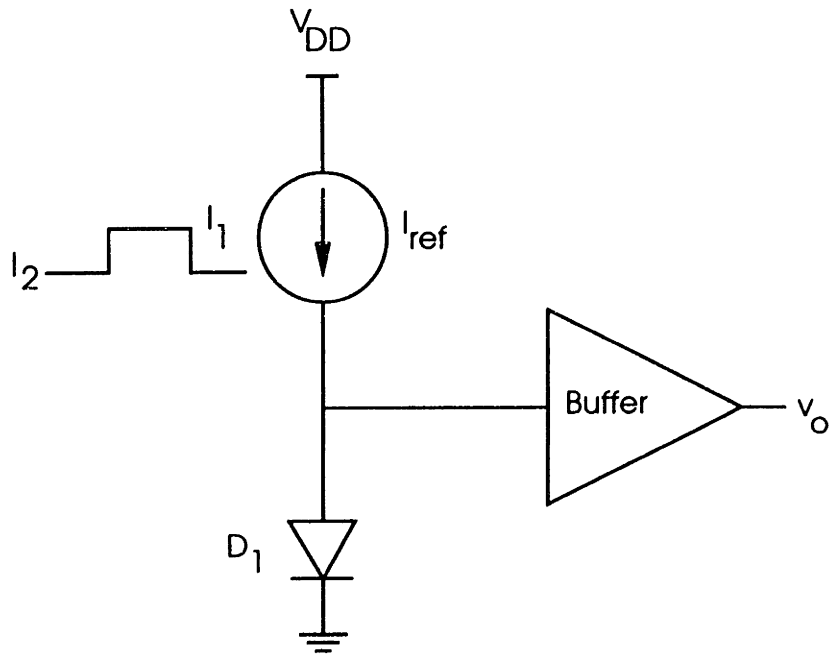


Figure 3.2: Sensor independent measurement scheme

(mismatch in the B and α parameters). This is a more difficult problem to deal with since the nature of the problem depends on the actual implementation. Fundamentally, errors in B affect the measurement in the same way that errors in A do, as can be seen in equation 3.8. Errors in α, however, produce a nonlinearity in the measurement that may or may not be significant depending on the exact implementation used.

One way to avoid these mismatch errors is to use the same diode and the same current source for both "halves" of the difference circuit [51]. This method, shown in figure 3.2, eliminates the effect of mismatches between both the current sources and the diodes. The current source is square wave modulated in order to provide two diode voltages that can be subtracted as described above. When the current source output is I_1, the output voltage is v_{D1}. This voltage is sampled and stored. The current source output is then changed to I_2. This voltage is also sampled and subtracted from the stored voltage v_{D1}. The difference voltage follows equation 3.9 above. The ratio $\frac{B_1}{B_2}$ is equal to the ratio of the currents $\frac{I_1}{I_2}$. The areas A_1 and A_2 must be equal since the diode is the same for both measurements. There is no current source temperature coefficient

mismatch either since the same current source is used for both measurements. In short, the measurement has become sensor and current source independent.

The sensor independent scheme eliminates the mismatch error at the expense of other performance parameters, however. First, the fully differential nature of the difference-of-v_D circuit is lost. As a consequence, power supply noise now couples directly into the measured signal, unlike the differential case where such noise appears as a common mode signal and is rejected. This is true not only of the sensor but the buffer amplifier also--a single-ended amplifier will suffer more from power supply noise than a comparable fully differential amplifier. This power supply noise as well as substrate noise coupling through the silicon wafer prevent this method from being used for high resolution measurement with current technology. Consequently, the ''best of both worlds'' is a sensing circuit combining the sensor independent nature of the modulated square wave scheme with the fully differential nature of the difference-of-v_D scheme.

Such a scheme has been developed as part of this project. The technique, outlined in figure 3.3, is a ''chopped PTAT'' sensing circuit. It is so named because of the chopper technique used to acquire the signal. Instead of chopping a single signal, however, the ''chopped'' waveform is actually two signals--one from the first diode and one from the second. Each diode is configured in the sensor-independent scheme, where the signal of interest is generated by modulating the current through a single diode. In this case, the square waves of current used to excite the diodes are the same, but out of phase by 180°. The resulting differential voltage across both diodes is therefore a chopped temperature signal: it is the PTAT voltage that would have been generated using the traditional PTAT circuit. Now, however, the signals from each diode are independent of one another and there is no error due to diode mismatch. Since the same two current sources are used to excite both diodes, there is no error due to current mismatch either.

Chopping these signals not only retains the device-independent sensing, but also eliminates the effects of instrumentation amplifier offset, as in the traditional chopper

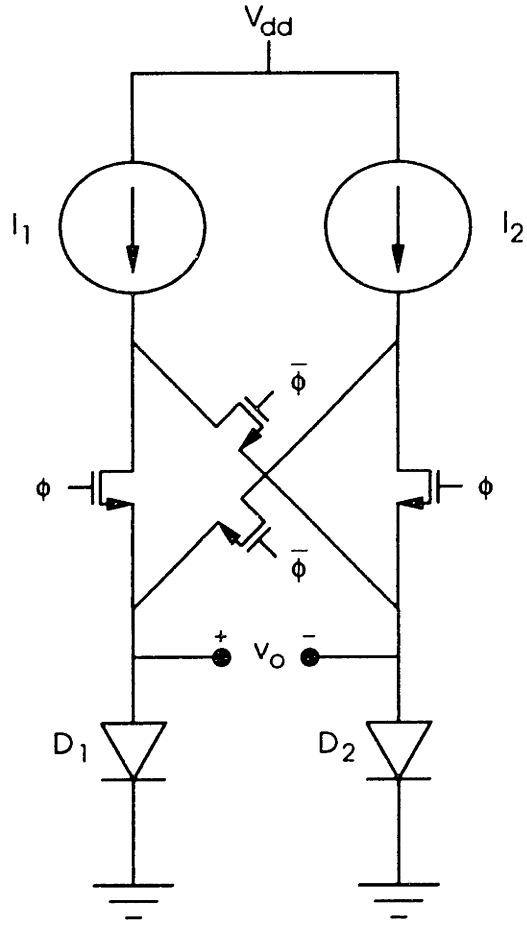


Figure 3.3: Chopped PTAT sensing scheme

amplifier. The chopping also reduces the noise requirements of the amplifier, as the signal of interest is modulated in the frequency domain to frequencies where $1/f$ noise is reduced. These effects can best be seen through a mathematical treatment of the problem.

The square wave of voltage that is developed across the first diode (D_1) is, over one period:

$$v_1(t) = \begin{cases} V_{TH} \ln\left(\frac{I_1}{I_{S1}}\right) & -\frac{T}{4} \leq t \leq \frac{T}{4} \\ V_{TH} \ln\left(\frac{I_2}{I_{S1}}\right) & t < -\frac{T}{4}, t > \frac{T}{4} \end{cases} \tag{3.11}$$

where I_1 and I_2 are the two currents used to modulate the diode voltage, I_{S1} is the diode saturation current, and T is the period of the square wave. The voltage across D_2

is, analogously:

$$v_2(t) = \begin{cases} V_{TH} \ln\left(\frac{I_2}{I_{S2}}\right) & -\frac{T}{4} \le t \le \frac{T}{4} \\ V_{TH} \ln\left(\frac{I_1}{I_{S2}}\right) & t < -\frac{T}{4}, t > \frac{T}{4} \end{cases} \tag{3.12}$$

It is important to note that the signal of interest, V_{TH}, is also a function of time. The Fourier transform of v_1 is easily derived from the transform of a square wave, taking into account the DC component and the amplitude scaling:

$$V_1(f) = V_{TH}(t) * \left[\frac{1}{2}\ln\left(\frac{I_1 I_2}{I_{S1}^2}\right) + \ln\left(\frac{I_1}{I_2}\right) \sum_{\substack{n=-\infty \\ n\ odd}}^{\infty} \frac{(-1)^{\frac{n-1}{2}}}{n\pi} \delta(f - \frac{n}{T})\right] \tag{3.13}$$

The transform of v_2 is similarly:

$$V_2(f) = V_{TH}(t) * \left[\frac{1}{2}\ln\left(\frac{I_1 I_2}{I_{S2}^2}\right) + \ln\left(\frac{I_2}{I_1}\right) \sum_{\substack{n=-\infty \\ n\ odd}}^{\infty} \frac{(-1)^{\frac{n-1}{2}}}{n\pi} \delta(f - \frac{n}{T})\right] \tag{3.14}$$

The transform of the differential output is the difference of the transforms $V_1(f)$ and $V_2(f)$, and one obtains as the final answer:

$$V_0(f) = V_{TH}(t) * \left[\ln\left(\frac{I_{S2}}{I_{S1}}\right) + 2\ln\left(\frac{I_1}{I_2}\right) \sum_{\substack{n=-\infty \\ n\ odd}}^{\infty} \frac{(-1)^{\frac{n-1}{2}}}{n\pi} \delta(f - \frac{n}{T})\right] \tag{3.15}$$

These functions are shown pictorially in figure 3.4: Figure 3.4(a) shows an arbitrary temperature spectrum $V_{TH}(f)$.[1] Modulation of the temperature signal across each of the sense diodes results in the spectra $V_1(f)$ and $V_2(f)$ as shown in figure 3.4(b) and (c). These spectra result from the convolution of the modulating square wave with the input temperature spectrum. The spectrum of the output signal, which is simply the difference between $V_1(f)$ and $V_2(f)$, is shown in figure 3.4(d).

The temperature signal of interest is modulated out to the fundamental and odd harmonic frequencies of the square wave. There is a scaling factor associated with each of the harmonics as one would expect; however, it is important to note that the

[1] Although on a microscopic scale the temperature spectrum is in reality very broadband, for this analysis the temperature signal of interest is the macroscopic (average) temperature, which changes much more slowly as a consequence of the spatial averaging. As a result, almost all of the energy in the spectrum will be contained in the low frequencies, and the broadband component is negligible.

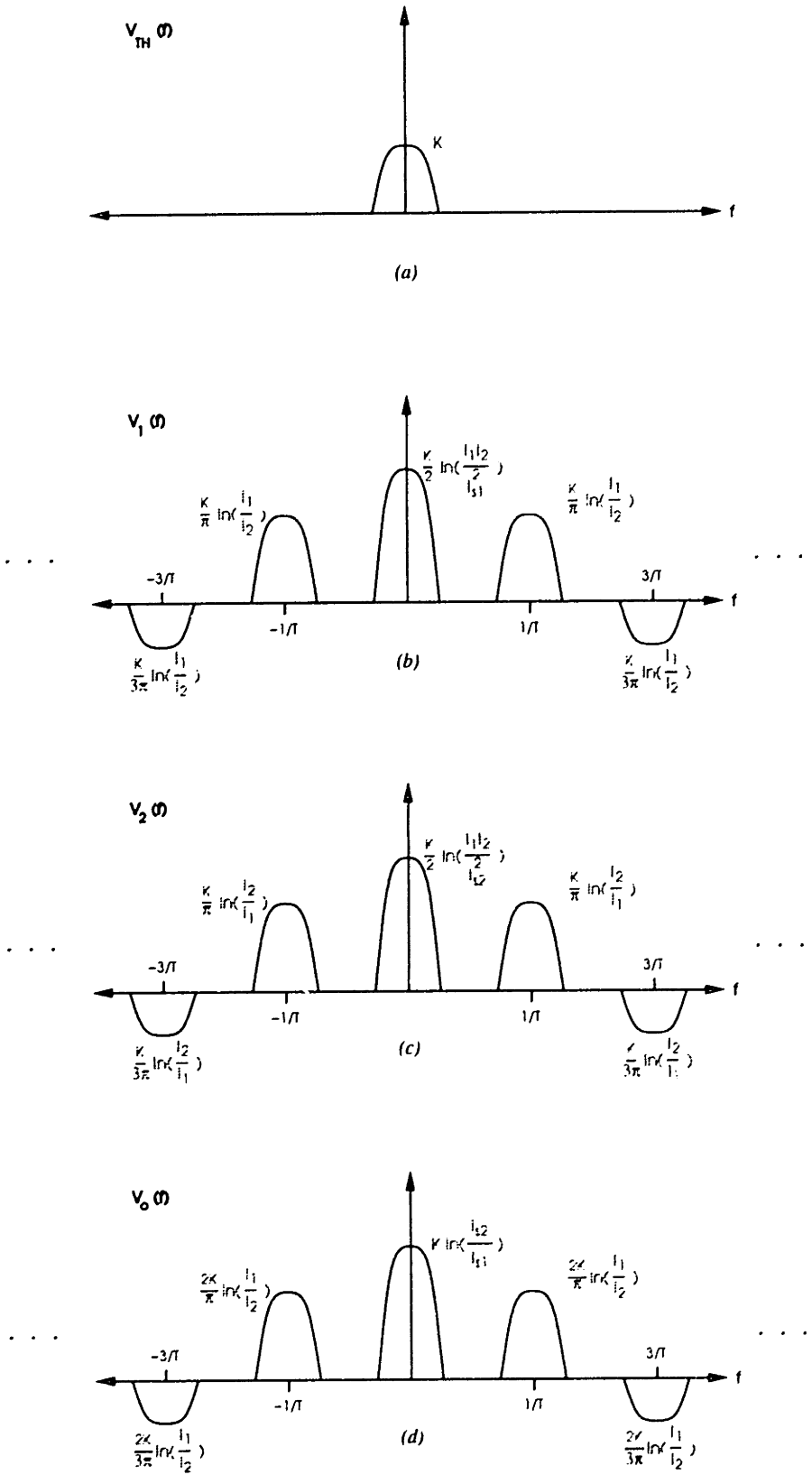


Figure 3.4: Chopped PTAT signal spectra

saturation currents of the two diodes do not influence the scale factors--they appear only in the DC term. It should also be noted that the saturation currents do not cancel one another, rather, they are cancelled in the spectrum of each of the component square waves. In other words, the saturation current I_{S1} cancels itself (as seen in the expression for $V_1(f)$), and similarly the saturation current I_{S2} cancels itself in the expression for $V_2(f)$. In short, all of the desirable characteristics are maintained: the device-independence, the fully differential signal, the linear temperature output, and the chopper modulation are all retained.

The difficulty with the chopped PTAT is that it is severely limited by the nonidealities of the switches. In a practical implementation of the circuit, the switching array would be realized using MOS transistors, and the noise associated with these devices cannot be neglected. In most switching applications, the steady-state current through the switches is zero; except for the transient currents associated with the switching operation itself, no current passes through the device. Consequently, the noise contributed to the circuit containing the switch is composed entirely of the thermal noise associated with the finite resistance of the switch.[2] This noise component is usually tolerable, and if not, the geometry of the switch can be changed to decrease this contribution. If the steady current through the switch is nonzero, however, then there is a second component to the switch noise, namely, the $1/f$ noise associated with the trapping of charge as it flows through the channel. The magnitude of this noise contribution tends to be much larger than the thermal noise, since the frequency range of interest tends to be very low.

The minimum channel length required to reduce the flicker noise in the switches to a tolerable level can be calculated rather easily. For a typical PMOS device (PMOS devices are studied here because they offer better $1/f$ noise performance than NMOS

[2]The $1/f$ noise is eliminated because there is no net flux of carriers from source to drain. The $1/f$ component usually results from the change in flux that results from carriers jumping into traps at the silicon/oxide interface. When there is no net flux, carriers that jump into and out of traps along the oxide/silicon surface change only the local potential profile in the vicinity of the trap but do not change the behavior at the terminals of the device.

devices), the flicker noise is given by:

$$\frac{v_n^2}{\Delta f} = \frac{K}{WLf^\alpha} \tag{3.16}$$

where K and α are process dependent constants. The total noise voltage over a bandwidth f_l to f_h is therefore:

$$\begin{aligned} v_n^2 &= \int_{f_l}^{f_h} \frac{K}{WLf^\alpha} df \\ &= \frac{K}{WL(\alpha-1)} \cdot \left(\frac{1}{f_l^{\alpha-1}} - \frac{1}{f_h^{\alpha-1}} \right) \end{aligned} \tag{3.17}$$

The equivalent drain current noise is simply $i_n^2 = g_m^2 \cdot v_n^2$:

$$\begin{aligned} i_n^2 &= g_m^2 \cdot v_n^2 \\ &= 2k' \left(\frac{W}{L} \right) I_D \cdot \frac{K}{WL(\alpha-1)} \cdot \left(\frac{1}{f_l^{\alpha-1}} - \frac{1}{f_h^{\alpha-1}} \right) \\ &= \frac{2k' I_D K}{L^2(\alpha-1)} \cdot \left(\frac{1}{f_l^{\alpha-1}} - \frac{1}{f_h^{\alpha-1}} \right) \end{aligned} \tag{3.18}$$

To achieve noise performance of the drain current at the N bit level, the total noise current must be less than $\frac{I_D}{2^N}$; using this, we can solve equation 3.18 for the minimum channel length required:

$$\begin{aligned} \left(\frac{I_D}{2^N} \right)^2 &= \frac{2k' I_D K}{L_{min}^2(\alpha-1)} \cdot \left(\frac{1}{f_l^{\alpha-1}} - \frac{1}{f_h^{\alpha-1}} \right) \\ L_{min} &= \sqrt{\frac{2^{2N+1} k' K}{I_D(\alpha-1)} \cdot \left(\frac{1}{f_l^{\alpha-1}} - \frac{1}{f_h^{\alpha-1}} \right)} \end{aligned} \tag{3.19}$$

For the fabrication process used for this project, the process constants are $K \approx 6 \times 10^{-22}\,\mathrm{V^2 m^2 Hz^{-1}}$, $\alpha \approx 1.1$, and $k' \approx 32.5 \frac{\mu A}{V^2}$ [52]. Since thermal processes are very slow, the bandwidth limits are approximately $f_l = .01$ Hz and $f_h = 1$ Hz: The lower limit corresponds to a period of 100 seconds--this is a worst-case limit which is on the order of the longest duration of a single measurement; the upper limit represents the fastest speed at which thermal signals could be expected to change. For a typical drain

current of approximately $10\,\mu A$ and 18.3 bit resolution (N = 18.3, corresponding to 1 m°C), equation 3.19 calculates a minimum channel length of approximately 800 μm--although switches of this size could be fabricated, they are unreasonably large for this application where chip area is a primary constraint. This also assumes that the entire error budget is used on this noise component; in practice, this would not be the case and even larger switches would be required. Even though the flicker noise can be reduced by increasing the size of the switches, the minimum area required to reduce the noise to tolerable levels would be too large to be practical. Consequently, although in theory the chopped PTAT circuit has many benefits, these benefits cannot be realized in practice. It is worth noting, however, that if the technology existed to make very low noise, current carrying switches, the chopped PTAT would indeed be a very desirable technique for temperature measurement.

A further limitation of the chopped PTAT is the noise contribution from the current sources themselves, since the noise from the sources is unaffected by the chopping. As a result, the low frequency noise of the current source transistors becomes significant in the same way that the noise from the switches does. Because dynamic performance of the sources is unimportant, however, there is more flexibility in the design of the current sources, and the noise can be reduced by employing low-noise design techniques.

One potential way of eliminating both the switch noise and current source noise is to move the point at which the chopping occurs. If the chopping somehow took place *before* the current sources (by chopping at the gates of the current source transistors, for example), then the low frequency noise from the current sources would be chopper modulated out of the band of interest. Furthermore, if the chopping is done in such a way as to eliminate steady currents from the switches, than the low frequency noise from the chopper switches could be eliminated also.

The problem with this line of thinking is that one ends up "chasing one's tail", that is to say that in trying to solve the low frequency noise problem this way other benefits of the circuit must be traded off. Once the chopping is moved further back

in the "signal path," there is no longer the I_S cancellation discussed earlier, since a single current source is tied to each sensing diode; the entire benefit of the chopping was that each current source is "sampled" by each sensing diode, and in that way the I_S mismatch could be eliminated. Furthermore, because this "sampling" is no longer being performed, the magnitudes of the currents during each switch cycle becomes important, since there will be ratio errors in *each* current source, not just in the ratio *between* the two current sources. Thus, most of the advantages of chopping are lost if the chopping is moved to some other place in the circuit.

3.2 A Low Noise, High Resolution Sensor

As has been demonstrated above, the major limitations with temperature measurement are primarily due to nonidealities in the excitation circuits--if the PTAT configuration could be excited with a low noise, very stable pair of current sources then the full capabilities of the PTAT could be exploited. Just such an excitation has been developed and is the core of the temperature measurement circuit used for this project. The fundamental mechanism used to both reduce the noise and increase the stiffness of the current sources is feedback--by "closing the loop" around the diode pair, many of the problems can be greatly reduced.

Conceptually, what is needed is some mechanism for sensing the excitation currents. This can be done in many ways: a simple and straightforward way would be to add a resistor to each leg of the PTAT and sense the voltage developed across it. Methods like this, however, fail on two counts: First, implicit in this scheme is that the value of the resistor is known to fairly high precision, which will certainly not be the case. Second, the scheme senses the current in each leg but does not directly control the *ratio* of the currents, which is the critical parameter of interest.

Two circuits designed to not only sense but also control the currents are shown in figures 3.5 and 3.6. The circuits are composed of three "blocks"--an operational amplifier, a differential pair, and a set of four diodes. One pair of diodes is used for

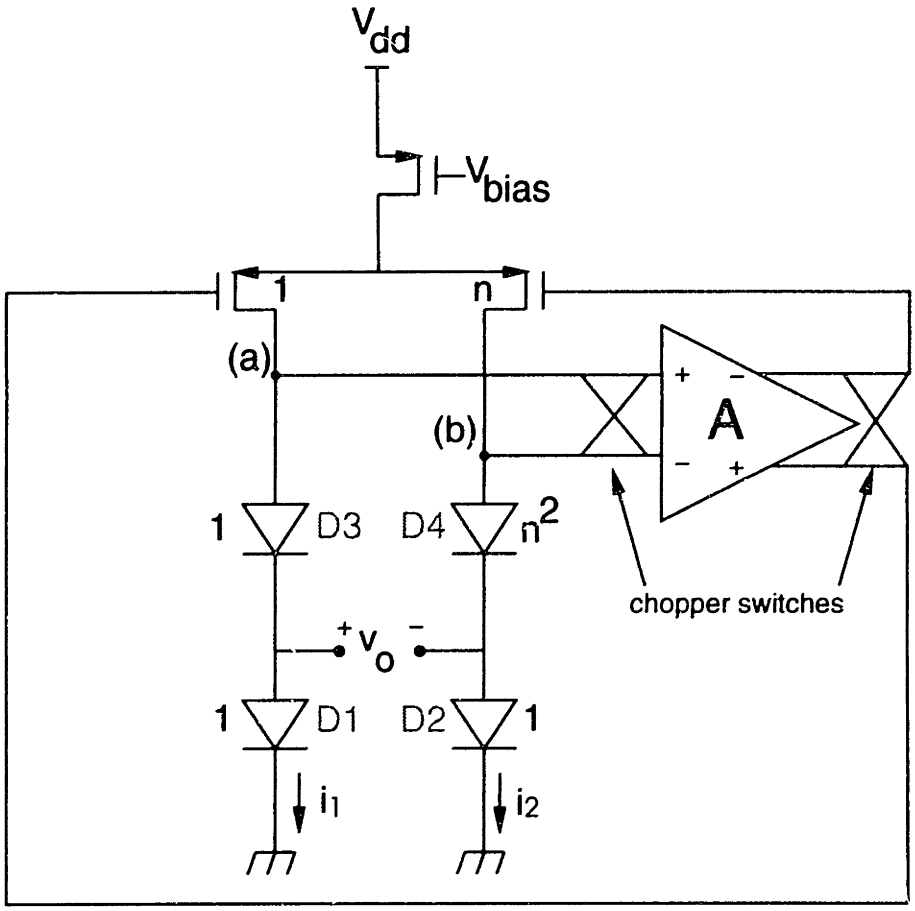


Figure 3.5: Feedback sensor circuit: Top loop

actual temperature sensing; the other is used for setting and monitoring the current ratio. The operational amplifier measures the error voltage generated by the current control diode pair, and generates a differential error voltage signal that is related to the deviation of the excitation current ratio from the desired value. The differential pair is used to convert the voltage output from the operational amplifier into the excitation currents, and to limit the total current through the sensing diodes. This differential pair is purposely mismatched--this mismatch is correlated to the desired current ratio in such a way that the gate drive on each transistor is equal when the current ratio is correct. In that case, the differential output voltage of the operational amplifier is nominally zero.

The only difference between the two circuits is which pair of diodes senses temperature and which pair controls the excitation currents. In the circuit of figure 3.5,

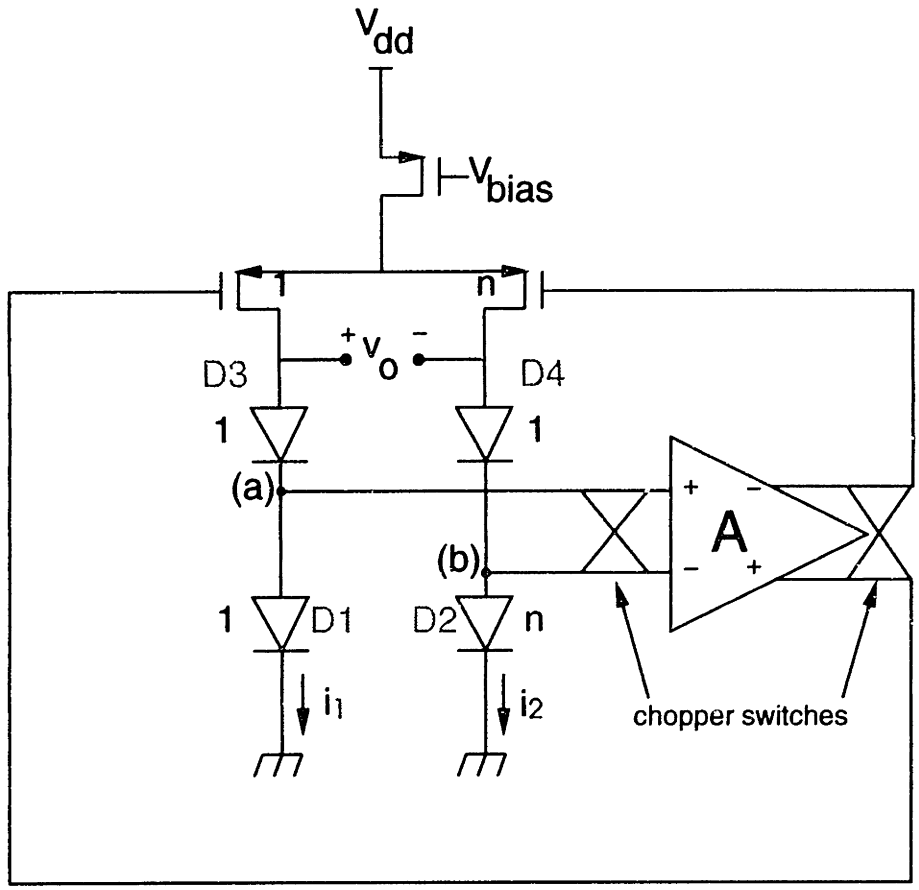


Figure 3.6: Feedback sensor circuit: Bottom loop

the grounded diodes are used for sensing. In this configuration, the voltages at the inputs to the amplifier are:

$$v_+ = 2V_{TH} \ln\left(\frac{I_1}{I_S}\right) \tag{3.20}$$

$$v_- = V_{TH} \ln\left(\frac{I_2}{I_S}\right) + V_{TH} \ln\left(\frac{I_2}{n^2 I_S}\right) \tag{3.21}$$

where I_1 and I_2 are the excitation currents through each leg, and I_S is the saturation current of a unit sized diode. When the loop is operating correctly, these two voltages are equal,[3] and the ratio $\frac{I_1}{I_2}$ can be found by setting $v_+ = v_-$. The resulting relationship

[3]For now it is assumed that the amplifier gain is infinite.

is

$$\boxed{\frac{I_2}{I_1} = n} \tag{3.22}$$

The excitation current ratio is therefore determined solely by the geometry ratio of the diodes, which is a very solid and stable ratio. Any mismatch in the diode area ratio due to normal process variation appears as a fixed ratio "offset" that can very easily be calibrated. Performing a simple two point calibration would show the slight ratio error, and all subsequent measurements could be appropriately adjusted for it. The important point is that the ratio is constant over temperature once it has been set by the process. With a fixed, temperature independent current ratio, the voltage across the two sense diodes is proportional to absolute temperature as described earlier.

The second circuit, figure 3.6, uses the bottom diodes for current ratio control and the top diodes for sensing. In this scenario, when $v_+ = v_-$, the relevant equation is:

$$V_{TH} \ln\left(\frac{I_1}{I_S}\right) = V_{TH} \ln\left(\frac{I_2}{nI_S}\right) \tag{3.23}$$

and it immediately evident that

$$\boxed{\frac{I_2}{I_1} = n} \tag{3.24}$$

as before. The chip area required to realize a given current ratio is significantly different in both circuits, however. In the top loop circuit, the diode area ratio necessary to produce a current ratio of n is n^2; the bottom loop circuit only requires an area ratio of n. This is quite a significant difference, since excitation current ratios on the order of 10:1 or greater are typical for PTAT sensors. For this reason, the bottom loop circuit is employed in the active needle system.

3.3 Sensor Nonidealities

According to the calculation above, the sensor ratio is constant over temperature, and the output voltage in each case is perfectly linear with temperature according to equation 3.9. Throughout all of the analysis, however, several ideal conditions were

assumed. In this section the effects of the nonidealities are considered. For simplicity, each of the effects is considered separately.

3.3.1 Temperature Coefficient Errors

In the initial derivation of the linear output of the PTAT, several temperature independent parameters were introduced: A, the proportionality constant for I_S; B, the proportionality constant for the excitation current; α, the temperature exponent of the excitation current; and β, the temperature exponent of the saturation current. Although different values for A and B are clearly necessary in each leg of the PTAT (since the excitation currents and/or the diode areas must be different to generate an output voltage), it was assumed that the values of α and β were equal in both legs of the circuit. Although this is a very good assumption, there is no guarantee that these parameters will indeed be equal. As a result, a nonlinearity will result in the output characteristic.

The nonlinearity can be examined mathematically if the derivation of equation 3.7 is performed assuming different α and β for each diode. In this case, the difference in the diode voltages is given by:

$$\begin{aligned} V_{D1} - V_{D2} &= V_{TH}\left[\ln\left(\frac{B_1}{A_1}\right) + (\alpha_1 - \beta_1)\ln(T) - \ln\left(\frac{B_2}{A_2}\right) - (\alpha_2 - \beta_2)\ln(T)\right] \\ &= V_{TH}\ln\left(\frac{B_1}{A_1}\cdot\frac{A_2}{B_2}\frac{T^{\alpha_1-\beta_1}}{T^{\alpha_2-\beta_2}}\right) \end{aligned} \tag{3.25}$$

Now define differential and average quantities for α and β as follows:

$$\alpha = \frac{\alpha_1 + \alpha_2}{2} \tag{3.26}$$

$$\Delta\alpha = \alpha_1 - \alpha_2 \tag{3.27}$$

$$\beta = \frac{\beta_1 + \beta_2}{2} \tag{3.28}$$

$$\Delta\beta = \beta_1 - \beta_2 \tag{3.29}$$

so that

$$\alpha_1 = \alpha + \frac{\Delta\alpha}{2} \tag{3.30}$$

$$\alpha_2 = \alpha - \frac{\Delta\alpha}{2} \tag{3.31}$$

$$\beta_1 = \beta + \frac{\Delta\beta}{2} \tag{3.32}$$

$$\beta_2 = \beta - \frac{\Delta\beta}{2} \tag{3.33}$$

Equation 3.25 can then be rewritten in terms of the average and difference parameters, and it is found that:

$$\begin{aligned} V_o &= V_{D1} - V_{D2} \\ &= V_{TH} \ln\left(\frac{B_1}{A_1} \cdot \frac{A_2}{B_2}\right) + V_{TH}\,(\Delta\alpha - \Delta\beta)\ln(T) \end{aligned} \tag{3.34}$$

This result shows that it is not strictly necessary that $\Delta\alpha = 0$ and $\Delta\beta = 0$ for the output to be linear. The correct, slightly looser constraint is that $\Delta\alpha = \Delta\beta$. In addition, since by definition all of the geometry and temperature coefficient parameters (α, β, A, B) are temperature independent, the nature of any nonlinearity that does occur is very well described, namely, it is of the form $T\ln(T)$. Thus, the general output voltage can be written as:

$$V_o = C_1 T + C_2 T \ln(T) \tag{3.35}$$

A simple two-point calibration can be used to determine the values of the unknown constants C_1 and C_2. This is the equivalent of the Steinhart-Hart relationship for thermistors [31].

3.3.2 Device Mismatch and Op Amp Gain Errors

Another potential source of error comes from the fact that the operational amplifier does not have infinite gain, and, therefore, the two input terminals v_+ and v_- are not necessarily equal, as was assumed in the derivation of the sensor output characteristic.

Thus, any nonzero differential voltage at the output of the operational amplifier introduces a direct voltage offset error (since the sensor diodes are no longer at a virtual ground) as well as an induced voltage error (since the current ratio is affected by the voltage difference). Both of these errors are offsets; however, the value of the offset depends on the open loop gain of the amplifier, which will certainly be temperature dependent. It is therefore critical that this error be held to a minimum.

The most straightforward way to insure that the op amp inputs are equal is to guarantee that the differential op amp output voltage is nominally zero. If this is the case, then the op amp inputs will be equal independent of the op amp gain. Since the desired current ratio is $1{:}n$ as derived above, this condition can be brought about by appropriately modifying the geometry ratios of the sensor current control differential pair transistors. If the differential op amp output is zero, then the V_{GS} of both differential pair transistors is the same. Thus, we have:

$$I_1 = \frac{k'}{2}\left(\frac{W}{L}\right)_1 (V_{GS} - V_T)^2 \tag{3.36}$$

$$I_2 = \frac{k'}{2}\left(\frac{W}{L}\right)_2 (V_{GS} - V_T)^2 \tag{3.37}$$

Since we know that the ratio $\frac{I_2}{I_1} = n$, it follows immediately that for zero differential op amp output it is necessary that:

$$\frac{\left(\frac{W}{L}\right)_2}{\left(\frac{W}{L}\right)_1} = n \tag{3.38}$$

as indicated in figure 3.6.

Of course, this eliminates to first order errors associated with the finite gain; however, normal process variation will guarantee that the geometry ratios are never exactly $1{:}n$. In addition, normal process variation in the device matching (V_T mismatch) will require a nonzero differential op amp output voltage. Therefore proper consideration of the op amp gain error requires accounting for the nonidealities of the differential pair transistors also. Hence both of these issues are treated together in this section.

The best way to examine these effects is to consider the errors as small perturbations from the ideal situation. Initially, assume that the differential output of the op amp is zero, and that there is both $\left(\frac{W}{L}\right)$ and V_T mismatch in the differential pair transistors. Also assume that the current error induced by the nonzero differential voltage difference at the op amp input is negligible.[4] With these assumptions, it can be shown (see appendix B) that the differential voltage error at the differential pair inputs is approximately

$$\Delta V \approx -\Delta V_T + \frac{c}{n}\sqrt{\frac{I_o}{2nk'\left(\frac{W}{L}\right)_1}} \tag{3.39}$$

where n is the desired (ideal) current ratio, c is the ratio mismatch between $\left(\frac{W}{L}\right)_1$ and $\left(\frac{W}{L}\right)_2$, and ΔV_T is the threshold mismatch in the differential pair.

There are several important things to note. First, the error voltage is linear with both ΔV_T and c. Second, the error contribution due to the threshold mismatch is on the same order as the threshold mismatch itself. Finally, the error due to $\left(\frac{W}{L}\right)$ ratio mismatch is roughly the gate drive times the percentage ratio error (in fact, it is a factor of $\sqrt{n}$ less than that). Thus, for typical process mismatch parameters, the error voltage will indeed be small. For a current ratio of 10, for example, with $\left(\frac{W}{L}\right)_1 = \frac{25}{3}$, $I_o = 44\,\mu A$ (corresponding to $I_1 = 4\,\mu A$ and $I_2 = 40\,\mu A$, ΔV_T = -10 mV[5] and c = .1, one finds that ΔV = 10.9 mV.

This error voltage is divided by the open loop gain of the amplifier when it is reflected back to the amplifier inputs. This is the actual error component in the measurement, and it appears as an offset in the differential measurement. This offset by itself is not significant since *absolute* temperature measurement is not important. What is critical, however, is that the drift of this offset with temperature is small enough to avoid interfering with the actual temperature signal. This drift component will clearly be strongly correlated with the gain drift of the op amp since the error voltage is the ΔV computed above divided by the open loop op amp gain. Thus, the way to minimize

[4]The validity of this assumption will be examined later.

[5]Because the differential pair is purposely mismatched, the sign of ΔV_T is important. The error is largest when ΔV_T is negative, as discussed in Appendix B.

this error is to use an op amp with a very high open loop gain, so that the initial offset error is negligible. If this is the case, then the drift of the offset with temperature will itself be negligible, and the entire error can be considered just another noise source.

It is therefore possible to compute a minimum required open loop gain for the op amp. As stated above, the offset error in the measurement is given by:

$$v_{os} = \frac{\Delta V}{A} \tag{3.40}$$

where A is the open loop gain of the amplifier. The drift in this voltage is the actual error. Since the exact nature of the drift with temperature is unknown (it is a function of the exact circuits used, layout topologies, etc), it is assumed that this offset drifts by 10% over the temperature range of interest.[6] The error of interest, therefore, is:

$$\Delta v_{os} = 0.1 v_{os} = \frac{\Delta V}{10A} \tag{3.41}$$

or, alternatively, the minimum required op amp gain is

$$A \geq \frac{\Delta V}{10 \Delta v_{os}} \tag{3.42}$$

This is one of the design constraints of the operational amplifier.

3.3.3 Noise Considerations

The third major error contribution is due to device noise. Throughout the above analyses it has been assumed that the MOSFETs, diodes, and the op amp are all noiseless. Clearly this is not the case, and, in fact, the device noise could become a dominant error source because of the high resolution requirements of the sensor. Figure 3.7 shows the sensing circuit with all noise sources present. Each noise source is discussed below.

3.3.3.1 MOSFET Noise

The noise in an MOS device is due to two physical mechanisms: First is the thermal noise associated with the uncertain nature of the electron drift velocity. This noise

[6]The actual drift is probably much less, but for worst-case purposes the 10% estimate should suffice.

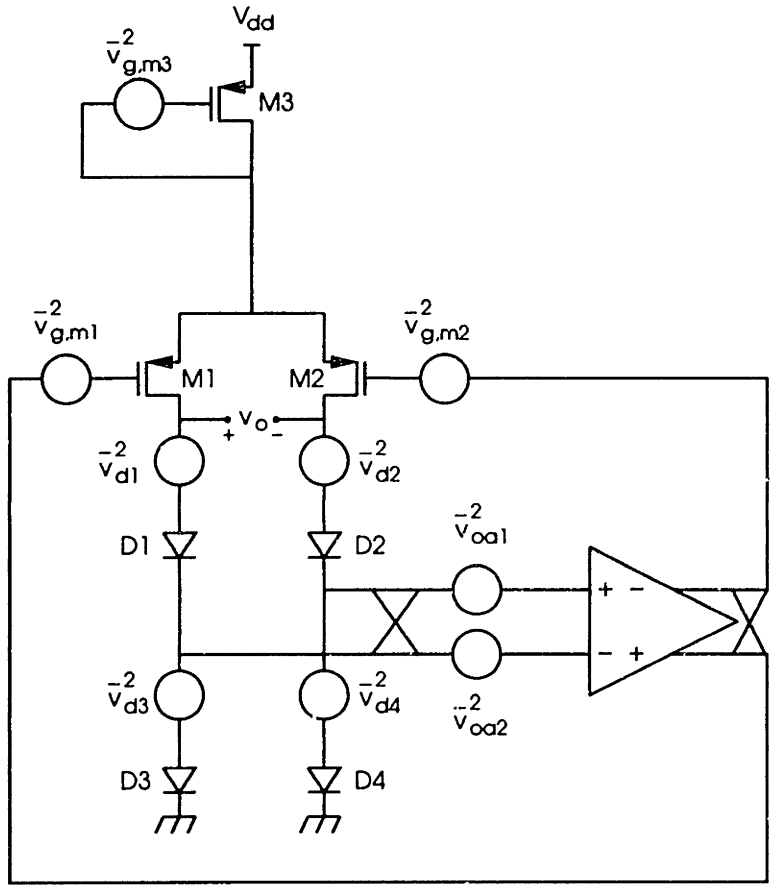


Figure 3.7: Sensing circuit with noise sources

component is very wideband and is "white." Second is the flicker ($1/f$) noise caused by carriers in the channel getting trapped in the gate oxide. Because of its $1/f$ nature, the noise spectral density of the flicker noise is very large at low frequencies and small at high frequencies. The total noise of the MOSFET is the sum of the two components [53]:

$$\frac{\overline{v_g^2}}{\Delta f} = \frac{\gamma 4kTg_{d0}}{g_m^2} + \frac{K}{WL\overline{f^\alpha}} \tag{3.43}$$

where $\overline{v_g^2}$ is the equivalent gate input noise spectral density (in $\frac{V^2}{Hz}$), k is Boltzmann's constant ($1.38 \times 10^{-23}\frac{J}{K}$), T is the absolute temperature (in Kelvin), α is the flicker noise exponent (≈ 1), g_m is the transconductance of the device, g_{d0} is the device channel conductance at $V_{ds} = 0$, and K is the flicker noise coefficient, a process dependent

parameter.[7] The thermal noise coefficient γ is bias dependent, and varies from $\gamma = 1$ at $V_{ds} = 0$ to $\gamma = \frac{2}{3}$ in saturation. It is important to note that both the thermal and flicker noise decrease with increasing device area ($\frac{W}{L}$ ratio)--the noise contribution of the MOS devices can be controlled by appropriate device sizing.

The total noise of the MOS transistor over a bandwidth from f_l to f_h is found by integrating equation 3.43 over the interval and taking the square root of the result:

$$v_{n,tot} = \sqrt{\int_{f_l}^{f_h} \left\{ \frac{\gamma 4kTg_{d0}}{g_m^2} + \frac{K}{WLf^\alpha} \right\} df} \tag{3.44}$$

$$v_{n,tot} = \sqrt{\frac{\gamma 4kTg_{d0}}{g_m^2}(f_h - f_l) + \frac{K}{WL(1-\alpha)}\left[f_h^{1-\alpha} - f_l^{1-\alpha}\right]}$$

where it has been assumed that $\alpha \neq 1$. In the case where $\alpha = 1$, the expression reduces to:

$$v_{n,tot} = \sqrt{\frac{\gamma 4kTg_{d0}}{g_m^2}(f_h - f_l) + \frac{K}{WL}\ln\left(\frac{f_h}{f_l}\right)} \tag{3.45}$$

The total noise of each of the three MOS transistors can be computed using the above formula. The noise of the individual devices must then be referred to the sensor output. For each of the differential pair devices, the noise is divided by the open loop gain of the amp when it is referred to the output of the sensing circuit. This reduction occurs because of the feedback action: any perturbation caused by the noise source stimulates a response from the loop, which will act to counteract the perturbation. The noise from the current source transistor is, in theory, not reflected in the output noise at all, as it appears as a strictly common mode noise. This will only occur if the differential pair device ratio and the current setting diode ratio are exactly equal. For analysis purposes, the circuit was simulated using typical mismatches in geometry (effective versus drawn channel lengths) to determine an approximate attenuation factor of 2×10^{-4}. The total noise contribution due to the differential pair and biasing transistors is therefore:

$$v_{n,in} = \frac{v_{n,tot,1}}{A} + \frac{v_{n,tot,2}}{A} + (2 \times 10^{-4})v_{n,tot,3} \tag{3.46}$$

[7]The dependence of the flicker noise on C_{ox} is lumped into K for this analysis.

3.3.3.2 Operational Amplifier Noise

The input referred noise of the operational amplifier $\overline{v_{oa}^2}$ is a potentially significant source of error, since noise at the operational amplifier inputs is directly reflected in the sensor output voltage. This results from the use of the operational amplifier inputs to force a virtual ground point for the measurement; any noise on the virtual ground point will show up as a direct error in the output voltage. Unlike the MOS current control devices which benefit from the high gain feedback, high resolution measurement is not possible unless the noise contribution of the amplifier is below the measurement resolution.

Because of the high low-frequency flicker noise associated with MOS devices, chopper modulation is used to move this noise out of the frequency range of interest. The modulation also eliminates the (temperature dependent) offset voltage of the op amp. Chopper modulation involves multiplying the op amp input signal by a square wave of frequency f_c prior to amplification. The signal is then demodulated following the amplification. In this way, the op amp noise is frequency shifted by f_c, since the op amp noise is added in after the initial modulation. The signal, however, is unaffected by the process since it is shifted up and then back by f_c. Thus the amplifier noise that is introduced into the signal is the noise at and around f_c. If f_c is chosen so that it is above the flicker noise corner frequency, the only noise contribution from the op amp will be the input referred thermal noise. This process is shown pictorially in figure 3.8.

Because the circuit is fully differential, implementing chopper modulation is a very straightforward process. Switches are introduced at the input and the output that alternately switch the input terminals at a frequency f_c. This is equivalent to multiplying the input signal by a square wave that alternates between +1 and -1. This process is repeated at the output (using the same clock signals to prevent phase problems) to demodulate the signal. This circuit is shown in figure 3.9.

The noise contributed by the op amp is therefore limited to its input referred thermal noise. Since this is strongly dependent on the op amp topology, an exact formulation

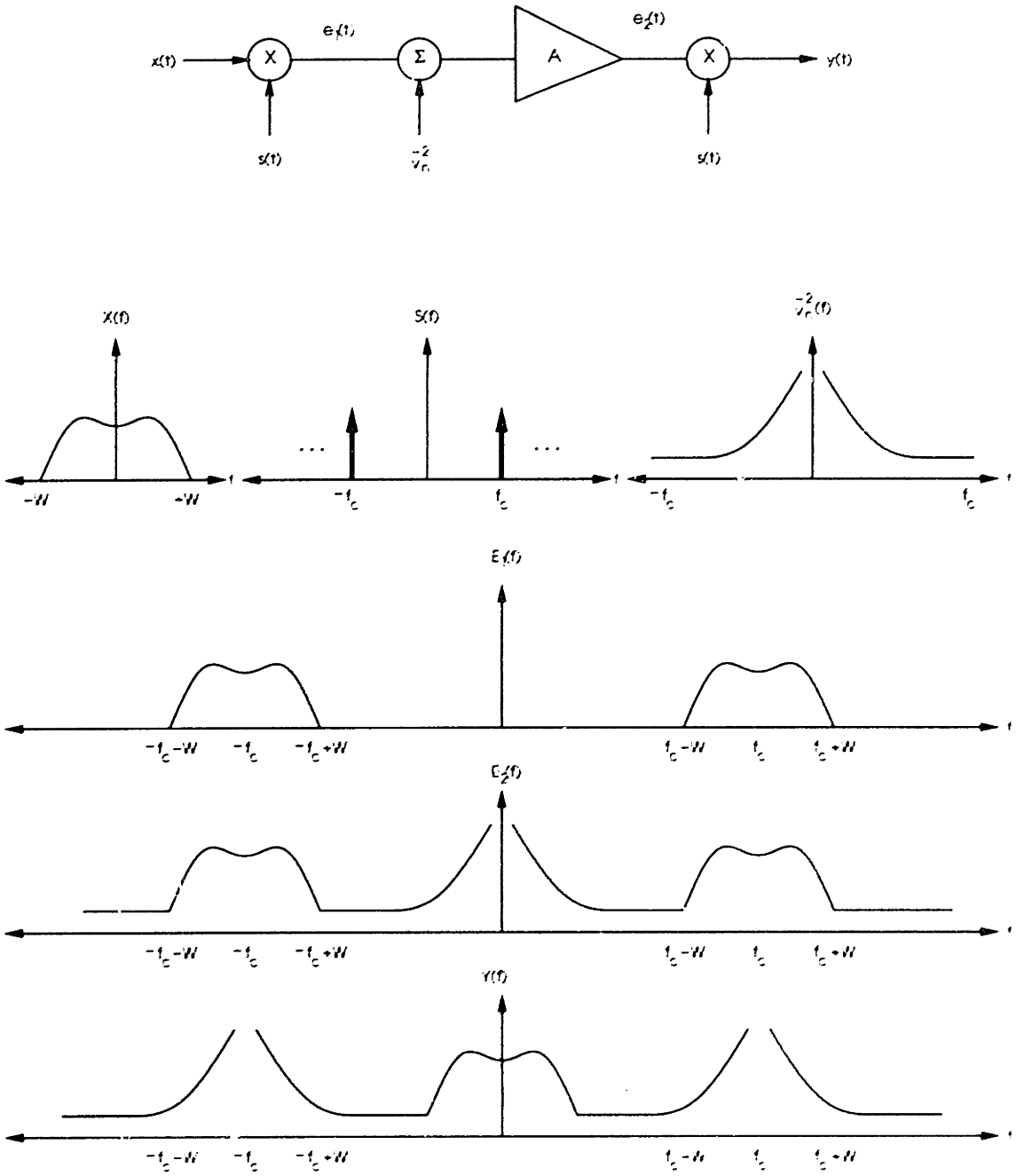


Figure 3.8: Chopper modulation

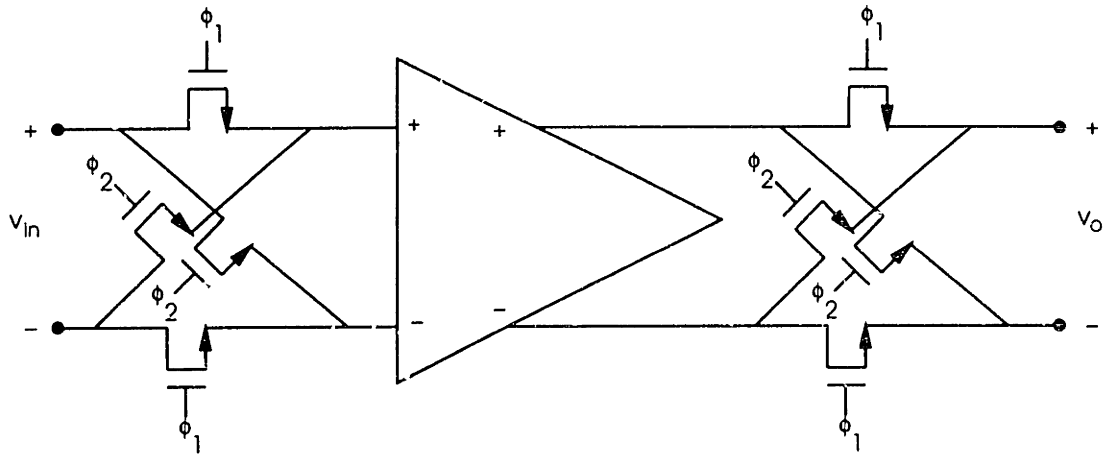


Figure 3.9: Chopper modulation circuit

of this error source is not given here. The important conclusions, however, are that the $1/f$ noise component is eliminated by chopper modulation, and that the thermal noise is unattenuated by the loop and appears directly at the sensor output. This is the noise component that must be minimized in the design of the operational amplifier.

