Investigations of Visual Transduction and Motion Processing

Tobias Delbrück

Ph.D. Thesis
Computation and Neural Systems Program
Caltech, 1993

Investigations of Analog VLSI Visual Transduction and Motion Processing

Thesis by

Tobias Delbrück

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1993

(defended November 4, 1992)

ACKNOWLEDGEMENTS

I want to thank **Max Delbrück**, **Carver Mead**, and **Misha Mahowald** for showing me what it means to be a scientist. I also want to thank **Manny Delbrück**, **Shih-Chii Liu**, and **Janet Manning** for encouragement and emotional support.

I thank **Christof Koch**, **Pietro Perona**, **Rod Goodman**, **Scott Fraser**, and **Carver Mead** for serving as my thesis committee. I also wish to thank **Christof Koch** for the early years of patronage as a CNS student, and **David Van Essen** for the middle years. I thank **Jim Fox** and **David Van Essen** for the opportunity to participate in their monkey cortical physiology experiments.

I also want to thank, from Mead's lab, **Buster Boahen**, **Rahul Sarpeshkar**, **Ron Benson**, and **Lloyd Watts**, who made the last year OK, and **Mass Sivilotti**, **Steve Deweerth**, **John Lazzaro**, **Mary Ann Maher**, and **Dave Gillespie** who earlier made it an enjoyable, interesting, and productive place to be. **Mass Sivilotti**, **John Lazzaro**, and **Dave Gillespie** wrote most of the CAD tools that we used to design the chips. I thank **Dan Naar** and **Dick Lyon** of Apple for their help in setting up the Macintosh data collection environment. I thank the people at Tanner Research, especially **John Tanner**, for the use of their layout tool L-Edit. I thank **Chuck Neugebauer** for many interesting discussions about silicon technology. I thank all of the above people for providing sounding boards for ideas, and for the enjoyable group meetings every week. I also acknowledge the extensive administrative support of **Helen Derevan**, **Jim Campbell**, **Calvin Jackson**, and **Donna Fox**. In each chapter, I have included a separate acknowledgment for the specific work described in that chapter.

Finally, I acknowledge the financial support of the Office of Naval Research, the System Development Foundation, Hewlett Packard, and the National Institute of Health. Chip fabrication was provided by the DARPA foundry service MOSIS, without which none of this work would have been possible.

ABSTRACT

This thesis is a detailed description of a neuromorphic visual-motion processing chip and its component parts. The chip is the first two-dimensional silicon retina with a full set of direction-selective, velocity-tuned pixels. The architecture for the chip is based on the biological correlation-type motion detector, with the addition of a novel spatiotemporal aggregation. All the processing on the chip is analog and occurs in parallel. Novel, on-chip, continuous-time, adaptive, logarithmic photoreceptor circuits are used to couple temporal image signals into the motion processing network. These continuous-time photoreceptor circuits have also been used in a wide variety of other vision chips. The photoreceptor circuits center their operating point around the history of the illumination, simultaneously achieving high sensitivity and wide dynamic range. The receptor circuits are characterized and analyzed carefully for their temporal bandwidth and detection performance. Noise properties are analyzed, resulting in a simple and intuitive understanding of the limiting parameters. Novel adaptive elements are described that are insensitive to light-generated minority carriers. Novel measurements are presented of the spectral response properties of phototransducers that can be built in ordinary CMOS or BiCMOS processes. A novel nonlinear circuit that measures similarity and dissimilarity of signals is described and characterized. These bump circuits are used on the motion chip to extract the motion energy signal, and have also been used in other chips in numerous ways.

CONTENTS

CHAPTER

1

INTRODUCTION

The world's fastest supercomputers still cannot perform the computation of a single housefly brain. If they could, we would see them driving cars for us. Look at these statistical comparisons: 1 kW to run a supercomputer that weighs 10^5 g. 1 μW to run a fly brain that weighs less than 1 mg [1]. Of course, a fly cannot predict the weather, cannot solve quantum chromodynamics, cannot play chess. A fly is *not* a general-purpose computer, and the bit-error rate is far from zero. A fly *is* a *special-purpose* computational device, designed to deal in *real time* with *imprecise* sensory input. These simple observations have convinced me that it is worthwhile to try building what Carver Mead calls neuromorphic electronic systems [2]. I believe that within my lifetime, we will be able to build artificial nervous systems that approach the computational power and efficiency of flying insects, and that we will achieve this goal by following the paradigms of the only truly functional systems we know about.

Over the past few years, we have made tremendous progress in the construction and demonstration of neuromorphic analog VLSI visual processing systems. The constraint of the analog VLSI medium—particularly the problems of finite wire and of susceptibility to circuit offsets—have driven the development of new ideas about representation and architecture. Our studies have revealed a rich class of new circuit primitives. I like to think of these circuit primitives as a vocabulary, and of the process of circuit and system synthesis as a form of speech.

This thesis describes a novel silicon system that processes visual motion information, in a simplified form of the early parts of the visual system of flying insects. The component parts of the motion chip are like words in our expanding silicon vocabulary. The thesis also discusses, in great detail, the component parts of the chip. These details will be interesting to people who want to build working systems of their own.