The noise from the input chopper switches does not contribute to the total sensor noise since this noise appears as gate voltage noise on the switches. As stated earlier, there is no $1/f$ drain noise component from these switches because they carry no current. The same reasoning applies to the thermal noise: Random thermal motion of carriers in the channel changes the local potential profile in the channel under the oxide surface. This requires a compensating change in the gate potential, so from a terminal point of view the fluctuations appear as a voltage noise source on the gate. Because the drain current is zero when the switch transients have settled, this changing potential does not appear at the source or drain. Since noise at the gate node is not coupled to the sensor output, this noise is not a consideration when computing the sensor output noise.[8] Similar reasoning applies to the noise from the output chopper switches. If this were a consideration, however, this noise is divided by the open loop gain of the op amp when it is referred to the sensor output, so this noise component can be neglected.

[8]This result was verified through simulation.

3.3.3.3 Diode Noise

The last major noise source is due to the diodes themselves, both the ones used for sensing and the ones used for current ratio control. Because these diodes are located directly at the sensor output nodes, noise from these devices is unattenuated by the loop. As a result, these diodes become the dominant noise generator of the entire sensor.

As with the MOS devices, the dominant source of noise in the diodes is the flicker noise; although the $1/f$ noise is lower in bipolar devices, it still dominates the total device noise over the very low frequency range of interest here. The general functional form of this diode noise is:

$$\frac{v_d^2}{\Delta f} = \frac{K_d I_d}{A f^{\gamma}} \tag{3.47}$$

where K_d is the process-dependent flicker noise coefficient, I_d is the diode current, A is the diode area, and γ is the flicker noise exponent. The total noise over a frequency range f_l to f_h is therefore:

$$v_d^2 = \int_{f_l}^{f_h} \frac{K_d I_d df}{A f^{\gamma}} \tag{3.48}$$

$$= \frac{K_d I_d}{A(1-\gamma)} \left[\frac{1}{f_h^{\gamma-1}} - \frac{1}{f_l^{\gamma-1}} \right] \tag{3.49}$$

for an individual diode. Since four diodes are used in the sensing circuit, the total noise contribution due to the four diodes is the sum (in the mean-square sense) of the individual contributions, and

$$v_{diodes,tot} = \sqrt{v_{d1}^2 + v_{d2}^2 + v_{d3}^2 + v_{d4}^2} \tag{3.50}$$

The exact values for each of these sources is determined by the design choices. Once these choices have been made, the expected noise contribution can be calculated from equations 3.49 and 3.50 and the noise parameters of the fabrication process. Typical values are in the tens of nanovolts per root-Hertz; the values from the implementation for this project are given at the end of the chapter.

3.4 Circuit Implementation

The performance of the temperature sensing circuit described above is governed by several parameters that the designer is free to choose. Although the architecture does not change, the circuit implementation of the operational amplifier in particular will be strongly dependent on these choices. The geometries of the other devices in the sensor will also be affected by the design goals. This section describes the circuit implementation realized in the active needle system. First, the system-level design choices are discussed. The operational amplifier circuit is then presented. At this point, with the entire sensor specified at the device level, the limitations of the sensor can be quantified based on the analysis presented in the previous section.

3.4.1 Design Parameters

3.4.1.1 Current Ratio

The sensor topology used in this project is the bottom loop implementation. The current ratio is 10:1. This ratio was chosen as a compromise between sensitivity and area: According to equation 3.10, the sensitivity of the sensor is directly proportional to the logarithm of the current ratio and therefore increases when the ratio increases. Because of the logarithmic dependence, the current ratio must increase exponentially to achieve any significant increases in sensitivity. Very large ratios are difficult to implement, however, since the minimum area required for the smaller diode is fixed by the process design rules. Consequently, the ratio must attained by increasing the size of the larger diode, which quickly becomes impractically large: An area ratio of 1000:1, for example, gives a sensitivity that is only 3 times greater than a sensor with an area ratio of 10:1, even though the area has increased by a factor of 100.

Furthermore, from a noise point of view it is better to *avoid* using minimum sized diodes, since the device noise is lower at larger areas. The noise contribution from the diodes is primarily determined by the smaller of the sensor devices, as will be shown

in section 3.7 below. Since the sensitivity increases logarithmically with the current ratio (and hence device area) and the noise increases with the square root of the device area (as the smaller device is made even smaller), it is actually worse to increase the current ratio by shrinking the smaller diode. One can envision a sensor in which, say, the smallest diode is 32 times the minimum size, and the ratio setting diode is the largest practical size. Say this results in a current ratio of 10:1. If the sensitivity is then maximized by shrinking the smaller ratio setting diode to the minimum size, which gives a ratio of 320:1, the resulting increase in sensitivity (a factor of 2.5) is countered by a corresponding increase in the device noise (by a factor of 5.7).

The 10:1 ratio chosen represents a compromise between temperature sensitivity and device noise. With the area limitations presented by the needle geometry, a ratio of 10:1 allows for fairly large device areas (and correspondingly low noise) and an acceptable temperature sensitivity of 198 nV/m°C.

3.4.1.2 Absolute Currents

Although the feedback controls the ratio of the currents I_1 and I_2, the actual magnitude of each of the currents is controlled by the differential pair bias current, since this controls the sum of the currents $I_1 + I_2$. The choice of currents is determined primarily by noise considerations: As the currents are made smaller, the signal to noise ratio of the diodes decreases. This can be seen from the diode noise equations. The noise current spectral density is given by

$$\overline{i_n^2} = 2qI_d\Delta f + \frac{KI_d}{f}\Delta f \tag{3.51}$$

The total noise current over a frequency range $[f_l, f_h]$ is therefore:

$$i_n^2 = 2qI_d(f_h - f_l) + KI_d \ln\left(\frac{f_h}{f_l}\right) \tag{3.52}$$

$$i_n = \sqrt{I_d} \cdot \sqrt{2q(f_h - f_l) + K \ln\left(\frac{f_h}{f_l}\right)} \tag{3.53}$$

And the signal to noise ratio is:

$$SNR = \frac{I_d}{i_n} = \sqrt{\frac{I_d}{2q(f_h - f_l) + K \ln\left(\frac{f_h}{f_l}\right)}} \tag{3.54}$$

The excitation currents must therefore be large enough to insure that the signal to noise ratio is high enough to maintain 1 m°C resolution; using the equations above it is found that the critical minimum current is approximately 1 μA.

The tradeoff is that the power dissipation in the sensing diodes must be low enough to avoid measurement errors due to thermal artifacts. The actual thermal problem is quite complex, but experimental measurements have shown that the total current must be kept below approximately 100 μA to avoid thermal artifact problems. As a result, the bias current selected is 44 μA; this way, a current ratio of 10:1 can be maintained (with currents of 4 and 40 μA) while keeping the total current comfortably below 100 μA and the lowest current greater than the 1 μA critical minimum.

3.4.1.3 Differential Pair Geometry

The geometry of the differential pair transistors is perhaps the least critical of the design choices. There are several issues governing the choice of device size. First, the devices must be sized so that the ratio of the $\frac{W}{L}$ ratios is equal to the current ratio n. Second, the devices must be large enough so that their noise contribution is low enough to maintain the desired resolution. Third, the devices must be small enough so that the total area consumed is not prohibitively large. Finally, the devices should be sized so that the swing of the differential pair is reasonable; i.e., so that the differential pair does not saturate when the input differential voltage is less than 100 mV or so.

All of these constraints can easily be met; the device sizes chosen are 25/3 and 250/3 respectively. The ratio of the device sizes is clearly 10:1. The noise contribution is negligible because of the feedback in the circuit. The area consumed by the devices

is far less than the available area. The swing provided by the differential pair with these device sizes is close to 200 mV. In short, the geometry chosen is more than adequate to meet the system requirements.

3.4.2 The Operational Amplifier

The operational amplifier used to implement the sensing scheme directly controls the final temperature resolution, primarily because the amplifier gain governs the attenuation of the MOSFET noise. Therefore, no discussion of the sensor implementation is complete without a knowledge of the operational amplifier design used. This section describes in detail the topology used for the temperature sensor on the active needle. It should be noted that in order to optimize the usefulness of this amplifier it was designed to function not only in the sensor but also in the preamplifier and the oversampled modulator. As a result, the specifications that govern the design choices represent a conglomeration of the various requirements for each of the different subsystems.

3.4.2.1 Electrical Design Specifications

From the standpoint of the system, the most critical design specifications are the open loop gain (since that controls the MOSFET noise and device mismatch attenuation in the sensor), the settling time (since full settling is critical to the modulator) and the input referred noise (since that also contributes to the sensor output noise). In the error analysis above, it was shown that the required op amp gain as a function of the device mismatch is

$$A \geq \frac{\Delta V}{10\Delta v_{os}} \tag{3.55}$$

As stated above, the sensitivity of the circuit for a 10:1 current ratio is 198 nV/m°C; if 10% of the error budget is allocated to this error source, this requires that $\Delta v_{os} <$ 19.8 nV, or, solving equation 3.55 for A, $A \geq 1.9 \times 10^4$. As will be shown later, this DC gain is also adequate for the preamplifier and the modulator; the DC gain specification is therefore $A > 2 \times 10^4$.

The bandwidth specification is controlled more by the requirements of the modulator, since the modulator loop operates at frequencies much higher than the temperature sensor signal frequencies of interest. The signal frequency range of interest is 0.01-1 Hz. For an oversampling ratio on the order of 25,000, typical of the ratio used by the modulator, this translates into a modulator clock frequency of at least 50 kHz. The operational amplifier outputs must settle completely within half of a cycle of the modulator clock. Since this settling occurs when the amplifier is connected in feedback, the relationship between the settling time with and without feedback must be examined. As detailed in Appendix C, the settling time constant under capacitive feedback is given by

$$\Upsilon = \frac{\tau(C_1' + C_2')}{(A+1)C_2' + C_1'} \tag{3.56}$$

where C_1' is the input capacitance, C_2' is the feedback capacitance, and the open loop behavior of the amplifier is described by the single pole model

$$a(s) = \frac{A}{\tau s + 1} \tag{3.57}$$

If completely linear settling is assumed (for purposes of approximation only), then the constraint on the settling time constant Υ is given by:

$$e^{-\frac{T}{2\Upsilon}} \leq \frac{1}{2^N} \tag{3.58}$$

$$\Rightarrow \Upsilon \leq \frac{T}{2N\ln(2)} \tag{3.59}$$

where T is the modulator clock period ($20\,\mu s$) and N is the desired number of bits of resolution (18.3). Equating 3.56 with 3.59 gives the relationship between the dominant pole location of the amplifier and the settling requirement under feedback:

$$\tau \leq \frac{T[(A+1)C_2' + C_1']}{2N(C_1' + C_2')\ln(2)} \tag{3.60}$$

Assuming $A = 20,000$, and maximum expected loading ($C_1' = 10$ pF, $C_2' = 10$ pF) the dominant pole time constant is 7.9 ms, which corresponds to a pole location at 20.2 Hz.

The unity gain bandwidth under these conditions is 404 kHz. The bandwidth, therefore, must be greater than 404 kHz.[9]

In order for the above approximation to be valid, the op amp must avoid ringing and slewing, which places further constraints on the design. Clearly, in order to guarantee that the settling is overdamped (since underdamped behavior would add ringing to the op amp response), the phase margin of the amplifier must be $\geq 60^{\circ}$. The issue of slewing is more difficult to account for since it is not necessary (and is quite power wasteful) to design the amplifier in such a way that there is essentially no slewing. Instead, it is better to design the amplifier so that any slewing occurs over a short enough time period so that the amplifier can fully settle in the desired time interval. In the case of this op amp, full settling must occur within $10\,\mu s$ as described above. In terms of slew rate, the maximum change that can possibly occur at the output of the modulator op amps is 1.5 V. If it is assumed that all slewing takes place in the first $1\,\mu s$ (i.e., 10% of the total settling time), this would translate into a slew rate of at least $1.5\ V/\mu s$. With a 10 pF worst-case load, this means that the output current in each leg of the amplifier must be at least $7.5\,\mu A$.

The output swing, although not a critical performance parameter, nonetheless plays an important role in the system characteristics, since clipping of the amplifier outputs will degrade the performance of the modulator significantly. Since the maximum change in output for either leg is only .75 V (due to coefficient scaling--see chapter 4), an output swing of approximately ± 1 V is required to guarantee that clipping will not occur. In terms of the output swing in the fully differential sense, this translates into an output swing of 2 V.

The last performance parameter that must be specified is the input referred noise, since that noise contributes directly to the sensor output noise level. As was discussed earlier, for a temperature resolution of 1 m°C, it is necessary to resolve 198 nV at the sensor output, which means that the sensor output noise must be kept below this level.

[9]Clearly it is advantageous to maximize the bandwidth; this number merely specifies the lower limit on what would be acceptable performance.

Based on the expected noise contributions from the various sources, it was decided that 10% of this tolerable noise limit ($\approx 20\,\mathrm{nV}$) would be budgeted for the op amp noise contribution. Such a low noise limit was not considered excessively stringent since it was assumed that chopper modulation would be used to eliminate the low frequency $1/f$ noise, as discussed above.

The other characteristics of the op amp are not as critical and therefore no specifications for them were developed, although "guidelines" for the different parameters can be stated. Since the power source for these circuits will be a battery and all circuits will be fully differential, excessive power supply rejection ratios are not required. The fully differential nature of all the circuits also relaxes the common mode rejection tolerance,[10] although it is clear that all of the rejection ratios should be maximized. There is no input bias or offset current because of the zero gate current of the MOS devices. The power dissipation should be as low as reasonably possible to minimize the thermal artifact that results from the on-chip power dissipation; for this reason, the total supply current was limited to $100\,\mu A$ (excluding the current source reference itself, which is shared among all of the op amps). Finally, and perhaps most importantly, the area of the amplifier should be as low as is practically feasible, since the geometry of the needle limits the available active area tremendously. The area per amplifier was therefore limited to $.2\,\mathrm{mm}^2$, or, equivalently, a total device active area of approximately $10{,}000\,\mu m^2$.[11]

3.4.2.2 Topology

The performance goals described above are not so stringent as to force the use of a specific topology. However, because the settling time is critical to the success of

[10] In fact, both the power supply and common mode rejection ratios are infinite in an ideal differential circuit. Although the fabricated circuit will certainly not be ideal, these ratios can only be evaluated by approximating these processing mismatches. As a result, the numbers generated will be strongly process dependent, and will therefore be inherently somewhat unreliable.

[11] A factor of twenty was selected because the total device active area does not account for source/drain areas, contact areas, or interconnections. This ratio was determined from a sampling of other existing circuit layouts.

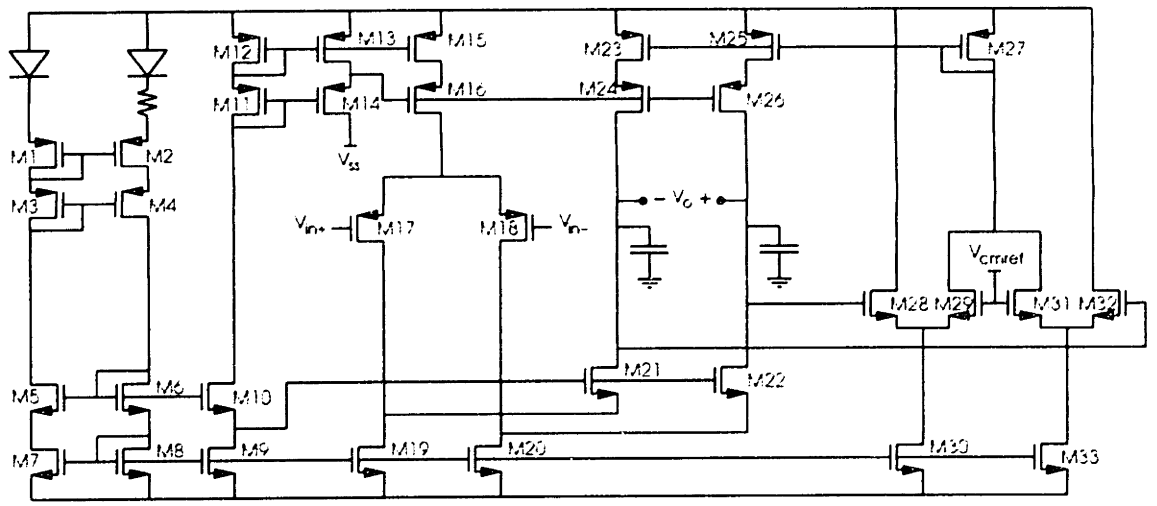


Figure 3.10: Folded cascode op amp topology

the modulator, the folded cascode topology is used. This topology generally results in moderate gain, moderate bandwidth systems with high phase margin. In addition, because the compensation network is tied between the output and analog ground, the power supply rejection ratios from both power supplies are superior. Furthermore, the compensation capacitance can be provided by the capacitive load, eliminating the need for (and the chip area of) two additional capacitors. This all comes at the expense of gain; the folded cascode typically realizes a lower gain than a comparable two-stage design. Since only moderate gain and bandwidth are required, however, this is an acceptable tradeoff.

The circuit is shown in figure 3.10, with the device geometries given in table 3.1. PMOS input transistors were originally selected because of their superior flicker noise performance, the subsequent addition of chopping circuitry obviated this consideration. Additionally, fully isolated PMOS devices are available from the fabrication process (see chapter 6). A self-biasing current source is used as the reference current for the biasing of each stage. Simple capacitive compensation is used to stabilize the amplifier and to provide the desired high phase margin and single pole response. Finally, improved cascoding is used throughout the design to maximize signal swing. [50]

Table 3.1: Device geometries

Device	Width (microns)	Length (microns)	Device	Width (microns)	Length (microns)
M1	100	3	M18	480	4
M2	100	3	M19	60	6
M3	100	3	M20	60	6
M4	100	3	M21	40	3
M5	40	3	M22	40	3
M6	8	3	M23	41	6
M7	40	3	M24	40	3
M8	40	3	M25	41	6
M10	40	3	M26	40	3
M9	40	3	M27	40	3
M11	12	3	M28	5	20
M12	60	3	M29	5	20
M13	60	3	M30	20	3
M14	60	3	M31	5	20
M15	60	3	M32	5	20
M16	60	3	M33	20	3
M17	480	4			

3.4.2.3 Differential Mode Characteristics

The gain of the op amp can be calculated using the simplified circuit shown in figure 3.11 and the corresponding small signal half circuit shown in figure 3.12; one quickly finds that:

$$\frac{v_o}{v_{in}} = \frac{-g_{m1}r_{oc}r_{ol}(1+g_{m2}r_{o2})}{r_{ol}+r_{o2}+r_{oc}(1+g_{m2}r_{o2})} \tag{3.61}$$

where r_{oc} and r_{ol} are defined in figure 3.12. From this result it is seen that the gain is directly proportional to the input stage transconductance. Furthermore, the "effective" resistance multiplying this transconductance is, approximately, the parallel combination of $r_{oc}(1+g_{m2}r_{o2})$ and r_{ol}. This represents an important limit of the topology: higher gain cannot be realized merely by increasing r_{ol} or r_{oc}, because the effective resistance will be governed by the lower of the two resistances. Increasing the transconductance of the input devices will increase the gain, but this increase is limited by r_{oc}, which

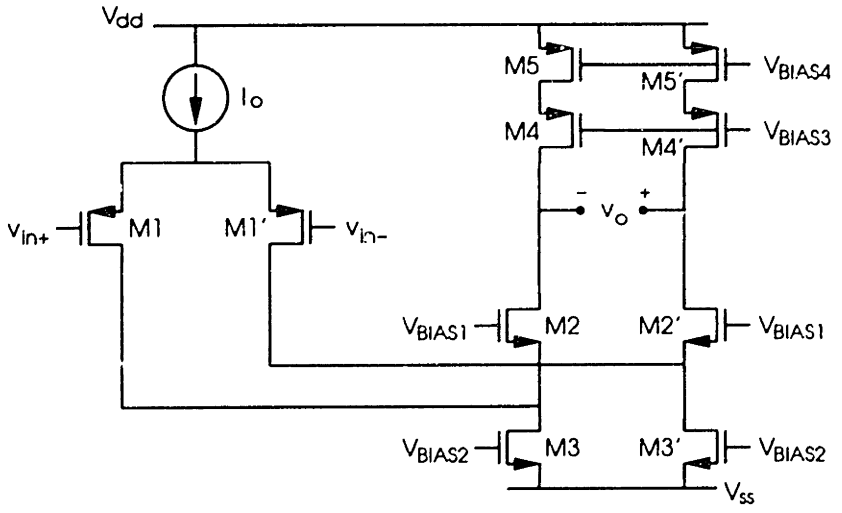


Figure 3.11: Simplified schematic of folded cascode

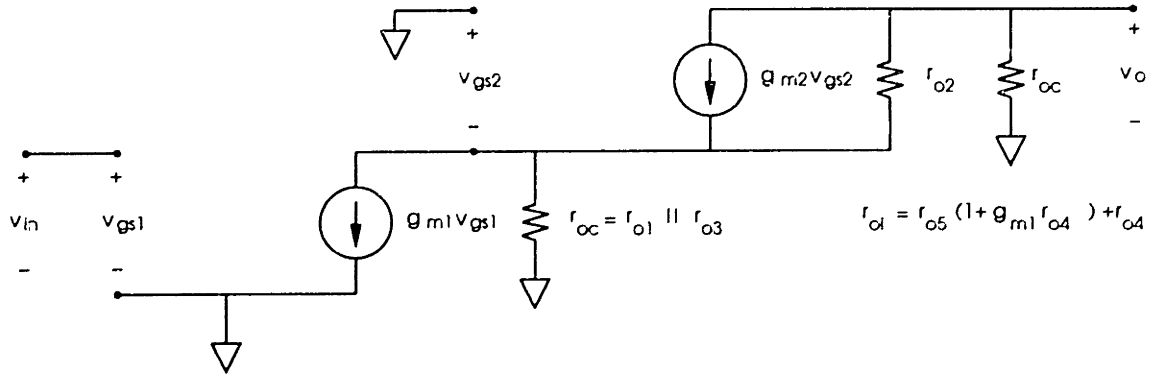


Figure 3.12: Small signal model of folded cascode

will decrease as the transconductance is raised; in other words, the gain increases significantly only when the intrinsic gain of M1 increases. Clearly, the same can be said for M2, since it's intrinsic gain appears directly in the gain equation.

Therefore, the key to realizing the highest possible gain is to operate the critical signal path devices in their moderate inversion region, for it is at that point that the intrinsic gain (or, equivalently, the open circuit voltage gain) of the device is maximized [54,55]. Mathematically, the open circuit voltage gain A_{oc} of a MOSFET is $g_m r_o$, where, in strong inversion,

$$g_m = \sqrt{2\mu C_{ox}\left(\frac{W}{L}\right) I_d} \tag{3.62}$$

and

$$r_o = \frac{1}{\lambda I_d} \tag{3.63}$$

where μ is the channel mobility, C_{ox} is the oxide capacitance per unit area, and λ is a parameter related to the change in effective channel length as a function of the drain voltage. The open circuit voltage is therefore:

$$A_{ol} = g_m r_o \tag{3.64}$$

$$= \sqrt{\frac{2\mu C_{ox}\left(\frac{W}{L}\right)}{\lambda^2 I_d}} \tag{3.65}$$

from which it is clear that the intrinsic gain goes up as the drain current through the device *decreases*. At some point, however, decreasing the drain current brings the device out of the saturation region and into the subthreshold region. When this occurs, the transconductance becomes:

$$g_m = \frac{I_d}{V_{TH}} \tag{3.66}$$

where V_{TH} is the thermal voltage ($\frac{kT}{q}$). The open circuit voltage gain becomes constant and independent of the drain current, and further decreases in the drain current do not affect the intrinsic gain. The bandwidth of the device, however, decreases as the drain current is reduced, since there is less current available to charge the device capacitances. This implies that there is some critical drain current at which the intrinsic gain and bandwidth are maximized; this point is clearly in the moderate inversion region where the transition between subthreshold and strong inversion occurs. Thus, the optimal operating point is in the moderate inversion region.

The dynamics of the amplifier can also be found from the small signal model. Using the method of open circuit time constants one finds that the dominant pole frequency of the amplifier is:

$$f_{p1} = \frac{1}{2\pi C_L \{r_{o4} \| [r_{o2} + r_{o1}(1 + g_{m2} r_{o2})]\}} \tag{3.67}$$

where r_{o4} is the output resistance of the current source load of the second stage,

$r_{oc} = r_{o1}||r_{o3}$, and C_L is the load capacitance. The second pole, which will control the phase margin, is at a frequency:

$$f_{p2} = \frac{r_{o2} + r_{oc}(1 + g_{m2}r_{o2})}{2\pi C'_{L2} r_{oc} r_{o2}} \tag{3.68}$$

where C'_{L2} is the capacitance at the drain of M1. It is clear from the first equation that the dominant pole frequency is essentially controlled by the $r_{ol}C'_L$ product. For large enough values of r_{ol}, however, the pole frequency becomes independent of the load resistance. Most importantly, however, the pole locations are controlled by many of the same parameters that control the gain, which implies that the gain and the bandwidth are generally not independently controllable. This can be seen by multiplying equation 3.61 with equation 3.67 to give the effective gain-bandwidth product:

$$\begin{aligned} G \cdot BW &= \frac{g_{m1} r_{oc} r_{ol}(1 + g_{m2}r_{o2})}{r_{ol} + r_{o2} + r_{oc}(1 + g_{m2}r_{o2})} \cdot \frac{1}{2\pi C'_L \{r_{ol}||[r_{o2} + r_{oc}(1 + g_{m2}r_{o2})]\}} \\ &= \frac{g_{m1} r_{oc}(1 + g_{m2}r_{o2})}{2\pi C'_L (r_{o2} + r_{oc}(1 + g_{m2}r_{o2}))} \end{aligned} \tag{3.69}$$

If $g_{m2}r_{oc} \gg 1$ then this product is approximately $\frac{g_{m1}}{C_L}$, which indicates that the dominant controlling parameter for the op amp is the input stage transconductance.

It is clear that the dominant pole location of the amplifier is governed by the RC' time constant at the outputs, (namely, the product of the output resistance and the load capacitance) since the load capacitance will dominate all of the device capacitances. If first order settling is assumed, the amplifier time constant τ for this topology becomes:

$$\tau = C'_L \{r_{ol}||[r_{o2} + r_{oc}(1 + g_{m2}r_{o2})]\} \tag{3.70}$$

The settling time predicted by equation 3.56 is therefore related to the characteristics of the folding transistors and their loads. Unlike the gain, the transconductance of the input stage does not strongly influence the settling time; the only effect the input stage has on the settling time is through r_{oc}.

Since the amplifier is essentially current-output, it is somewhat incorrect to characterize the op amp as having a specific slew rate; it is more appropriate to discuss

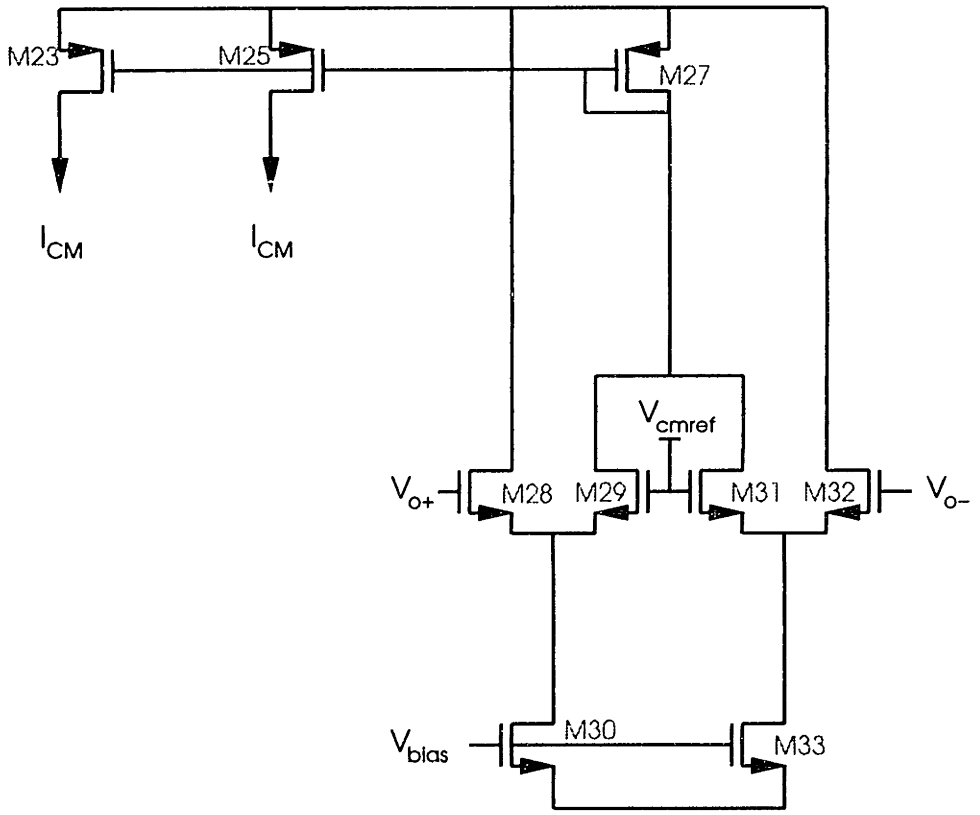


Figure 3.13: Op amp common mode feedback circuit

the amplifier's maximum output current I_{max}, since the slew rate $S = I_{max}/C_L$. The slew rate therefore decreases as the load capacitance increases. For a fixed load capacitance, a higher slew rate requires an increase in the current in the output stage, and, in general, the larger the output current the faster the outputs can slew. The current limit is clearly the bias current in the output legs I_{bo}; when the input is overdriven, one of the outputs is sourcing I_{bo} while the other is sinking I_{bo}. Consequently, meeting the slew rate specification (or the corresponding output current specification) forces the output stage bias currents to be at least as large as the desired output current, which in this case is 7.5 μA.

3.4.2.4 Common Mode Feedback

Because the amplifier is fully differential, circuitry must be provided to stabilize the output common mode level. This is done using devices M_{27}-M_{33}; this section of the circuit is shown in figure 3.13. The outputs of the amplifier are sensed by two differential pairs, each of which has one input tied to a common mode reference voltage (the desired common mode output level). With no differential output, the current in each diff pair splits equally, and devices M_{28} through M_{32} all carry the same current. The drain currents of M_{29} and M_{31} are added; this current is then mirrored to the output stage. When a purely differential signal is present at the amplifier outputs, the currents through M_{29} and M_{31} are no longer equal, but their sum remains constant, and no adjustment in the output common mode level is made. When the output common mode level changes, the sum of the currents changes, which in turn changes the output stage bias current so that common mode level is restored to its desired value.

Mathematically, the two equations governing the differential pairs are:

$$I_{28} + I_{29} = I_b \tag{3.71}$$

$$I_{31} + I_{32} = I_b \tag{3.72}$$

where I_b is the bias current in each diff pair. From this it is clear that the sum of I_{29} and I_{31} is:

$$I_{29} + I_{31} = 2I_b - I_{28} - I_{32} \tag{3.73}$$

If it is assumed that the amplifier output swing is low enough (or that the transconductance of the diff pair devices is low enough), then the currents I_{28} and I_{32} are approximately

$$I_{28} \approx \frac{I_b}{2} + g_{m28}\Delta V_{o+} \tag{3.74}$$

$$I_{32} \approx \frac{I_b}{2} + g_{m32}\Delta V_{o-} \tag{3.75}$$

where the ΔV are defined as the difference between the each output and the desired

common mode level. Since all four devices are sized and biased identically, g_{m28} = g_{m32} = g_m; substituting equations 3.74 and 3.75 into equation 3.73 gives:

$$I_{29} + I_{31} = I_b - g_m(\Delta V_{o+} + \Delta V_{o-}) \tag{3.76}$$

The current is therefore a function of the common mode (average value) at the amplifier outputs. When the common mode level at the output increases, the bias current in the output stage is decreased, which acts to restore the common mode level to its setpoint. Conversely, when the output common mode level decreases, the output stage current is increased to negate the change.

Device geometries for the common mode feedback circuit are selected to minimize area while making sure that the loop will operate correctly when large differential signals are present. The differential pair devices are very long and narrow ($\frac{5}{20}$) so that their transconductance is very low (which widens the range of linear operation of the loop). Smaller W/L ratios are not used because of the increased gate drive required; the increased V_{gs} that results can cause a loss of compliance of the differential pair bias current source. When this occurs, the loop will also fail to function linearly. The $\frac{5}{20}$ geometry used increases the operating range of the common mode feedback without increasing the gate drive requirements too much. The current mirror devices are made very small ($\frac{20}{3}$) to save area; scaling to the desired output current levels is done by increasing the size of the PMOS loads on the amplifier output stage. The $\frac{W}{L}$ ratio of the bias current device is fixed by the current source reference; 3 μm channel lengths are used to "stiffen" the current sources without requiring additional cascode devices. The dynamic properties of the common mode feedback loop formed by these devices are shown in figure 3.14.

The advantage of this common mode stabilization circuit is that it is very compact, which is critical to an application such as this where chip area must be minimized. The disadvantage is that the assumption that the change in current is linearly related to the change in common mode level does not hold true over the entire differential

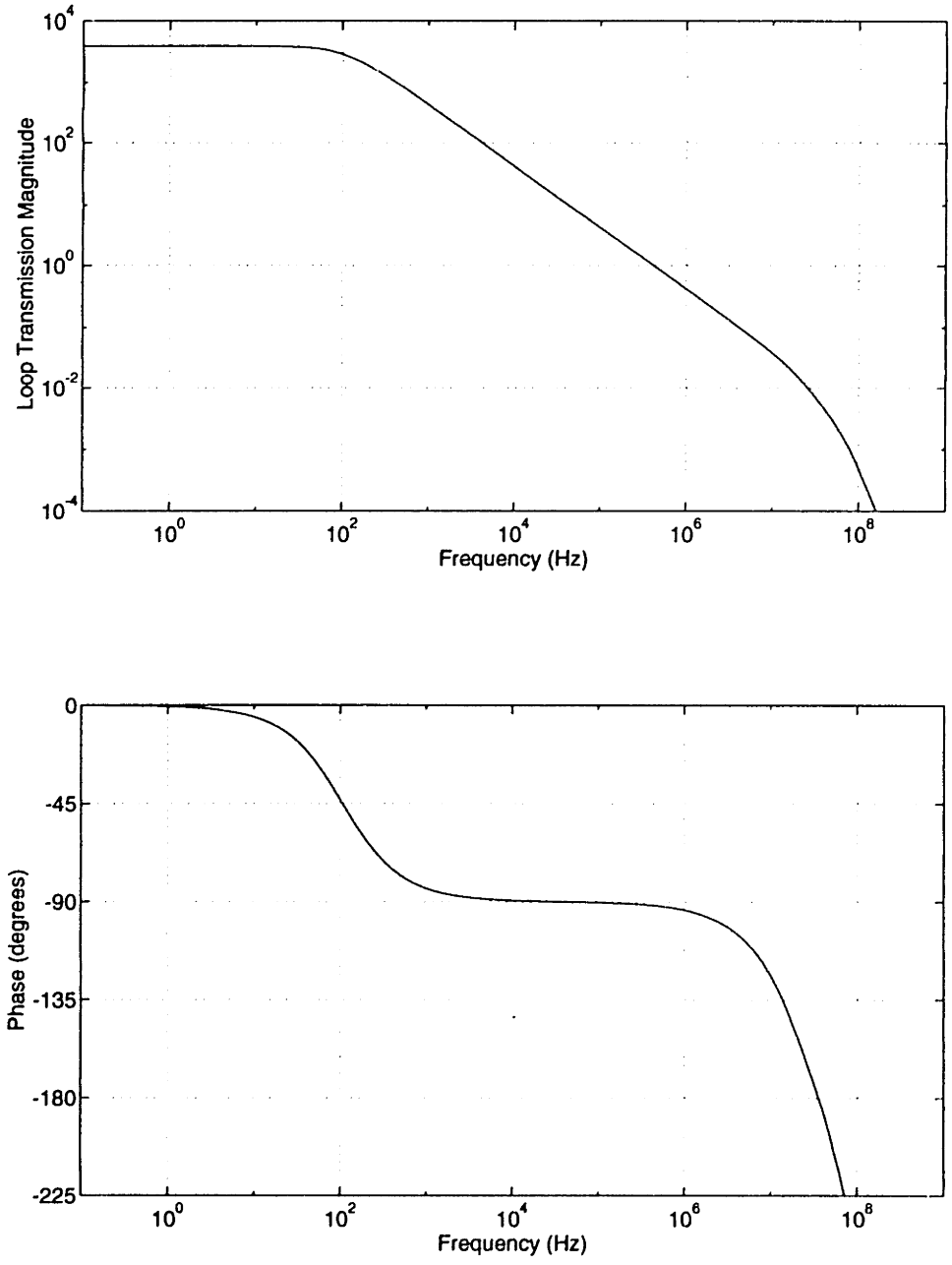


Figure 3.14: Common mode feedback loop transmission, HSPICE simulation

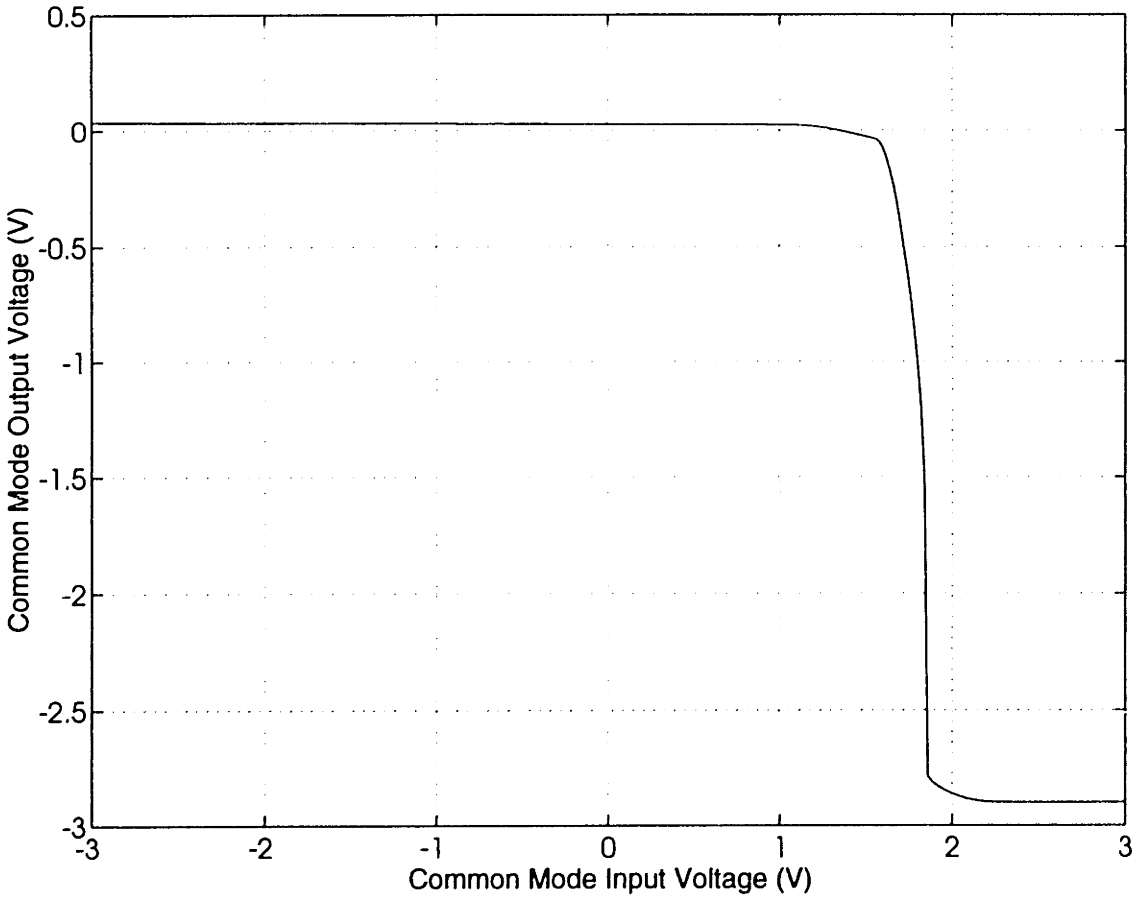


Figure 3.15: Output common mode voltage vs. input common mode voltage

swing range of the amplifier. As a result, the common mode level at the output changes slightly at the extremes of swing, and the feedback only operates well over a limited range of common mode input voltages. This is shown in figure 3.15, which shows the common mode level at the output as a function of the input common mode. This simulation assumes that the amplifier is operated from supplies of ± 3 V and that the desired output common mode level is 0 V. Capacitor-refresh schemes exist that do not have this problem [56], but the tradeoff is drastically increased chip area, since four additional capacitors are required. Furthermore, the switched capacitor circuit suffers from capacitive coupling of the clock signals, which generates an AC common mode signal at the amplifier outputs. This is only a consideration for the operational amplifier used in the sensor, however. Regardless, the continuous scheme used here is compact,

has no feedthrough problems, and works very well if large differential signals are avoided at the amplifier output.

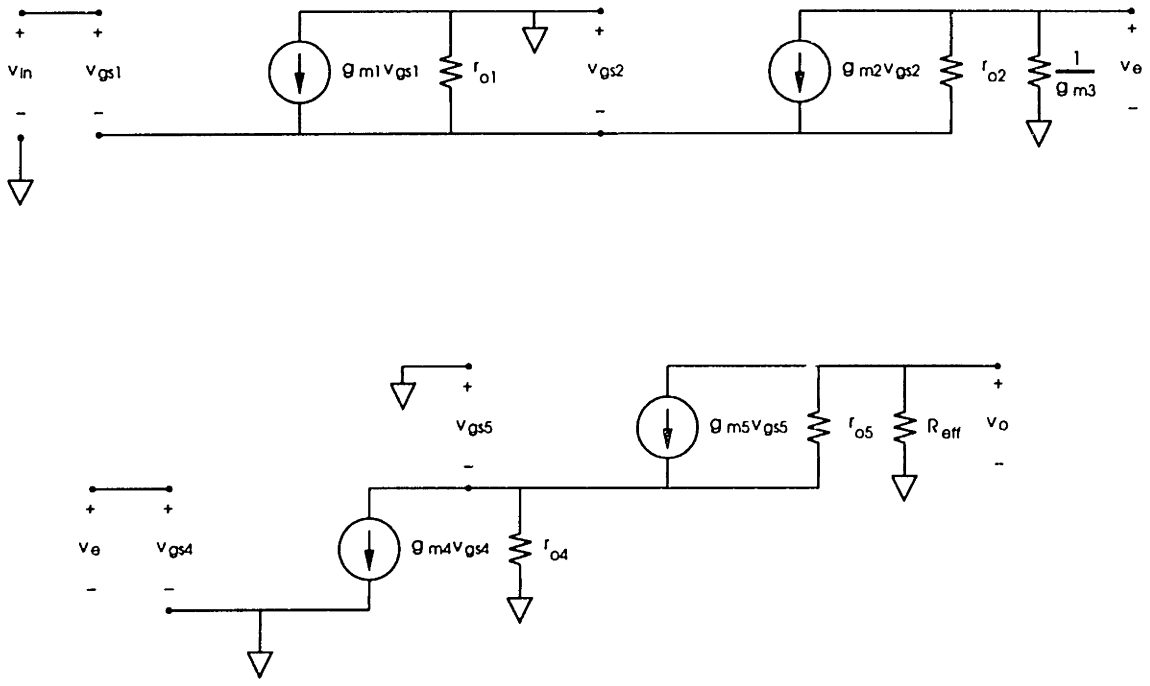


Figure 3.16: Common mode feedback small signal model

Finally, the dynamics of the common mode feedback must be examined to insure that the loop is in fact stable. The loop transmission of the circuit can be found by breaking the loop at the input to the feedback differential pairs and studying the small signal common mode half circuit of the resulting system. This circuit is shown in figure 3.16, where the folding transistor and the input stage have been replaced by an effective load resistance R_{eff}. The DC loop gain of the system is:

$$\begin{aligned}\frac{v_o}{v_i} &= \frac{-g_{m25}r_{o25}R_{eff}(1+g_{m26}r_{o26})}{r_{o26}+R_{eff}+r_{o25}(1+g_{m26}r_{o26})} \cdot \frac{g_{m28}r_{o28}(1+g_{m28}r_{o28})}{g_{m28}r_{o28}+2g_{m27}r_{o28}+2g_{m27}g_{m28}r_{o28}^2+1} \\ &\approx \frac{-g_{m28}g_{m25}R_{eff}}{2g_{m27}} \end{aligned} \tag{3.77}$$

The dominant pole of the loop is clearly the one formed by the load capacitance and the resistance at the output node, which is the same as the dominant pole for the amplifier f_{p1} derived above. The second pole of the system is due to the gate to source

capacitances of the diff pairs; this pole is at a higher frequency than the second pole of the differential transfer function because the device capacitance is relatively small, as is the resistance at the diff pair common source node. Thus, for this topology, if the DC gain of the common mode loop is lower than the DC differential gain of the amplifier, then the common mode loop must be stable.[12] The bandwidth of the common mode loop, however, will be lower than the bandwidth of the differential loop by a factor equal to the ratio of the two DC gains. The optimal operating condition in terms of bandwidth, therefore, is when the two DC gains are equal; however, it is not a necessary condition, and in fact for both power and area reasons it is more desirable to operate the common mode loop at a slightly lower bandwidth.

3.4.2.5 Design Budget/Predicted Performance

The analysis above indicates that the key parameter controlling performance of the amplifier is the input stage transconductance. For this reason, a large part ($40\,\mu A$) of the current budget of $100\,\mu A$ was used to bias the first stage. The output stage bias current is dictated by the output current requirement; only slightly more current than is theoretically necessary ($10\,\mu A$ per leg) is used because higher output stage current lowers r_{ol}, which in turn lowers the differential gain. The remaining $40\,\mu A$ is used for the common mode feedback circuitry; such a large fraction was allocated for this task in order to insure that the common mode response of the amplifier would be adequate.

The current reference, shown in figure 3.17, is a self-biased ΔV_{be}-based current source, with a simple startup circuit that guarantees that the circuit biases to the desired operating point. The voltage generated across resistor R is equal to the difference of the voltages across diodes D_1 and D_2, which are forced to operate at equal currents by the current mirror formed by M_5-M_8. As a result, the current in each leg of the reference structure is:

$$I_{ref} = \frac{V_{TH} \ln(n)}{R} \tag{3.78}$$

[12]The assumption, of course, is that the differential loop transmission is stable also.

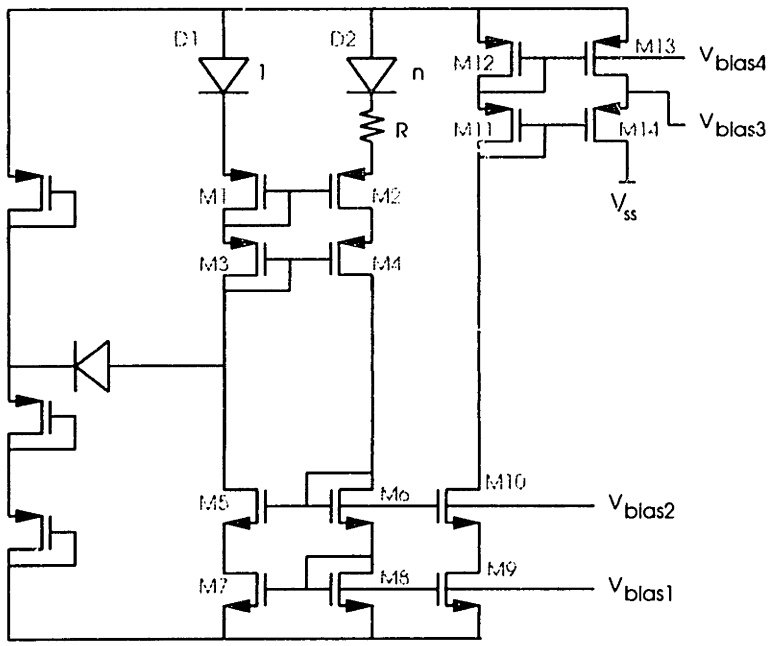


Figure 3.17: Op amp bias current reference circuit

where n is the ratio of the junction areas of the diodes. In the circuit used, R is chosen such that $I_{ref} = 40\,\mu A$, so that the current source is "stiff" enough to serve as a reference for several amplifiers. Finally, transistors M_9-M_{14} are used to generate the actual bias voltages from the current reference; improved cascoding is used to maximize signal swing.

Because the circuit is self-biased, there are two stable point of operation. The first is the one governed by equation 3.78 above. The second is the trivial case in which all currents are zero. When power is applied to the system, it is not possible to predict to which stable point the circuit will settle. In order to force the circuit to bias to the nontrivial operating point, a startup circuit (shown in the dashed box in figure 3.17) is added to the design. The three gated-diode MOSFETs are used to generate a voltage of $V_b = \frac{2}{3}(V_{dd} - V_{ss})$. Without the startup circuit, the voltage at the drain of M3 goes to the positive rail if the bias circuit powers up in the trivial zero current condition. With the startup circuit, this cannot happen; the startup diode conducts as the voltage on the

drain of M3 rises above V_b. This forces a current through the biasing circuit, which drives the circuit to the operating point specified by equation 3.78. As this occurs, the voltage on the drain of M3 drops and the startup diode turns off, which disconnects the startup circuit from the biasing. Consequently, the startup circuit guarantees operation at the proper stable point without affecting normal operation of the bias circuit.

The size of the input stage devices M_{17} and M_{18} is $\frac{480}{4}$; this is done so that their transconductance is large. The channel length of $4\mu m$ is chosen so that the output resistance of the devices does not load down the drain nodes of the circuit, since a low impedance at those nodes would limit the amplifier gain as discussed above. For the same reason, long channel devices are used for the input stages loads, M_{19} and M_{20}, which are $\frac{60}{6}$ (remember that these devices operate at a larger current than the input devices, since they must absorb the current from both the input and output legs). The folding transistors, M_{21} and M_{22}, are $\frac{40}{3}$; the $3\,\mu m$ length is again chosen to keep the output resistance of the device high. A longer length is not required because the device is operating at a lower current than the other signal path transistors. The width is chosen so that the V_{gs} of the devices is small enough to insure that $V_{ds_{19,20}} > V_{dsat_{19,20}}$.

The output stage load is formed by M_{23}-M_{26}; cascoding is used to increase the load impedance. Because the current source reference generates the biasing through an improved cascode circuit, use of the cascode only compromises the output swing by ΔV instead of $V_T + \Delta V$. Three micron channel lengths are used to further increase the load resistance; the $40\,\mu m$ width is fixed by the output current desired and the reference current provided. The output resistance of the cascode is:

$$r_{ol} = r_{o24}\,(1 + g_{m23} r_{o23}) \tag{3.79}$$

which, for these numbers, yields a load resistance of $4.8 \times 10^8 \Omega$.

These device sizes and currents result in the following performance parameters, as computed using the formulae derived above: The DC open loop gain is 117,000. The first pole location is approximately 140 Hz, which, with this DC gain, predicts a differential

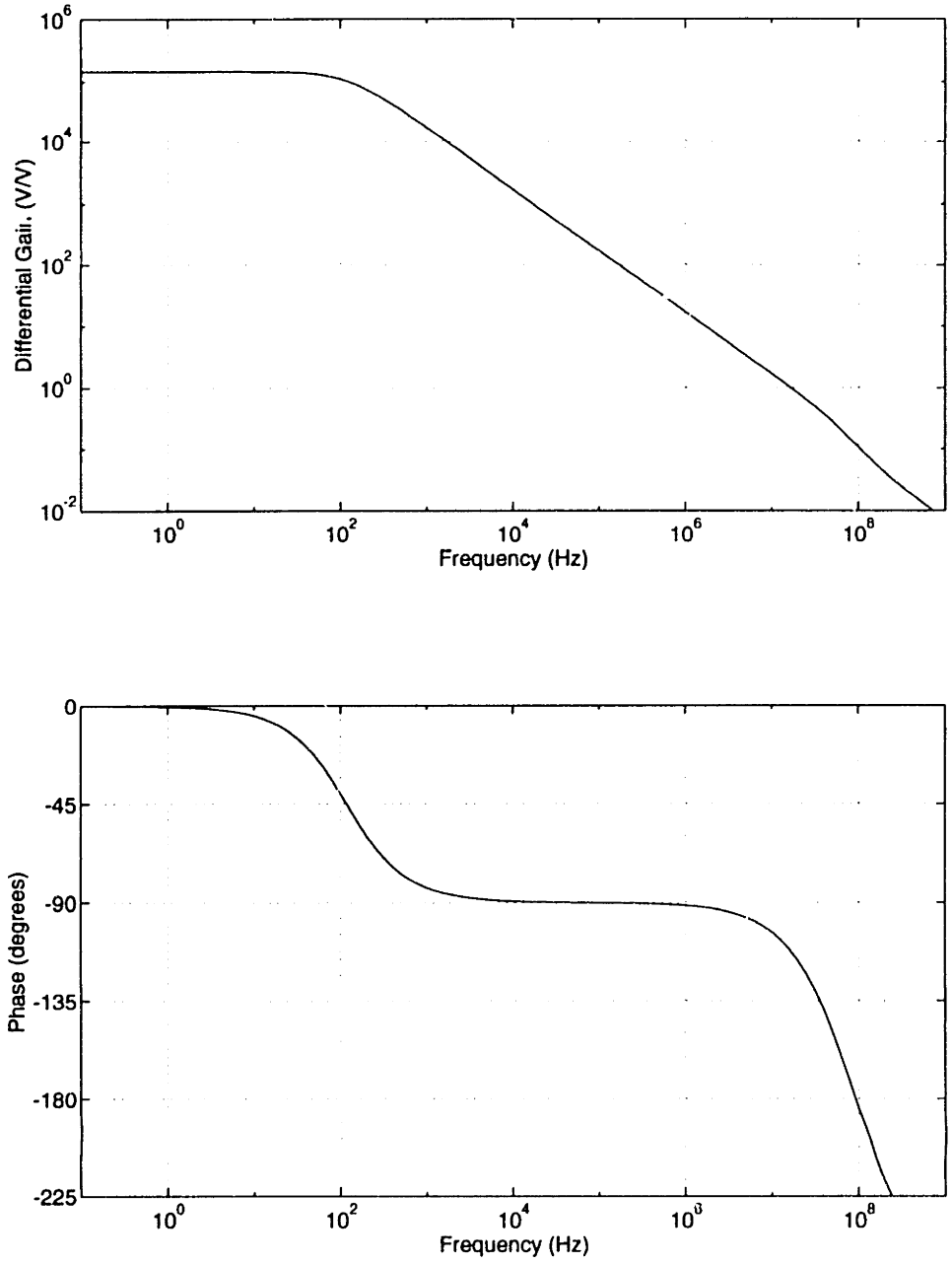


Figure 3.18: Differential loop transmission, HSPICE simulation

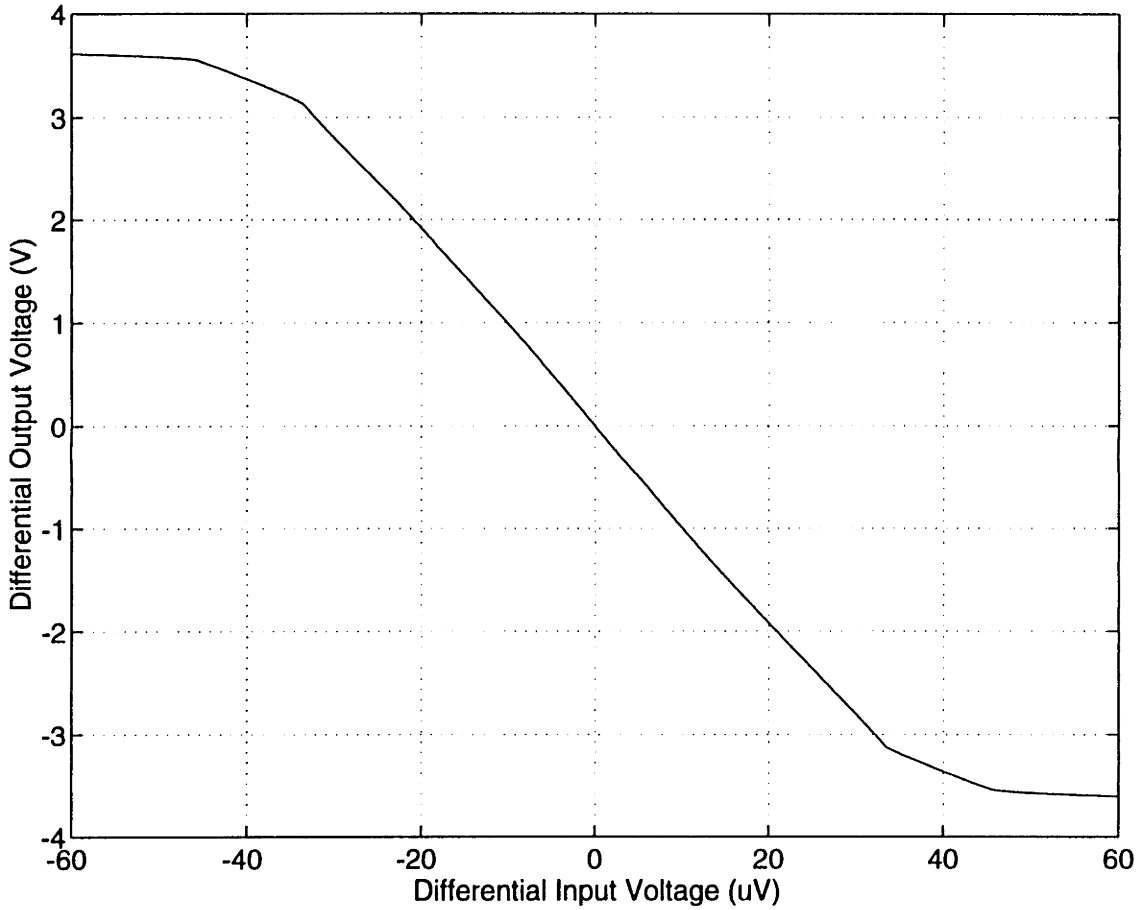


Figure 3.19: DC transfer characteristic, HSPICE simulation

unity gain bandwidth of 16.4 MHz. The second pole location is approximately 42 MHz; using a two pole model of the amplifier, this predicts a phase margin of 69 degrees, and a settling time (to .0001%) of 134 ns. The differential output current of 20 μA results in a slew rate of 4 V/μs into a 5 pF load. Table 3.2 summarizes these computed results as well as parameters obtained via HSPICE simulation. Other relevant simulation results, including Bode plots of the loop transmission, are presented in figures 3.18 and 3.19.

3.5 Stability Considerations

Since the sensor uses an active feedback scheme, the dynamics of the system must be examined to ensure that the sensor will operate as expected. Because there is no true

Table 3.2: Amplifier predicted performance summary

Parameter	*Value*
DC Gain	117,000
Unity Gain Bandwidth	16.4 MHz
Phase Margin	69°
Differential Output Swing	8 V
Output Current	$20\mu A$
Power Dissipation (ex. ref. current)	560 μW
Power Dissipation (incl. ref. current)	1.56 mW
Power Supply	6 V

electrical input, the stability analysis of the sensor is simply an analysis of the electrical feedback loop to determine whether the circuit maintains the desired operating point. The loop transmission can be determined by breaking the loop at any point, injecting a test signal, and examining the gain and phase relationship of the feedback signal generated by the loop. Since the loop is fully differential, both the common mode and differential mode behavior must be examined. This section explores the dynamics of the loop and the conditions under which the sensor will operate properly.

Consider the sensor as shown in figure 3.6. To determine the loop transmission, the feedback loop is broken at nodes (a) and (b). Note that the two loops that together form the differential system are not completely equivalent because of the current ratioing. Each "half loop" must therefore be considered separately, although it will be shown below that, as might be expected, the dynamic behavior of each loop is identical. First, consider the high current loop (the loop containing node (b)). A fictitious perturbation source v_{test} is applied to the noninverting input of the operational amplifier. The signal that results at node (b) is:

$$v_b = -\frac{V_{TH}}{I_2} g_{m2} a(s) v_{test} \tag{3.80}$$

The ratio of the device transconductance g_{m2} to the drain current I_2 to is equal to one-half of the gate drive of the transistor ΔV; using this result and solving for the

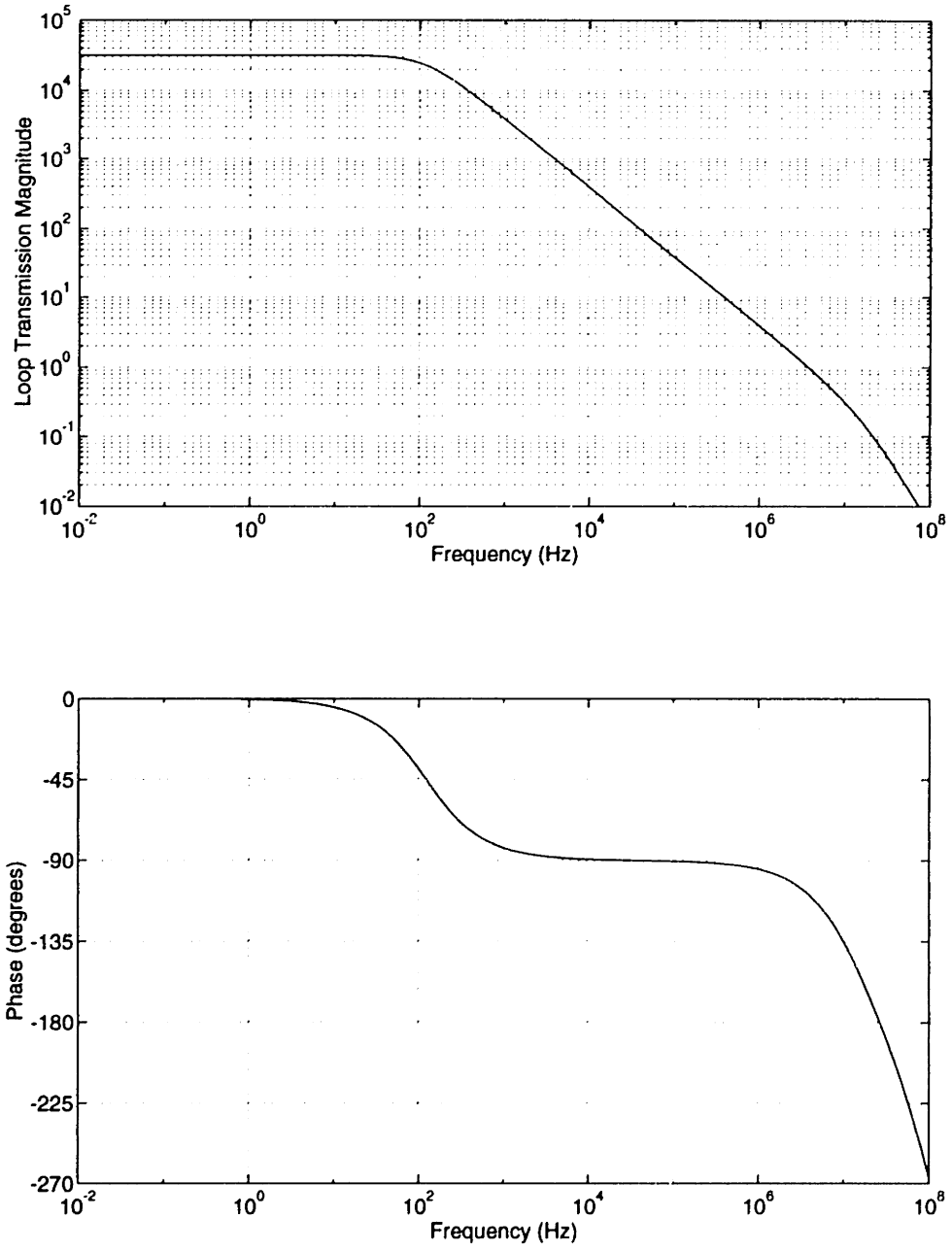


Figure 3.20: Sensor loop transmission bode plots

loop transmission gives:

$$\frac{v_b}{v_{test}} = \frac{2V_{TH}}{\Delta V} a(s) \tag{3.81}$$

Bode plots of this loop transmission for the circuits used in the active needle system are shown in figure 3.20. For this analysis it has been assumed that the effective transconductance of the leg of the differential pair formed by M2 is equal to the transconductance of M2, i.e., this assumes that the source node of M2 is a virtual ground. This represents a worst-case situation, since the small changes in the source node voltage that do occur act to reduce the effective transconductance. Furthermore, since the chopper modulation has no net effect on the signal, it is neglected in this analysis.

Several important conclusions can be drawn from equation 3.81. First, the dynamics of the feedback loop are similar to the dynamics of the operational amplifier, since the dominant pole of the amplifier is many orders or magnitude lower in frequency than the poles associated with the other components of the loop. Second, if the gate drive of the differential pair transistor is greater than $2V_{TH}$, the scale factor $\frac{2V_{TH}}{\Delta V}$ is less than one. Thus, if the gate drive is larger than apprximately 50 mV the loop is guaranteed to be stable if the operational amplifier is stable. Third, since the differential pair transistors are purposely mismatched so that the nominal differential output of the op amp is zero, the gate drives of the differential pair transistors are equal, and the dynamic behavior of each leg of the differential loop is identical, even though they operate at different currents.

The common mode dynamics are much more subtle, since, in theory, the operational amplifier common mode gain is zero and the common mode loop is always stable. In practice, the common mode gain is small but nonzero because of device mismatch. Consequently, detailed mathematical analysis of the common mode behavior offers no insight, since the results will depend on mismatch parameters that are a function of the fabrication process, circuit layout, etc. Qualitatively, however, it can be demonstrated that the circuit should be common mode stable.

Assume that the common mode level at the output of the op amp rises. This increase will cause a very small decrease in the magnitude of the currents I_1 and I_2 because of the finite output conductance of the current source. This small change in the currents will cause a small change in the common mode level at the input to the op amp. As long as this change in common mode level is within the common mode input range of the amplifier, the amplifier common mode feedback will act to restore the common mode level at the output of the amplifier. Since the output conductance of the current source and the common mode gain of the op amp are both relatively low, small perturbations in the common mode level should not result in significant common mode shifts throughout the circuit. The op amp common mode feedback should therefore be able to maintain common mode stability.

3.6 Preamplification

In order to prevent noise corruption of the temperature signal, the output of the sensor is amplified using a switched capacitor gain stage, shown in figure 3.21. As will be shown below, the circuit amplifies the temperature signal by a factor equal to the capacitor ratio. Noise from the amplification is reduced by using correlated double sampling. Finally, the switched capacitor approach produces an output signal that interfaces well with the sigma-delta modulator so that no additional interfacing circuitry is required.

3.6.1 Transfer Characteristic

The preamplifier gain is realized by charge transfer between the input capacitor and the feedback capacitor. Consider the input half-circuit formed by v_{i+} and v_{o+}. When ϕ_2 is active, the input capacitor C_1 holds a charge $Q_1 = C_1(v_{i+} - V_{CM})$, where V_{CM} is the quiescent common mode voltage at the output of the op-amp. The charge on the feedback capacitor C_2 is $Q_2 = C_2(V_B - V_{CM})$. The total charge stored is therefore

$$Q_{tot,2} = C_1(v_{i+} - V_{CM}) + C_2(V_B - V_{CM}) \tag{3.82}$$

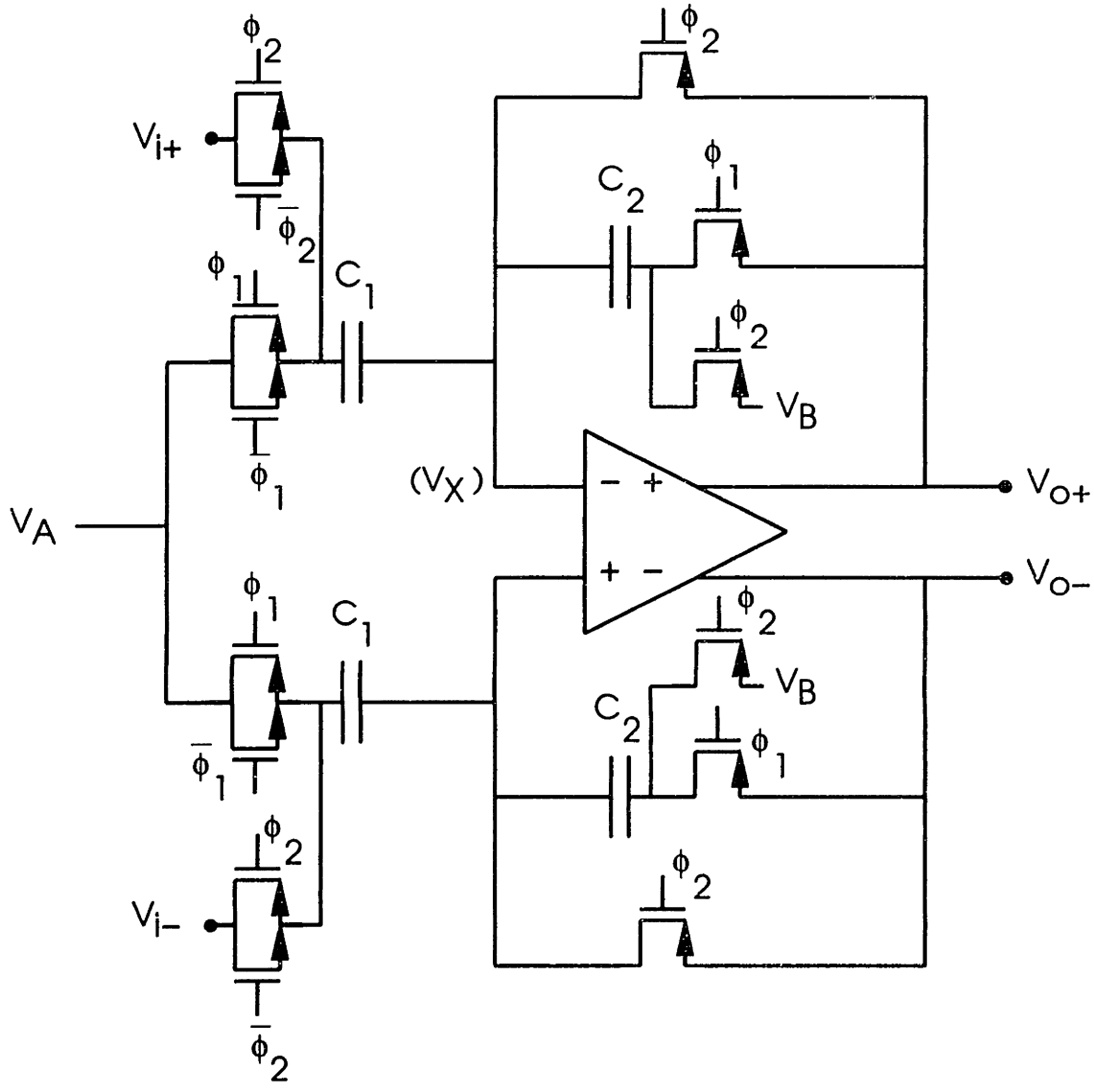


Figure 3.21: Switched capacitor preamplifier

When ϕ_1 is active, capacitor C_1 holds a charge $Q_1 = C_1(V_A - V_X)$, where V_X is the common mode voltage that develops at the amplifier inputs (see below). The charge on C_2 is $Q_2 = C_2(v_{o+} - V_X)$. The total charge is now

$$Q_{tot,1} = C_1(V_A - V_X) + C_2(v_{o+} - V_X) \tag{3.83}$$

Charge conservation requires that the total charge $Q_{tot,1}$ stored during phase 1 is equal to the total charge $Q_{tot,2}$ stored during phase 2:

$$C_1(v_{i+} - V_{CM}) + C_2(V_B - V_{CM}) = C_1(V_A - V_X) + C_2(v_{o+} - V_X) \tag{3.84}$$

from which it is found that

$$v_{o+} = \frac{C_1}{C_2}(v_{i+} - V_A) + V_B + \left(\frac{C_1 + C_2}{C_2}\right)(V_X - V_{CM}) \tag{3.85}$$

Clearly this analysis applies to the lower half circuit as well, such that

$$v_{o-} = \frac{C_1}{C_2}(v_{i-} - V_A) + V_B + \left(\frac{C_1 + C_2}{C_2}\right)(V_X - V_{CM}) \tag{3.86}$$

It directly follows that the differential output of the preamplifier is

$$v_{od} = v_{o+} - v_{o-} = \frac{C_1}{C_2}(v_{i+} - v_{i-}) \tag{3.87}$$

and the differential gain is

$$\frac{v_{od}}{v_{id}} = \frac{C_1}{C_2} \tag{3.88}$$

The differential gain is therefore determined by the ratio of the capacitor areas, which can be controlled well.

A more important observation to be made from this analysis, however, relates to the magnitude of the voltage V_X. Although the differential output is a function of the differential input alone, the magnitude of each output is a function of several voltages including V_A and V_B. These voltages cause the shift in the common mode level of the inputs during ϕ_1. In order for the amplifier to work correctly, it is essential that the

common mode level of the input remain within the allowed input range of the amplifier. From equations 3.85 and 3.86, the common mode voltage at the amplifier outputs is:

$$\begin{aligned} 2v_{o,cm} &= \frac{C_1}{C_2}(v_{i+} - V_A) + V_B + \left(\frac{C_1 + C_2}{C_2}\right)(V_X - V_{CM}) \\ &+ \frac{C_1}{C_2}(v_{i-} - V_A) + V_B + \left(\frac{C_1 + C_2}{C_2}\right)(V_X - V_{CM}) \end{aligned} \tag{3.89}$$

The internal op-amp common mode feedback will maintain $v_{o,cm} = V_{CM}$. Substituting this into equation 3.89 and solving for V_X gives:

$$V_X = \left(\frac{C_1 + 2C_2}{C_1 + C_2}\right) V_{CM} - \left(\frac{C_1}{C_1 + C_2}\right)(V_A - v_{i,cm}) - \left(\frac{C_2}{C_1 + C_2}\right) V_B \tag{3.90}$$

where $v_{i,cm}$ is the common mode level at the signal input:

$$v_{i,cm} = \frac{v_{i+} + v_{i-}}{2} \tag{3.91}$$

In the case where the $V_A = V_B = v_{i,cm} = V_{CM}$, $V_X = V_{CM}$ and the circuit amplifies normally. If this is not the case, however, and the voltage V_X exceeds the range of common mode input voltage for which the op amp operates normally, the op-amp will no longer function correctly and, most likely, the amplifier will saturate. This, of course, renders the circuit useless as a preamplifier.

The choice of the bias voltages V_A and V_B is therefore critical to proper circuit operation when $v_{i,cm} \neq V_{CM}$. Since V_A is weighted much more strongly than V_B (as seen in equation 3.90), it is used as the mechanism for controlling V_X, and V_B is fixed at V_{CM}. This gives:

$$V_X = V_{CM} + \left(\frac{C_1}{C_1 + C_2}\right)(V_A - v_{i,cm}) \tag{3.92}$$

While it is not necessary that V_A equal the common mode input level, it must be selected such that the common mode feedback of the op amp does not fail. From figure 3.15, it is clear that the amplifier fails when the input common mode level is approximately 1.5 V above V_{CM}. The common mode level at the inputs must therefore be less than V_{CM} + 1.5 Volts. In this particular application, the exact common mode level at the

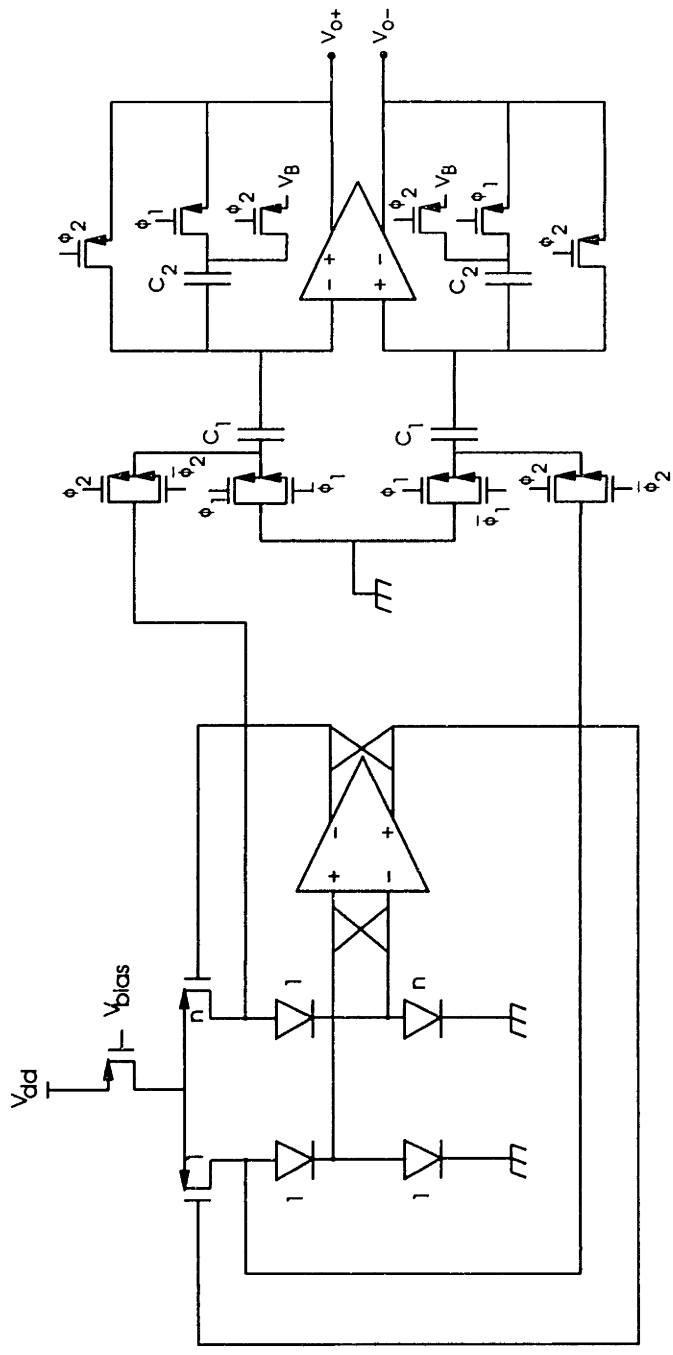


Figure 3.22: Proper sensor/preamplifier connection

inputs is unknown; however, it is approximately equal to two diode drops above a known supply voltage.[13] Proper amplifier operation can be obtained by tying V_A to the same known supply voltage used as the reference potential for the sensor, as shown in figure 3.22. In this case, $V_A = v_{i,cm} - 1.2$ Volts, where it is assumed that the diode drops are approximately 0.6 V. Consequently,

$$V_X = V_{CM} - 1.2\left(\frac{C_1}{C_1 + C_2}\right) \tag{3.93}$$

from which it is clear that for positive capacitor values $V_X < V_{CM}$. This connection ensures that the common mode range of the sensor maps into the valid common mode range of the preamplifier. This is an important conclusion since setting V_A equal to the op amp common mode reference voltage (as is usually done) does *not* guarantee that the ranges will overlap correctly for proper operation of the gain stage; in fact, in this case it places V_X at or above the common mode input range limit. Setting V_A equal to the sensor reference potential *does* guarantee proper behavior in this case, as is clear from equation 3.93.

3.6.2 Noise Performance

The preamplifier uses correlated double sampling [57] to significantly reduce the effects of the op amp noise, which is the major noise source in most amplification circuits. The basic principle behind this technique is that the noise signal is sampled twice and subtracted during each amplification cycle; if the noise signal is not changing significantly between the sampling intervals, then the difference in the samples is nearly zero and the noise is effectively cancelled.

Figure 3.23 shows the simplified half circuit equivalent of the preamplifier. For simplicity, the common mode biases have all been set to 0 V; it should be clear that this is unimportant to the analysis. A noise source is shown at the inverting input to the operational amplifier to model the op amp noise. The fundamental clock period is T:

[13] This should be apparent from the circuit diagram of the temperature sensor.

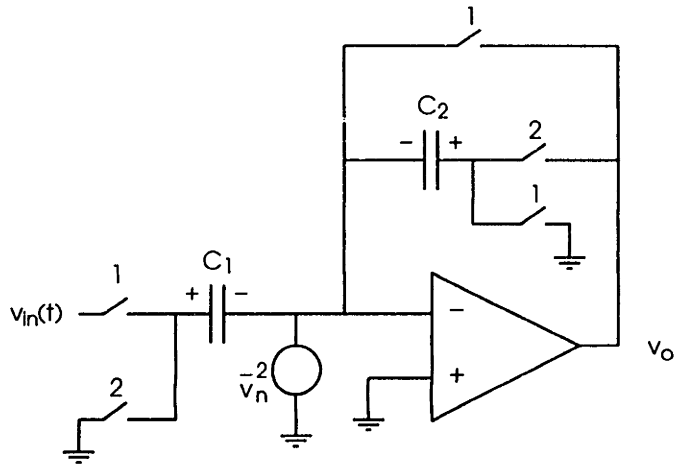


Figure 3.23: Simplified half circuit for noise analysis

The switches marked "1" are closed at times $t = nT$ and those marked "2" are closed exactly one half-period later, at times $t = nT + \frac{T}{2}$, where n is an integer. The switches are all considered ideal, and are closed for a negligibly small amount of time, so that the switches can in essence be considered samplers.

The total charge in the system during each of the clock phases is given by:

$$Q_{tot,1}(nT) = C_1[v_{in}(nT) - v_n(nT)] - C_2 v_n(nT) \tag{3.94}$$

$$Q_{tot,2}(nT + \frac{T}{2}) = C_2[v_o(nT + \frac{T}{2}) - v_n(nT + \frac{T}{2})] - C_1 v_n(nT + \frac{T}{2}) \tag{3.95}$$

Charge conservation on each cycle requires that $Q_{tot,1}(nT) = Q_{tot,2}(nT + \frac{T}{2})$: Setting equations 3.94 and 3.95 equal and solving for the output voltage v_o yields:

$$v_o(nT + \frac{T}{2}) = \frac{C_1}{C_2} v_{in}(nT) + \left[\frac{C_1}{C_2} + 1\right]\left[v_n(nT + \frac{T}{2}) - v_n(nT)\right] \tag{3.96}$$

From which it is clear that the output voltage consists of the amplified input (the first term on the right hand side) as well as a sampled noise component (the second term). This second term also demonstrates the effects of the correlated double sampling; if the noise signal is changing slowly enough such that $v_n(nT + \frac{T}{2}) \approx v_n(nT)$, then the second term is zero and the noise is eliminated.

The correlated double sampling is therefore most effective on low frequency signals (like $1/f$ noise), and in fact will completely eliminate the effects of any DC offsets in the op amp. As the noise signal frequency increases, less and less of the noise is cancelled, so broadband signals (like thermal noise) are less effectively eliminated. The effect of the correlated double sampling as a function of frequency can be demonstrated mathematically by assuming a purely sinusoidal noise source. For example, assume the noise signal is:

$$v_n(t) = A\cos(2\pi f t) \tag{3.97}$$

The noise term (ignoring the scale factor for the moment) is therefore:

$$\begin{aligned} v_n(nT + \frac{T}{2}) - v_n(nT) &= A\cos(2\pi f nT + \frac{\pi fT}{2}) - A\cos(2\pi f nT) \\ &= -2A\sin(2\pi f nT + \frac{\pi fT}{2})\sin(\frac{\pi fT}{2}) \end{aligned} \tag{3.98}$$

Because of the sampling process, this is a discrete-time signal; call it $e[n]$. The mean-square power spectrum of this signal is given by:

$$< e[n] >^2 = \lim_{m\to\infty}\left\{\frac{1}{2m+1}\sum_{n=-m}^{m}\left[-2A\sin(2\pi f nT + \frac{\pi fT}{2})\sin(\frac{\pi fT}{2})\right]^2\right\} \tag{3.99}$$

which can be simplified and evaluated using the exponential definition of the sine function, from which one finds that:

$$< e[n] >^2 = \begin{cases} 4A & f = \frac{n}{T},\ n \text{ odd} \\ 2A\sin^2(\frac{\pi fT}{2}) & \text{otherwise} \end{cases} \tag{3.100}$$

The mean-square value of $v_n(t)$ is:

$$\begin{aligned} < v_n(t) >^2 &= \frac{1}{T}\int_0^{2\pi} A\cos(\omega t)dt \\ &= \frac{A}{2} \end{aligned} \tag{3.101}$$

The effective "gain" of the noise as a function of frequency is therefore:

$$G = \begin{cases} 8 & f = \frac{n}{T},\ n \text{ odd} \\ 4\sin^2(\frac{\pi fT}{2}) & \text{otherwise} \end{cases} \tag{3.102}$$

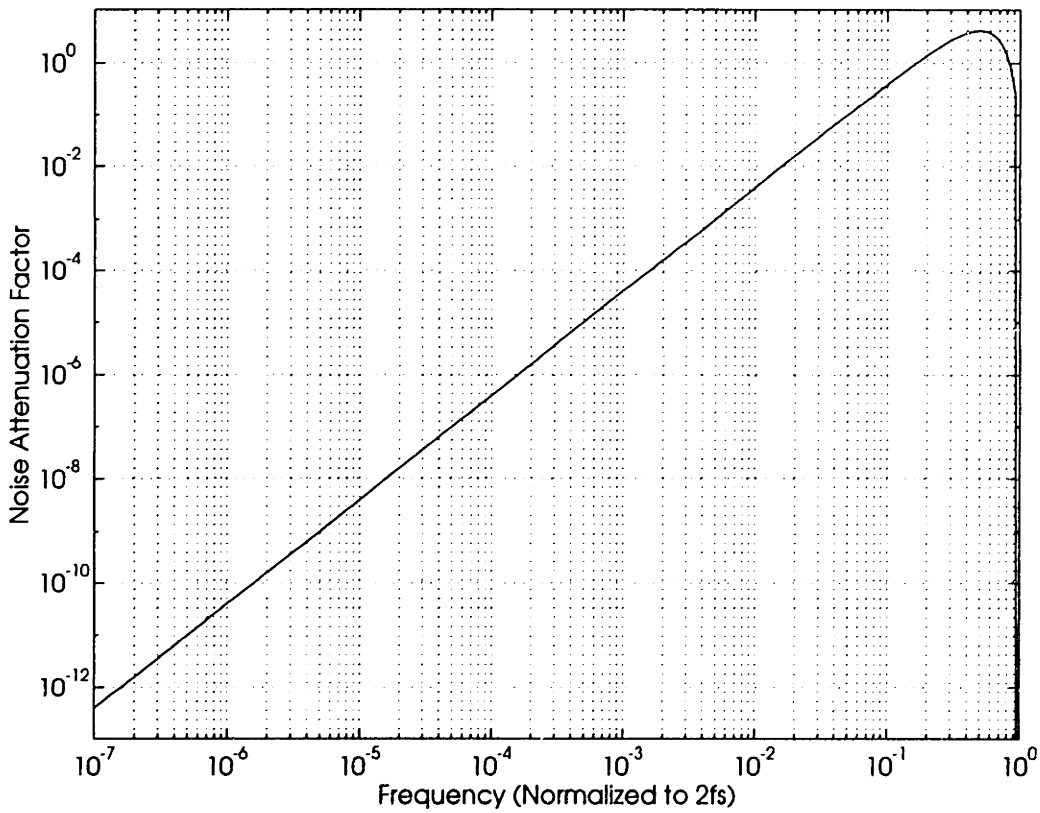


Figure 3.24: Gain factor "G" vs. frequency

Qualitatively, the gain factor G has the expected behavior: At low frequencies ($f \ll \frac{1}{T}$), $G \approx 0$, since the noise signal is quasi-static and is almost completely cancelled by the subtraction process. At frequencies near the sampling frequency, the change in the signal amplitude between the sampling at nT and $nT + \frac{T}{2}$ is large; G is higher there. At frequencies that are exact odd multiples of the sampling frequency, the amplitude change between samples is maximum, as is G. Conversely, even multiples of the sampling frequency are aliased to DC by the sampling process, and G goes to zero at each of those points. The gain factor G over the frequency interval $0 \leq f \leq 2f_s$ is shown in figure 3.24; since G depends only on the ratio of $\frac{f}{f_s}$, the frequency axis has been normalized to $2f_s$.

Since the noise bandwidth is significantly higher than the sampling frequency,

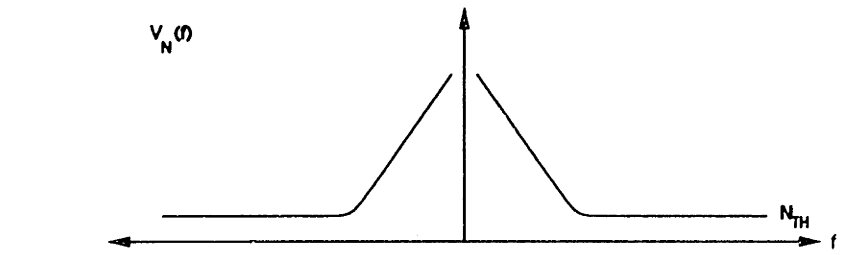
aliasing of the noise occurs. Because the noise cancellation takes place as part of the sampling process, it is not apparent whether to consider the effects of aliasing *before* or *after* accounting for the correlated double sampling noise shaping factor G. The noise is clearly shaped by G, but it is important to note that this shaping occurs *after* the noise is aliased by the sampling process. If the aliasing were to occur after the noise shaping, then there would be noise energy at even multiples of the sampling frequency, which cannot be the case since these frequencies are completely indistinguishable from DC signals and therefore the noise energy at these frequencies must be zero.

The spectra corresponding to this process is shown in figure 3.25. Figure 3.25(a) shows the general shape of the op amp input noise; at low frequencies, $1/f$ noise dominates. At higher frequencies, thermal noise dominates. The effective noise bandwidth f_{th} is very large, on the order of 10^{12} Hz. When this noise spectrum is sampled at f_s, significant aliasing occurs; this aliased spectrum is shown in figure 3.25(b). The $1/f$ component is present, but is essentially swamped out by the much larger aliased thermal noise component. As a result, the spectrum looks mostly white. The correlated double sampling noise shaping spectrum is shown in figure 3.25(c); peaks occur at odd multiples of the sampling frequency and nulls occur at even multiples. When this spectrum is multiplied by the noise spectrum of figure 3.25(b), the spectrum in figure 3.25(d) results.

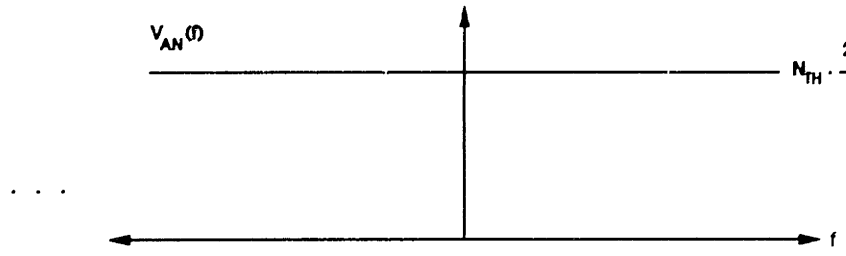
Mathematically, since the noise generated by the op amp is dominated by the noise from the input MOSFETs, it has two primary components, one from the $1/f$ noise and the other from the thermal noise. As discussed earlier, this noise is:

$$\frac{\overline{v_{oa}}^2}{\Delta f} = \frac{8kT_o}{3g_m} + \frac{K}{WLf^{\alpha}} \tag{3.103}$$

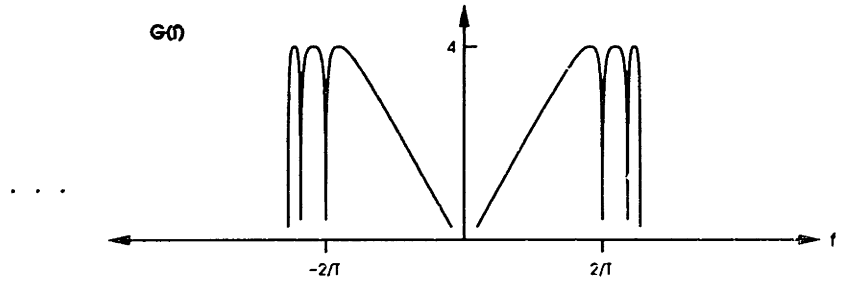
where T_o is used to represent the absolute temperature instead of T to avoid confusion with the clock period T. The $1/f$ noise dominates at low frequencies but becomes negligible above frequencies of several tens of kilohertz, at which point the thermal noise dominates. The bandwidth of this thermal noise is quite large, typically on the



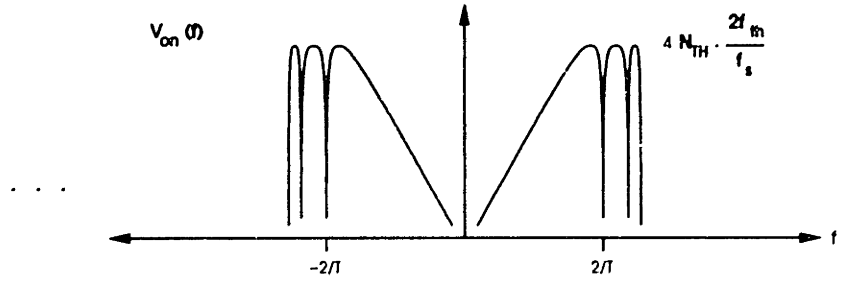
(a)



(b)



(c)



(d)

Figure 3.25: Correlated double sampling spectra

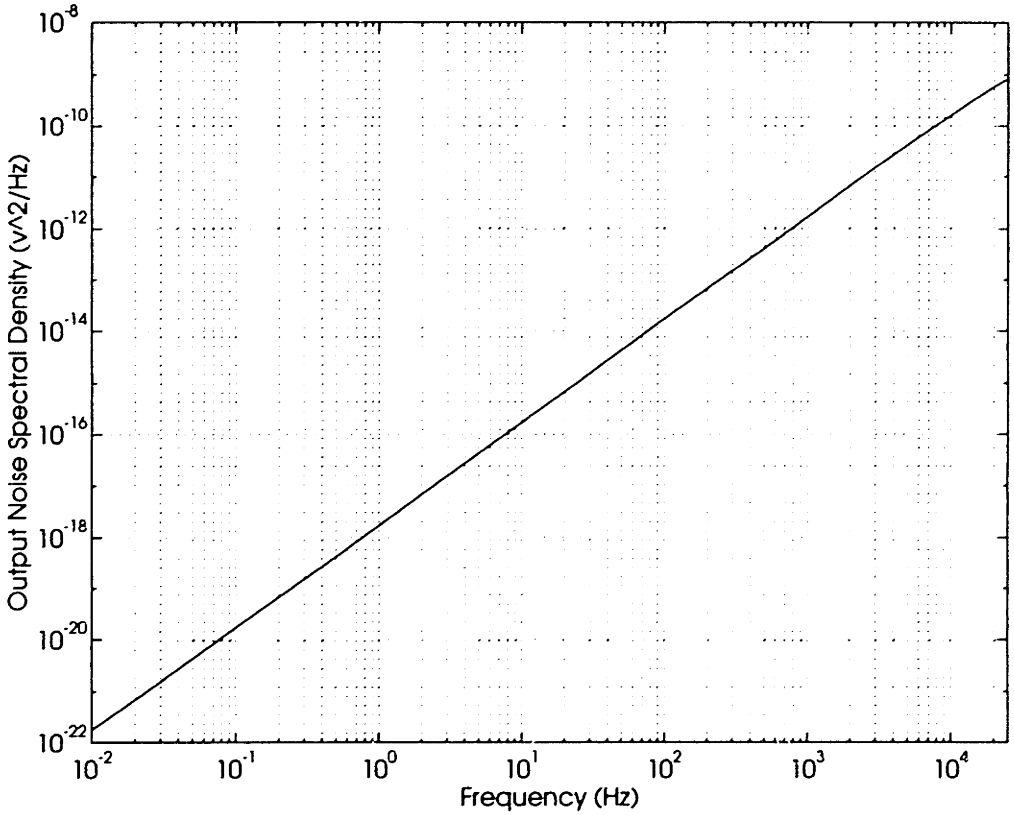


Figure 3.26: CDS preamp output noise spectral density

order of 10^{12} Hz;[14] this is the component that gets significantly aliased, since the clock frequency for this system is on the order of 50 kHz. Thus, the aliased noise spectral density, defined over the frequency range of $[0, \frac{1}{t}]$ is:

$$\frac{\overline{v_{n,a}}^2(f)}{\Delta f} = \sum_{n=0}^{f_{th}T} \overline{v_{oa}}^2(f + \frac{n}{T}) \tag{3.104}$$

where f_{th} is the thermal noise bandwidth. For simplicity, it was assumed that the noise abruptly drops to zero at this frequency. The noise spectral density at the output of the amplifier due to the op amp noise is the product of equations 3.104 and 3.102:

$$\frac{\overline{v_{o,n}}^2(f)}{\Delta f} = \sin^2(\frac{\pi f T}{2}) \cdot \sum_{n=0}^{f_{th}T} \frac{8kT_o}{3g_m} + \frac{K}{WL(f + \frac{n}{T})^\alpha} \tag{3.105}$$

[14]Clearly the noise cannot have an infinite bandwidth, as this would imply infinite noise energy.

Table 3.3: Sensor noise sources

Error Source	*Value (nV)*
Device Mismatch/Op amp gain	9.3
MOS Diff Pair	0.3
Op Amp	12.0
Small low current diode	58.0
All other diodes (each)	18.0
Preamplification	5.3
Total	67.8

The special case for G need not be considered since this function is only defined over the above stated frequency interval. This output noise spectral density function is shown in figure 3.26, where $WL = 1920\,\mu m^2$, $g_m = 0.3\,\mathrm{m}\mho$, $T = 20\,\mu s$, and the typical values for the noise parameters stated earlier were used. The values for WL and g_m are based on the devices actually employed in the fabricated circuit. It is clear from this figure that over the bandwidth of interest, $0.01 - 1$ Hz, the noise spectral density is less than $5.3 \times 10^{-9}\,\frac{\mathrm{V}}{\sqrt{\mathrm{Hz}}}$; integrating over this bandwidth gives a total noise contribution of 5.3 nV.

3.7 Summary: Predicted Sensor Performance

With the sensor circuit completely specified, it is now possible to calculate the expected resolution of the sensor using the results from sections 3.3 and 3.6.2. The error due to mismatch and finite op amp gain is described by equation 3.41; with an assumed mismatch of 10 mVand the op amp parameters derived above, this error source is 9.3 nV.

The output noise contribution from the differential pair transistors used in the sensor circuit is calculated from equations 3.46 and 3.45. The flicker noise coefficient for the PMOS devices manufactured in the process used for this project is K =

$6 \times 10^{-10} \mu m^2 V^2 \text{Hz}$,[15] and the noise exponent is $\alpha = 1.1$. Thus, over a bandwidth of .01 - 1 Hz (the lower limit corresponds to the duration of a measurement; the upper limit is the maximum signal frequency of interest), the total gate-referred noise is 6.84 μV from the 25/3 device, 2.16 μV from the 250/3 device, and 1.1 μV from the 100/30 current source device. When this noise is referred to the sensor output using the open loop amplifier gain of 1.17×10^5, the total output referred noise contribution from the MOS devices is 0.3 nV, which is less than 0.2% of the desired minimum detectable signal.

The noise contribution from the operational amplifier is limited to the input referred thermal noise as described above. This is primarily the thermal noise from the input transistors and their current source loads, since the noise from other devices and kT/C noise from the compensation in the op amp is reduced by (at least) the input stage gain when referred to the op amp input. Because the thermal noise is white, the total noise contributed is the thermal noise spectral density multiplied by the bandwidth of interest, approximately 1 Hz. Thus, the total noise at the sensor output due to the op amp is given by:

$$v_{oa} = \sqrt{\frac{\gamma 4kT g_{d0,17}}{g_{m17}^2} + \left(\frac{g_{m19}}{g_{m17}}\right)^2 \left(\frac{\gamma 4kT g_{d0,19}}{g_{m19}^2}\right)} \tag{3.106}$$

where g_{m17} is the transconductance of the op amp input stage devices and g_{m19} is the transconductance of the active load devices as shown in figure 3.10. Since both devices are operated in saturation, $\gamma = \frac{2}{3}$. As given in table 3.1, the input device M17 has $\frac{W}{L} = \frac{480}{4}$, which gives $g_{m17} = .286\,\text{m}\mho$ and $g_{d0,17} = .646\,\text{m}\mho$; for the active load, $\frac{W}{L} = \frac{60}{6}$, which gives $g_{m3} = .175\,\text{m}\mho$ and $g_{d0,19} = .339\,\text{m}\mho$, and the total noise contribution is 12 nV.

The noise from the diodes used for sensing and current ratio control is computed from equations 3.49 and 3.50. For the process used to fabricate these devices, the

[15]Technically, the units of K should be $\mu m^2 V^2 \text{Hz}^{\alpha-1}$ to reflect the correct units for $\overline{v_n^2}$ when $\alpha \neq 1$. Although K is stated in $\mu m^2 V^2 Hz$ it is implicit that the units are properly adjusted for the cases where $\alpha \neq 1$.

flicker noise coefficient is $K_d = 1.6 \times 10^{14}\, V^2 \mu m^2\, Hz^{\gamma-1}$ and the noise exponent is $\gamma = 1.03$.[16] The diode areas are maximized in order to reduce the noise as much as possible; the smallest diode used (the "unit" diode) is 961 μm^2, which contributes noise of 58 nV at the bias current of 40 μA used in the system. The other diode in the high current leg contributes only 18 nV since it is 10 times larger. The low current leg of the sensing circuit uses two of the unit diodes, but at a current that is only 4 μA, so the noise contributed from each diode is also 18 nV.

The only remaining noise contribution is from a preamplification circuit used to boost signal levels; its input referred noise directly corrupts the sensor output. As will be shown below in section 3.6.2, this component is 5.3 nV over the frequency range of interest. The total noise referred to the sensor output is the mean square sum of each of the individual noise components; using the numbers derived in this section, this total noise is 67.8 nV, which, for this system, corresponds to an error of approximately 0.34 m°C. Table 3.3 summarizes all of the noise sources.

[16]These values are extrapolated from actual diode noise measurements.

Chapter 4

Analog-to-Digital Conversion

One of the major advantages of the active needle system is the fully digital transfer of information onto and off of the needle probes. The consequence of this, however, is that some type of analog to digital conversion circuitry must be present on each of the sensor chips. This conversion increases both the complexity of the sensor chips and the total chip area, but the ability to communicate with the needle using a completely digital interface and the elimination of signal corruption from long analog signal lines are advantages that far outweigh these drawbacks. This chapter describes the oversampling delta-sigma analog-to-digital (A/D) conversion technique used on the temperature sensor chips. First, the fundamental principles behind the technique are presented. This is followed by a description of the modulator design used on the active needle and the performance that can be expected from this system. The chapter concludes with an examination of the circuit implementation of the modulator and a discussion of the limitations of the system.

4.1 Principles of Oversampled Data Conversion

4.1.1 Background: Conventional A/D Conversion

Traditional analog to digital conversion techniques involve performing one data conversion for each sample of the input signal, resulting in a one-to-one correspondence

between the input samples and the digital output data. For high resolution systems, however, this type of conversion becomes very difficult, since finer and finer quantizations of each input sample are required. Because of finite word length effects, there is inherently an uncertainty in the digital output: Each digital output word D_i is associated with a *range* of analog input samples (call this range $[a, b]$). In other words, if the digital word is re-converted into an analog signal, the difference between the original analog input and the reconstructed analog signal will be nonzero. In theory, this error difference will have a mean value of zero, and will be uniformly distributed over the interval $[\frac{a-b}{2}, \frac{b-a}{2}]$. This uncertainty can be viewed as the equivalent of a noise component present on the analog input signal. If the input signal is sufficiently random, this *quantization noise* looks uniform, and can be modelled as a white noise source that is injected at the input to the A/D converter.

Thus, the fundamental digital resolution of the input signal that can be attained in these systems is equal to the resolution of the converter itself. This has two major implications for these systems. First, the initial conversion of the analog signal to a digital signal *must* be at the highest resolution level required by any downstream digital system; further digital processing of the signal cannot extract any information below the quantization noise threshold, which is determined solely by the initial conversion. Second, integrated circuit implementations of these high resolution systems must perform precise digitization on-chip, which in turn requires precision components that are often difficult if not impossible to fabricate.

4.1.2 Oversampled A/D Converters

Oversampled modulators approach the conversion process as one of signal processing rather than signal quantization: The input samples are not considered in isolation, but are viewed as part of a continuous set of samples of an input signal. The basic framework of the oversampled A/D converter is shown in figure 4.1. The analog signal is first low pass filtered to prevent aliasing in the sampling process. This filtered

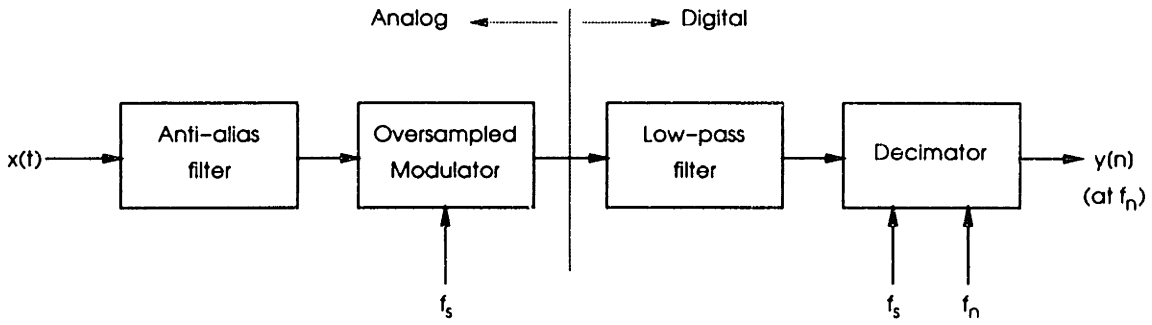


Figure 4.1: Oversampled A/D conversion architecture

signal is then sampled by the modulator (which is typically some form of interpolative coder) at a frequency much higher than the Nyquist frequency f_n of the filtered input signal, hence the term "oversampling." The output of the modulator is a digital data stream corresponding to a coarse quantization of the input. As will be shown below, the modulator is a feedback system, and the coarse quantization is a function of the previous input samples as well as the current input sample; it is *not* a quantization of the input sample. This digital data stream is then downsampled and digitally processed (filtered) to produce the high resolution digital result at f_n. Thus, although the modulator generates a digital signal from an analog one, the "real" conversion is performed in the digital domain; the modulator merely encodes all the information necessary for the conversion into the high speed, coarsely quantized output. This reduces the conversion process to a simpler problem of digital signal processing. This is the fundamental principle behind the oversampling technique: that signal processing can be more easily performed in the digital domain. The most stringent performance requirements must be met in the digital domain, where they are much easier to satisfy. A more rigorous discussion of all aspects of the the oversampling process can be found in [58].

4.1.3 Modulators

As was described in the previous section, the primary purpose of the modulator is to "encode" the input signal information into a digital data stream in such a way that it can be extracted by the digital processor. The key component in oversampled A/D

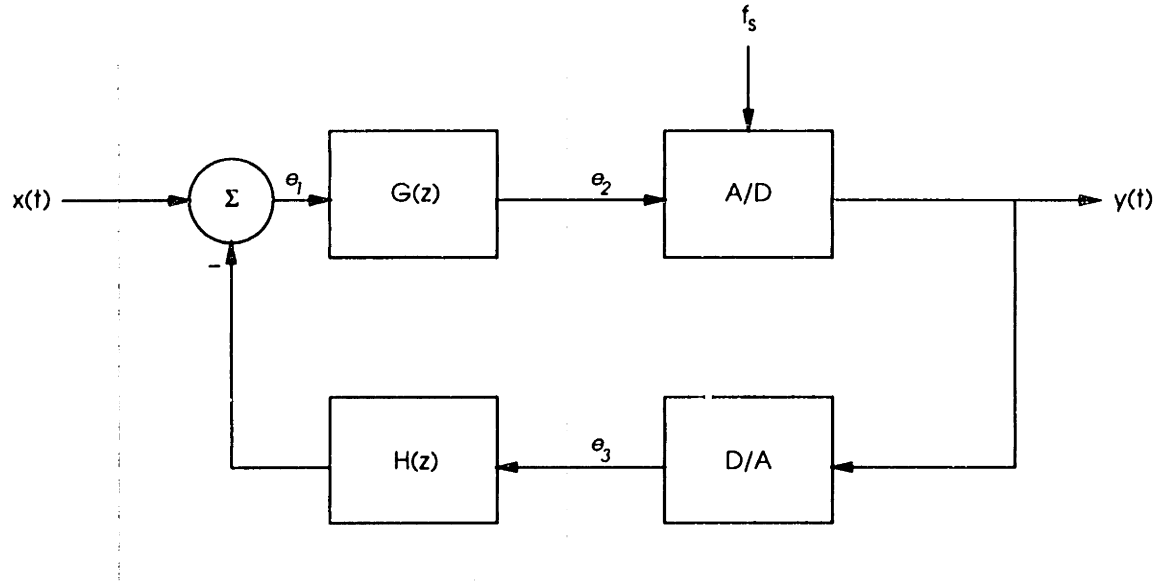


Figure 4.2: General modulator loop

converters is therefore the modulator, since this is where the initial data conversion takes place. For the purposes of the active needle system, this is the component of the converter that must be implemented on-chip; the additional downsampling and filtering of the digital data can be performed on the host personal computer. Thus, the discussion here focuses on the modulator only.

One basic modulator architecture is shown in figure 4.2. In the most general case, the area in the dashed box is unknown and can be replaced by an arbitrary transfer function $T(z)$ that takes the signal $x(t)$ and $e_3(t)$ as inputs and produces the A/D input signal $e_2(t)$ as output. The forward path consists of an arbitrary transfer function $G(z)$ and an A/D (comparator) that produces the digital output signal. The feedback path has a D/A converter that regenerates an analog signal from the quantizer output and another arbitrary transfer function $H(z)$. Since this loop represents a sampled-data system, there is also a clock frequency (or, equivalently, an output data rate) f_s and a corresponding clock period $T = \frac{1}{f_s}$ associated with it. Samples of the (analog) input signal are taken and one quantization is performed every T seconds.

The quantization error for the circuit in figure 4.2 (and for oversampled modulators

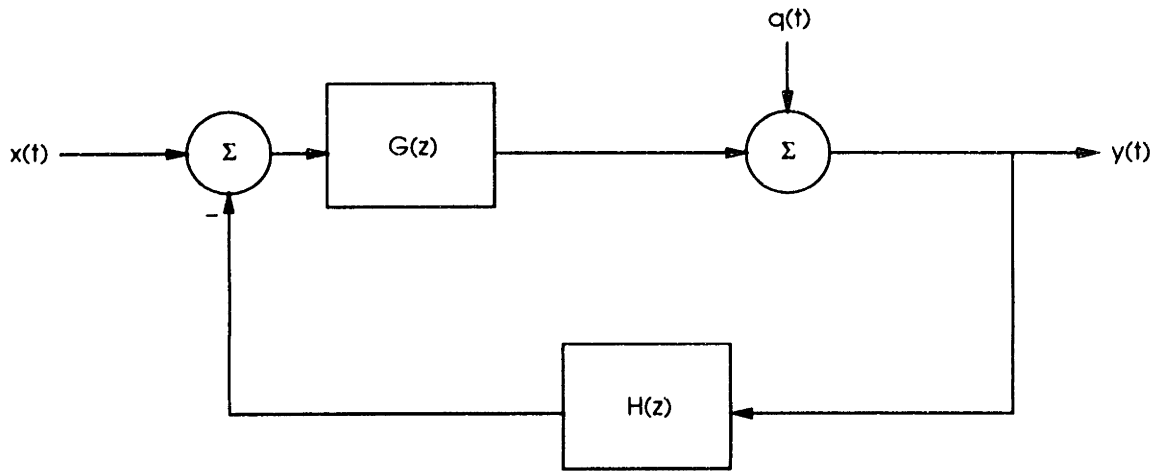


Figure 4.3: Linearized model of modulator loop

in general) results from the finite word length of the quantizer in the forward path. Ideally, the signal at node e_3 would be a (possibly) delayed version of the signal at node e_2, and for linear analysis purposes the quantizer/converter combination could be replaced by a linear gain element. During the quantization process, however, information about the signal at e_2 is lost in exactly the same way as described in the previous paragraphs, and the magnitude of the difference between the quantizer input and the output of the D/A converter is not necessarily zero. This is the quantization noise source for oversampled modulators; for analysis purposes this can be modelled as an injected noise source as shown in figure 4.3. [1]

There are several important points to note about this model. First, it is clear that the signal e_2 is not completely independent of the input signal, and, therefore, the quantization error is not completely uncorrelated with the input as an ideal white noise source would be. The degree to which the input and the quantization noise are correlated is dependent on the transfer functions $G(z)$ and $H(z)$ and the exact topology used to implement the loop. Thus, modelling the quantization noise as an additive white noise source is, strictly speaking, erroneous [59,60]. For analysis purposes, however, it is

[1]There should also be a gain element associated with this linearization of the quantization and reconversion process, but with proper choice of loop coefficients this gain can always be made equal to one.

assumed that the correlation is weak between the input and the noise source so that the white noise approximation is valid. Second, unlike traditional A/D converters, the quantization noise and the input signal are injected into the system at *different* nodes that are separated by the transfer function $G(z)$. The transfer functions from input- and quantization noise-to-output therefore differ only by a factor of $G(z)$: It is clear from figure 4.3 that the system function of this loop is:

$$Y(z) = \frac{G(z)}{1 + G(z)H(z)} \cdot X(z) + \frac{1}{1 + G(z)H(z)} \cdot Q(z) \tag{4.1}$$

Alternatively, the system function can be written as

$$Y(z) = \frac{G(z)}{1 + G(z)H(z)} \cdot [X(z) + Q_i(z)] \tag{4.2}$$

where $Q_i(z)$ is the equivalent input quantization noise. Comparing the two equations, it is clear that

$$Q_i(z) = \frac{Q(z)}{G(z)} \tag{4.3}$$

The effect of the quantization noise is therefore *shaped* by the transfer function $G(z)$.[2] Thus, even though the quantization noise itself is white, the effect of the quantization noise on the digital bit stream can be made to vary as a function of frequency by appropriately choosing $G(z)$. This is an important observation, since it is this noise shaping that permits high resolution data conversion using these modulators even though the in-loop quantization is quite coarse. Unlike the traditional converters, the burden of the extraction of the full resolution digital data falls entirely in the digital domain. Implicit in this result is that the need for very high precision analog components on-chip is eliminated, which significantly simplifies the microfabrication.

Clearly, the extent to which the noise is shaped is controlled by the transfer function $G(z)$; the order of this polynomial is the *order* of the modulator. Less apparent, but equally important, is the effect of the ratio of the sampling frequency f_s to the input

[2]In the most general case in which there are multiple feedback paths between $y(t)$ and $G(z)$, the noise shaping will be a function of both $G(z)$ and $H(z)$.

signal bandwidth W. This *oversampling ratio* is another degree of freedom that controls the resolution of the conversion. In discrete time Fourier space, the total noise power is uniform. As the oversampling ratio is increased, the bandwidth of the input signal in the digital domain (i.e., in discrete time Fourier transform space) is reduced, and the "slice" of quantization noise that gets added in to the signal is lowered. Thus, the amount of quantization noise in the final conversion can be arbitrarily reduced by increasing the oversampling ratio. It follows directly that the resolution of the system is increased as the oversampling ratio is raised. As the oversampling ratio approaches infinity, the width of the slice approaches zero. The quantization noise contribution therefore approaches zero and the conversion becomes, theoretically, perfect.

Mathematically, the effect of oversampling on the noise can be demonstrated by looking at what happens to the frequency spectrum of the noise. If the quantizer step size is Δ, and the quantization noise is white, the quantizer error has an equal probability of lying anywhere in the interval $\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]$, and the mean square error is:

$$e = \frac{\Delta}{\sqrt{12}} \tag{4.4}$$

The noise spectral density of the modulation is given by [61]:

$$|N(\omega)| = e\sqrt{\frac{2}{f_s}}\left(2\sin\left(\frac{\omega}{2f_s}\right)\right)^L \tag{4.5}$$

where f_s is the sampling frequency and L is the order of the modulator. For oversampling ratios greater than 2, the RMS noise can be approximated by [61]:

$$n_o = \frac{\Delta}{\sqrt{12}} \cdot \frac{\pi^L}{\sqrt{2L+1}}\left(\frac{2f_o}{f_s}\right)^{L+\frac{1}{2}} \tag{4.6}$$

where f_o is the bandwidth of the input signal. From this formula it is clear that for every doubling of the oversampling ratio the noise falls by $6L+3$ dB. Note also that the improvement is also a function of L, and, as might be expected, higher order modulators show better noise characteristics than lower order ones at the same oversampling ratio.

4.2 Basics of Modulator Design

The discussion above shows qualitatively how the oversampling process works, but does not address the more fundamental issue of modulator design. Like traditional A/D converters, the design of the modulator depends heavily on the target application (resolution) and the properties of the input signal (bandwidth, range). Because the principles of oversampled A/D converters are based more on signal processing issues than on single-point sampling techniques, it is not apparent what the significant design constraints are and how they can be satisfied. This section outlines some of the more important points that must be considered when choosing a modulator and presents the method of modulator design used for the active needle implementation.

4.2.1 Important Design Criteria

Oversampled A/D converters can be very powerful, as shown above. The use of oversampling techniques for data conversion is not without its own difficulties, however. One of the major problems is caused by the correlation between the quantization noise and the input signal. Although the noise shaping itself does not change, when the quantization noise is not independent and white, the character of the shaped noise can be changed dramatically. Consequently, it is of paramount importance that the design of the modulator system ensures some level of randomization to reduce the correlation between the input and the quantization noise. This is typically accomplished by increasing the complexity of the transfer functions $G(z)$ and $H(z)$; in general, higher order modulators result in less correlation between the input signal and the quantization noise.

A second factor that influences the design of modulators is the linearity of the quantization process in the loop. Although the general loop is shown with a comparator (single-bit quantizer), this is not a requirement for modulators. Higher-order (multi-bit) quantization can just as easily be used; this will reduce the injected quantization noise

because of the lower uncertainty in the quantized output. Higher resolutions can be attained at lower oversampling ratios because of this additional reduction in the noise. A further benefit is that the multi-bit approach whitens the quantization noise. The loop would therefore require an n-bit D/A converter in the feedback path. Nonlinearity in this D/A converter can be modelled as an equivalent nonlinearity in the function $H(z)$, which will clearly cause a nonlinearity in the output. Thus, the increased resolution from the multibit quantizer can only be realized if the multibit D/A is very linear. From a circuit point of view, the multibit approach increases the complexity of the design and the chip area required for the modulator. Although the single-bit quantization requires a higher oversampling ratio for a desired resolution when compared to multibit approaches, the advantage of using a single-bit quantization is that the D/A conversion process in the feedback loop is inherently linear, and that the circuits are relatively simple and low-area. The D/A, for example, is simply a wire.

A third, more practical design issue is that of quantizer overload. Throughout the entire discussion, it has been assumed that the error associated with the quantization is bounded by the quantization interval size. This translates into a requirement that the input signal to the quantizer never exceeds the input range of the quantizer. If this is not the case, the error produced by the quantization process increases rapidly as the quantizer input signal exceeds the range of the quantizer. The equivalent quantization noise therefore increases dramatically and the resolution of the system drops.

A similar overload problem must be avoided in the operational amplifiers used to implement the modulator. Implicit in the analysis is that the functions $G(z)$ and $H(z)$ are linear and time invariant. When op amp saturation occurs, the linearity condition is violated and the performance of the system may change significantly. In short, the circuit realization of the system must ensure that each of the component blocks performs as expected over the range of anticipated inputs.

Finally, there is the issue of modulator stability. In equation 4.6 above, it was shown that from a strictly noise shaping perspective higher order modulators are better.

Unfortunately, as the order of the system is increased beyond 2, the modulator loop will be unstable unless careful design is used. When $G(z)$ is of order 3 or higher, there are at least three poles in the feedback system, which can produce large negative phase shifts that will cause instability. Thus, use of higher order ($L > 2$) modulators must be approached carefully, and the pole locations must be examined to insure that they all lie within the unit circle in the z plane. Also associated with this potential instability is the overload situation from the previous paragraph--in addition to the larger noise that results from overload, in higher order modulators this overload changes the effective loop characteristics and may produce instability. In most cases, the modulator cannot recover from this overload even if the input signal returns to a valid signal level. For this reason, higher order modulators are almost always implemented with a reset capability so that the modulator can be zeroed in the event that instability develops.

4.2.2 Design Methodology

Given a desired resolution and bandwidth, the first step in modulator design is to select the modulator order and oversampling ratio, and to decide whether a multibit approach is necessary. Typically, the oversampling ratio is a power of 2 (this simplifies the decimation), and is made as large as is feasible given the input signal bandwidth and the characteristics of the devices that will be used to realize the circuits. Audio signals, for example, have a 20 kHz maximum bandwidth; for system clocks on the order of 10 MHz, the maximum oversampling ratio would be approximately 256.[3] Once selected, the oversampling ratio can be used to compute the minimum required modulator order using equation 4.6; for this computation, the modulator is usually assumed to be single-bit to take advantage of the inherent linearity and simplicity of this case. Alternatively, the modulator order could be chosen based on desired shaping characteristics, and the formula could be used to determine the minimum oversampling ratio required.

[3]The Nyquist frequency is 40 kHz, and 256*40 kHz = 10.24 MHz.

For low frequency signals, the single-bit approach will always suffice, and is greatly preferred because of the inherent linearity of the quantization. For high frequency signals, if the numbers computed for the single-bit case turn out to be too stringent or unrealistic, a multibit approach must be considered. The corresponding reduction in the quantization level spacing Δ reduces the quantization noise and eases the requirements on L and f_s. As should be evident from this discussion, there is no one universally "correct" approach for selecting the modulator order, the oversampling ratio, and the number of quantization levels in the modulator. They are all coupled to the overall performance through equation 4.6. Often times several different implementation options are available for a given set of specifications; the designer is given several degrees of freedom. The key point is to realize that there are tradeoffs involved as discussed above. No matter which approach is used, however, the important point is that the oversampling ratio and the modulator order should be the first parameters selected for the design; the selection of the number of quantization levels will fall out of this calculation.

The next step in the design process is the selection of a topology based on the desired modulator order. This depends largely on the desired quantization noise shaping, which in turn specifies the desired $G(z)$.[4] Recall equation 4.3; this equation states that the shaping is inversely related to $G(z)$. At frequencies where $|G(z)|$ is large, the quantization noise is low. Conversely, if $|G(z)|$ is small at some frequency, the quantization noise will be higher at that frequency. For the signals considered here, namely, low frequency signals bandlimited to f_o, rejection of the quantization noise in the signal band implies that $G(z)$ must be low pass for the signal, or, equivalently, high pass for the quantization noise. This will shape the noise so that most of the noise energy falls outside of the baseband, where it can be removed in the digital domain by appropriate digital filters.

The desired function $G(z)$ can then be determined using standard filter design

[4]And $H(z)$ if the feedback is distributed. In this case, $H(z) = 1$ was assumed for convenience, and because it accurately reflects the modulator implementation described here.

techniques, subject to the constraint that the order of the filter must be equal to the order of the modulator, since by definition the order of the modulator is equal to the order of $G(z)$. The design of the filter can be done using whatever method is most convenient; well described filters such as Butterworth or Chebychev are typically employed. No matter what method is used, a function $G_1(z)$ results; realization of this $G_1(z)$ in the loop produces the desired noise shaping.

In terms of the system topology, the typical realization of this low pass filter function uses integrators in the forward path, since the magnitude of the frequency response of an integrator is inversely proportional to frequency. An integrator-only system, however, puts all of the poles of $G(z)$ at $z = 1$, which results in a strictly monotonic noise shaping characteristic and instability when the order of the system is higher than 2. Feedforward and feedback coefficients are added between the integrators to control the pole and zero locations, respectively. In this way the frequency response can be customized to the particular modulator requirements. This particular aspect of modulator design is discussed in greater detail in the literature [62,63].

No matter where the feedback and feedforward paths are placed, the end result is a topology for which $G(z)$ can be expressed as a rational function in z; call this function $G_2(z)$. The coefficients of each of the terms in $G_2(z)$ will be a function of the feedback and feedforward coefficients. The problem then reduces to one of matching the desired filter characteristic $G_1(z)$ to the topology transfer function $G_2(z)$. Since $G_1(z)$ and $G_2(z)$ are of the same order, this can be accomplished by individually matching each of the polynomial coefficients in the two functions. This results in a set of linear equations relating the feedback and feedforward coefficients to the filter coefficients. Solving this system of equations for the feedback and feedforward coefficients gives the values required to produce the desired noise shaping. At this point, the system is completely specified and attention is turned to the circuit realization of the individual system blocks.

4.3 System Level Design

The target temperature resolution of the active needle system is 1 m°C over a temperature range of 30-50°C, or 303.15-323.15 K. In terms of the required resolution of the modulator, this translates into 1 part in 323,150, or 18.3 bits. The bandwidth of the temperature signal of interest is approximately 1 Hz. The modulator must therefore produce a digital output from which a 1 Hz signal can be extracted at the 19 bit level. This section discusses the system topology of the modulator that is used to satisfy these requirements.

The choice of modulator order was made first, and was based on the very low-frequency nature of the input signal. With such slowly varying inputs, the quantization noise in low order modulators is highly correlated with the input signal as was discussed above. In order to reduce this correlation, a fourth order modulator topology was selected.[64] For linearity, and because of chip area constraints, single-bit quantization is used. From equation 4.6, the required oversampling ratio can be computed: For 19 bit resolution, the quantization noise needs to be attenuated by 18 bits over its normal RMS value. Rewriting equation 4.6 in terms of the normal RMS quantization noise gives:

$$n_o = e \cdot \frac{\pi^L}{\sqrt{2L+1}} \left(\frac{2f_o}{f_s}\right)^{L+\frac{1}{2}} \tag{4.7}$$

For an additional q bits of noise rejection in the baseband, it is necessary that

$$\frac{n_o}{e} = 2^{-q} = \frac{\pi^L}{\sqrt{2L+1}} \left(\frac{2f_o}{f_s}\right)^{L+\frac{1}{2}} \tag{4.8}$$

Solving this equation for the oversampling ratio $\frac{f_s}{2f_o}$ gives

$$OSR_{min} = \left(\frac{\sqrt{2L+1}}{2^q \pi^L}\right)^{-\frac{1}{L+\frac{1}{2}}} \tag{4.9}$$

With $L = 4$ and $q = 18$, this formula predicts a minimum required oversampling ratio of $\approx$35. For the specifications above, this translates into a requirement that the circuits operate at 70 Hz. The choice of operating frequency is therefore completely flexible; in

order to avoid unnecessarily complicating the circuit design, a very conservative clock frequency of 65.536 kHz was chosen. This results in an oversampling ratio of 32,768 and places no severe design constraints on the hardware. It should be noted that with this clock frequency, the desired resolution should in theory be attained for a wider range of frequencies since the minimum requirements have been greatly exceeded, or, correspondingly, that the clock frequency can be reduced significantly while still meeting the resolution specification.

Having chosen the parameters of the modulator, the system topology was selected. The general architecture used is of the type presented in [64]. Because the modulator is fourth order, feedforward coefficients are required to alter the pole locations and stabilize the loop. Feedback coefficients are not used since the input signal bandwidth is very low and the additional hardware required to implement the feedback zeros would unnecessarily increase the overall chip area. The resulting architecture is shown in figure 4.4.

The linear model for this topology replaces the quantizer with a linear unity gain element. If the integrators are implemented as switched-capacitor elements, the transfer function of each integrator block will be

$$I(z) = \frac{Cz^{-1}}{1 - z^{-1}} \tag{4.10}$$

where an integrator scale factor C has been included for completeness, since C will be used to prevent quantizer overload. The overall transfer function from the quantization noise to the output then becomes

$$\frac{Y(z)}{Q(z)} = \frac{(1-z^{-1})^4}{1+(A_1C-4)z^{-1}+(A_2C^2-3A_1C+6)z^{-2}+(A_3C^3-2A_2C^2+3A_1C-4)z^{-3}+(A_4C^4-A_3C^3+A_2C^2-A_1C+1)z^{-4}} \tag{4.11}$$

and the problem now reduces to one of choosing filter coefficients to match the desired noise shaping characteristic.

MATLAB^{TM}, a commercially available system analysis software package, was used to find the appropriate filter coefficients. The process is an iterative one, in which

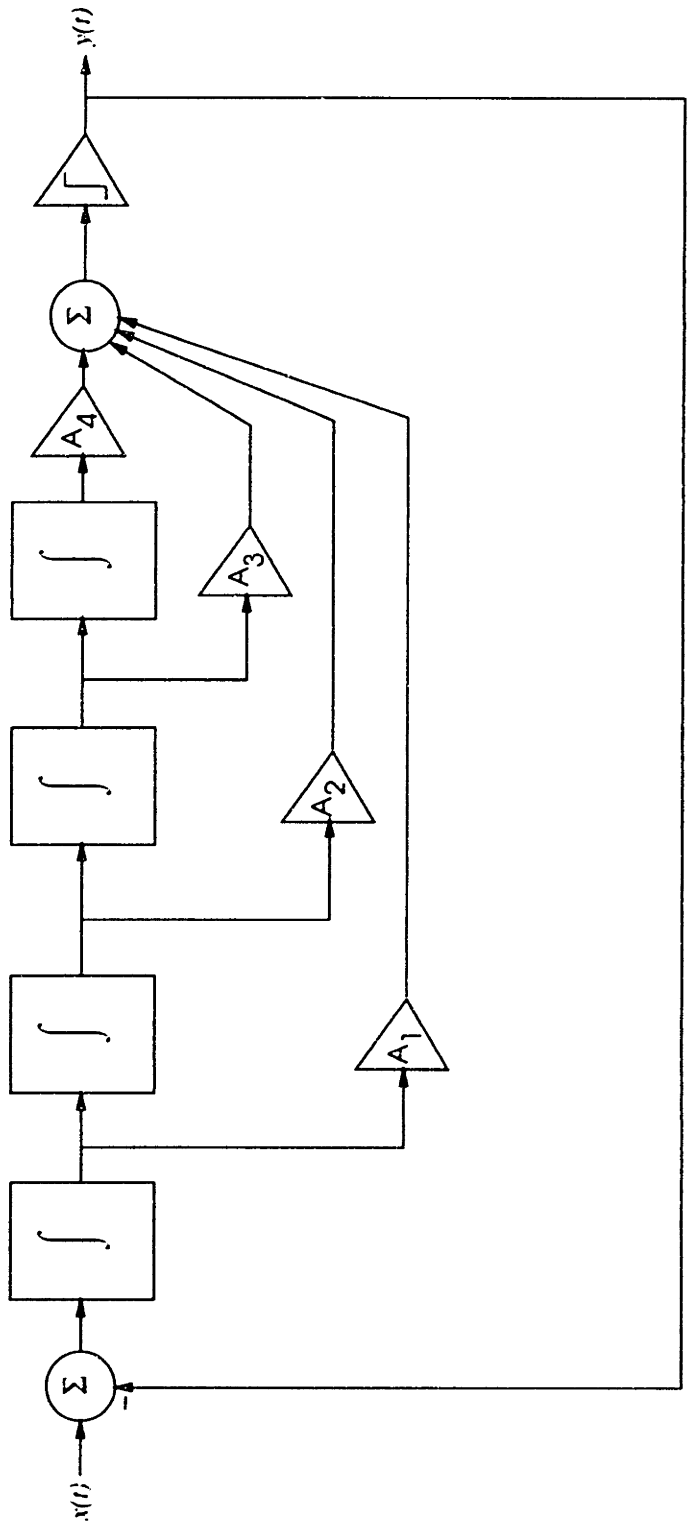


Figure 4.4: Fourth order modulator topology

Table 4.1: Fourth order modulator system parameters

Parameter	Value
C	.125
A_1	3.067
A_2	4.5952
A_3	3.9072
A_4	1.5813

a filter is designed, the coefficients are mapped into the required A coefficients using equation 4.11, and the resulting system is simulated (using a custom simulation program [65]) to verify that quantizer overload does not occur. A second assumption that is also verified is that the "gain" of the quantizer is indeed one, as was assumed. This is performed as outlined in [66], which discusses in detail the implications if the gain is not unity. The integrator scale factor is adjusted to eliminate any quantizer overload, and the cutoff frequency of the filter is adjusted to bring the gain of the quantizer closer to one. For this modulator, a Chebyschev type II filter design algorithm was used; after several iterations, all of the constraints were satisfied. The resulting system parameters are given in table 4.1.

The pole-zero plot of the quantization noise transfer function is shown in figure 4.5; the corresponding quantization noise response is shown in figure 4.6. As expected, the four zeroes at $z = 1$ (DC) produce the sharp drop in quantization noise as the frequency is lowered. The poles are uniformly spaced on an arc centered at $z = 1$, which is characteristic of Chebychev type II filters. The effect of the high oversampling ratio can also be seen in the frequency response: For performance at the 19 bit level, the modulator must have quantization noise attenuation of at least 104 dB over the baseband of interest. In this case, the baseband of interest is so low that it is below the limits of the graph. At the lowest frequency shown, the noise rejection is already greater than 150 dB, or more than a factor of 100 less than the requirement. Thus, this modulator design more than adequately meets the target specifications.

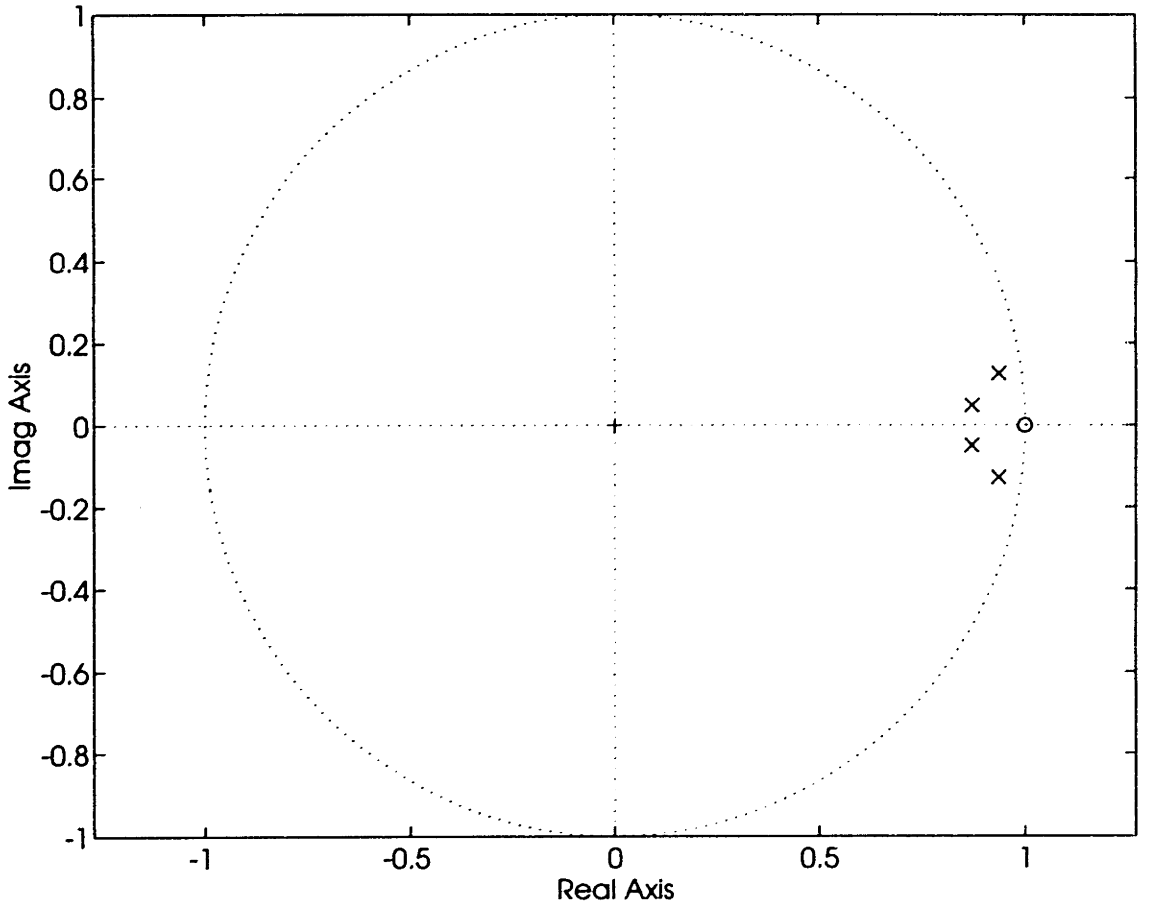


Figure 4.5: Fourth order modulator: Quantization noise pole-zero diagram

4.4 Circuit Implementation

The circuit realization of the modulator requires only three basic blocks: Integrators, summers, and a quantizer. Although there are a number of specific circuits that can perform each of these functions, the approach taken in this design was to use switched-capacitor implementations. The switched-capacitor approach was selected for several reasons: First, a large body of switched-capacitor circuits have already been extensively studied, which provides a large "library" of circuits that can be used "off the shelf" to meet the requirements of the modulator. Second, the critical specifications of switched-capacitor circuits are usually determined by capacitance ratios (rather than resistor ratios). Since capacitance ratios can be well controlled, these circuits

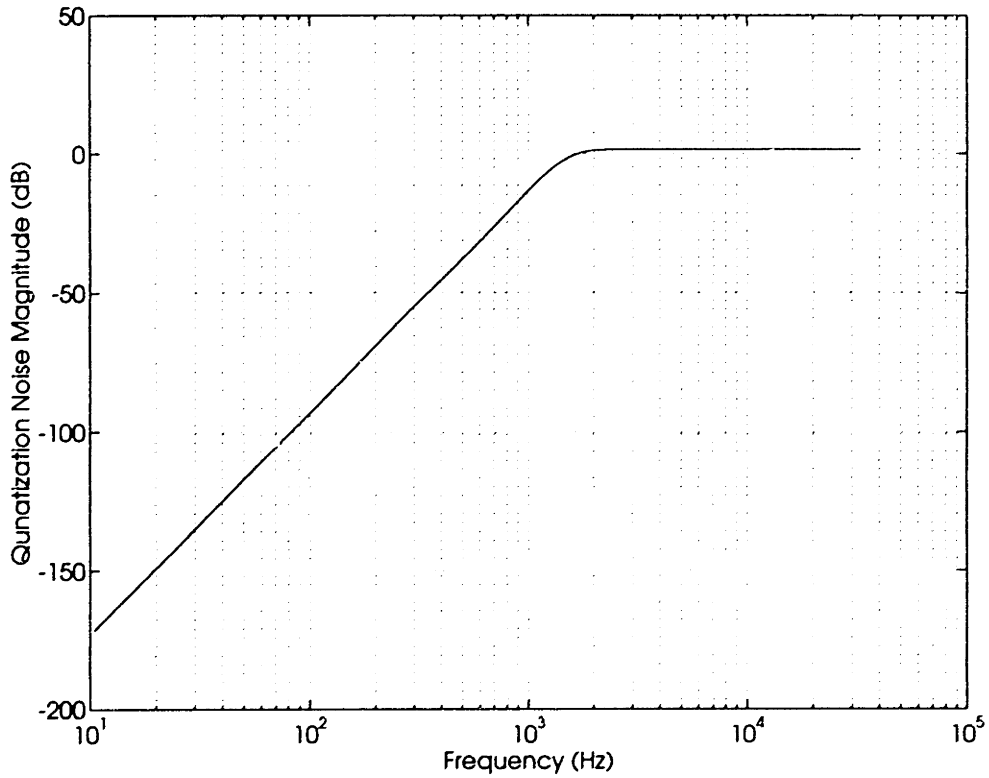


Figure 4.6: Fourth order modulator: Quantization noise shaping

are particularly well suited for microfabrication. Finally, these circuits are inherently discrete-time in nature, so the transformation from the block diagram realization to a working circuit implementation is simplified. This section examines the circuits for each of the three blocks required to construct the modulator system developed in the previous section.

4.4.1 Integrator

The switched capacitor integrator design used is shown in figure 4.7. It is a fully differential, parasitic insensitive, noninverting topology. Delayed clocking is used to prevent feedthrough nonlinearities from the input onto the sampling capacitor C_1 [56]. Transmission gates are used only where necessary to save chip area; single channel PMOS gates are used otherwise. The operational amplifier at the core of the circuit is the same one designed for use in the temperature sensor, which simplifies the chip

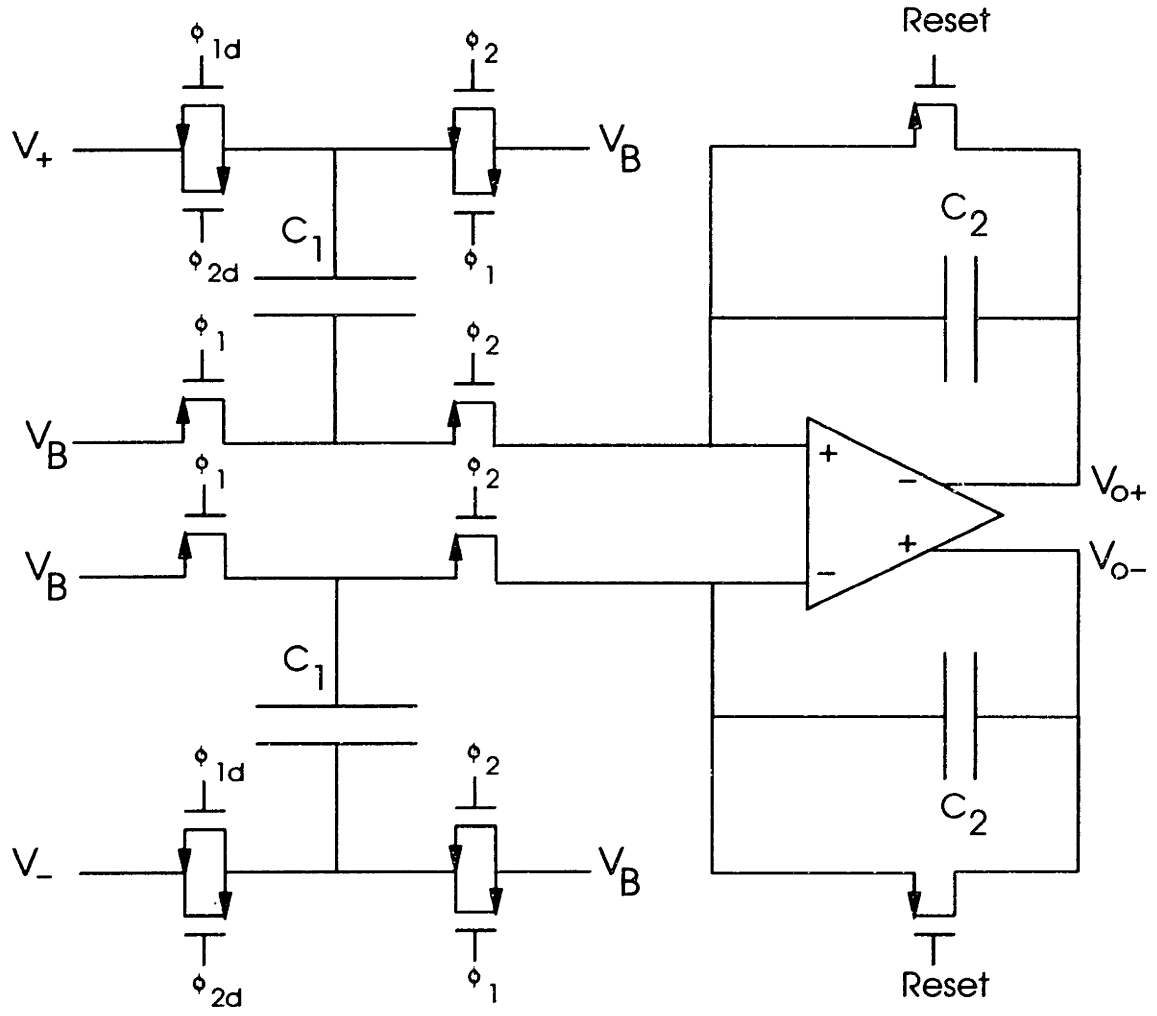


Figure 4.7: Switched capacitor integrator

layout and obviates the need for two separate op amp designs. The transfer function of the integrator is easily derived by looking at the charge balance during each clock cycle. First, look at the behavior of the top half of the circuit. When ϕ_1 is active, the charge on C_1 is

$$Q_{C1,1} = C_1(V_+ - V_B) \tag{4.12}$$

The charge on C_2 is given by

$$Q_{C2,1} = \frac{C_2 V_{op}}{2} \tag{4.13}$$

where V_{op} is the previous value of the differential output. When ϕ_2 is active, the charge on each of the capacitors is

$$Q_{C1,2} = C_1(V_B - V_{CM}) \tag{4.14}$$

$$Q_{C2,2} = C_2(V_{o+} - V_{CM}) \tag{4.15}$$

where V_{CM} is the nominal common mode level of the op amp. Equating the total charge during each phase and solving for V_{o+} yields:

$$V_{o+} = \frac{C_1}{C_2}(V_+ - 2V_B + V_{CM}) + \frac{V_{op}}{2} \tag{4.16}$$

An analogous derivation for the bottom half of the circuit gives:

$$V_{o-} = \frac{C_1}{C_2}(V_- - 2V_B + V_{CM}) - \frac{V_{op}}{2} \tag{4.17}$$

and the equation relating the differential input to the differential output is:

$$V_{o+} - V_{o-} = \frac{C_1}{C_2}(V_+ - V_-) + V_{op} \tag{4.18}$$

from which it is clear that the circuit is an integrator, with a gain determined by the ratio of the capacitances C_1/C_2; in this particular implementation, an 8:1 ratio (C_1 = 1 pF, C_2 = 8 pF) is used to realize the integrator scaling factor of .125 that the system requires. No kT/C noise problems result from these values since the noise is reduced by the square root of the oversampling ratio [67]. The reset switch is provided so that the accumulated charge on the integrating capacitor can be zeroed out if overload occurs.

4.4.2 Summer

The implementation of the summing functions in the loop is done using two circuits that are slight modifications of the basic integrator topology above. The first summer required in the loop immediately precedes an integrator, so the summing function is combined with the first stage integration by providing additional input capacitors. This

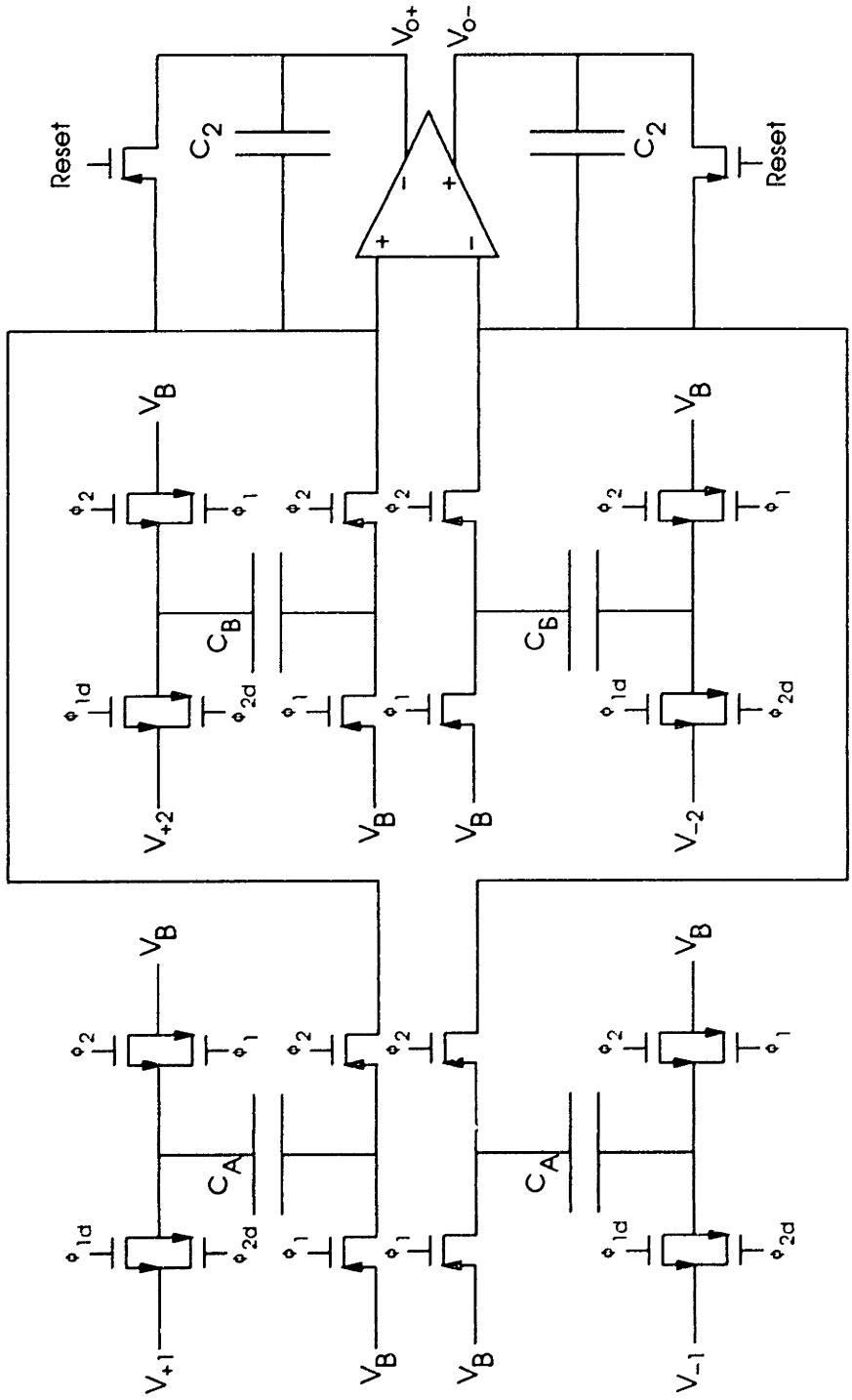


Figure 4.8: Switched capacitor integrator with input summing

summing-integrator topology is shown in figure 4.8. The derivation technique above can be used to show that the transfer function of this circuit is

$$V_{o+} - V_{o-} = \frac{C_A}{C_2}(V_{+1} - V_{-1}) + \frac{C_B}{C_2}(V_{+2} - V_{-2}) + V_{op} \tag{4.19}$$

where C_A and C_B are the two input capacitors as shown in the figure. The output is the integral of a weighted sum of the two inputs. The circuit therefore implements both the input summing and the first stage integration of the system, which saves one op amp (reducing the chip area and power consumption) over implementing the summer and integrator separately.

The second summer topology is a four-input summer that does not perform any integration, and realizes the summer immediately preceding the quantizer in the system block diagram. The circuit is shown in figure 4.9; only two inputs are shown for clarity. Additional inputs are realized by replicating the section in the dashed box for each additional input. The design is almost identical to the summing integrator; the only difference is that the switch around the feedback capacitor is clocked on ϕ_1. Consequently, the charge on C_2 is emptied each clock cycle, forcing $V_{op} = 0$. The output voltage on a given cycle therefore represents only the weighted sum of the current values of the inputs. As before, the weighting is controlled by the capacitors C_A, C_B, etc. Thus, the A coefficients can be realized by appropriately ratioing the input capacitors. This eliminates the need for any separate gain blocks to implement the loop coefficients. Thus, all of the blocks (except the quantizer discussed below) can be implemented using only the three switched-capacitor circuits presented, which are all variants of the same basic topology. Combining the various functions of the loop saves area and power: The loop requires only 5 op amps instead of the 10 that might be required if each block were completely separate.

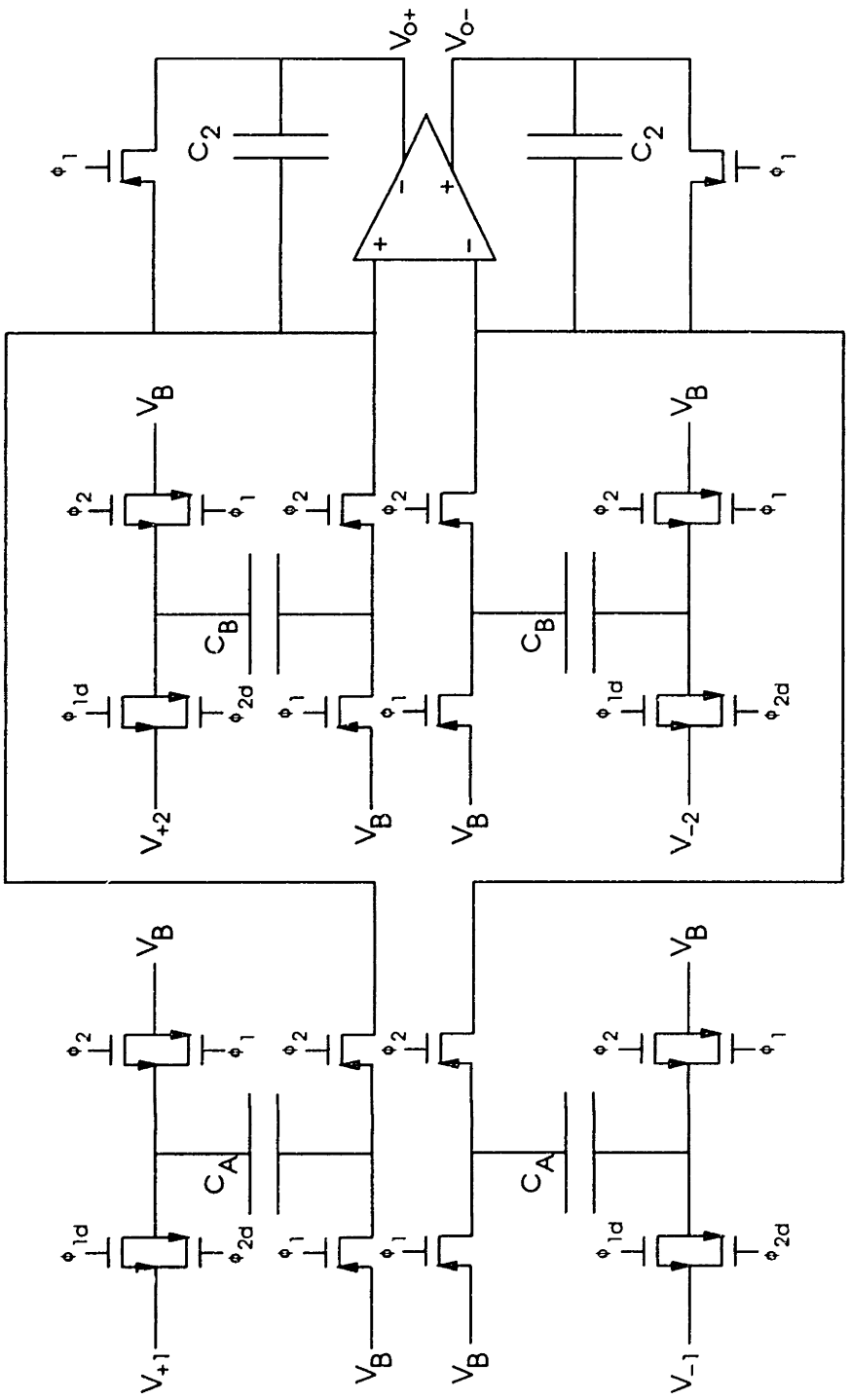


Figure 4.9: Switched capacitor summer

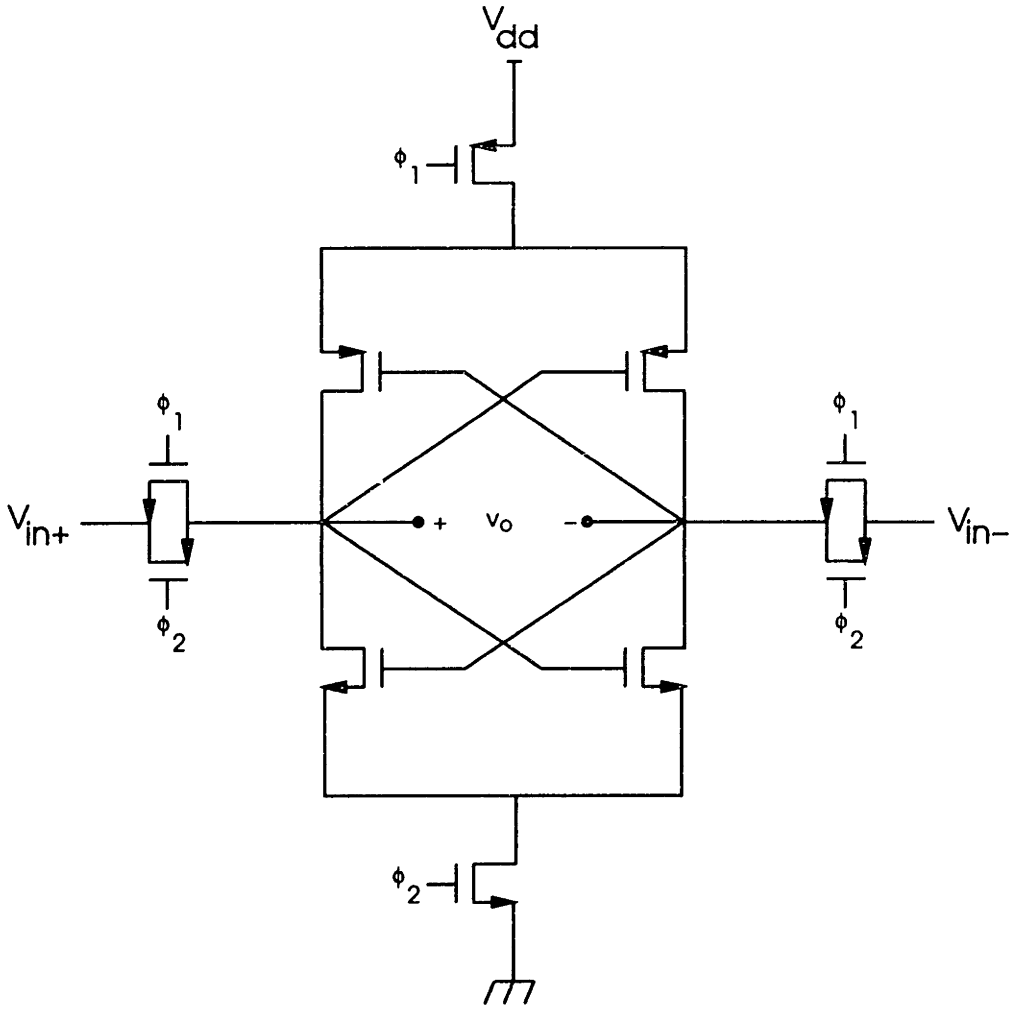


Figure 4.10: Fully differential comparator

4.4.3 Comparator and D/A

The last circuits required for the modulator are the A/D and D/A. In the single-bit design used here, the A/D is a comparator, and the D/A is a simple CMOS buffer. Because the circuits are fully differential, the comparison in the A/D can be carried out in a very straightforward way, and the A/D and D/A can be performed in one step. The comparator threshold is zero volts differentially, regardless of the common mode level of the circuits. The quantizer design is shown in figure 4.10, and is a clocked bistable latch. When ϕ_1 is active, the capacitances (from the devices and the parasitics) at the output nodes are charged with the two output voltages of the differential summer.

When ϕ_2 is active, the inputs are disconnected from the circuit and power is applied to the latch. Because of the positive feedback of the circuit, the two outputs are driven to opposite rails. The output node that started with a higher voltage is driven to the positive rail; the other output is driven to the negative rail. The output is therefore $+V_{dd}$ if the differential input voltage is positive, and 0 if the differential input is negative. If the inputs are exactly equal, the circuit is metastable and the output is theoretically zero. In reality, any noise present on the output nodes will drive the output off of the metastable point. Because of this, the probability of the system being in the metastable state decreases rapidly as the clock frequency is decreased. In this particular implementation, the clock frequency is so low that the metastability can be ignored. The D/A conversion is performed by a basic CMOS inverter; the purpose of this buffer is to generate an analog feedback signal that is strongly pinned at either the positive or negative rail. The inversion of the signal is of no consequence since the system is fully differential.

4.5 Summary

This chapter has presented the basic modulator design used in the active needle temperature sensing system. The basic concepts of oversampling A/D conversion were discussed, with an emphasis placed on the characteristics of these converters that make them especially well suited for monolithic realizations of high resolution converters. A simple design methodology was outlined in an effort to bring some understanding to the modulator design process. Finally, the modulator designed for use with this system was presented, both on the system and circuit levels.

The emphasis has been placed on presenting only the information about these converters required to understand the design of the modulator that was implemented. The discussion is intended to serve as a backdrop for this purpose, and is not meant as a tutorial in modulator design. Several in-depth discussions of the principles of oversampling sigma-delta converters and the myriad of issues associated with their design can be found in the literature [58,62,66,64]. These and other sources have been

cited throughout. The available literature on these converters is quite vast; readers that wish to delve further into this topic should consult the references cited for a more rigorous, in-depth analysis of the behavior of these systems.

Chapter 5

The Digital Controller/Interfacing

The coordinated transmission of information on and off of the needle is in some sense the most critical part of the project, since even the most accurate sensors are useless if one cannot access their data. For this project, a special interfacing chip was designed to perform all of the off-needle communication "overhead" necessary to use the active needle. Thus, the digital control chip acts as the "brain" of the entire system: The chip receives instructions (in this case a sensor number) from the host computer, decodes the instructions, activates the appropriate sensor, and transmits data from the selected sensor, tagged appropriately, back to the computer. A block diagram of the functions performed by this chip is shown in figure 5.1. This chapter describes the details of operation of this chip, including the communications protocol used for both signal transmission and reception, the architecture used to realize these functions, and the design of the gates and higher-level digital circuits employed.

5.1 Communications Protocol

The basic communications requirements of the controller are very clearly defined. First, it must receive instructions from the host computer. In order to keep the number of wires coming off of the needle to a minimum, these instructions are received serially. Second, measurement data must be communicated from the needle to the host computer; again, this transmission is serial. Finally, the controller must translate the instructions

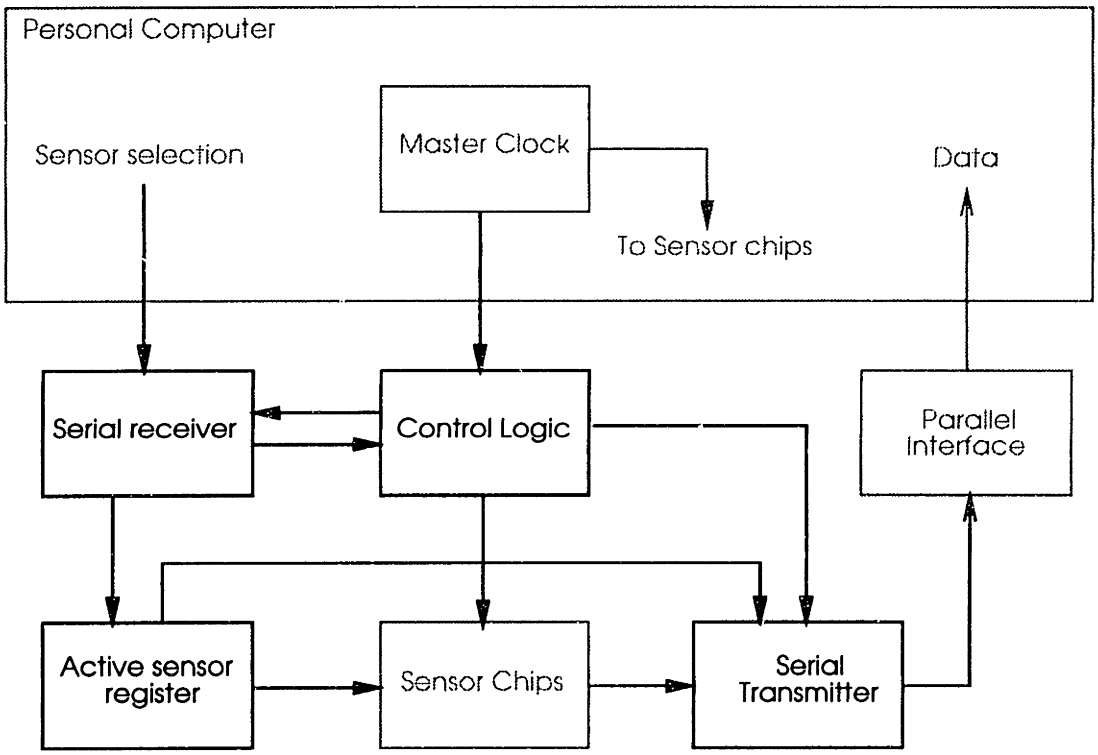


Figure 5.1: Control chip block diagram

received from the host computer into the appropriate "needle-bus" control signals for the individual sensor chips.

Input signals from the host personal computer are transmitted using a straightforward serial protocol. In the default state, the serial input line is held at a logic LOW, indicating the idle condition. A transmission is initiated by the presence of a logic HIGH (a start bit) on this line for exactly one clock cycle. Following the start bit, exactly four instruction bits are received, one per clock cycle. This packet is then followed by a logic LOW stop bit that returns the line to the idle state regardless of the last bit transmitted. This packet protocol is shown in figure 5.2. In the present implementation, the four instruction bits represent the number of the sensor (0 through 15) that is to be activated for a measurement, since the measurement of temperature is strictly passive. Future controller implementations may use a longer instruction word that would allow more sophisticated active measurement control; for the purposes of this system, and

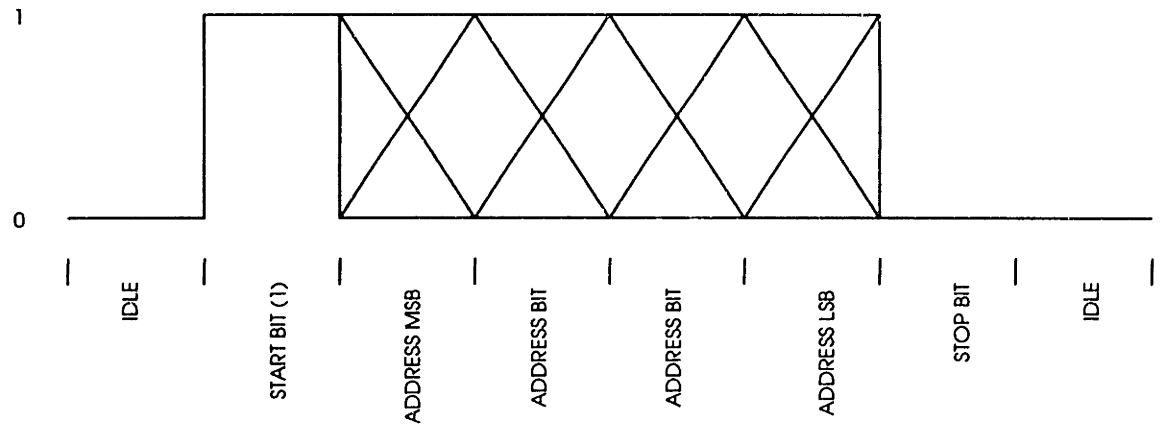


Figure 5.2: Needle input packet format

initial verification of the system architecture, the simple protocol suffices.

The output data format is not as rigid, since the duration of a measurement is determined by the user. When a new instruction word is sent to the input receiver circuits, the word is also latched into the output shift register. Exactly 10 clock cycles after the start bit is transmitted, the active sensor number is transmitted back to the host computer. This transmission not only verifies the reception of the instruction, but also acts as a marker to indicate that the data that follows is from the transmitted sensor. The controller then passes the output bits from the modulator on the selected sensor directly back to the host computer for processing and display. The length of this data transmission is unlimited, and continues until a new instruction is received. In this way the timing of any measurements can be controlled by the host computer: Each sensor can be polled on a regular basis or any one sensor can be monitored for a long time.

5.2 System Architecture/FSM Operation

The system architecture used to realize these functions is shown in figure 5.3. It consists of an input system (a D-flip flop and a serial-to-parallel shift register) for receiving information from the host computer, a latch for storing the current active sensor, an output shift register for sending data back to the host computer, and a

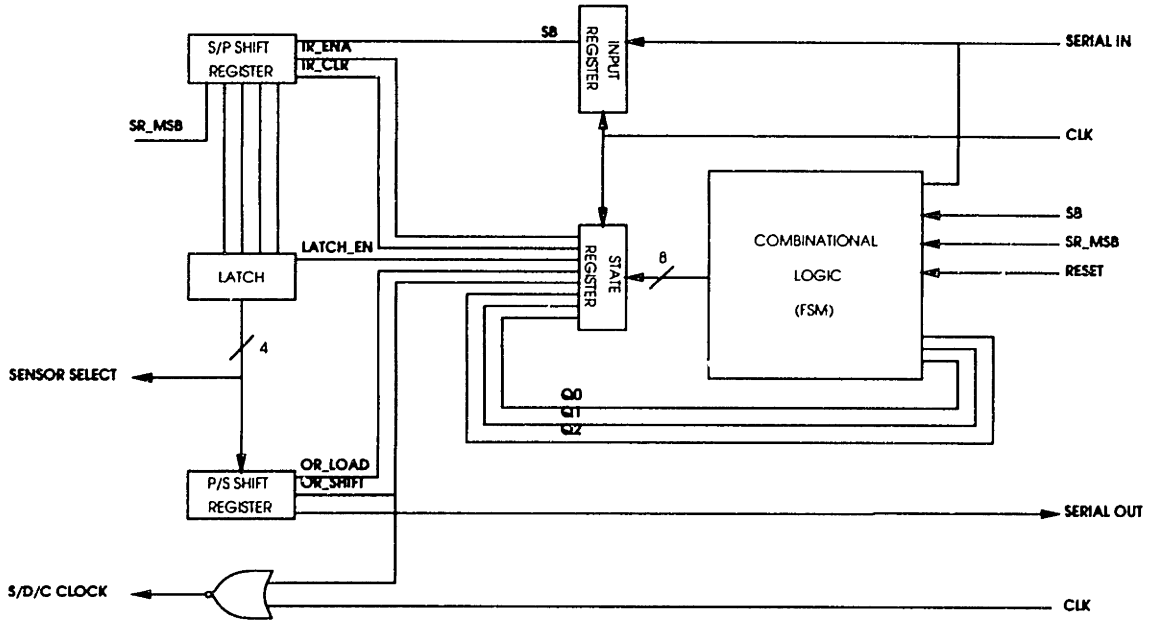


Figure 5.5: Digital controller logic diagram

simple microcontroller (a combinational logic "program" and a D-flip flop "program counter") for controlling each of the other subsystems.

Operation of the controller is straightforward: In the default state, the system latches a bit from the serial input data stream into the input shift register. The fifth (most significant) bit is monitored by the microcontroller for the presence of a logic HIGH start bit that would indicate the beginning of an instruction transmission from the host computer to the controller. When this bit is received, the microcontroller enables the sensor latch, which latches the sensor number onto the needle's internal bus so that the proper sensor is selected. The internal clock on the needle bus is enabled, the output register is enabled, and the latched data is clocked into the output shift register so that it can be transmitted back to the host computer to indicate the start of data transmission.

In order to realize these functions, a finite state microcontroller is used to generate appropriate control signals. The finite state machine that performs the desired actions is shown in figure 5.4. In the default state (**000**), the controller is looking for the presence of the start bit as described above. This is also the state to which the system goes when the FSM is reinitialized by activating the RESET line from the host computer. When the

start bit is detected, the system moves into state **001**, which enables the input register for exactly 5 clock cycles, until the start bit has been shifted into the most significant bit position. The system then passes into state **010**, where the input register is disabled, and the needle bus latch is enabled; state **011**, where the latched data is loaded into the output shift register; and state **100**, where the controller continually receives data from the selected sensor and transmits it directly back to the host computer. The system remains in this state until a new start bit is detected.

5.3 Logic Circuits

All of the logic used on the controller is derived from the three basic CMOS gates shown in figure 5.5.[1] Positive logic is used throughout. The first circuit is the two transistor inverter in which the output is the logical inverse of the input. The second circuit is the NAND gate, where the output is LOW only when both inputs are HIGH. The last circuit is the NOR gate, in which the output is HIGH only when both inputs are LOW. Each of these gates has two important properties: First, for "reasonable" input voltages, the voltage levels at the output go from rail to rail; logic HIGH output are pulled all the way up to the positive rail, and logic LOW outputs are pulled all the way down to the negative rail. This is important because this insures that cascaded gates will function properly even if the inputs voltage levels are not at either rail, i.e., the valid input logic level ranges (and the noise margins) are very wide. Second, the static power dissipated in each gate is zero; any power dissipated by the gates is strictly dynamic, and is associated with the charge transfer required to switch logic levels. The average power dissipated by the gates, therefore, is related to the frequency at which the logic levels switch. In this particular application this is a tremendous benefit because the operating frequency is relatively low, which reduces the power consumption and the associated thermal artifact.

[1] It can be proved that any combinational logic function can be realized by an appropriate combination of these three basic gates.

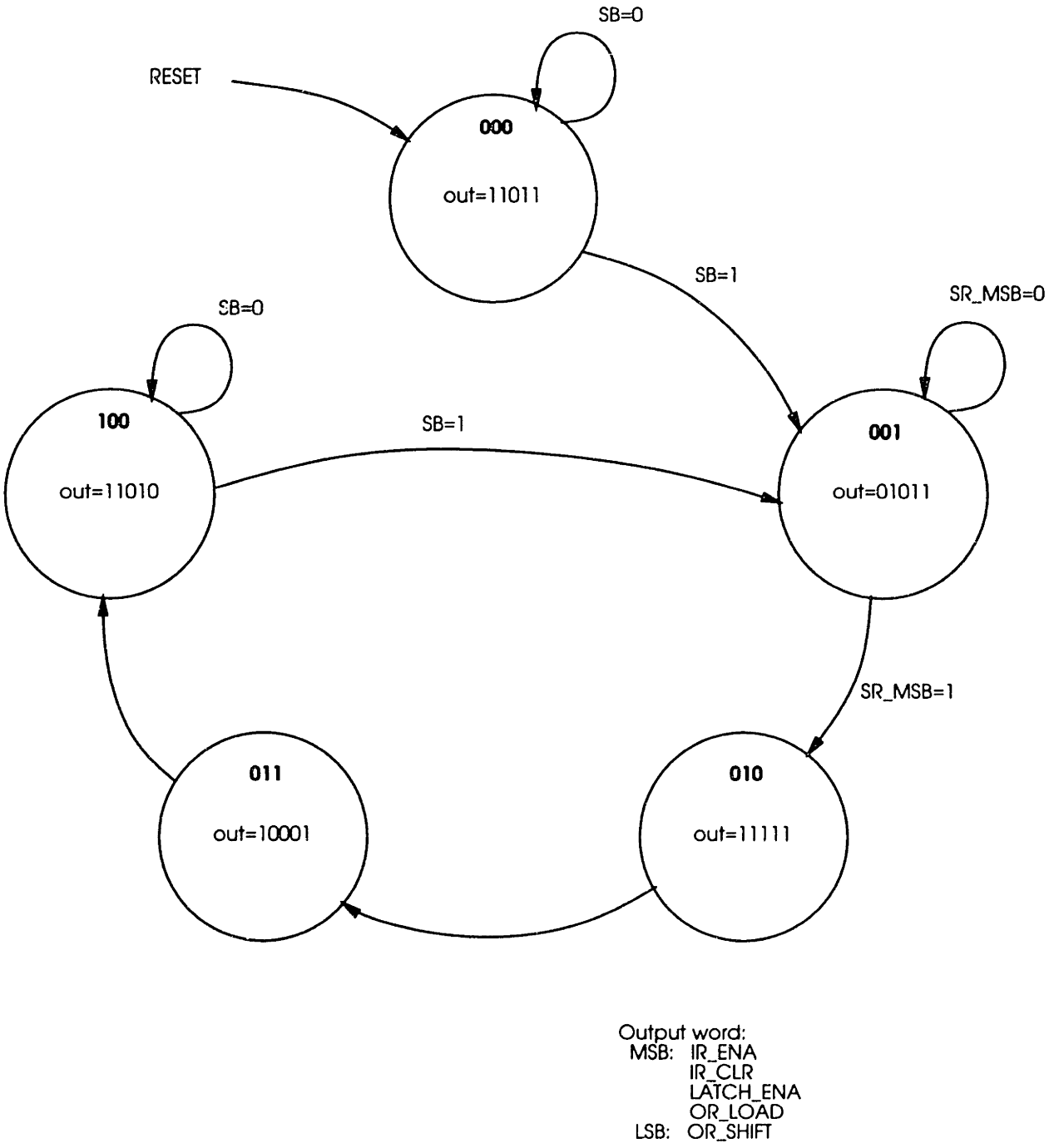


Figure 5.4: State diagram of the controller FSM

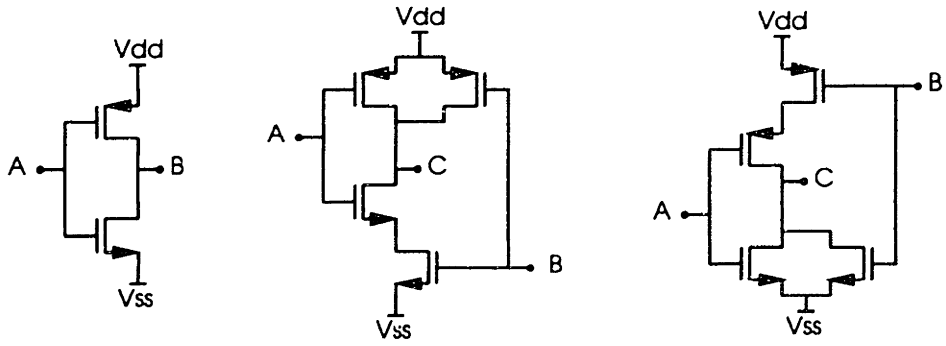


Figure 5.5: Three basic CMOS logic gates

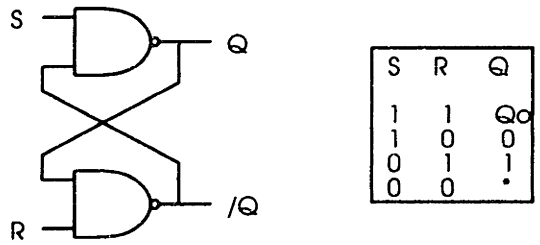


Figure 5.6: Basic S-R bistable latch

The higher order logic functions are generated using several intermediate hierarchical logic functions, each of which will be described briefly here; a more extensive look at these systems can be found in [68]. For simplicity, a "bottom up" approach will be used, showing how each more complex block is constructed from the simpler blocks. The first circuit, constructed from two NAND gates, is the bistable latch or S-R flip-flop, which forms the core of the other flip-flops and shift registers. The logic diagram and truth table are given in figure 5.6. With the addition of two more NAND gates, this circuit becomes an S-R latch (figure 5.7), where the output responds to the input state only when the clock input is high. By connecting the R input to the inversion of the S input, the circuit becomes a transparent, active HIGH enable D-type latch.

This circuit is not satisfactory for the state register, since changes in the inputs that occur when the clock input is HIGH are immediately reflected at the output, or, equivalently, the "hold time" for this flip-flop is equal to half the clock period. This is unacceptable for this architecture; the register must be edge triggered, where the input is

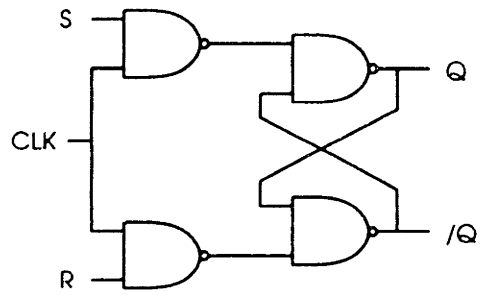


Figure 5.7: S-R flip flop with enable (Clock)

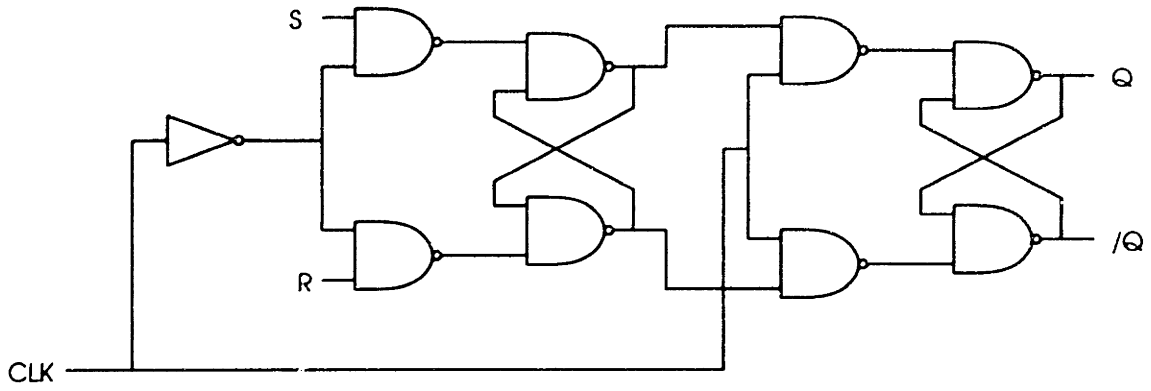


Figure 5.8: Edge triggered S-R flip flop

clocked through to the output on either the rising or falling edge of the clock, so that the hold time is greatly reduced. A positive edge-triggered S-R flip-flop can be constructed from the transparent latch as shown in figure 5.8. The positive edge-triggering ensures that the outputs of the flop change only when a LOW to HIGH transition occurs on the clock input. As before, a D-type flop is realized by tying the R input to the inversion of the S input. It is this basic D-flop that is used to construct the state register for the microcontroller.

The shift registers are constructed using the basic positive edge-triggered S-R flop. The input serial to parallel register is formed by cascading the S-R flops as shown in figure 5.9, which shows four bits of the 5-bit register. Operation of the shift register is straightforward: Each serial input bit is latched at the input and is rippled through the register on successive clock edges. The circuit is essentially five cascaded D-flops; the S-R flops are used because both the Q and $\overline{Q}$ are available as outputs, and using those

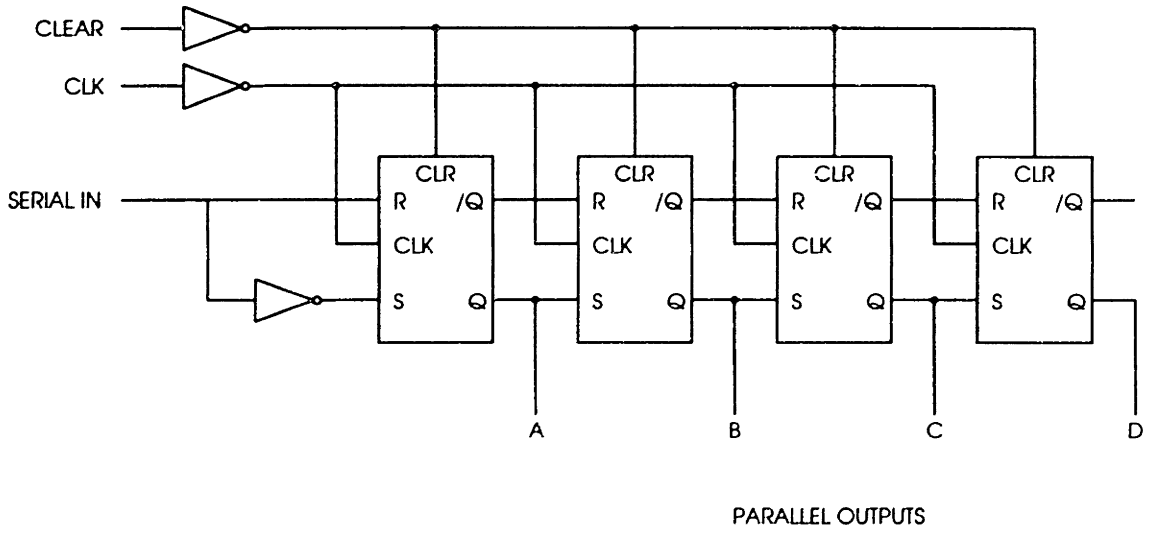


Figure 5.9: Serial-to-parallel shift register

directly as inputs to the next stage avoids the need for an additional inverter for each stage. The single inverter at the input is necessary because the inversion of the input is not provided. Since for this application an active low CLEAR input is required, one was provided as shown in the figure.

The final logic function required, the parallel to serial shift register shown in figure 5.10, is slightly more complicated because a LOAD operation must be included. The shifting operation is essentially the same as before: On each clock cycle, each bit ripples through the S-R flops; the serial output is the output of the last register in the chain. The loading operation uses two NAND gates to drive the (active low) PRESET and CLEAR inputs of each flop, depending on whether or not the parallel input bit present is HIGH or LOW. This method is used because it avoids the use of additional logic at the S and R inputs to each register, and because using the PRESET and CLEAR inputs changes the state of the flops regardless of the S and R inputs. In this way, the LOAD operation is completely independent of the current state of the internal registers and depends only on the parallel inputs and the state of the SHIFT/LOAD input.

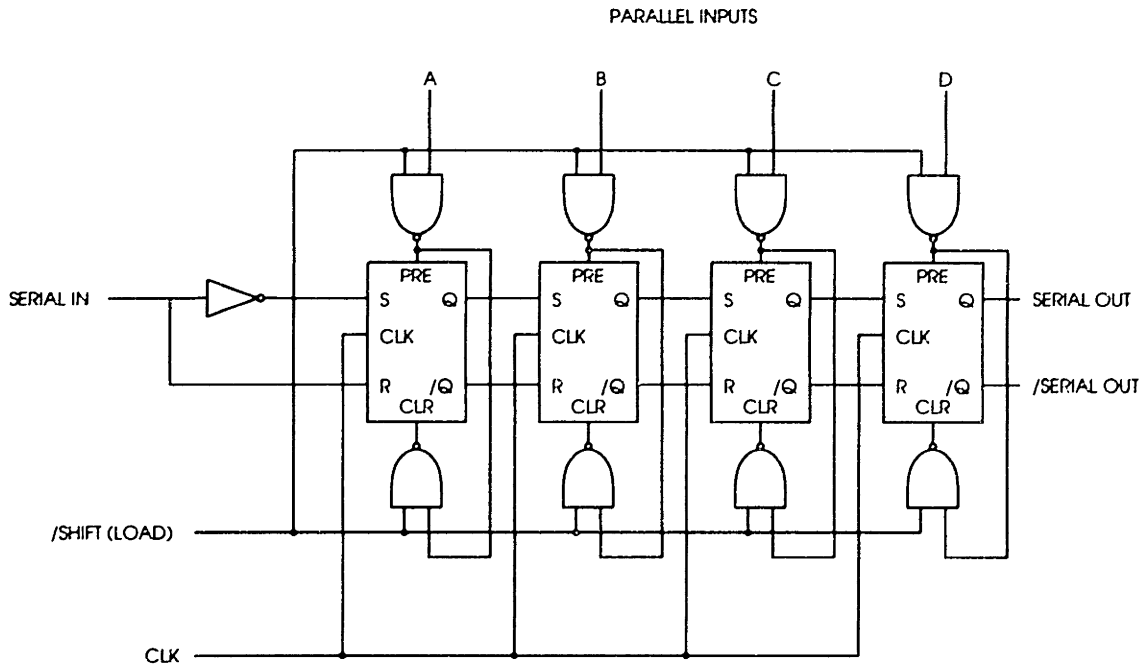


Figure 5.10: Parallel-to-serial shift register

5.4 Personal Computer Interface

The previous sections outlined the implementation of the microcontroller chip that coordinates the sensors on the needle and communicates the information to the personal computer for processing. For simplicity the interface consists of only two information transmission lines (one for sending instructions to the controller and one for sending data back to the computer), several power supply lines, and a clock line. For a typical configuration in which the system clock (and hence the data rate) is several tens of kilohertz, the serial port on a typical personal computer is too slow to keep up with the controller. Consequently, the parallel port on the computer is used for data transmission, as shown in figure 5.1, since much higher data rates can be attained using that port. This section outlines the inner workings of the parallel port interface.

Fundamentally, all that is needed to perform serial-to-parallel or parallel-to-serial conversion is a shift register and some timing circuitry. This is the core of the interface. The data lines on the parallel port are used for bidirectional communication. The

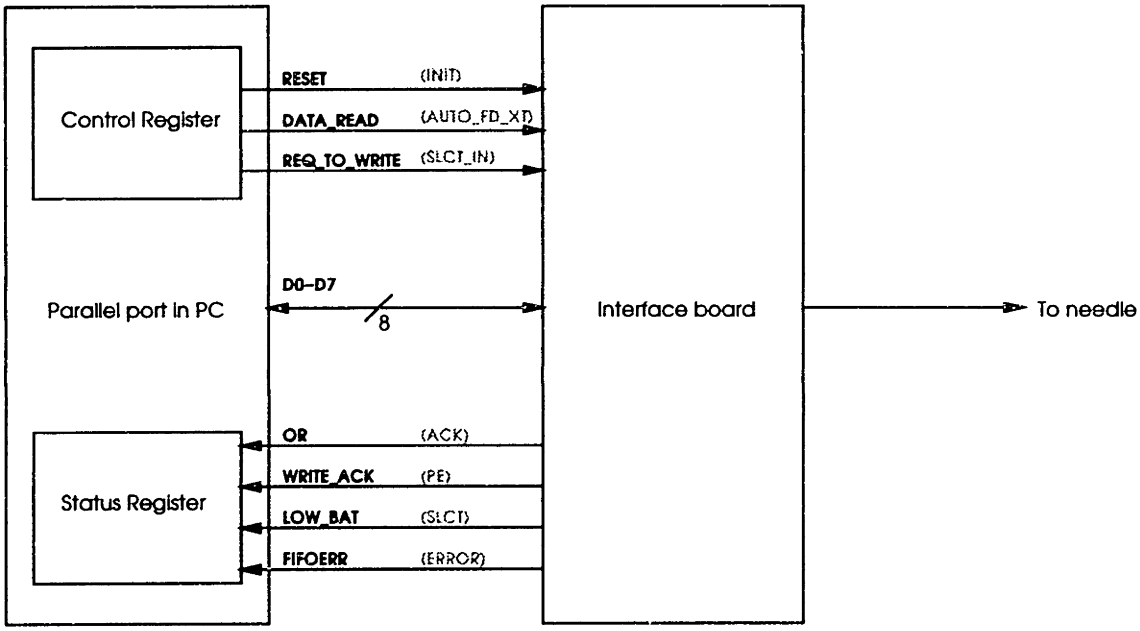


Figure 5.11: Parallel port interface configuration

control register of the parallel port is used for handshaking signals from the computer to the interface. The port status lines are used by the interface for handshaking to the computer. This configuration is shown in figure 5.11. The IBM-PC standard names for the signal lines are indicated in gray.

In order to allow the required bidirectional data flow through the parallel port, a simple handshaking protocol is used. For instruction transmission to the needle, the personal computer first places the instruction code on the data lines, then requests control of the port by asserting the REQ_TO_WRITE line. The interface board then asserts WRITE_ACK to acknowledge the write request and to indicate that it has read the instruction code from the data lines. The computer then deasserts the REQ_TO_WRITE line; this is acknowledged by the interface by deasserting WRITE_ACK. This sequence causes the instruction word to be loaded into a parallel-to-serial shift register that is preloaded with the start and stop bits required by the controller chip. Once the shift register has been loaded, the instruction is shifted directly to the needle.

Unlike the instruction transmission, the interface allows the personal computer to read data without handshaking so that the data transmission rate is maximized.

As the interface board receives data from the needle, it is loaded into an eight bit serial-to-parallel shift register so that every eight serial bits is sent to the personal computer as a single byte. Each 8 bit byte from the register is stored in a first-in, first-out (FIFO) buffer so that immediate transmission to the personal computer is not necessary. This is essential to proper operation of the interface: The personal computer is not always immediately able to receive data because of the many housekeeping tasks that it must perform. When data is waiting in the FIFO, the OR (output ready) status line is asserted. When the CPU on the personal computer is available, it can read as much accumulated data as it would like from the FIFO. This is done by toggling the DATA_READ line, which clocks data directly out of the FIFO. When the FIFO has emptied, the OR status line from the interface is deasserted. Data transmission rates of approximately 1.2 Mbits/sec have been attained using this scheme.

Several additional features are built into the interface to ensure proper operation. Because the instruction codes in this implementation are merely sensor numbers, the maximum instruction code is 15. Since this is represented with 4 bits, the other 4 bits of the instruction byte can be used for special functions. One bit is used to enable a loopback capability on the interface. This feature is provided so that the interface can be tested automatically by PC-resident software. When this bit is asserted, the interface reads the data from the computer, performs the parallel-to-serial conversion as if the data were going to the needle, then reads in the same serial data as if it were received from the needle. In other words, the PC can test the system by writing an instruction to the interface, reading data from the FIFO, and checking that the data read is the same as the instruction written. A second bit of the high order nibble of the instruction byte is used to reset the FIFO should an error occur. The two remaining bits are not used at this time.

Another feature of the interface is the extensive status reporting. In addition to the WRITE_ACK line used for handshaking and the OR line for FIFO status, other lines are provided for FIFO overflow (FIFOERR) and battery exhaustion (LOW_BAT). The

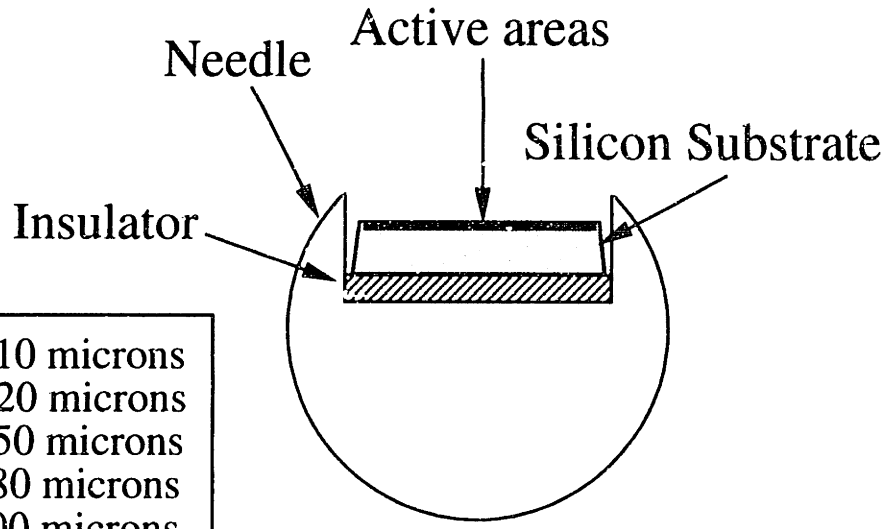
FIFOERR signal is critical, since assertion of this line means that the computer has taken too long to read data, so the FIFO has overfilled and subsequently missed valid data coming from the needle. If this signal was not provided, the personal computer would continue reading data from the interface without realizing that data has been lost.

The low battery indicator is provided so that PC-resident software can monitor the status of the needle and the interface, appropriately notifying the end-user that the batteries that power both the needle and the interface must be changed. This also avoids the masking of "bogus" data as valid. The signal is internally produced on the interface by the voltage regulators that generate the supply voltages for the needle and the interface board logic. Both the FIFOERR and LOW_BAT lines are latched, so that even transient error conditions are reported. The error must be specifically addressed and the interface explicitly reset by the PC-resident software in order for the condition to be cleared.

Chapter 6

Fabrication and Packaging

Certainly in the medical application of integrated circuit technology the manufacturing and packaging issues are at least as critical to success as the design of the circuits. From a silicon microfabrication point of view, the large aspect ratio of the needle places unusual constraints on the layout, namely, the circuits must be long and very narrow. Additionally, there are special processing requirements necessitated by the circuits used on the chip. Finally, there is the biocompatibility issue, which requires not only that the needle not injure the patient, but also that the biofluids not contaminate the circuit. Once fabricated, the assembly of the chips into the final needle structure presents more challenges. The mechanical stresses associated with insertion and removal of the needle force a reexamination of standard packaging techniques. Interchip bonding becomes difficult because of the narrow chip width. A cable must be attached to the probe so that it can be connected to the personal computer. This chapter studies these and other salient points in two parts: First, the microfabrication process and issues associated with chip manufacturing are discussed. The chapter concludes with an examination of the post-microfabrication packaging and assembly process.



Needle diameter	710 microns
Channel width	620 microns
Channel depth	250 microns
Chip width	580 microns
Chip length	7500 microns

Figure 6.1: Needle cross section

6.1 Silicon Microelectronic Processing

6.1.1 Layout Considerations

The layout of the circuits is constrained primarily by the requirement that the chips be narrow enough to fit on a 22 gauge needle, which is 710 microns in diameter. The length constraint is much less restrictive; in order to be compatible with processing equipment in the MIT Integrated Circuits Laboratory, the chips must be less than 1 cm long. The density of measurements, however, is directly related to the length of each sensor chip; therefore, the objective is to minimize the length of the chips while still satisfying the width constraint.

In order to reasonably fit on the needle, the total width of the circuits should be approximately 580 microns. With a 620 micron wide channel, this allows for 20 microns of sidewall on each side of the needle. This extra "gap space" is critical because the sawing process will not produce completely vertical sidewalls on the chips. Failure to account for this sloping sidewall would clearly prevent the chips from fitting properly into the needle groove. This arrangement is illustrated in figure 6.1. For the purposes of layout, this 580 micron allowable width must include not only the circuits themselves

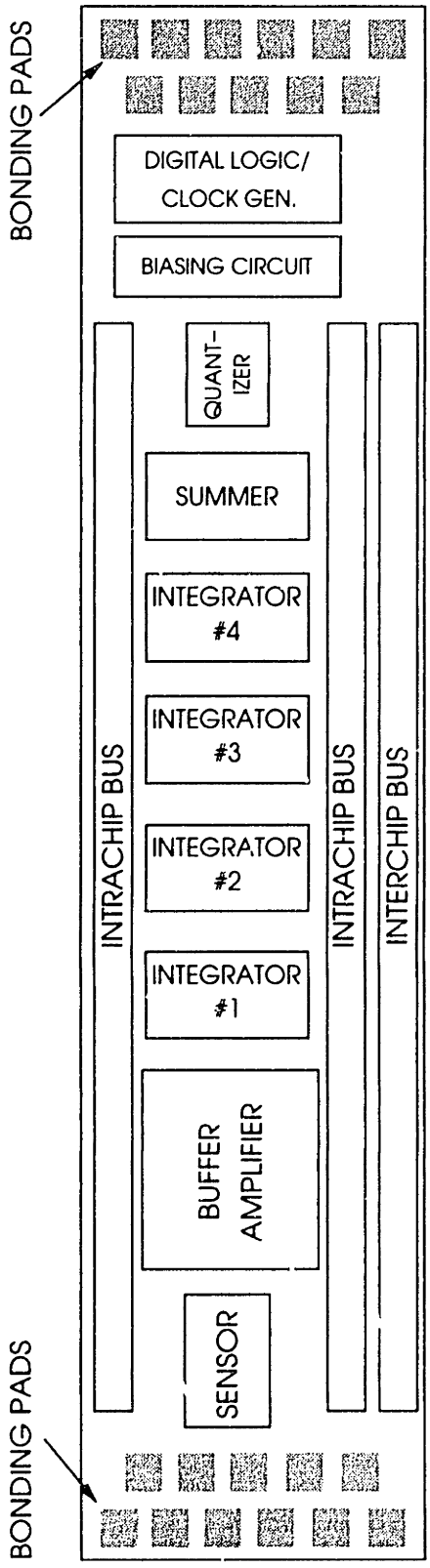


Figure 6.2: SDC chip layout

but also all of the interconnect and bussing from chip to chip. For this reason, the layout was partitioned into four slivers, two for intrachip bussing, one for interchip bussing, and one for the circuits themselves. In this way layout symmetry of the fully differential circuits could be maintained. It is important to note, however, that because of the severe width constraint common centroid geometry layout techniques could not be used, since there is simply not enough space. The layout of the SDC chips is shown in figure 6.2. The four slivers are apparent. Lengthwise, the system is partitioned into blocks following the signal flow. These blocks are identified in the figure. Finally, at each end of the chips are the bonding pads for daisy-chaining several sensors on a single needle.

6.1.2 Process Development

The process used to fabricate the circuits is the BioCMOS process developed at MIT. This flow is a derivative of the CCD/CMOS process developed by Dr. Craig Keast [69]. The BioCMOS process is a two level polysilicon, one level metallization CMOS process with buried channel MOSFETs, non-optimized NPN bipolar transistors, and silicon nitride passivation for biocompatibility. The detailed process flow can be found in appendix A. A technical description of most of the process flow can be found in [69], since the process described there is a subset of the BioCMOS process. The technical details of the enhancements made to the CCD/CMOS process, namely, the addition of a non-optimized NPN bipolar transistor and the biocompatible coating, are described below.

6.1.2.1 The Non-optimized NPN Bipolar Transistor

For this project, it was necessary to have a fully isolated diode available, since these devices are used both in the temperature sensing circuit and the current source reference structure. In an ordinary twin-well CMOS process, it is theoretically possible to develop

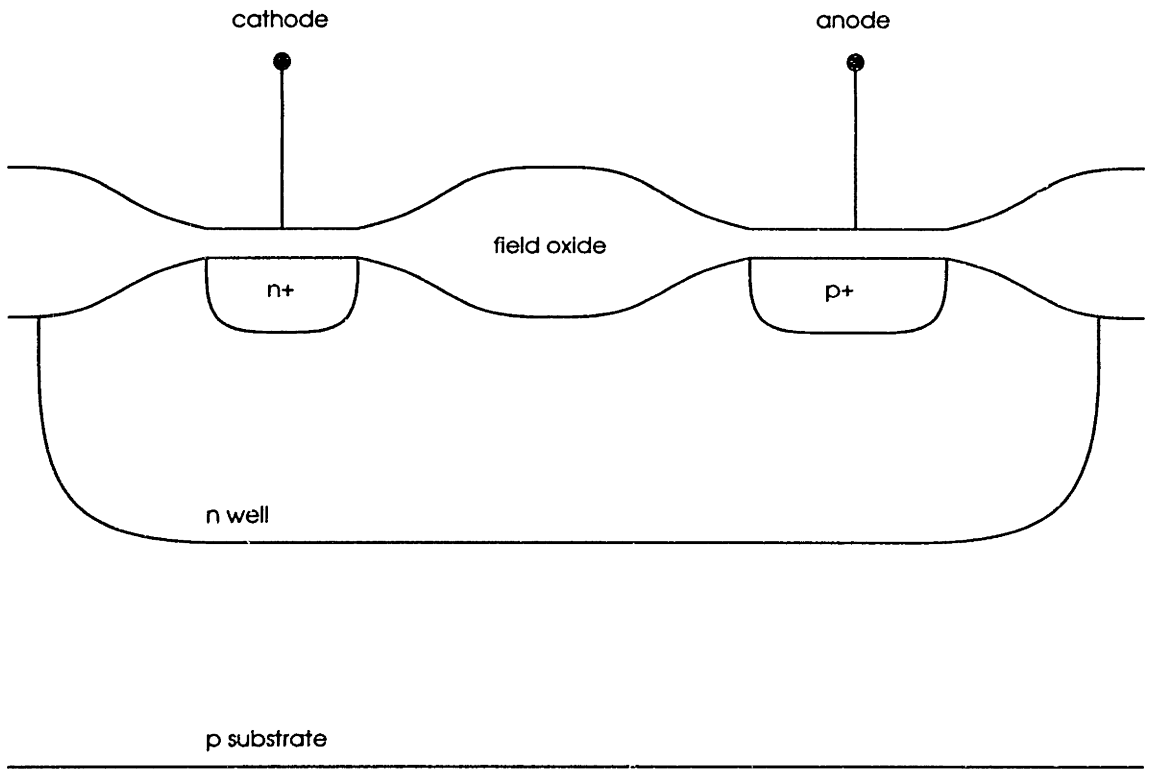


Figure 6.3: P+/N diode structure

such a structure, using, for example, an n-well and a p+ source/drain implant.[1] This structure is shown in figure 6.3. The diode action occurs at the junction formed by the n-well, p+ interface. The diode terminals are as shown in the figure.

This structure works quite well as an isolated diode as long as the negative terminal of the diode is held at the same potential as the substrate. When this is the case, then the junction formed by the n-well, p-substrate interface plays no role in circuit behavior. If the negative terminal is raised to a potential above the substrate potential, however, this junction begins to play a very important role. The two terminal structure continues to behave as a diode as expected, but the field generated by the reverse bias on the n-well, p substrate junction begins to capture carriers injected from the n-well, p+ interface. In other words, a parasitic PNP transistor is formed. The base of this device is the negative terminal of the diode; the emitter is the positive terminal, and

[1]Throughout this section, a p-type substrate is assumed. As a result, all n-wells are isolated.

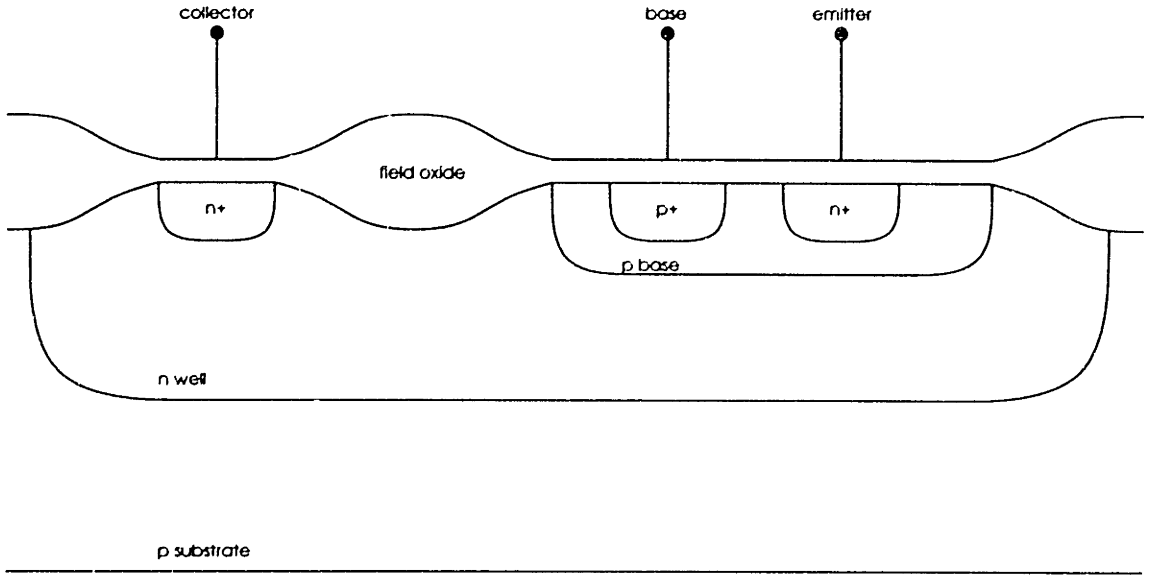


Figure 6.4: Triple-diffused NPN structure

the collector is the substrate. Consequently, the current lost through the substrate is roughly the current through the diode multiplied by the current gain β of the parasitic device. This dramatically increases the power dissipation of the circuit in addition to causing malfunctions.

The simplest way to truly isolate the diode and allow both the positive and negative terminals to "float" is to use a four-layer, triple diffused structure as shown in figure 6.4. The structure that results is actually an NPN bipolar device: The first diffusion defines the collector, the second defines the base, and the third defines the emitter as shown. Now, to create a truly floating diode the transistor is diode connected by tying the base and collector together. As before, the well (now collector)-substrate junction is reverse biased; however, there is no longer injection of carriers across the base-collector junction, so the parasitic PNP device is off and no current is lost to the substrate.

Integrating this triple-diffused NPN structure into the CMOS process must be done with care, so that the MOS device characteristics are not affected. In particular, this means that no additional high temperature diffusion steps can be added to the

process. In the context of the CMOS process flow, two of the three ion implantation steps required for the NPN device are already present: The collector can be formed with the n-well implant, and the emitter can be formed with the n+ source/drain implant. The only additional implant step required is the base formation, which requires a more careful implantation than either the collector or the emitter. This implant is done as a separate mask step following the MOS threshold and punchthrough adjustment implants. Subsequent high temperature steps (gate oxidations, etc) provide the necessary implant anneal.

Although the collector can be formed simultaneously with the n-wells, it is better if a separate collector implantation is done. This is because of the extremely light doping of the n-wells; when this doping is used to form the collectors, the collector resistance of the resulting NPN bipolar is so high that the transistor is saturated for any reasonable current bias. Numerically, with an average n-well phosphorus doping of approximately $9 \times 10^{14}\,\mathrm{cm}^{-3}$, corresponding to an ion dose of $2 \times 10^{12}\,\mathrm{cm}^{-2}$, the resistivity is about $5\ \Omega - cm$. The final well depth after processing is approximately $2.3\ \mu m$. The resulting nominal collector resistance is on the order of $22\ k\Omega$. With this high a collector resistance, a diode connected transistor would be limited to a current of approximately $25\ \mu A$ before the device would saturate.

As a result, a separate collector formation implant is performed. This implant is in essence an "incremental" implant; all wells, both for the MOS devices and the bipolar devices, receive the baseline well implant. A separate mask step is then used tc cover all of the MOS wells while leaving the bipolar wells (collectors) exposed. A second implant is then performed on top of the baseline well dose. This second implant is then annealed with the MOS wells during the well drive-in step.

The heavier doping not only lowers the resistivity of the well, but also increases the well depth, further lowering the collector resistance. In order to guarantee that the collector resistance would not saturate the device for the current levels used in the circuits, the resistance should be on the order of $1\ k\Omega$. Using commonly available

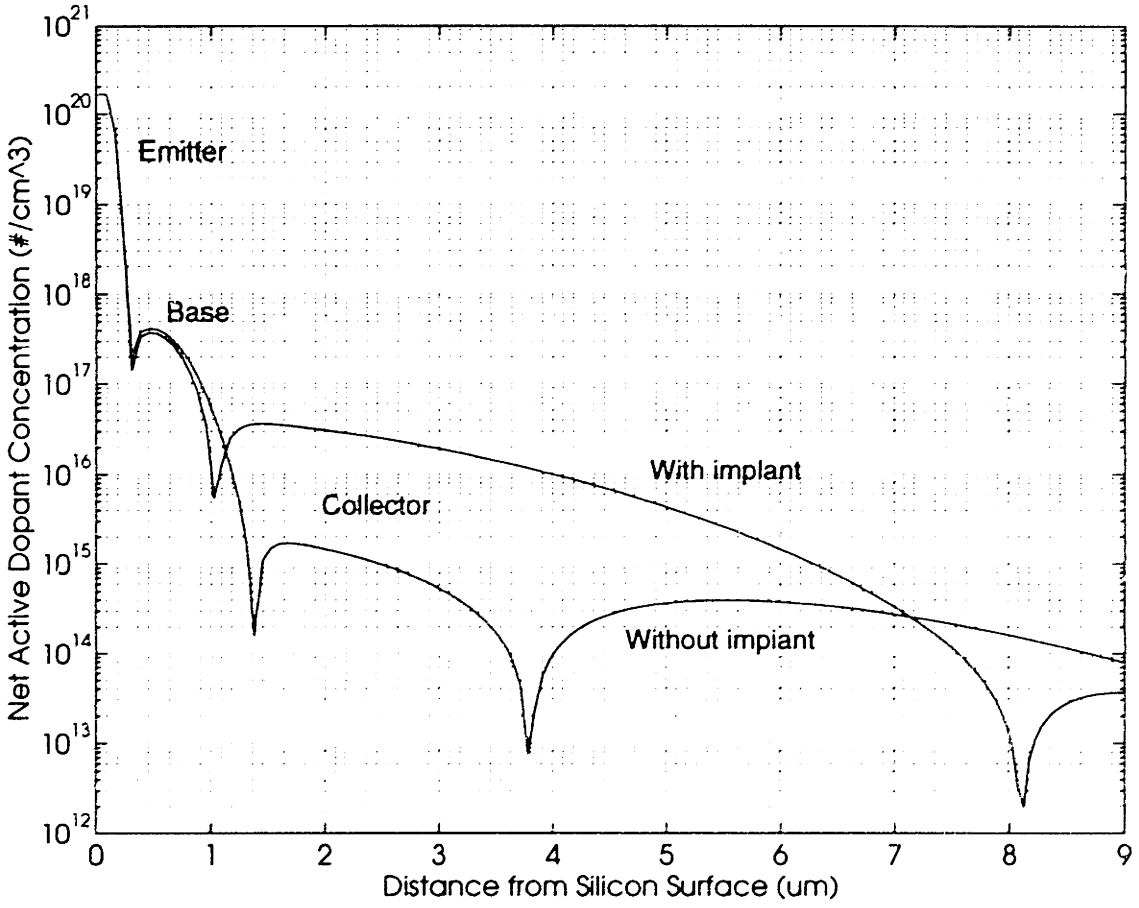


Figure 6.5: NPN doping profile with and without collector implant

charts of resistivity versus doping [70] and typical geometries, one finds that a final well doping on the order of $1 \times 10^{16}\,\mathrm{cm}^{-3}$ is required to produce a collector resistance on the order of $1\,k\Omega$. This corresponds to a well implant dose of approximately $2.2 \times 10^{13}\,\mathrm{cm}^{-2}$. Since the baseline well dose is $2 \times 10^{12}\,\mathrm{cm}^{-2}$, the additional collector implant dose used is $2 \times 10^{13}\,\mathrm{cm}^{-2}$. The energy of the implant is the same as the baseline energy, namely, 180 keV. The resulting collector resistance for a typical geometry device is approximately $720\,\Omega$. The effects of this additional implant can be seen in figure 6.5, which shows the NPN doping profiles both with and without the additional collector implant.

The base implant parameters must be selected with some care, since this implant

determines most of the relevant performance parameters of the resulting transistor. In order to avoid affecting the MOS devices, this implantation is done just prior to the first level gate oxide growth, through the sacrificial gate oxide, as discussed above. Any drive-in of the base, therefore, is performed by the subsequent high temperature steps; since there are no significant long high temperature steps after the field oxidation, it becomes important that the base implantation is done as deeply as possible into the silicon. For this reason an implant energy of 190 keV was selected.[2] The implant species is boron since a p-type doping is required.

The implant dose was selected based on several factors. For low implant doses, corresponding to a more lightly doped base region, the base becomes more narrow and the ratio of emitter to base doping becomes larger. As a result, the current gain of the device becomes larger. There are several disadvantages of a lighter dose, however. First, the output resistance of the device is lowered, since a larger fraction of the base-collector space charge region extends into the base (higher base width modulation). Second, the base-collector reverse bias necessary to produce punchthrough of the base is lowered, since the base is narrower and more modulated by the base-collector bias. Finally, the base resistance of the device is raised, because the base is both narrower and more lightly doped. This last effect is the most critical to this project, as it increases the deviation of the I_C vs. V_{BE} from the ideal.

Heavy doping of the base alleviates most of these problems. At higher doping levels the β of the transistor is lowered because the base is wider and the ratio of emitter to base doping falls. The base resistance decreases because the base is wider and because the resistivity of the base decreases with increasing doping. However, because the β of the transistor is also falling, there is no real advantage of having a lower base resistance (i.e., the base current goes up because of the falling β, so the $I_B R_B$ product remains roughly the same). Most importantly, however, is the lateral diffusion at higher doping levels--the minimum geometry of the devices must be increased so that the base dopant

[2]This energy is the maximum ion energy that can be produced with the ion implanter at MIT.

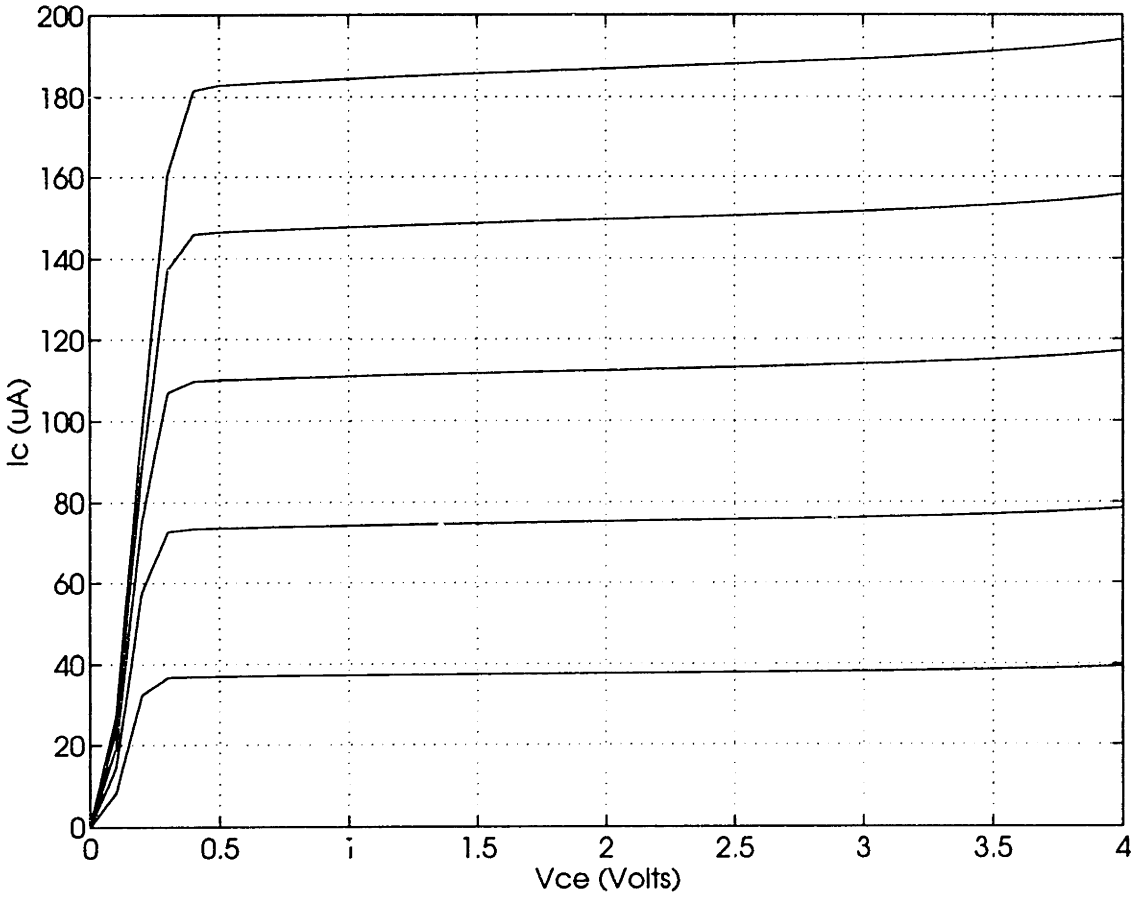


Figure 6.6: Vertical NPN transistor, typical measured output characteristic ($I_b = 1\mu A \rightarrow 5\mu$A in $1\mu A$ steps)

does not laterally diffuse through the collector well, resulting in a base-substrate short.

Simulation using SUPREM3 [71] was used to select the correct compromise between light and heavy doping. The results demonstrated that implant doses in the range of $10^{13} - 10^{15}\,\mathrm{cm}^{-2}$ will provide satisfactory results. Since the effects of lateral diffusion could not be studied, a dose closer to the low end of the range, $3 \times 10^{13}\,\mathrm{cm}^{-2}$, was used. The resulting vertical NPN structure as simulated by SUPREM3 is shown in figure 6.5. The output characteristic of a fabricated NPN device is shown in figure 6.6. Most important for this project is the behavior of the device when diode connected ($V_{BC} = 0$). The $I - V$ characteristic in this case is shown in figure 6.7.

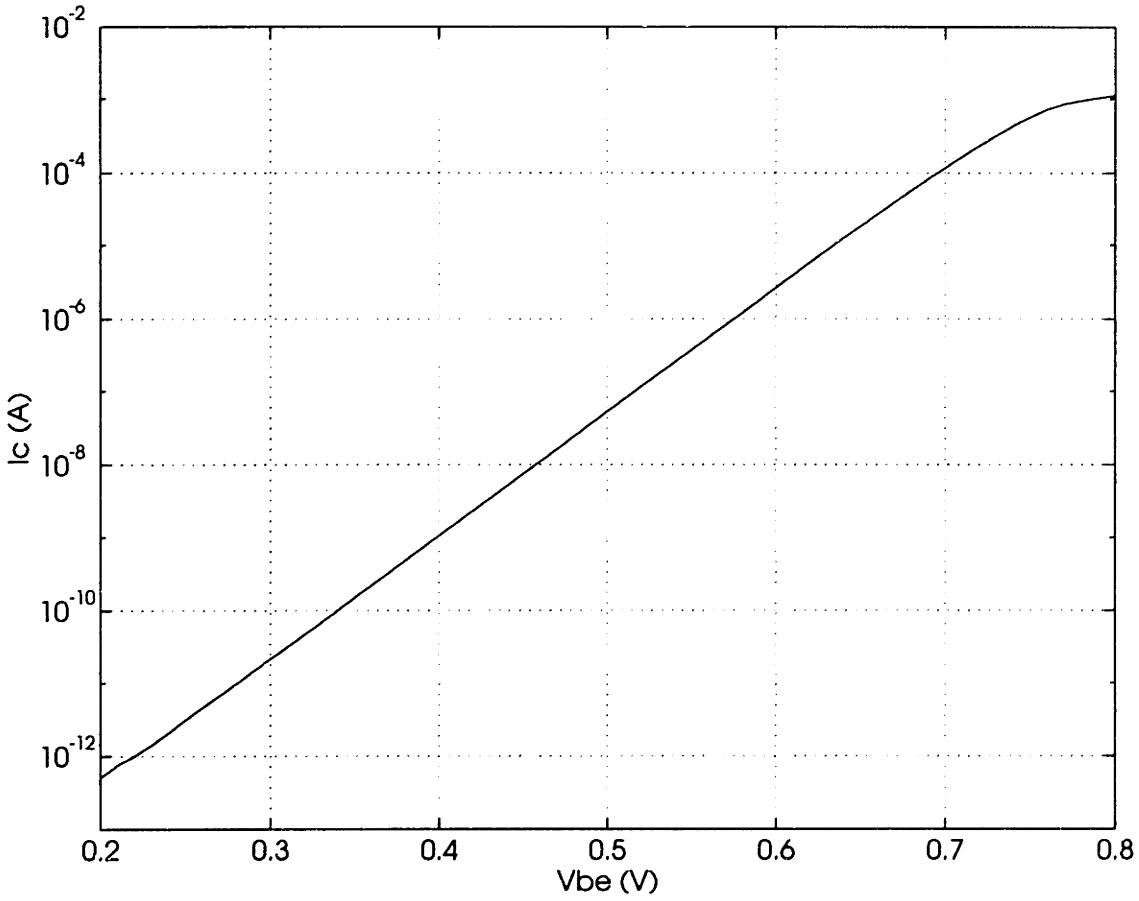


Figure 6.7: I-V characteristic of diode connected vertical NPN

6.1.2.2 Biopassivation

As discussed earlier, passivation of the circuitry is critical since the chips will be operated in the very hostile physiologic environment. The requirements of the passivant are clear: First, the passivating material must be able to block diffusion of impurities from the body to the chips; since the primary (and fastest diffusing) impurity found in the body is the sodium ion, diffusion of sodium through the passivant can be used as a benchmark for its effectiveness. Second, the passivant must be a very good electrical insulator, not only to prevent short circuits on the chips, but more importantly to guarantee that little or no current can flow from the electrically active circuit through

the patient. Finally, because the passivation is at the end of the fabrication sequence, it is critical that deposition of the material is performed at temperatures below 450°C to prevent an additional thermal dose from affecting the already-completed devices.

The material that meets all of these requirements rather handily is silicon nitride. It is ideally suited for this system, and has been shown to effectively isolate devices from the physiologic environment over extended periods [72,73]. It has been shown experimentally [74] that sodium penetration in silicon nitride is less than 100 $\mathring{A}$. The bulk resistivity of the film is $10^{15}\Omega \cdot cm$ [75]. The thermal requirements are satisfied by depositing the film using plasma enhanced chemical vapor deposition, which can be performed reliably at temperatures as low as 250°C. Holes in this coating are opened only to the bonding pads for interchip connections; these areas are later sealed as described below.

In low pressure, high temperature CVD the chemical reaction that forms the film is driven by the elevated temperature. Plasma enhanced CVD works by using the plasma energy to drive the gaseous reaction to form the film. Silane (SiH_4) and ammonia (NH_3) are used as the ambient gases; although nitrogen gas (N_2) could be used (and is actually used as a carrier gas), it is not, because the energy required to dissociate the nitrogen molecule (945 kJ/mol [76]) is significantly higher than the dissociation energy of ammonia (356 kJ/mol [77]). The plasma causes the breakup of both molecules, resulting in the formation of silicon, nitrogen, and hydrogen ions. The free ions then react to form the silicon nitride that is deposited on the wafer surface.

There are several parameters that affect the quality of the film. The first is the ambient temperature of the deposition chamber. At lower temperatures, the defect density of the film increases, reducing the effectiveness as a passivant. At deposition temperatures of 300°C and above, however, Kern and Rosler [74] observed that these defects were essentially eliminated The second important parameter is the deposition pressure; generally, the lower the pressure during deposition, the denser the film. Higher film density further improves the passivant properties. Third is the plasma

Table 6.1: PECVD silicon nitride film parameters

Parameter	*Value*
Silane (SiH_4) Flow	24 sccm
Ammonia (NH_3) Flow	40 sccm
Nitrogen (N_2) Flow	50 sccm
Chamber Pressure	300 mTorr
RF Power	90 W (125 mW/cm^2)
Dep. time (1μm)	65 min

power density. As with the lower pressure, high power densities result in a better, denser film. In addition, the deposition rate increases with increasing plasma power.

The fourth and most critical process parameter is the silane/ammonia gas flow ratio. As discussed above, the plasma dissociates each of the molecules to produce silicon, nitrogen, and hydrogen ions. The gas ratio determines the relative concentrations of each ion. Since hydrogen must be present (from both the silane and the ammonia), it is inevitable that the resulting film will contain a certain amount of hydrogen in addition to silicon and nitrogen, and in fact, rather than a film that is entirely Si_3N_4 (with a Si/N ratio of .75), the deposited film is of the form $Si_xN_yH_z$, with a typical Si/N ratio of .8-1 [74]. The amount of hydrogen in the film is evaluated by looking at the index of refraction (which decreases with increasing hydrogen concentration) and the infrared absorption spectrum (which clearly shows the presence of Si-H bonds). Better films minimize the amount of hydrogen incorporated and have indices of refraction between 2.0 and 2.1.

As part of this project, a plasma enhanced CVD nitride deposition process was developed. The film deposition parameters and measured film properties are given in table 6.1. The FTIR spectrum of the film is given in figure 6.8. As can be seen from the spectrum, there is indeed a small amount of hydrogen present in the film. The index of refraction verifies that this amount is relatively small (indices of between 1.6 and 1.8 were obtained for more hydrogen-rich films). Although a thickness of

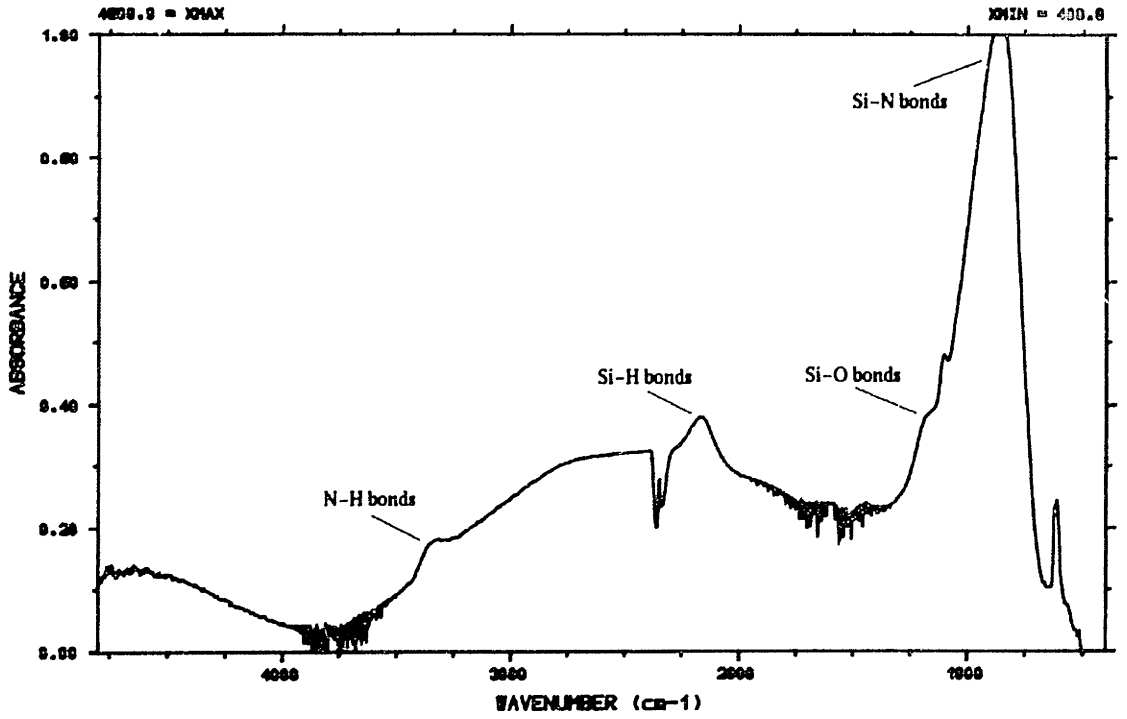


Figure 6.8: FTIR spectrum of PECVD silicon nitride film

several hundred Angstroms would suffice, the film thickness of 10,000$\mathring{A}$ minimum was used to guarantee an extremely strong barrier, and to ensure good step coverage over the underlying 10,000$\mathring{A}$ metal layer. Stress cracking of the film does not occur because of the lower stress of the PECVD nitride ($\approx 5 \times 10^9 \frac{dyn}{cm^2}$ vs. $1.5 \times 10^{10} \frac{dyn}{cm^2}$ for LPCVD nitride [78]). Circuits passivated with this film have been operated in a fluid environment for several hours with no sign of circuit degradation.

6.2 Probe Assembly

The microfabrication of the circuits takes place on 4-inch silicon wafers; once the silicon nitride passivation process is performed, the wafer-level fabrication is complete. At this point, the chips must be assembled into a complete needle. The process involves three basic steps: Grinding, sawing, and bonding. This section discusses each of these steps in detail, outlining how the circuits are taken from the silicon processing facility

(wafer level) to the needle structure (chip level).

Once the wafers leave the clean room facility, the first post-processing step is the thinning of the wafers from a manufactured thickness of 500 μm to the 200 μm thickness required for the needle; this is performed first because it is a wafer-level operation. There are two primary methods for accomplishing this: Chemical etching or mechanical grinding of the silicon from the backside of the wafer. Chemical etching is avoided because of the elaborate process required to protect the completed circuits on the front of the wafer. The front side is first protected with a special plastic tape. The wafers are then sent to a commercial wafer grinding facility where high precision machinery is used to grind 300 μm of silicon from the backside of the wafer. No backside polishing is done following the coarse grinding, since the larger surface area of the roughened backside improves adhesion to the needle substrate. Once the wafers are returned from the grinding facility, an acetone soak is used to remove the tape from the front side of the wafer. The wafers are then rinsed with methanol to remove the residue that acetone typically leaves on the wafer surface. The now-thinned wafers are ready to be diced.

The wafers are diced using a diamond blade silicon saw. Because of the extremely small size to which the wafer must be diced, the wafers are once again mounted on the special tape; this time the tape is placed on the back side of the wafers. The wafers are sawn using standard procedures. The only subtlety involved is the blade height above the saw chuck, which must be carefully selected because the wafers have been thinned. Typically, the blade cuts approximately 100 μm above the saw stage. This corresponds to a cut of 400 μm through a standard 500 μm wafer. In this application, the blade-stage gap used is 425 μm, corresponding to a cut depth of 125 μm through the 200 μm wafer--the remaining 300 μm is required because that is the thickness of the tape on the back of the wafer. Once sawn, the wafer is gently stressed along the cut lines and the dice separate (but remain adhered to the backside tape). The wafers are not sawn completely through because the mechanical stress of the sawing process can

pull the sawn dice off the tape.

When the sawing is completed, the tape is removed using the same acetone soak and methanol rinse described earlier, only this time, the dice are stored in a methanol filled, sealed petri dish, to prevent the accumulation of dirt on the surface of the wafer during storage. At this point, the dice are ready for mounting in the needles, which are solid 22 gauge needles that have had the channel already milled in them. The needles are cleaned in methanol to remove any particles that may be in the channel. A thermally insulating, FDA-approved epoxy, BA-F DA2, is then used to coat the inside of the channel. The silicon slivers are then carefully mounted into the channel end-to-end. The epoxy is cured to fix the chips in place.

At this point, the only assembly task remaining is the chip-to-chip bonding. Because of the narrow width of the dice, the bonding pads are quite close to one another, and extreme care must be used to make the bonds. To assist the process, thin wire (0.7 mil diameter) is used. A special bonding tip is also employed--the tip is designed for the thinner wire, and for a very small bond footprint. In this way the tail of the bonds do not interfere with one another. Each chip is bonded to its adjacent dice. The small microribbon cable used to connect the needle to the "outside world" is then ball bonded to pads on the digital controller. Gold wire is used for the interchip connections, as these wires have shown extremely low leakage, even under electrical stress [79]. Gold also bonds well to aluminum, the standard metallization used in the fabrication process. The bond areas are then coated with H77 epoxy; soak testing on this epoxy as a bond area coating has demonstrated its suitability [80]. The epoxy, through surface tension, completely encapsulates the bonding area. Once cured, the needle is ready for the final overcoating.

6.3 Final Coating

Although the circuits themselves have been passivated with silicon nitride, there is no guarantee that there will be uniform coverage of the chip sidewalls by the epoxy.

In addition, the bond wires themselves are uncoated and are must be protected. The complete packaging of the system, therefore, is critically important to the success of the project, as the needle will be operated in the hostile physiologic environment as discussed earlier. There are a number of issues that must be addressed to insure that the system will function in this environment: First, there are diffusional impurities, most notably sodium, that will degrade device performance or cause complete failure if allowed to come in contact with the wafer surface. Then there are the effects of the aqueous environment itself, which can cause corrosion of the metal interconnect layer on the surface of the chips in addition to large ionic leakage currents [81]. Mechanical stresses on the needle during insertion and removal also stress the silicon substrates, possibly causing catastrophic system malfunction. Finally, as discussed earlier, short circuits due to aqueous solutions bridging interconnects pose a threat not only to the system but also to the patient.

The performance goals outlined above can be restated in terms of essential mechanical and chemical properties of the passivant. These requirements have been outlined at length [82]; those of particular importance to this system are summarized here. Foremost is the protection of the circuit from moisture. This requires hermeticity in the seal, impermeability of the coating to water vapor, and good adhesion of the passivant to the substrate. Also necessary is a biocompatible coating, namely, a coating must not only prevent metallic corrosion (as mentioned above) but its ability to do so must not be degraded over time by the substances in the body. Contamination of the body by the coating must also be prevented; in other words, sterilization of the coating is also essential. From a mechanical standpoint, the passivant must not be too rigid or too flexible so that it can absorb the transient stresses. This is not contradictory, even though it may appear so. Some rigidity is needed to prevent mechanical stress on the chips themselves, and some flexibility is needed to prevent microcracks and the eventual failure of the coating that results from them. Finally, there are the long-term reliability issues, namely mechanical, electrical, and thermal stability: The

effectiveness of the coating must not significantly degrade over time as a result of normal mechanical, electrical, or thermal stress.

A material that satisfies all of these requirements is Teflon^{TM}, which provides integral protection for any areas that are not appropriately passivated already, or any pinhole defects that may exist in the underlying passivation layers. This film by itself has been shown to be an adequate barrier to a saline environment [83]; the combination of the silicon nitride, the biocompatible epoxy, and this film should provide failsafe protection of the circuits from the tissue and vice-versa. The film is deposited using a vapor coating technique, and is done at a commercial facility (Precision Coatings, Inc). Once coated, the needle assembly process is complete and the probes are ready to use.

Chapter 7

Results and Conclusions

The previous chapters have discussed in detail the design and fabrication of the active needle system. Those chapters emphasized the design goals, and the thought process behind the procedures used for the various designs, process modifications, and assembly techniques. This chapter examines the measured performance of the temperature sensing system. First, the measured performance of the operational amplifier that forms the core of the system is studied. The results of temperature measurements using the system are then presented, including data that demonstrates the capabilities of the system. The chapter concludes with a summary of the project, and suggestions for future research.

7.1 The Operational Amplifier

Since the operational amplifier is the basic building block upon which the temperature sensor, the preamplifier, and the analog modulator are all based, quantification of its behavior is critical. For this reason, op amp measurements were made prior to evaluation of the entire system. This section explains the experimental test setup and the results of the measurements taken with this test system.

7.1.1 Test Setup

Five parameters were used to describe the performance of the amplifier: the DC gain, the bandwidth, the offset voltage, the output current, and the power dissipation. The DC

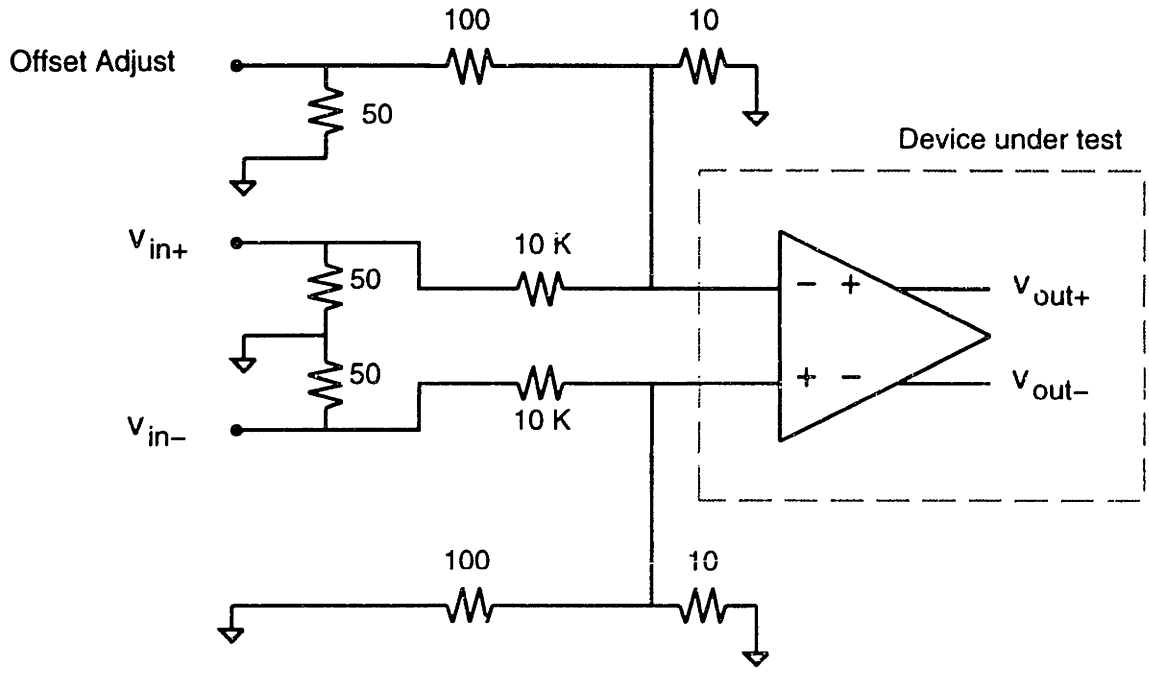


Figure 7.1: Op amp DC gain measurement setup

gain is most important since it determines the noise attenuation in the temperature sensor as was described in chapter 3. The bandwidth ultimately determines the maximum data rate of the system, since it is used in all of the component subsystems. The output current is a measure of the slewing ability of the amplifier; this measurement is preferred over the slew rate since the latter is a function of the load capacitance for a folded cascode topology where the output load capacitance is also the compensation capacitance. Because the resolution of the modulator is compromised when the switched capacitor integrators do not fully settle within half of a clock cycle, the op amp output current controls the maximum clock frequency at which the system can be operated at full resolution, even if the small signal bandwidth is large. The power dissipation controls the amount of self-heating that occurs in the system as described in Chapter 2. Finally, the offset voltage is a measure of the process matching, which affects the resolution of the temperature sensor as was described in section 3.3.2. To facilitate testing, these measurements were all taken using a general-purpose, configurable operational amplifier test board designed and developed at MIT [84].

The DC gain measurement was made using the configuration shown in figure 7.1. Because of the very high output resistance of the amplifier, active probes with an input resistance of $10^{12}\,\Omega$ were used to measure the voltage at the output nodes. Prior to making the measurement, the attenuation of the voltage divider at the amplifier inputs was measured, since small series resistances could significantly alter the expected division. The measured DC gain was approximately 100 dB, as predicted by the design.

The bandwidth was measured two different ways. First, a measurement of the 3 dB point was made using the DC gain setup. The measured location of the dominant low frequency pole was 89 Hz. Using the single pole model, this predicts a bandwidth of 8.9 MHz. Of course, the second pole of the amplifier will affect performance at frequencies below 8.9 MHz, so this number represents an upper bound on the bandwidth. The second measurement was the direct examination of the unity gain bandwidth. For this measurement the voltage dividers at the op amp inputs were removed. A very small signal amplitude (50 mV) was used to prevent slew rate limiting, since the output current is low (see below). The total output capacitance of the test setup was quantified prior to the measurement and was found to be 19.2 pF. Under these conditions, the unity gain bandwidth of the amplifier as extrapolated from measurements was 2.2 MHz.

The offset voltage of the amplifier was measured on several different dice so that an average value of the offset could be calculated. The offset of each amplifier was measured using the DC gain test configuration. One of the amplifier inputs was grounded; a precision programmable voltage source (Data Precision Model 8200) was tied to the other input. The voltage at that input was then varied until the output voltage was zero. The average value of the offset was 2.4 mV.

The output current was measured two ways using the DC gain test setup. The first, approximate measurement was made by applying a moderate frequency, large signal sine wave to the amplifier inputs. The large signal amplitude and the higher frequency (10 kHz) caused slew rate limiting at the amplifier outputs. The slope of the output triangle wave was measured; this value, with the knowledge of the capacitive

Table 7.1: Measured op amp performance

Parameter	*Value*
DC Gain	90,000
Bandwidth (19.2 pF load)	2.2 MHz
Output Current	17 μA
Offset Voltage	2.4 mV
Power dissipation	1.1 mW
Power supply	6 V

loading, permitted calculation of the output current. The second measurement was a direct measurement, in which the amplifier outputs were purposely loaded with $10\,k\Omega$ resistors so that the current limiting would occur. The output current was easily calculated from the peak voltage measured at each output using Ohm's law. Both measurements produced approximately the same result: The maximum output current in each leg was $8.5\,\mu A$, for a total differential output current limit of $17\,\mu A$. This is slightly below the design value of $20\,\mu A$ but is still above the required output current of $15\,\mu A$.

Finally, the power dissipation was measured by monitoring the power supply currents. The power dissipation of a single op amp is 1.1 mW. This includes the power consumed by both the current source reference circuit and the operational amplifier itself. The consumption is slightly lower than expected, but this is due to process variation in the current source reference resistor. The design calls for a $2\,k\Omega$ resistor; the actual resistance, computed from sheet resistance measurements of the fabricated chip, is approximately $2.2\,k\Omega$. This 10% increase in resistance reduces the overall power consumption by the same percentage by changing the value of the reference current source. This conclusion is also borne out by the output current measurement, which also shows a 10% reduction below the design value.

Overall, the amplifier measurements match well with the design specifications. Table 7.1 summarizes the measured performance of the op amp; figure 7.2 graphically

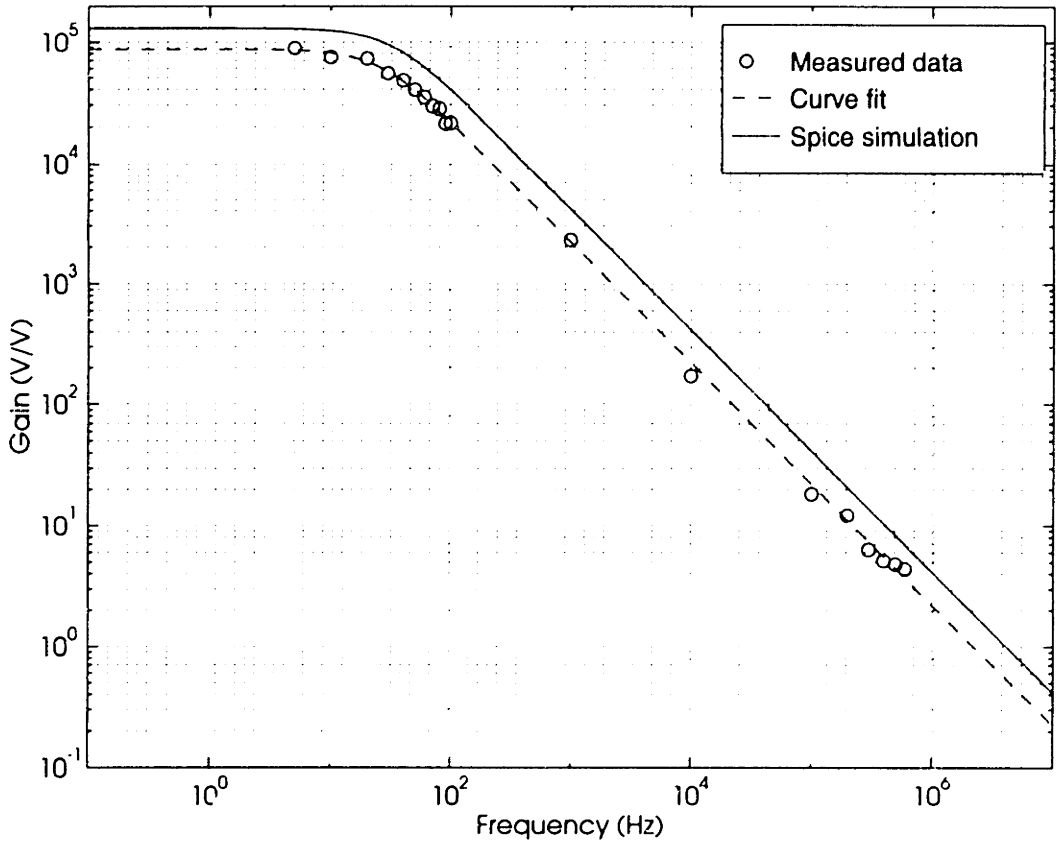


Figure 7.2: Measured vs. predicted op amp AC performance

compares the AC performance of the op amp with a Spice simulation of the design. Although the correlation between the predicted and actual performance is high, the op amp simulation is consistently more optimistic in its gain prediction than the measurements bear out. The measured curve is shifted slightly downward from the simulation prediction. This is due to inaccuracies in the device models used for the simulation, namely, the output conductance of the devices is slightly underpredicted by the models. The device models used were derived using a parameter extraction algorithm on a limited set of measured device sizes. Agreement of the extracted models with other devices is quite good but errors of 5-10 percent per device can be expected.

7.2 Single-Point Temperature Measurements

Ultimately, the measure of the success or failure of the chip is dictated by the measured digital temperature output. All of the design choices were governed by the desired temperature resolution of 1 m°C at 1 Hz. This section presents some temperature measurements made with the fabricated chips. First the temperature testing setup and experimental method is described. This is followed by a detailed examination of the temperature tests to determine the limits of the system.

7.2.1 Experimental Setup

A stable environment for temperature measurement was created using a circulating water bath. The temperature of the bath was controlled using a Techne TU-19 heater/controller. The chips themselves were mounted in 40 pin dual in-line packages for testing. Two thermistors, calibrated to ±3 m°C, were attached to the ceramic chip carrier so that the actual carrier temperature could be monitored; since the thermal conductivity of the silicon/ceramic system is much larger than any other material used in the system, the carrier temperature is an excellent measure of the chip temperature. The two thermistors were separated by a known fixed distance; this allowed for quantification and correction for any thermal gradients that might be present due to temperature stratification in the bath.

The chip/thermistor assembly was placed in a small waterproof plastic bag that was sealed around the communication cable and immersed in the water bath. The plastic "capsule" was clamped under the surface of the water to prevent motion during the measurements and flotation caused by residual air in the bag. A specially built cable connected the sensor chip in the bag to the parallel interface board. Custom software on a 486-based personal computer was used to communicate with the interface board. Digital data generated by the chip under test was passed to the interface board, where it was reformatted and passed in to the parallel port on the

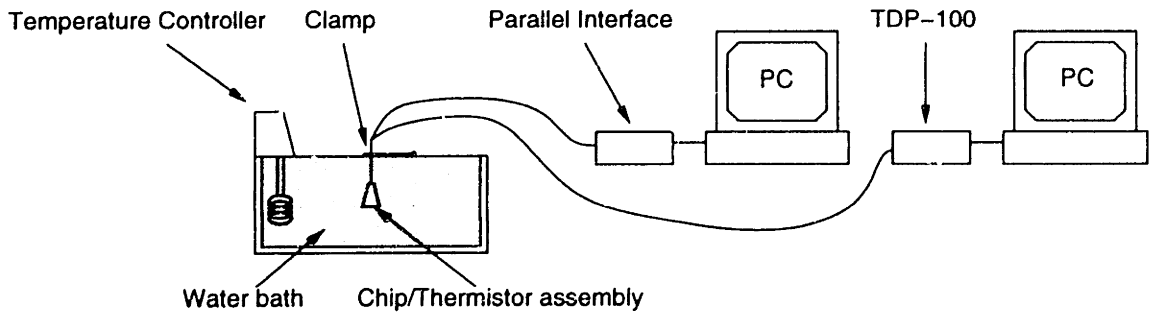


Figure 7.3: Temperature testing setup

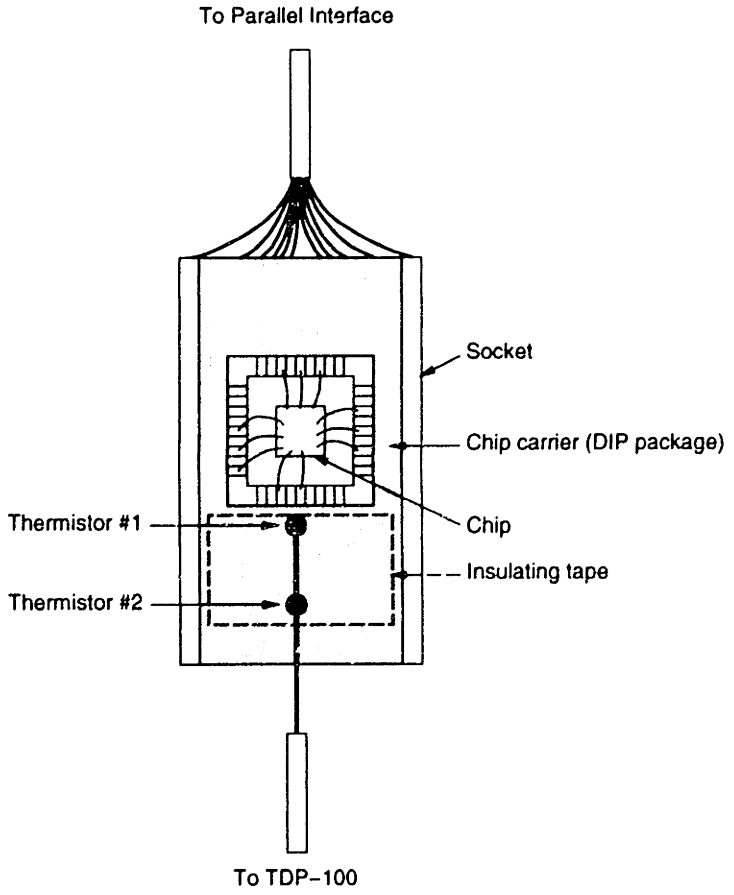


Figure 7.4: Chip/thermistor assembly

personal computer. Temperature recordings from the thermistors were taken using a thermistor-based temperature measurement system of resolution 3 m°C (TDP-100, Thermal Technologies Inc., Cambridge, MA) calibrated against a precision thermometer traceable to NIST. This system was connected to a second personal computer, where the measurements were displayed. Figure 7.3 diagrams the test configuration; figure 7.4 shows a close up view of the chip/thermistor assembly.

Under normal circumstances, data from the sensors would be processed in real-time by a digital signal processor resident on the personal computer. In order to evaluate the sensor system without this capability (which is under development) an alternative scheme was developed since timing and space limitations on the personal computer made continuous long-duration measurement of the output bit stream impossible: The maximum block of continuous data that could be acquired by the computer was approximately 2 megabits, corresponding to a measurement time of approximately 30 seconds. For this reason, an accumulate-and-dump function was performed on the output bits to reduce the space and memory requirements on the personal computer end. Every 3000 bits were accumulated (summed) and dumped to a file. Depending on the clock frequency used, this action is equivalent to a downsampling to 4, 8, or 16 times the Nyquist frequency. The prefiltered data was then stored and analyzed at the conclusion of an experiment. Several data analysis algorithms were used depending on the particular experiment; these procedures and the effects of the accumulate-and-dump function are discussed below.

7.2.2 Sensor Calibration

Because of the desire for a completely digital interface to the sensor chips, the actual analog temperature signal of interest is never directly sensed. Instead, this signal is amplified then modulated; the resulting digital bits are recorded, and are used to extract the temperature signal. It is possible to associate with each stage of processing scale factors and offsets; although in theory these values are known from the design, in

practice they vary from chip to chip due to random process fluctuations. Clearly there is a need to quantify the relationship between the output bits and the temperature signal in order to make accurate determination of temperature possible.

Although any individual bit in the modulator output data stream cannot be uniquely mapped to temperature, it is a basic property of the modulator that the output pulse density (i.e., the average value of the bits) is correlated to the modulator input. This input is the amplified version of the temperature voltage signal, which is in turn related to the physical temperature by equation 3.9. Mathematically, the relationship between the temperature and the output pulse density is:

$$T = \frac{qG}{k\log(n)} \cdot (D - D_{mid}) \tag{7.1}$$

where D is the output pulse density, G is the modulator D/A reference voltage divided by the preamplifier gain, n is the sensor excitation current ratio, and D_{mid} is the pulse density that occurs for zero input voltage.[1] Ideally, for this system, G = 20, n = 10, and D_{mid} = 0.5. In practice, none of these values are known with certainty. If they are treated as unknowns, it is clear that a two point calibration is required, from which D_{mid} and $\frac{G}{\log(n)}$ can be extracted.[2]

The calibration is therefore performed by measuring the output pulse density when the sensor is placed in a known, fixed temperature environment. These two points determine a line

$$T = mD + b \tag{7.2}$$

Matching terms in equation 7.2 with equation 7.1 gives:

$$\frac{G}{\log(n)} = \frac{mk}{q} \tag{7.3}$$

[1]This is required since the output bits are 1 or 0 and the system is fully differential. When the differential input is zero, the modulator generates an equal number of high and low bits, resulting in a pulse density of 0.5, not 0.

[2]It is not possible from this relationship and it is not necessary to find G and n independently in order to completely determine the pulse density-temperature relationship, since they can be lumped into a single scale factor.

$$D_{mid} = \frac{-b}{m} \tag{7.4}$$

These values are then used to generate the actual sensed temperature from the filtered and downsampled modulator data, which is a measure of the instantaneous pulse density. When more than two calibration points are available, the above procedure can still be employed; in this case, the values of m and b are the slope and intercept of the best fit (in the least-squares sense) line to the measured temperature points.

7.2.3 Results

The first test performed measured the DC temperature characteristics of the sensor over the biological temperature range of interest. To do this, the water bath temperature was fixed and bits were recorded over a long time period, approximately 30 minutes. The chip carrier temperature during this time was monitored by the thermistor probes. At the end of the 30 minute period, the average value of the bits D was computed and the steady temperature of the two thermistors was noted. The approximate chip temperature T_c was computed by extrapolation from the two thermistor measurements. This D, T_c pair was one data point. The water bath temperature was changed, and the chip temperature allowed to reach steady state; this process took approximately 1 hour (due to the air in the bag containing the chip) and was verified by the thermistors. This measurement process was repeated several times at different temperatures to produce a total of 6 data points.

The calibration scheme described above was then employed to calculate the calibration constants $\frac{G}{\log(n)}$ and D_{mid}. For the purposes of calculating the equivalent temperature sensor output voltage, it was assumed that n = 10 since this cannot be independently determined. The equivalent approximated temperature sensor output voltage v_o was calculated from the pulse density, the calibration constants, and the assumed value of n.

The results of this experiment are shown in figure 7.5. As expected, the nominal

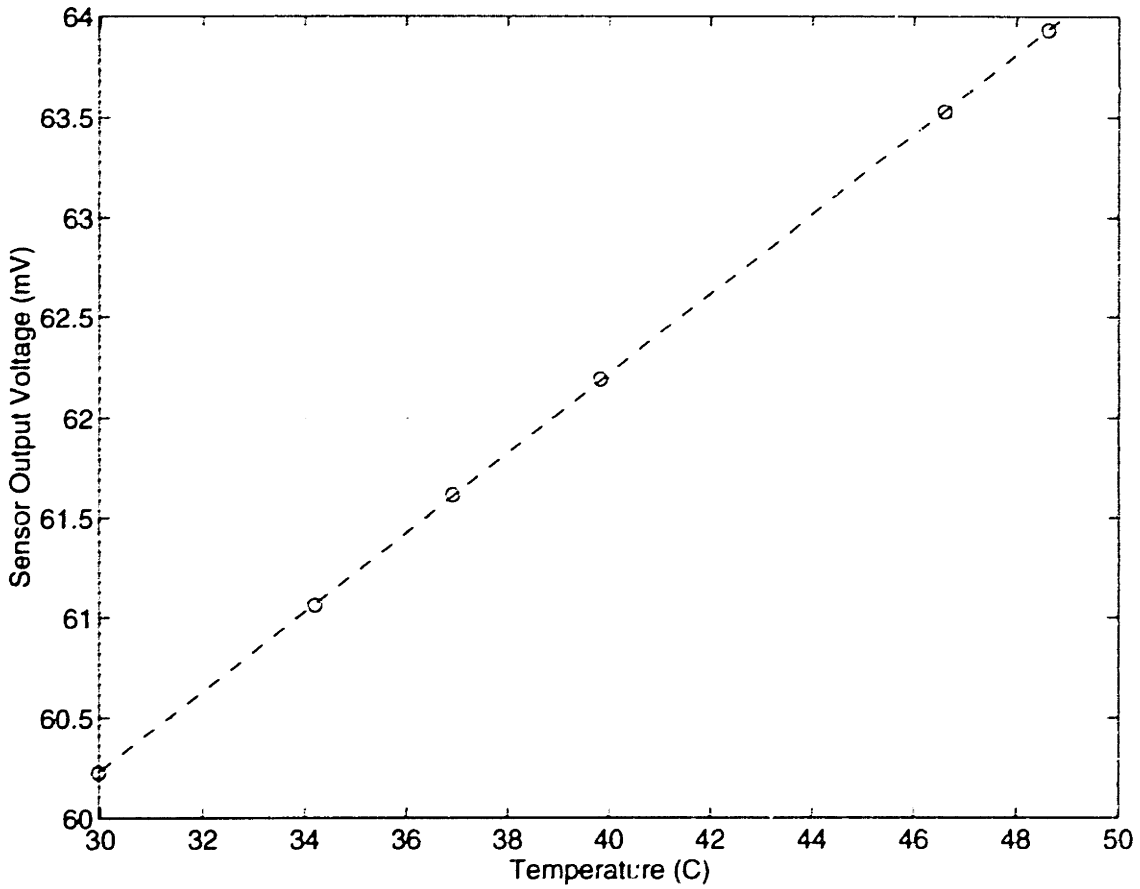


Figure 7.5: Measured temperature sensor output characteristic

sensor output voltage is approximately 62 mV ($\approx \frac{kT}{q} \ln 10$ at $T = 313.15$ K). Although calculations from the best-fit line show linearity of approximately 35 m°C over the 30-50 °C temperature range, this value is subject to error due to the ±3 m°C uncertainty in both thermistor measurements and, correspondingly, in calculation of the actual chip temperature from the thermal gradient. Another source of error is the water bath temperature controller, which controls the temperature to ±10 m°C at its interrogation point. For this reason no conclusive figure on the inherent linearity of the circuits can be determined from this experiment. The results nonetheless demonstrate that the general behavior of the system is as expected.

The deviation from the ideal line can also be viewed in terms of a "distortion" figure. To derive such a figure, the best-fit line parameters are first calculated. The

residuals from this line (error between predicted and actual values at each data point) are then fitted to a higher order polynomial, such that the output voltage can be expressed as:

$$v_o = a_0 + a_1 T + a_2 T^2 + \cdots + a_n T^n \tag{7.5}$$

where n is made large enough so that the resulting polynomial fit is "good" enough, i.e., that it matches the actual values to within some small error ϵ[3], and a_0 and a_1 are the parameters of the best-fit line. In essence, a constrained polynomial fit is performed. The total "distortion" in the characteristic can then be calculated as the mean-square sum of the ratio of the higher-order coefficients to the linear coefficient:

$$d = \sqrt{\sum_{j=2}^{n} \left(\frac{a_j}{a_1}\right)^2} \tag{7.6}$$

Applying this method to the data shown in figure 7.5 with $n = 4$ yields a "distortion" of 0.053%. It is important to note that while this figure is a measure of the nonlinearity in the output it is *not* the exact equivalent of the total harmonic distortion figure usually associated with similar calculations. This is because of the inability to apply a sinusoidally varying temperature to the system, which would be required to calculate the total harmonic distortion.

The second test was the most critical and was used to evaluate the fundamental thermal resolution of the system as designed. Typically, the performance limits of oversampled modulators are verified by applying a sine wave input to the system. The Fourier transform of the output bit stream shows a sharp peak that rises out of the system noise floor at the sine wave frequency. This not only verifies proper operation but also quantifies the noise floor of the system.

Unfortunately, for the thermal system presented here it is not possible to apply a sinusoidal input, as that would require a sinusoidally varying temperature, which is difficult if not impossible to produce at the required resolution. Instead, a thermal

[3]Clearly if n is equal to one less than the number of data points, $\epsilon = 0$ and the fit is perfect. In practice, the function usually converges fast enough so that $n = 4$ or 5 is usually sufficient.

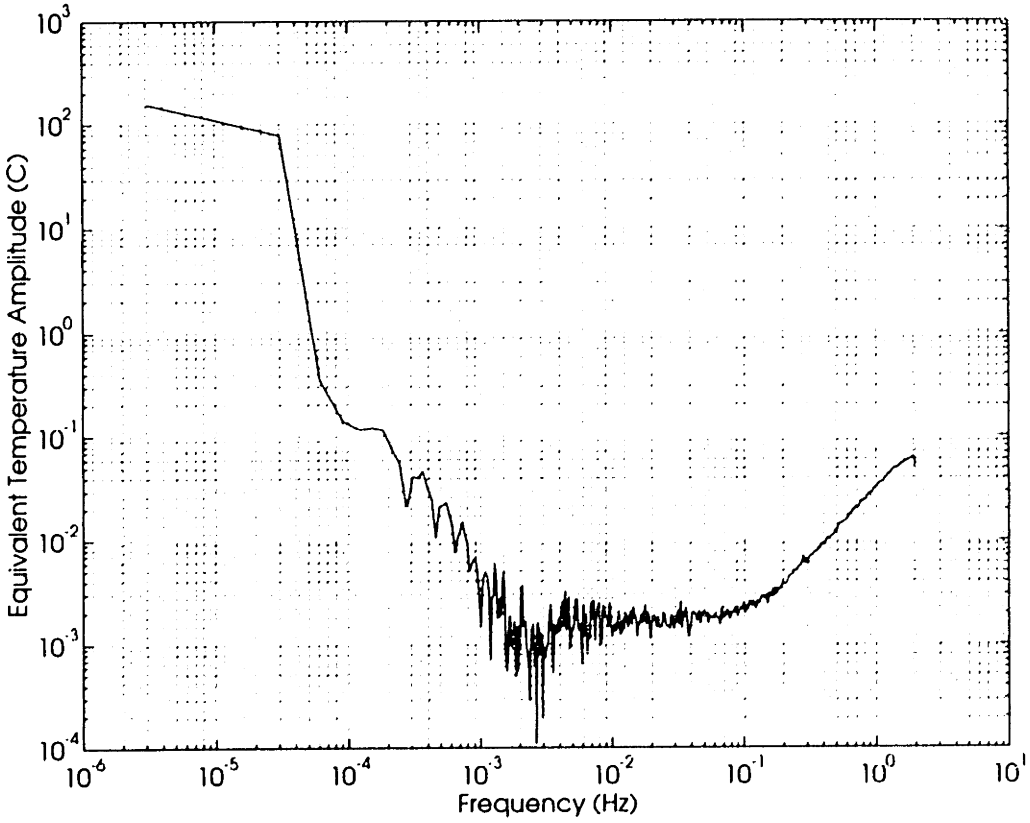


Figure 7.6: 90 min temperature step experiment output spectrum

"staircase" function was applied to the system. After allowing initial equilibration of the test setup, the temperature of the water bath was stepped periodically over the course of a day. An attempt was made to keep the magnitude of the applied steps relatively uniform, but this proved difficult due to the various thermal interactions between the components of the experimental setup and the testing environment. For this experiment, one of the most important parameters was the time interval at which the steps were applied, since the spectral purity of the staircase function depends on the periodicity in the applied steps. Two separate tests were performed using periods of 60 and 90 minutes between steps to insure that the period of the steps did not affect the system noise performance.

The results of these tests are shown in figures 7.6 (90 minute steps) and 7.7 (60

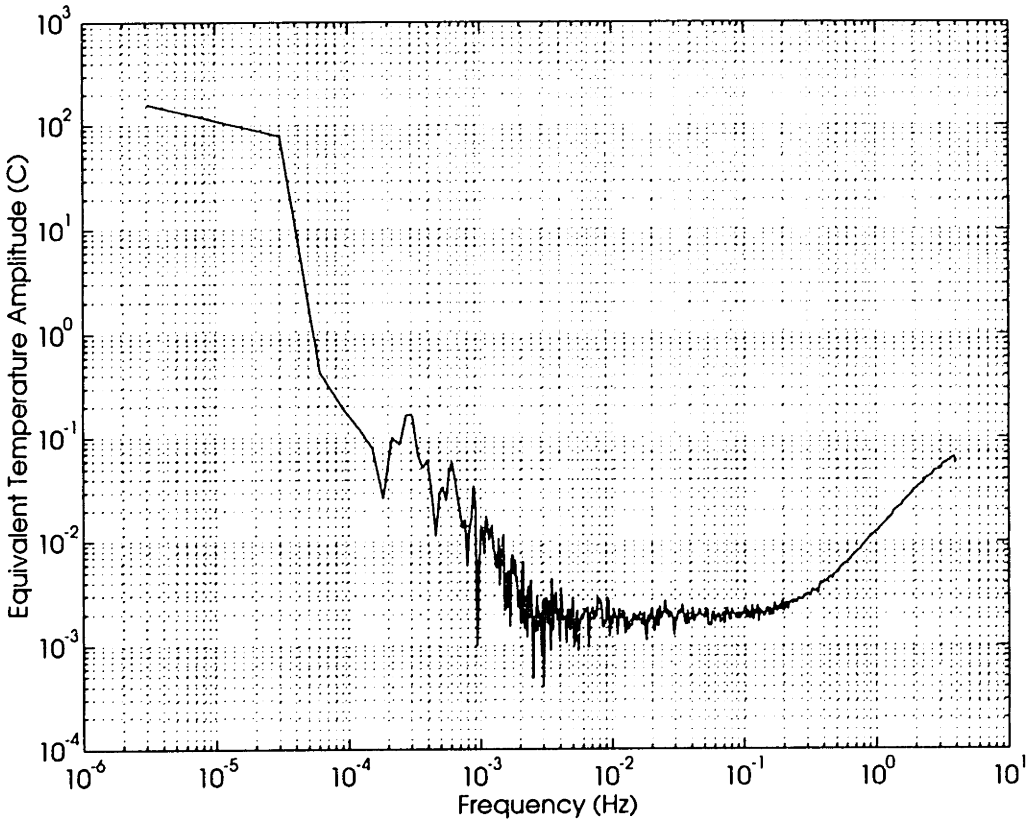


Figure 7.7: 60 min temperature step experiment output spectrum

minute steps). For reference, figure 7.8 shows the spectrum of a noiseless staircase with 60 minute steps; the transform for the 90 minute steps would have the same shape with a different periodicity of the humps. The Fourier transform of the measured output bits should show the spectrum of the input staircase function superimposed on the system noise spectrum. A comparison of figure 7.7 with figure 7.8, for example, clearly shows the ''humps'' that occur at the fundamental step frequency (.28 mHz for the 60 minute steps, .19 mHz for the 90 minute steps) and harmonics. At higher frequencies, the magnitude of the harmonics is very small and the transform is dominated by the noise floor of the system. From this test, the resolution of the test chips is approximately 2.5 m°C.

The rise in noise at the higher end of the output spectrum is due to the combined

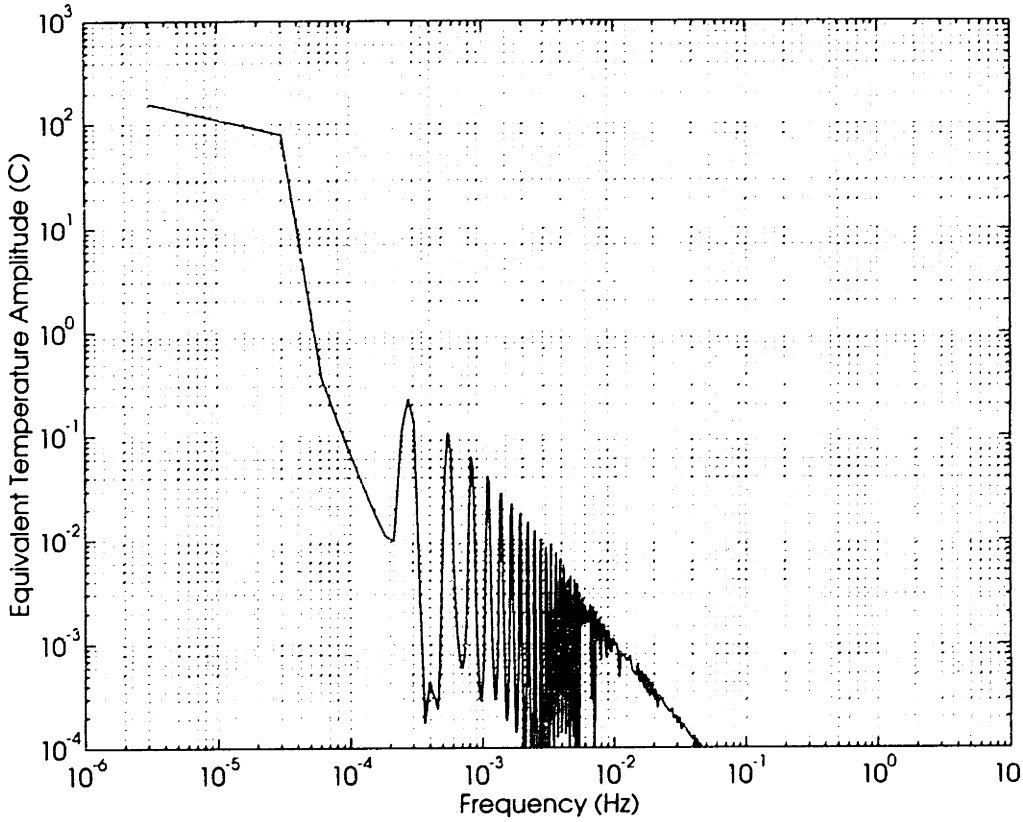


Figure 7.8: Spectrum of "noiseless" 60 minute step experiment

effect of the accumulate-and-dump function and the noise shaping of the modulator. When the data is initially processed by the modulator, quantization noise is injected into the system as described in chapter 4. When the signal (sampled at f_s) is resampled at a lower output frequency f_r, the high frequency quantization noise is folded into the spectrum of the resampled signal. If f_r is a submultiple of f_s, then the noise in the vicinity of f_r and its harmonics folds into the signal band; conversely, noise that is maximally distant from f_r and its harmonics is folded into the high end of the resulting spectrum. The spectrum of the accumulate-and-dump function is given by:

$$H(e^{j\omega}) = \frac{\operatorname{sinc}\left(\frac{\pi f}{f_r}\right)}{\operatorname{sinc}\left(\frac{\pi f}{f_s}\right)} \tag{7.7}$$

This function has nulls at f_r and all its harmonics. Consequently, the quantization noise

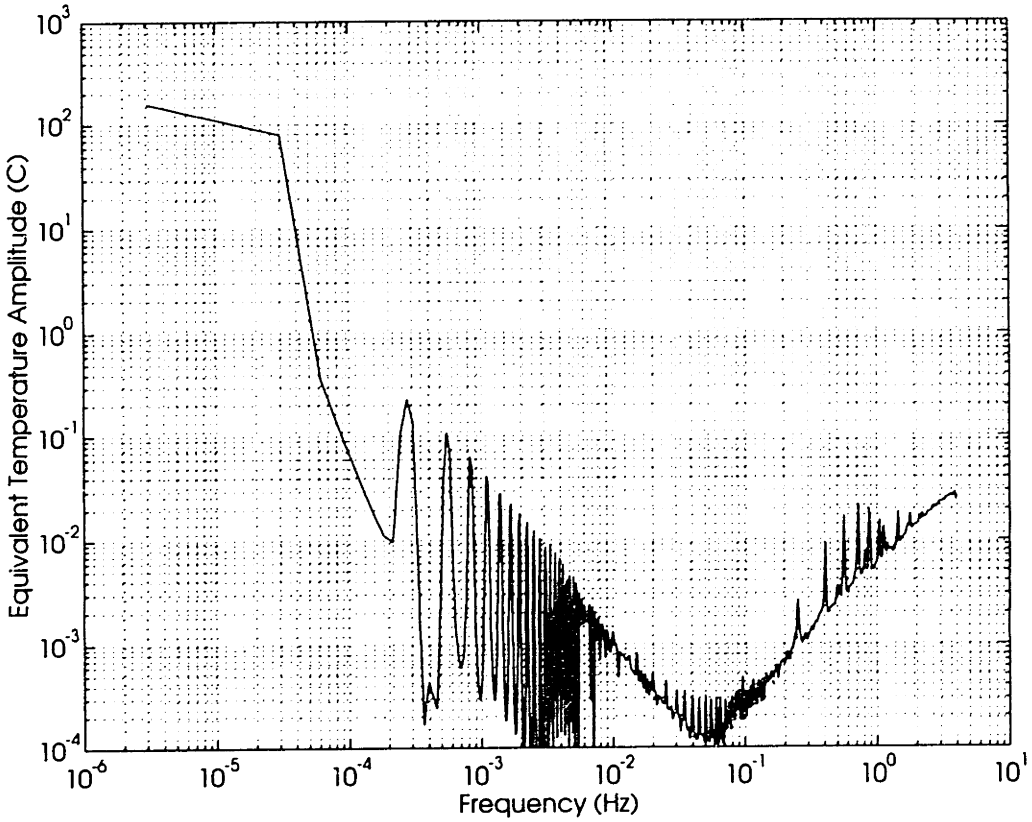


Figure 7.9: Output spectrum of ideal 60 minute step signal when processed by modulator and accumulate-and-dump function

is not folded into the low end of the signal band, but *is* folded into the high end of the output spectrum. Reference [61] presents a more complete discussion of the effects of the accumulation-and-dump function.

This effect is illustrated graphically in figure 7.9, which shows the spectrum of the signal that results when the noiseless temperature signal (60 minute steps) is processed through the modulator and the output bits are accumulated-and-dumped as they were in the actual measurements. The few noise spikes seen in the spectrum are due to artifact from the computer simulation and should be ignored. Comparison of figure 7.7 with figure 7.9 more clearly demonstrates the system noise floor at approximately 2.5 m°C.

7.3 Needle-based Temperature Measurements

Once the behavior of the sensors in isolation had been quantified, it was critical to verify these results in a needle-based system to insure that the system would perform as expected in its target application. To this end, a single-sensor system consisting of a temperature sensor and a digital controller chip was constructed, using the techniques discussed in chapter 6. Since this was a prototype system, the chip-to-chip bonds were potted, but the controller-to-cable bonds were not. This was done so that the needle interface could be studied at the point of entry to the needle. The constructed needle was then completely coated in a clear water-resistant epoxy; the vapor Teflon coating was not applied because its opaque characteristics would not allow optical inspection of the needle following water exposure.

The temperature testing setup employed for the needle measurements was similar to that used for the single-sensor packaged part measurements. To assist in generating a constant temperature environment, the needle was clamped to a slab of steel. Two thermistors were mounted to the opposite side of the steel slab in approximately the same location as the needle temperature sensor. The slab assembly was then placed in the water bath using a weighted bag as described in section 7.2.1. As before, the temperature was monitored using a TDP-200 system, and accumulated-and-dumped data from the needle was recorded on the personal computer driving the needle. The experiment took place over a time period of approximately 6 hours.

The system was evaluated using the same temperature-stepping technique employed for the packaged sensors. For this experiment, hourly temperature steps were used. The results from this temperature step experiment are shown in figure 7.10, which shows the output spectrum of the accumulated-and-dumped temperature signal. As before, peaks in the spectrum occur at the step frequency of .28 mHz and harmonics. At the higher end the effect of the accumulate-and-dump function can be seen.

Unlike the packaged part measurements, however, the system noise floor is not

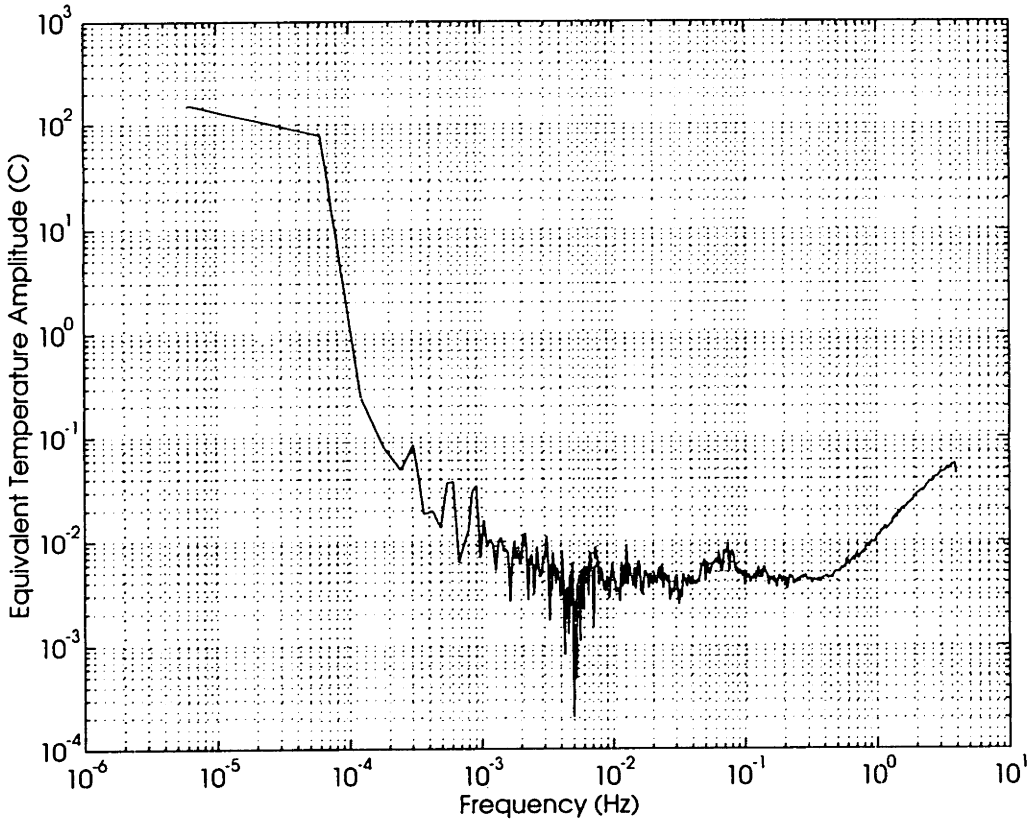


Figure 7.10: Measured output spectrum, single sensor needle, 60 min steps

entirely flat; over most of the middle frequencies, the noise floor is approximately 4 m°C, but a slight hump is present at approximately .07 Hz. This is due to water exposure that occurred early in the experiment that was not discovered until the experiment was completed. When the bag was immersed into the water bath, a small portion of the top seal was inadvertently submerged. Since this top seal is not water-resistant, water from the bath leaked into the bag; because of the orientation of the probe in the bag, it was calculated that the sensor chip was immersed for several hours. Eventually, the water level in the bag reached the exposed bond wires at the cable/chip interface and the needle failed, as was seen on the personal computer. The bag was removed from the bath, at which time it was discovered that the bag was full of water and the sensor had been immersed for a significant time period.

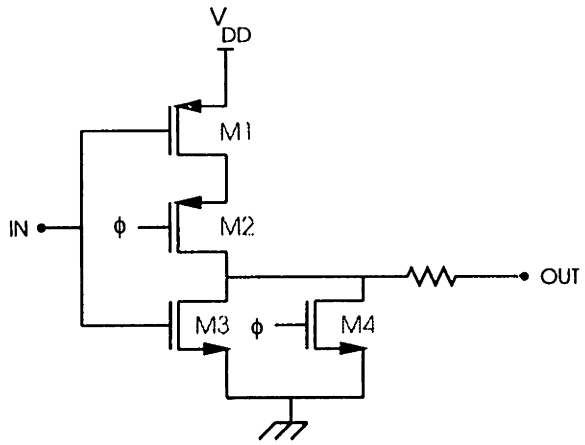


Figure 7.11: Output driver circuit

The hump that occurs in the noise floor is the temperature controller maintaining the bath temperature. The hump peaks at approximately 10 m°C, the limit of the temperature controller. Although the system noise floor of 4 m°C is still quite low, it is slightly higher than the measurements from the packaged parts and most likely reflects the effects of the improved coupling through water: Since the water was directly coupled to the sensor, the "low pass filtering" effect of the air in the bag was lost, and temperature fluctuations in the bath are more strongly reflected in the sensor measurements.

Although multiple-sensor needles were constructed, they were not functional due to a bus contention problem in the needle interface. The system is designed for one-at-a-time sensor measurement, i.e., at any given moment, only one sensor is actively communicating with the digital controller. In theory, the other sensors are entirely powered down, except for a very small logic circuit that monitors the sensor address lines from the controller to see if activation of the sensor is desired. In this way, the data communications line with the controller is controlled exclusively by the active sensor. Unfortunately, a bug in the interface caused the other sensors to load down this line even when in an inactive state. As a result, the active sensor could not drive the line to a valid logic high that could be properly detected by the digital controller.

This problem is illustrated in figures 7.11 and 7.12. The first figure shows the

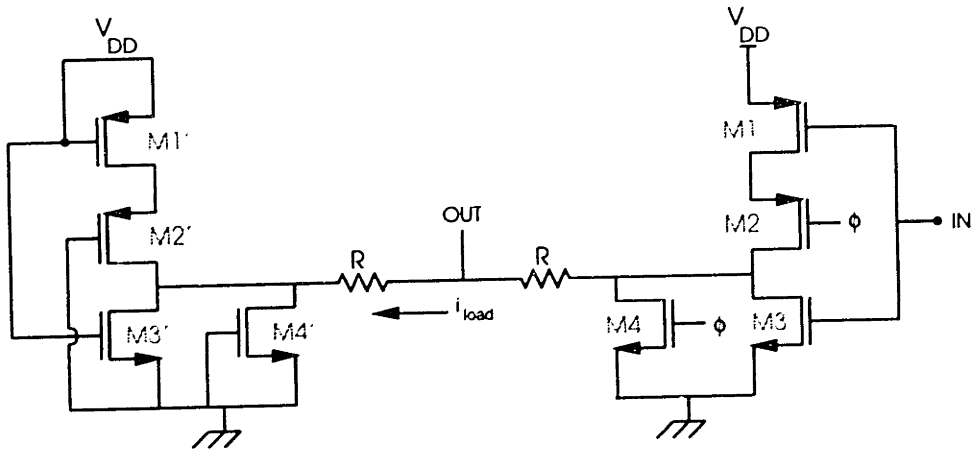


Figure 7.12: Output driver with "inactive" load circuit

output driver circuit present on each sensor chip. Devices M1 and M3 are both $\frac{100}{2}$; switch devices M2 and M4 are $\frac{5}{2}$. When the circuit is active, the output node is the inversion of the input (when the clock line is LOW) or zero (when the clock line is HIGH). This scheme is used so that the output is forced to a valid logic state at all times since, strictly speaking, there is no guarantee that the input will be in a valid logic state when the comparator of the modulator is being precharged. The two switch transistors (M2 and M4) disconnect the inverter and pull the line to ground (logic LOW) during this period. During the other half of the cycle when the input is valid M2 is switched on to connect the inverter and M4 is switched off to allow the inverter to assume its proper state. When the sensor is inactive, the supervisory circuits on the sensor are supposed to ground the clock lines, the input, and the power to the output circuit so that it does not load down the output node. In reality, due to a layout error the supervisory circuit is not correctly shutting down the signals--the clock line is grounded, but the power supply is not turned off. Since the supervisory circuit ties the input to the supply during shutdown (as it was designed to do), the input is being held at the supply also.

This results in the connection shown in figure 7.12, which shows the connections to the output line when one active and one "inactive" sensor are present. For completeness, the small polysilicon output resistors of the drivers are included. Instead

of the sensor being "transparent" to the active sensor, the circuit is active, and the output node is loaded by M3′. When the input is HIGH, this is not a problem, since the output is being driven LOW by the active sensor. When the input is LOW, however, M3′ competes with the active output driver device M1 that is attempting to pull the output HIGH. Because the magnitude of the threshold voltage for PMOS devices in the BioCMOS process is larger than that of the NMOS devices and the channel electron mobility is higher than the channel hole mobility, M3′ is the stronger device, and the output node does not reach a valid logic HIGH. The digital controller polling this node is not able to detect a logic HIGH and the output is always LOW. This behavior has been experimentally verified to the extent possible, since not all of the output driver nodes are available for testing. Measurements of the output current from the active sensor have indeed shown that during attempts to drive a logic HIGH the output current increases to several hundred microamps while the output node voltage does not exceed approximately 1 Volt. When the second ("inactive") sensor is removed from the output node, the driver performs as designed and the system works correctly. Multi-sensor needle temperature measurements were therefore not possible, although this limitation is strictly an interface problem and is not due to any inherent limitation in the sensing system or the active needle architecture.

7.4 Conclusions

An integrated circuit system for biomedical temperature measurement has been presented. The complete system specification has been discussed: At the highest level, the "active needle" system architecture is a generalized framework for multiparameter biomedical parameter characterization. Although the focus of this work has been a temperature probe, it has been shown that because of its all-digital nature the architecture itself is not limited solely to temperature measurement--any integrated "smart sensor" that can be manufactured with the appropriate interfacing circuitry can be used with the system as presented here. In addition, the temperature sensing system as designed

is suitable for use in a perfusion measurement system.

At a lower level, the complete circuit description of the system has been presented. First the low noise, high resolution temperature sensor that forms the core of the temperature measurement system was discussed. This includes a detailed analysis of the theory of operation, noise performance, and advantages over existing temperature measurement techniques. The switched-capacitor gain stage that amplifies the differential temperature signal was then presented, with emphasis placed on its correlated double sampling low frequency noise cancellation. The last of the major sensor chip systems, the analog modulator that performs the on-chip digitization of the amplified temperature signal, was studied in detail, at both the linear system and circuit levels. Finally, the digital controller chip that acts as the "brain" of the system was presented; this included a discussion of the control algorithm (finite-state machine) as well as the circuit implementation. Together, these circuits form a low noise, high resolution temperature smart sensor that has an experimentally demonstrated resolution of 3 m°C and a linearity of approximately .012% over the 30-50°C biomedical temperature range.

From a manufacturing standpoint, the special fabrication and packaging needs of the active needle system were discussed. The microelectronic fabrication process modifications required to realize the circuits required for the system were presented, along with experimental characterization of the resulting "BioCMOS" process. This was followed by a detailed examination of the needle assembly process, beginning with the thinning of the wafers to 200 μm from the standard 500 μm wafer thickness and ending with the final vapor Teflon coating of the entire probe. In short, the research presented in this document encompasses the entire engineering effort, from bare silicon wafers to ready-for-use needle probes.

The results from this research demonstrate the potential of microelectronics to further improve clinical medical technology and assist the medical community in their efforts to fight disease. The use of microelectronic technology helps reduce size, reduce cost, increase functionality, and improve measurements because of the

tremendous amount of signal processing that can be done at the measurement site. As clinical methodology improves, the need for better instrumentation to assist in the clinic grows. By meeting this need with microelectronic instrumentation and novel packaging techniques, the overall efficacy of medical science is improved.

7.5 Suggestions for future work

Although the experimental measurements from this system are excellent, many lessons have been learned over the course of this research. These lessons, when applied to future smart sensor probes, can improve the measurement or simplify many of the most difficult manufacturing steps. For this reason, several possible improvements to the active needle system as described here are presented.

Clearly the first issue that needs to be addressed is the output driver circuit that prevents multi-sensor needles from working properly. This problem, described above, requires a layout fix, or, more rigorously, a redesign of the output driver circuit to be a true high impedance load when the sensor is inactive. With such a circuit in place, the sensing system could be put in a "standby" state where each section of the sensor chip could be powered down while leaving the output driver powered (but in the high-impedance state). Multi-sensor operation would therefore no longer require switching of the main power supplies on each of the sensor chips, and loading of the output line would be avoided.

One of the most straightforward improvements that can be made to this system is a reduction in the power consumption. Initially, the power supply voltage was selected because of the estimated system interface requirements; when the system was completed, however, the need for a 6 Volt interface had been obviated by the parallel communications interface. With the elimination of this constraint, the power supply voltage could be reduced, perhaps to 3 Volts or less. Additionally, the circuits that are used on the sensor chips could be redesigned for lower power, most notably the

operational amplifier, since that circuit is repeated 7 times on each sensor chip.[4] This would not only extend the life of the power supply batteries, but, more importantly, it would significantly reduce the temperature artifact as discussed in Chapter 2.

A second improvement would be a redesign of the modulator from a fourth-order system to a third-order one. The original motivation for using a fourth-order modulator was concern over the correlation between the quantization noise and the input temperature signal, since the temperature signal is very low frequency. This resulted in a very conservative modulator design. As was shown in Chapter 4, the quantization noise shaping of the fourth-order system based on the linear model of the modulator is more than adequate to meet the requirements of the active needle system, and, in fact, a third-order system would also meet the noise shaping requirements. The advantages of the third-order system would be a reduction in total sensor chip power consumption (since one less op amp would be required) as well as a reduction in the total sensor chip length. This second benefit would allow for an increased density of sensors and thus better spatial temperature resolution.

Along that same line, the digital controller could be redesigned for lower area by eliminating some of the excess capabilities of the logic circuits. At the time of the original design, the controller was purposely designed with fully flexible logic circuits; all registers, for example, were designed with a preset and clear capability. However, most of the excess functionality is not used by the controller and could be eliminated without changing the behavior of the finite state machine. The excess area of the controller could instead be used to expand its capabilities to handle more complex instructions from the personal computer, or to implement a direct serial interface to the personal computer, now that high speed serial ports are available for personal computers. This would eliminate the need for the parallel interface that currently feeds the needle data to the personal computer.

From a manufacturing standpoint, the most significant improvement that could be

[4]There is one op amp in the sensor, one in the preamplifier, and five in the modulator.

made would be a reduction in the total number of bonding pads on both the sensor chips and the digital controller. The current design requires very fine wire bonding capabilities in order to place the bonds accurately on the bonding pads. Any reduction in the number of essential bond wires would allow expansion of the size and spacing of the remaining pads, which would simplify the bonding requirements. The reduction in the number of pads could be achieved in a number of ways, including on-chip generation of supply voltages, multiplexing of lines, and redesign of the "sensor enable" decoding circuits. The greater simplicity in manufacturing would lower the probe assembly time and increase overall probe yield by reducing chip loss in the bonding process.

The combined effect of these changes would be to further improve the measurement capabilities of the active needle system; in essence, these suggestions are the lessons learned to date that will form the basis of the next generation probes. This document has addressed the "first generation" of the active needle, and is certainly not to be interpreted as the end goal of this research path. Instead, it is just a beginning, a demonstration that the techniques presented here can be used to realize an *in-vivo* microelectronic instrument that has practical clinical application. Ultimately, it is hoped that the work presented here will lead to a continual evolution and improvement in this area, so that the full potential of the active needle system and the active needle techniques can be realized. By continually improving and adapting the system to the needs of the medical community the true benefit of this work, namely, saving lives and improving clinical efficacy, can be achieved.

Appendix A

BioCMOS Process Flow

Step #	Description

1. Stress Relief Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	100	Dry O_2
950	30	N_2

2. LPCVD Nitride Deposition
Deposition Temperature = 800 °C

3. Photolithography: Well Definition
Mask CPW

4. Plasma Etch Silicon Nitride

5. N-Well Ion Implant

Element	*Energy (keV)*	*Dose*
Phosphorus	180	2×10^{12}

6. Resist Ash

7. Photolithography: NPN Definition
Mask CBJ

8. NPN Collector Ion Implant

Element	*Energy (keV)*	*Dose*
Phosphorus	180	2×10^{13}

9. Resist Ash

10. N-Well Cover Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	30	Dry O_2
950	175	Wet O_2
950	30	Dry O_2
950	30	N_2

11. Nitride Wet Etch

12. P-Well Ion Implant

Element	*Energy (keV)*	*Dose*
Boron	30	1.5×10^{12}

13. Well Drive-In

Temp (°C)	*Time (min)*	*Gas*
1150	900	Dry O_2
1150	30	N_2

14. Well Oxide Wet Etch

15. Stress Relief Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	100	Dry O_2
950	30	N_2

16. LPCVD Nitride Deposition
Deposition Temperature = 800 *°C*

17. Photolithography: Active Area Definition
Mask CD

18. Plasma Etch Silicon Nitride

19. Photolithography: P-Field
Mask CPF

20. P-Field Ion Implant

Element	*Energy (keV)*	*Dose*
Boron	70	1×10^{13}

21. Resist Ash

22. Photolithography: N-Field
Mask CNF

23. N-Field Ion Implant

Element	*Energy (keV)*	*Dose*
Phosphorus	40	3×10^{12}

24. Resist Ash

25. Field Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	30	Dry O_2
950	175	Wet O_2
950	30	Dry O_2
950	30	N_2

26. Nitride Wet Etch

27. Stress Relief Oxide Etch

28. Dummy Gate Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	35	Dry O_2
950	30	N_2

29. Photolithography: NMOS Adjustment
Mask CNT

30. N-Punchthrough, Threshold Adjust Ion Implants

Element	*Energy (keV)*	*Dose*
Boron	100	6×10^{11}
Boron	30	1.5×10^{12}

31. Resist Ash

32. Photolithography: PMOS Adjustments
Mask CPT

33. P-Punchthrough, Threshold Adjust Ion Implants

Element	*Energy (keV)*	*Dose*
Phosphorus	180	6×10^{11}
BF_2	50	2.3×10^{12}

34. Resist Ash

35. Photolithography: NPN Base
Mask CBB

36. NPN Base Ion Implant

Element	*Energy (keV)*	*Dose*
Boron	180	3×10^{13}

37. Resist Ash

38. Dummy Gate Oxide Wet Etch

39. First Level Gate Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	35	Dry O_2
950	30	N_2

40. LPCVD Polysilicon Deposition

41. Phosphorus Deposition

Temp (°C)	*Time (min)*	*Gas*
925	60	Dry O_2, N_2
925	15	O_2
925	10	N_2

42. Frontside Resist Coat

43. Phosphorus Glass Wet Etch

44. Backside Polysilicon Plasma Etch

45. Resist Ash

46. Phosphorus Glass Wet Etch

47. Photolithography: Polysilicon
Mask CQ

48. Resist Flow

49. Polysilicon Plasma Etch
Modified for thin polysilicon

50. Resist Ash

51. First Level Gate Oxide Wet Etch

52. Second Level Gate Oxidation

Temp (°C)	*Time (min)*	*Gas*
950	22	Dry O_2
950	30	N_2

53. LPCVD Polysilicon Deposition

54. Phosphorus Deposition

Temp (°C)	*Time (min)*	*Gas*
925	60	Dry O_2, N_2
925	15	O_2
925	10	N_2

55. Phosphorus Glass Wet Etch

56. Photolithography: Polysilicon
Mask CP

57. Polysilicon Plasma Etch

58. Resist Ash

59. Photolithography: P+ Source/Drain
Mask CPP

60. P+ Source/Drain Ion Implant

Element	*Energy (keV)*	*Dose*
BF_2	30	7×10^{15}

61. Resist Ash

62. Pirhana Clean

63. Photolithography: N+ Source/Drain
Mask CNP

64. N+ Source/Drain Ion Implant

Element	*Energy (keV)*	*Dose*
Arsenic	90	7×10^{15}

65. Resist Ash

66. Pirhana Clean

67. Reoxidation

Temp (°C)	*Time (min)*	*Gas*
900	30	Dry O_2
900	30	N_2

68. Junction Drive

Temp (°C)	*Time (min)*	*Gas*
950	15	Dry O_2
950	15	N_2

69. Low Temperature Oxide Deposition

Thickness = 1000Å

70. BPSG Deposition

4% Phosphorus, 4% Boron

71. BPSG Flow

Temp (°C)	*Time (min)*	*Gas*
925	15	Dry O_2

72. Frontside Resist Coat

73. Backside Oxide Wet Etch

74. Backside Polysilicon Plasma Etch

75. Backside Oxide Wet Etch

76. Resist Ash

77. Photolithography: Contact

Mask CC

78. Contact Plasma Etch (Oxide)

79. Resist Ash

80. Pirhana Clean

81. HF Contact Hole Clearing Dip

4 min in 50:1 HF

82. Metal Deposition

Al, 1% Si, Thickness = 1μm

83. Photolithography: Metal

Mask CM

84. Metal Plasma Etch

85. Resist Ash

86. Contact Sinter

Temp (°C)	*Time (min)*	*Gas*
375	10	Forming Gas

87. PECVD Nitride Deposition

88. Photolithography: Bonding Pads

Mask CG

89. Resist Flow

90. Nitride Plasma Etch

91. Resist Ash

92. Nitride Stress Relief Sinter

Temp (°C)	*Time (min)*	*Gas*
375	10	Forming Gas

Appendix B

Derivation of Temperature Sensor Mismatch Error

Define the relevant voltages and currents as shown in figure B.1. Note that the gated-diode current source in figure 3.6 has been replaced by an ideal current source; it has been implicitly assumed that v_s changes by such a small amount that the corresponding changes in I_o are insignificant.

Thus, with zero differential output from the op amp, there is an (incorrect) current ratio n_Δ, where

$$n_\Delta = \frac{I_2}{I_1} = \frac{\frac{k'}{2}\left(\frac{W}{L}\right)_2(-v_s - V_{T2})^2}{\frac{k'}{2}\left(\frac{W}{L}\right)_1(-v_s - V_{T1})^2} \tag{B.1}$$

The currents I_1 and I_2 are the *initial* currents through each leg, with zero differential input voltage to the diff pair. Now assume that a differential voltage is applied to the diff pair so that the current ratio is restored to n, the desired value. Essentially there has been a change in current in each leg; define the changes as ΔI_1 and ΔI_2. The new (correct) current ratio is therefore:

$$n = \frac{I_1 + \Delta I_1}{I_2 + \Delta I_2} \tag{B.2}$$

Equations B.1 and B.2, combined with the additional constraint that

$$I_1 + I_2 = I_o \tag{B.3}$$

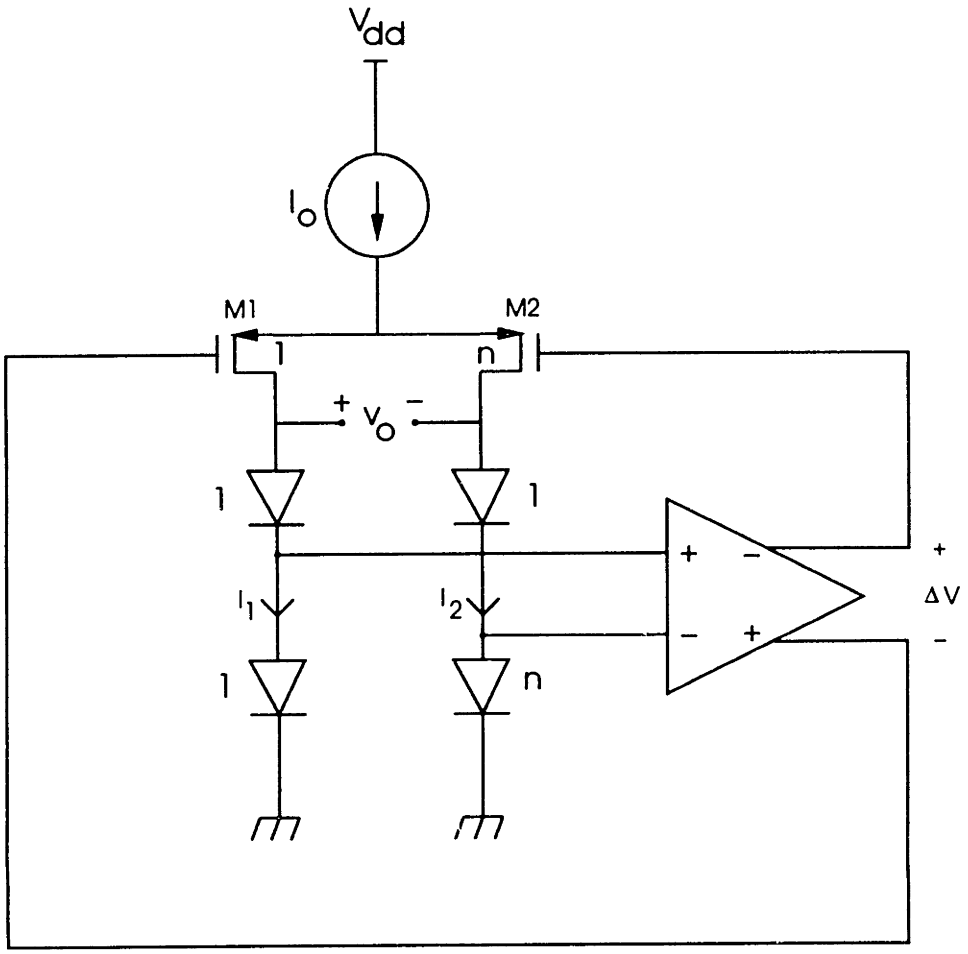


Figure B.1: Circuit for error analysis

results in an equation in terms of the known values n, n_Δ, and I_o:

$$\frac{n_\Delta}{n_\Delta + 1} I_o + \Delta I_1 = \frac{n}{n_\Delta + 1} I_o + n\Delta I_2 \tag{B.4}$$

Because the differential pair is purposely mismatched, the magnitude of the voltage required to restore the desired current ratio depends on whether the starting ratio is too low or too high. To see this, consider the differential pair shown in figure B.2. Figure B.3 shows the current ratio as a function of differential input voltage for this circuit with the voltages and currents defined as shown. Notice that the magnitude of the voltage that causes a +10% error is not the same as the voltage that causes a -10% error.

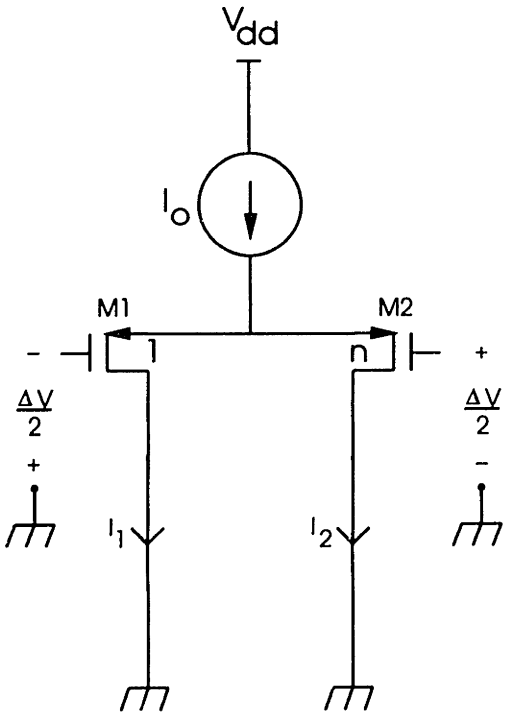


Figure B.2: Unbalanced differential pair

Therefore, in order to look at the worst-case (largest) ΔV for a given set of mismatch parameters, it is assumed that $n > 1$ and $n_\Delta > n$. Since it has been assumed that the ΔV required is small, the current changes can be approximated as:

$$\Delta I_1 = \frac{g_{m1}\Delta V}{2} \tag{B.5}$$

$$\Delta I_2 = -\frac{g_{m2}\Delta V}{2} \tag{B.6}$$

Substituting these equations into equation B.4 and solving for ΔV gives:

$$\Delta V = \frac{2\,[I_2 - nI_1]}{n \cdot g_{m1} + g_{m2}} \tag{B.7}$$

This equation gives ΔV as a function of only n, I_1, and I_2, since the g_ms are themselves functions of the currents.

The values of I_1 and I_2, however, are a function of the mismatches in $\left(\frac{W}{L}\right)$ and V_T. Thus it is necessary to relate the currents to the mismatch. Since I_1 and I_2 are defined as the currents in each leg with zero differential input, the only unknown is v_s; this can

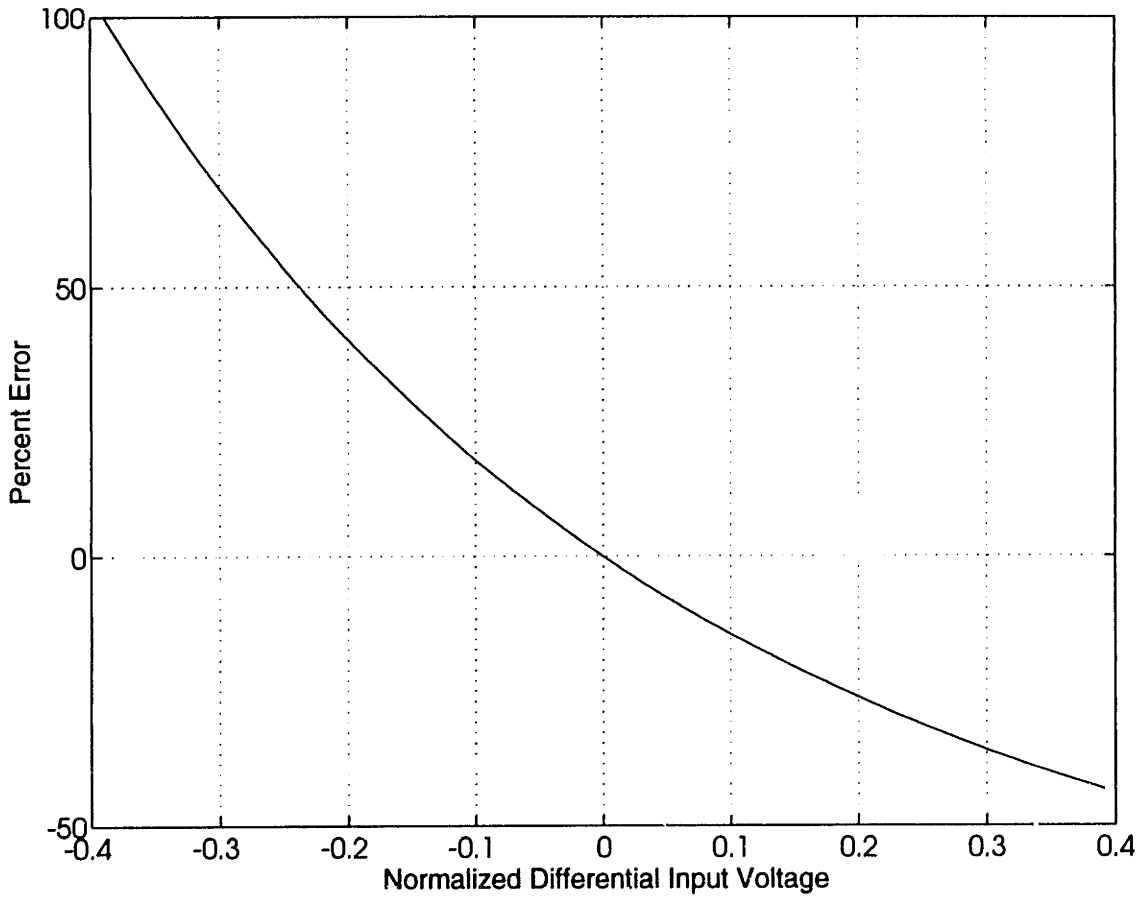


Figure B.3: Normalized current ratio vs. differential input voltage

be found using equation B.3 above:[1]

$$v_s = \frac{\sqrt{2k'I_o\left[\left(\frac{W}{L}\right)_1+\left(\frac{W}{L}\right)_2\right]-k'^2\left(\frac{W}{L}\right)_1\left(\frac{W}{L}\right)_2[V_{T1}-V_{T2}]^2}}{k'\left[\left(\frac{W}{L}\right)_1+\left(\frac{W}{L}\right)_2\right]} - \frac{\left[V_{T1}k'\left(\frac{W}{L}\right)_1+V_{T2}k'\left(\frac{W}{L}\right)_2\right]}{k'\left[\left(\frac{W}{L}\right)_1+\left(\frac{W}{L}\right)_2\right]} \tag{B.8}$$

As before, the following mismatch parameters can be defined:

$$V_T = \frac{V_{T1}+V_{T2}}{2} \tag{B.9}$$

$$\Delta V_T = V_{T1}-V_{T2} \tag{B.10}$$

[1]The common mode input is assumed to be zero; clearly including a common mode voltage only shifts the value of v_s but does not affect the currents I_1 and I_2.

$$\frac{\left(\frac{W}{L}\right)_2}{\left(\frac{W}{L}\right)_1} = n + c \tag{B.11}$$

Note that the mismatch in device geometry is treated as a ratio error rather than a differential error. This alternative representation is used because of the geometry ratio that is built in to the system; what is important is not the effect of the geometry mismatch on the individual currents but the effect on the ratio of currents. Therefore this mismatch is represented as a deviation c from the ideal ratio n. With these definitions, it is clear that:

$$V_{T1} = V_T + \frac{\Delta V_T}{2} \tag{B.12}$$

$$V_{T2} = V_T - \frac{\Delta V_T}{2} \tag{B.13}$$

$$\left(\frac{W}{L}\right)_2 = (n + c)\left(\frac{W}{L}\right)_1 \tag{B.14}$$

and v_s can be rewritten in terms of $\left(\frac{W}{L}\right)_1$, n, I_o, V_T and the mismatch parameters ΔV_T and c:

$$v_s = \sqrt{\frac{2I_o}{k'(n+c+1)\left(\frac{W}{L}\right)_1} - \frac{(n+c)\Delta V_T^2}{(n+c+1)^2}} + \frac{\Delta V_T}{2}\left(\frac{n+c-1}{n+c+1}\right) - V_T \tag{B.15}$$

Back substituting this expression for v_s into the MOSFET current equations gives:

$$I_1 = \frac{k'}{2}\left(\frac{W}{L}\right)_1 \cdot \left[\frac{\Delta V_t(n+c)}{n+c+1} + \sqrt{\frac{2I_o}{k'(n+c+1)\left(\frac{W}{L}\right)_1} - \frac{\Delta V_T^2(n+c)}{(n+c+1)^2}}\right]^2 \tag{B.16}$$

$$I_2 = \frac{k'}{2}(n+c)\left(\frac{W}{L}\right)_1 \cdot \left[\frac{\Delta V_T}{n+c+1} - \sqrt{\frac{2I_o}{k'(n+c+1)\left(\frac{W}{L}\right)_1} - \frac{\Delta V_T^2(n+c)}{(n+c+1)^2}}\right]^2 \tag{B.17}$$

These expressions for I_1 and I_2 can be used to find the values of g_{m1} and g_{m2} (this is

straightforward and will not be shown here).[2] Substituting everything into equation B.7 gives the desired result, namely, ΔV as a function of the mismatch parameters:

$$\Delta V = \frac{n+c}{n+c+1} \cdot \left\{ \frac{-\Delta V_T^2(n^2 - nc - c - 1) - 2\Delta V_T(n+1)\sqrt{\frac{2I_o(n+c+1)}{k'(\frac{W}{L})_1} - \Delta V_T^2(n+c)}}{-\Delta V_T(n+c)c + (2n+c)\sqrt{\frac{2I_o(n+c+1)}{k'(\frac{W}{L})_1} - \Delta V_T^2(n+c)}} \right.$$

$$\left. + \frac{\frac{2I_o c(n+c+1)}{k'(n+c)(\frac{W}{L})_1}}{-\Delta V_T(n+c)c + (2n+c)\sqrt{\frac{2I_o(n+c+1)}{k'(\frac{W}{L})_1} - \Delta V_T^2(n+c)}} \right\} \tag{B.18}$$

Although this equation represents the exact ΔV under the conditions stated, it certainly does not lend much insight into the behavior of this error. A more useful formula can be derived if it is assumed that $n \gg 1$ and $n \gg c$ and that the errors ΔV_T and c are small. In this case, the formula simplifies to:

$$\Delta V \approx -\Delta V_T + \frac{c}{n}\sqrt{\frac{I_o}{2nk'\left(\frac{W}{L}\right)_1}} \tag{B.19}$$

as used in the text.

[2]Care must be taken, however, to ensure that the expressions for g_{m1} and g_{m2} result in positive values, i.e., one must make sure that $\sqrt{x^2} = +x$, not $-x$.

Appendix C

Relationship Between Open and Closed Loop Settling Times

The relationship between the open loop dynamics of an operational amplifier and the settling time under capacitive feedback can be determined using the half circuit model shown in figure C.1, which depicts the general case in which there are arbitrary input and feedback impedances. The open loop dynamics of the operational amplifier are described by the transfer function $a(s)$. Assuming that the input resistance of the op amp is infinite, simple network analysis shows that the transfer function of the circuit is:

$$\frac{v_o}{v_i} = \frac{-Z_2}{\frac{(Z_1+Z_2)}{a(s)} + Z_1} \tag{C.1}$$

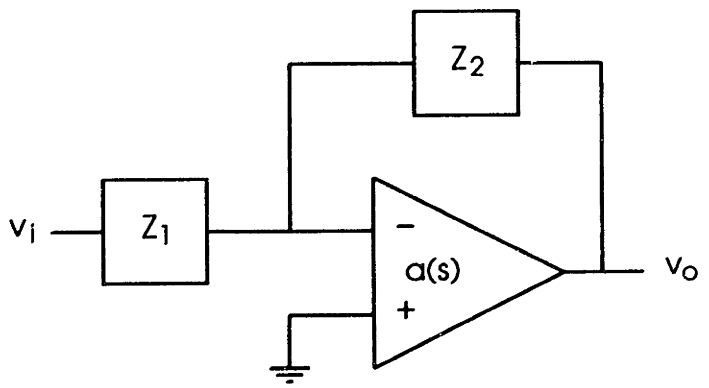


Figure C.1: Circuit for settling time constant calculation

If the feedback is capacitive, the general impedance can be replaced by the impedance of the capacitors:

$$Z_1 = \frac{1}{C_1 s} \tag{C.2}$$

$$Z_2 = \frac{1}{C_2 s} \tag{C.3}$$

If the operational amplifier dynamics can be described by a single dominant pole, then the transfer function $a(s)$ can be replaced by

$$a(s) = \frac{A}{\tau s + 1} \tag{C.4}$$

where A is the DC gain of the op amp and τ is the time constant of the dominant pole. Substituting equations C.2, C.3, and C.4 into C.1 and simplifying gives:

$$\frac{v_o}{v_i} = \frac{\frac{-C_1}{C_2 + \frac{C_1 + C_2}{A}}}{\frac{\tau(C_1+C_2)}{(A+1)C_2+C_1} s + 1} \tag{C.5}$$

The form of this equation is the same as equation C.4, where the numerator represents the DC closed loop gain of the connection, and the time constant of settling is:

$$\Upsilon = \frac{\tau(C_1 + C_2)}{(A+1)C_2 + C_1} \tag{C.6}$$

which gives the desired relationship between the time constant of the open loop dominant pole and the time constant of settling under feedback.

References

[1] Haim I. Bicher, Fred W. Hetzel, Taljit S. Sandhu, Stanley Frinak, Peter Vaupel, Michael D. O'Hara, and Terrence O'Brien. Effects of hyperthermia on normal and tumor microenvironment. *Radiology*, 137:523--530, November 1980.

[2] H. F. Bowman. Thermal dosimetry. In L. J. Anghileri and J. Robert, editors, *Hyperthermia in Cancer Treatment*, volume 2, chapter 9, pages 155--177. CRC Press, Inc., Boca Raton, FL, 1986.

[3] J. Overgaard. Influence of extracellular pH on the viability and morphology of tumor cells exposed to hyperthermia. *Journal of the National Cancer Institute*, 56(6):1243--1250, June 1976.

[4] Leo E. Gerweck. Modification of cell lethality at elevated temperatures: The pH effect. *Radiation Research*, 70:224--235, 1977.

[5] Leo E. Gerweck, Edward L. Gillette, and William C. Dewey. Killing of chinese hamster cells *in vitro* by heating under hypoxic or aerobic conditions. *European Journal of Cancer*, 10:691--693, 1974.

[6] Judith A. Power and John W. Harris. Response of extremely hypoxic cells to hyperthermia: Survival and oxygen enhancement ratios. *Radiology*, 123:767--770, June 1977.

[7] Leo E. Gerweck, Torbjoern G. Nygaard, and Margaret Burlett. Response of cells to hyperthermia under acute and chronic hypoxic conditions. *Cancer Research*, 39:966--972, March 1979.

[8] Mark W. Dewhirst, Edward J. Ozimek, Joseph Gross, and Thomas C. Cetas. Will hyperthermia conquer the elusive hypoxic cell? *Radiology*, 137:811--817, 1980.

[9] Eugene W. Gerner, William G. Connor, Max L. M. Boone, J. Daniel Doss, Eric G. Mayer, and Robert C. Miller. The potential of localized heating as an adjunct to radiation therapy. *Radiology*, 116:433--439, August 1975.

[10] J. M. Cosset, J. Dutreix, C. Haie, A. Gerbaulet, P. Janoray, and J. A. Dewar. Interstitial thermoradiotherapy: A technical and clinical study of 29 implantations performed at the Institut Gustave-Roussy. *International Journal of Hyperthermia*, 1:3--13, 1985.

[11] I. A. Brezovich and J. H. Young. Hyperthermia with implanted electrodes. *Medical Physics*, 8:79--84, 1981.

[12] I. A. Brezovich, W. J. Atkinson, and M. B. Lilly. Local hyperthermia with interstitial techniques. *Cancer Research*, 44 (Suppl.):4752s--4756s, 1984.

[13] C. T. Coughlin, E. B. Douple, J. W. Strohbehn, W. L. Eaton, B. S. Trembly, and T. Z. Wong. Interstitial hyperthermia in combination with brachytherapy. *Radiology*, 148:285--288, 1983.

[14] C. T. Coughlin, T. Z. Wong, J. W. Strohbehn, T. A. Colacchio, J. E. Sutton, R. Z. Belch, and E. B. Douple. Intraoperative interstitial microwave-induced hyperthermia and brachytherapy. *Int. Radiat. Oncol., Biol. Phys.*, (11):1673--1678, 1985.

[15] A. A. Puthawala, A. M. N. Seyd, K. M. A. Sheikh, S. Rafie, and C. S. McNamara. Interstitial hyperthermia for recurrent malignancies. *Endocurie, Hypertherm, Oncol.*, (1):125--131, 1985.

[16] P. R. Stauffer, T. C. Cetas, A. M. Fletcher, D. W. DeYoung, M. W. Dewhirst, J. R. Oleson, and R. B. Roemer. Observations on the use of ferromagnetic implants for inducing hyperthermia. *IEEE Transactions on Biomedical Engineering*, (BME-31):76--90, 1984.

[17] P. R. Stauffer, T. C. Cetas, and R. C. Jones. Magnetic induction heating of ferromagnetic implants for inducing localized hyperthermia in deep-seated tumors. *IEEE Transactions on Biomedical Engineering*, (BME-31):235--251, 1984.

[18] A. L. Burton, M. Hill, and A. E. Walker. The RF thermoseeds--a self-regulating implant for the production of brain lesions. *IEEE Transactions on Biomedical Engineering*, (BME-18):104--109, 1971.

[19] M. B. Lilly, I. A. Brezovich, and W. J. Atkinson. Hyperthermia induction with thermally self-regulated ferromagnetic implants. *Radiology*, (154):243--244, 1985.

[20] P. Fessenden, E. R. Lee, T. L. Anderson, J. W. Strohbehn, J. L. Meyer, T. V. Samulski, and J. B. Marmor. Experience with a multitransducer ultrasound system for localized hyperthermia of deep tissues. *IEEE Transactions on Biomedical Engineering*, (BME-31):126--135, 1984.

[21] P. P. Lele. rationale, technique and clinical results with scanned, focused ultrasound (SIMFU) system. In G. V. Kondraske and C. J. Robinson, editors, *Proceedings of the Eighth Annual Conference of the IEEE Engineering in Medicine and Biology Society*, page 1435, Dallas-Fort Worth, TX, 1986.

[22] G. K. Ogilvie, S. A. Goss, C. W. Badger, and E. C. Burdett. Performance of a multi-sectored ultrasound hyperthermia applicator and control system. results of animal studies *in vivo*. In *Proceedings of the 34th Annual Meeting of the Radiation Research Society*, Las Vegas, NV, 1986. abstract Bd-3.

[23] K. Hynynen, R. Roemer, D. Anhalt, C. Johnson, Z. X. Xu, W. Swindell, and T. Cetas. A scanned focused, multiple transducer ultrasound system for localized hyperthermia treatment. *International Journal of Hyperthermia*, 6(5):891--908, 1990.

[24] J. L. Hansen, B. A. Bornstein, G. K. Svensson, W. H. Newman, G. T. Martin, D. A. Sidney, and H. F. Bowman. A quantitative, integrated, clincal focused ultrasound system for deep hyperthermia. In L. J. Hayes, editor, *Advances in Biological Heat and Mass Transfer*. ASME, 1994. In press.

[25] H. F. Bowman, G. T. Martin, W. H. Newman, S. Kumar, C. Welch, B. Bornstein, and T. S. Herman. Human tumor perfusion measurements during hyperthermia therapy. In *Hyperthermic Oncology 1992, Volume 1 - Summary Papers Addendum, Proceedings of the Sixth International Conference on Hyperthermic Oncology*, volume 1, page A17, April 1992.

[26] H. F. Bowman. Estimation of tissue blood flow. In A. Shitzer and R. C. Eberhart, editors, *Heat Transfer in Medicine and Biology*, volume 1, chapter 9, pages 193--230. Plenum Publishing Corporation, 1984.

[27] T. A. Balasubramaniam and H. F. Bowman. Thermal conductivity and thermal diffusivity of biomaterials: A simultaneous measurement technique. *ASME Journal of Biomechanical Engineering*, 99:148--154, August 1977.

[28] H. F. Bowman, T. A. Balasubramaniam, and Monty Woods III. Determination of tissue perfusion from *in vivo* thermal conductivity measurements. In *Proceedings*

of the ASME Winter Annual Meeting. American Society of Mechanical Engineers, December 1977.

[29] J. W. Valvano, J. T. Allen, and H. F. Bowman. The simultaneous measurement of thermal conductivity, thermal diffusivity, and perfusion in small volumes of tissue. *Transactions of the ASME Journal of Biomechanical Engineering*, 106:192--197, August 1984.

[30] E. D. Macklen. *Thermistors*. Electrochemical Publications Limited, 1979.

[31] John S. Steinhart and Stanley R. Hart. Calibration curves for thermistors. *Deep-Sea Research*, 15:497--503, 1968.

[32] W. H. Newman, S. C. Summit, T. A. Balasubramaniam, and H. F. Bowman. In-vitro in-vivo measurement of low level tissue blood flow. In K. T. Yang, editor, *Collected Papers in Heat Transfer 1988*, pages 51--56. ASME Heat Transfer Division, 1988.

[33] R. N. Sengupta. A widely linear temperature to frequency converter using a thermistor in a pulse generator. *IEEE Transactions on Instrumentation and Measurement*, 37(1):62--65, March 1988.

[34] S. Natarajan. Widely linear temperature-to-frequency converters. *IEEE Transactions on Instrumentation and Measurement*, IM-24(3):235--239, September 1975.

[35] Motoaki Ikeuchi, Tomozo Furukawa, and Goro Matsumoto. A linear temperature to frequency converter. *IEEE Transactions on Instrumentation and Measurement*, IM-24(3):233--235, September 1975.

[36] Anwar A. Khan and R. Sen Gupta. A linear temperature-to-frequency converter using a thermistor. *IEEE Transactions on Instrumentation and Measurement*, IM-30(4):296--299, December 1981.

[37] Anwar A. Khan and R. Sen Gupta. A linear thermistor-based temperature-to-frequency converter using a delay network. *IEEE Transactions on Instrumentation and Measurement*, IM-34(1):85--86, March 1985.

[38] Gerard C. M. Meijer. Thermal sensors based on transistors. *Sensors and Actuators*, 10:103--125, 1986.

[39] Gerard C. M. Meijer and Kees Vingerling. Measurement of the temperature dependence of the $I_C(v_{be})$ characteristics of integrated bipolar transistors. *IEEE Journal of Solid State Circuits*, SC-15(2):237--240, April 1980.

[40] Gerard C. M. Meijer. A low-power easy-to-calibrate temperature transducer. *IEEE Journal of Solid State Circuits*, SC-17(3):609--613, June 1982.

[41] Gerard C. M. Meijer. An IC temperature transducer with an intrinsic reference. *IEEE Journal of Solid State Circuits*, SC-15(3):370--373, June 1980.

[42] G. C. M. Meijer, R. Van Gelder, V. Nooder, J. Van Drecht, and H. Kerkvleit. A three-terminal integrated temperature transducer with microcomputer interface. *Sensors and Actuators*, pages 195--206, 1989.

[43] David Van Maaren, Jan Klijn, and Gerard C. M. Meijer. An integrated micropower low-voltage temperature controlled oscillator. *IEEE Journal of Solid State Circuits*, SC-17(6):1197--1201, December 1982.

[44] Philip W. Barth and James B. Angell. Thin linear thermometer arrays for use in localized cancer hyperthermia. *IEEE Transactions on Electron Devices*, ED-29(1):144--150, January 1982.

[45] Phillip W. Barth, Sharon Lea Bernard, and James B. Angell. Flexible circuit and sensor arrays fabricated by monolithic silicon technology. *IEEE Transactions on Electron Devices*, ED-32(7):1202--1205, July 1985.

[46] H. Schaffer and G. Koeder. A sensitive all-silicon temperature transducer. *Sensors and Actuators*, 4:661--667, 1983.

[47] H. S. Carslaw and J. C. Jaeger. *Conduction of Heat in Solids*. Oxford University Press, London, first edition, 1947.

[48] Gregory T. Martin, personal communication.

[49] H. H. Pennes. Analysis of tissue and arterial blood temperatures in the resting human forearm. *Journal of Applied Physiology*, 1(2):93--122, 1948.

[50] Paul R. Gray and Robert G. Meyer, editors. *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, New York, NY, 1984.

[51] James K. Roberge, Personal communication.

[52] Blaine Jeffrey Gross, Personal communication.

[53] Aldert van der Ziel. *Noise in Solid State Devices and Circuits*. John Wiley & Sons, New York, 1986.

[54] Monica H. Choi. Ultra low power operational amplifier design. Master's thesis, Massachusetts Institute of Technology, May 1993.

[55] Eric Vittoz and Jean Fellrath. CMOS analog integrated circuits based on weak inversion operation. *IEEE Journal of Solid State Circuits*, SC-12(3):224--231, June 1977.

[56] Shujaat Nadeem. Design and implementation of fourth order modulator for 16-bit oversampled a/d converter. Master's thesis, Massachusetts Institute of Technology, May 1989.

[57] Marvin H. White, Donld R. Lampe, Franklyn C. Blaha, and Ingham A. Mack. Characterization of surface channel CCD image arrays at low light levels. *IEEE Journal of Solid State Circuits*, SC-9(1):1--13, February 1974.

[58] James C. Candy and Gabor C. Temes, editors. *Oversampling Delta-Sigma Data Converters: Theory, Design, and Simulation.* IEEE Press, Piscataway, NJ, 1992.

[59] R. M. Gray. Oversampled sigma-delta modulation. *IEEE Transactions on Communications*, COM-35:481--489, May 1987.

[60] R. M. Gray. Quantization noise spectra. *IEEE Transactions on Information Theory*, IT-36:1220--1244, November 1990.

[61] James C. Candy and Gabor C. Temes. Oversampling methods for a/d and d/a conversion. In James C. Candy and Gabor C. Temes, editors, *Oversampling Delta-Sigma Data Converters: Theory, Design, and Simulation*, pages 1--29. IEEE Press, New York, NY, 1992.

[62] Max W. Hauser. Principles of oversampling a/d conversion. *Journal of the Audio Engineering Society*, 39(1/2):3--26, January/February 1991.

[63] Bernhard E. Boser and Bruce A. Wooley. The design of sigma-delta modulation analog-to-digital converters. *IEEE Journal of Solid State Circuits*, SC-23:1298--1308, December 1988.

[64] K. C.-H. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini. A higher order topology for interpolative modulators for oversampling a/d converters. *IEEE Transactions on Circuits and Systems*, CAS-37:309--318, March 1990.

[65] *BDSIM, a flexible simulator for block diagrams,* written by Kenneth Szajda and Jennifer Lloyd, 1992.

[66] R. W. Adams, P. F. Ferguson Jr., A. Ganesan, S. Vinclette, A Volpe, and R. Libert. Theory and practical implementation of a fifth-order sigma-delta a/d converter. *Journal of the Audio Engineering Society*, 39(7/8):515--528, July/August 1991.

[67] Max W. Hauser and Robert W. Brodersen. Circuit and technology considerations for MOS delta-sigma a/d converters. *IEEE Proceedings of the International Symposium on Circuits and Systems '86*, pages 1310--1315, May 1986.

[68] Samuel C. Lee. *Digital Circuits and Logic Design.* Prentice Hall, Englewood Cliffs, NJ, 1976.

[69] Craig L. Keast. A CCD/CMOS Process for Integrated Image Acquisition and Early Vision Signal Processing. Master's thesis, Massachusetts Institute of Technology, September 1989.

[70] W. R. Thurber, R. L. Mattis, Y. M. Liu, and J. J. Filliben. The relationship between resistivity and dopant density for phosphorus- and boron-doped silicon. NBS Special Publication 400-64, National Bureau of Standards, May 1981.

[71] *Suprem-III User's Manual.*

[72] K. D. Wise and R. H. Weissman. Thin films of glass and their application to biomedical sensors. *Medical and Biological Engineering*, 9:339--350, 1971.

[73] David J. Edell. Coatings for protection of integrated circuits technical proposal. Grant application to the National Institues of Health, National Institute of Neurological and Communicative Diseases and Stroke Neural Prosthesis Program, 1987.

[74] Werner Kern and Richard S. Rosler. Advances in deposition processes for passivation films. *Journal of Vacuum Science and Technology*, 14(5):1082--1099, September/October 1977.

[75] Robert B. Comizzoli. Surface and bulk electrical conduction in low deposition temperature Si_3N_4 and Al_2O_3 films for silicon devices. *RCA Review*, 37(4):473--481, December 1976.

[76] K. P. Huber and G. Herzberg. *Molecular Spectra and Molecular Structure Constants of Diatomic Molecules*. Van Nostrand, New York, NY, 1979.

[77] P. Gray. *Q. Rev. (London)*, Volume 17, p. 441, 1963.

[78] A. K. Sinha. *Electrochem. Soc. Ext. Abstract*, (244), 629 (1976).

[79] David J. Edell, Stephen K. Burns, Carl V. Thompson, Joyce Palmer, and Lloyd D. Clark. Coatings for protection of integrated circuits. Eigth Quarterly Progress Report, July-September 1989, Contract NIH-NINCDS-N01-NS-7-2399.

[80] David J. Edell, Stephen K. Burns, Carl V. Thompson, Lisa P. Devaney, and Lloyd D. Clark. Coatings for protection of integrated circuits. Ninth Quarterly Progress Report, October-December 1989, Contract NIH-NINCDS-N01-NS-7-2399.

[81] Lyn Bowman and James D. Meindl. The packaging of implantable integrated sensors. *IEEE Transactions on Biomedical Engineering*, BME-33(2):248--255, February 1986.

[82] Wen H. Ko and Thomas M. Spear. Packaging materials and techniques for implantable instruments. *Engineering in Medicine and Biology Magazine*, 2:24--38, March 1983.

[83] Eduardo del Pino. Packaging considerations for an invasive probe with an array of thermal sensors. Bachelor's thesis, Massachusetts Institute of Technology, May 1991.

[84] Jennifer A. Lloyd, personal communication.

About the Author

Kenneth S. Szajda was born the son of a poor African-American sharecropper in Mobile, Alabama, on December 18, 1964. Ok, maybe not, but it seems that every biography you see on TV begins like that. Actually, I was born in Bronxville, NY, very early in the morning (12:24am) on that day. I am the third child of Richard and Cloe Szajda. They weren't poor sharecroppers, just good old middle class Americans. Not very exciting, but true.

After spending the first four years of my life in Riverdale, NY, eating free Stella D'Oro cookies (the woman in the house next to ours owned the company), the family packed up and moved to Rockaway, NJ, where I spent the next 14 years. I am a graduate of Katherine Dwyer Grammar School (1977), Copeland Middle School (1979), and Morris Hills High School (1983). They are all fine schools, and sometimes when I am in Rockaway I wonder what it would be like to return to the hallowed halls of these schools, where I was shaped from a formless blob of clay into a true academic who actually chose of his own free will to attend a difficult school like MIT.

My roots in Rockaway were torn up on September 2, 1983, when I boarded a People Express flight to Boston to begin my first semester at MIT. My education here has been tremendous. When I started school, the first clunky laser printers were tens of thousands of dollars, the hottest computer of the day was the VAX 11/780, and people still used typewriters to type papers. Now, laser printers are cheaper than a one-night hospital stay, the Alpha processor is pushing 300 MHz, and you can do all sorts of fancy word processing tricks, such as including a scanned picture of yourself in your own thesis. Since 1983, I have received three degrees from here (Bachelors, 1987; Masters, 1989; Ph.D., as soon as this document is turned in), and have lived in the Boston area for over 11 years. It is startling to think that in a very short while I will have been in Massachusetts longer than I was in Rockaway.

On the serious side, the focus of my work while at MIT has been electronics for medical applications, specifically instrumentation for tissue characterization for hyperthermic treatment of tumors. Both my Bachelor's and Master's degree theses focused on the design and construction of a board-level system for temperature and perfusion measurement. My doctoral research is, of course, presented in this document. I am a member of Eta Kappa Nu and Tau Beta Pi, and a student member of the Institute for Electrical and Electronics Engineers and the North American Hyperthermia Society. Anyone who is really that interested in finding out about my background can request a copy of my *curriculum vitae* by sending e-mail to kens@hotstuff.mit.edu.

So that's where I've been and where I am. Change is what life is all about, however, and it's time to look ahead to the future. Bon voyage.