A GUIDED TOUR

I will give here a brief preview of each part, stating novel results and summarizing content. The thesis is divided into two parts. Part 1 is about the individual components used in the system described in part 2. Each chapter is essentially self-contained and has its own introduction.

The first part starts with Chapter 2. This chapter is about visual transduction with analog silicon photoreceptors, and describes the device physics and circuit design of isolated, continuous-time, adaptive, logarithmic photoreceptors. The chapter gives detailed explanation, theory, and comparative measurement of various photoreceptor circuits, and will be useful to designers who need to generate well-conditioned information about the temporal structure of visual images. The receptor circuits incorporate several novel technical advances that substantially widen their dynamic range over previous devices. The analysis and measurement of the photoreceptor circuits is much more complete than in previous reports. I am particularly fond of the analysis and test of a novel theory of noise in logarithmic, subthreshold photoreceptors.

Several appendices to Chapter 2 further explore the device physics of visual transduction and electronic noise. Appendix 2a is a novel measurement of the spectral (i.e., color) responses of phototransducer devices that can be built in an ordinary CMOS fabrication process. The data in this appendix is a valuable reference for anyone who is interested in the interaction of light with silicon, and especially for people who use standard processes. Appendix 2b is a short description of a direct measurement of minority-carrier diffusion length, and of the effectiveness of guard structures designed to prevent light-generated minority carriers from interacting with other circuits. Questions about guard bars and minority-carrier diffusion length come up continually in the context of circuit design. Appendix 2c is about transistor noise from a phenomenological viewpoint. I show measurements of the spectrum of flicker noise, and investigate the dependence of flicker noise on bias current. I give a very brief, qualitative treatment of the theory of transistor noise, because of the already-vast literature on this subject. Electronic noise is often mentioned in the

abstract, is generally not understood very well, and is almost never measured—at least so far in the analog VLSI business.

Chapter 3 is about circuits that measure similarity and dissimilarity of signals—bump circuits. The similarity–dissimilarity computation is a fundamental nonlinear operation, and we have found it to be useful in a wide range of systems. In this thesis, it is used as the fundamental nonlinearity in the motion computation chip.

The second part of the thesis consists of Chapter 4, which is about a novel motion computing chip that uses the components described in Chapters 2 and 3. The motion architecture is inspired by known biological motion processing. The example system is the first functional implementation of a two-dimensional silicon retina with analog, correlation-type, motion detectors. It is also a beautiful example of how constraints imposed by the medium lead to new architectural ideas. The motion circuit uses a novel aggregation to combine image information over an extended spatiotemporal range. This aggregating motion computation leads to interesting new properties for the motion detector. In this chapter, I analyze the properties of the motion architecture, compare the analysis with measurements of the chip response, and discuss the properties of the chip in relation to other analog motion chips, and, briefly, in relation to biological motion computation.

Part 3 concludes this thesis with a very short Chapter 5, summarizing the general lessons from this work. Two postscripts list some miscellaneous technical details and list my publications while at Caltech.

I have received help from many people with this work, and I want to involve you in the ideas. From now on I will often use “we” either to involve you in the work or to indicate that the work was done with the interaction of other people, either physically or spiritually.

REFERENCES

1. K. Götz, “Course-control, metabolism and wing interference during ultralong tethered flight in Drosophila Melanogaster,” *J. exp. Biol.*, vol. 128, pp. 35–46, 1987.

2. C.A. Mead, “Neuromorphic Electronic Systems,” *Proceedings of the IEEE*, vol. 78, pp. 1629–1636, 1990.

1

THE BUILDING BLOCKS

C H A P T E R

2

PHOTOTRANSDUCTION

The engineering community knows by bitter experience that optimizing the initial transudction procoess is crucial. The body of this chapter is a detailed description and analysis of a sensitive, adaptive, continuous-time, logarithmic photoreceptor circuit. This receptor has proven itself useful in a range of vision chips that process time-domain image information. We have optimized the design of this receptor and carried through an analysis of its properties that goes farther than any other work we know of in the design of devices for continuous-time visual transduction.

The essence of the idea is that the photoreceptor centers its operating point around the history of the intensity, simultaneously achieving high sensitivity and wide dynamic range. Around each operating point, the receptor is a logarithmic detector with high gain that produces a robust signal in response to the small contrasts in typical scenes. Over long time-scales, the gain of the receptor is low, allowing operation over the wide range of scene intensities that are reality.

A logarithmic receptor is sensitive to *relative* changes in the intensity, not *absolute* intensity, and so it is useful for reporting about image *contrast.* Image contrast is due mostly to the reflectance of the physical surfaces (aside from shadows). Logarithmic receptors are sensitive to properties of the surface, and not the lighting conditions—that is why they are useful devices.

A purely logarithmic receptor seems at first sight like a good idea for the reasons just stated. In fact, most of the early work on silicon retinas labored under this misconception. The problem is

that the wide dynamic range of the logarithmic receptor is poorly matched to the small dynamic range of image intensities within a particular image. A simple logarithmic receptor reports signals that are no larger than the random offsets between adjacent receptors that are due to transistor mismatch. The solution to this problem is simple: Match the short–time-scale dynamic range of the receptor to the short–time-scale dynamic range within a particular scene, while at the same time making the long–time-scale gain of the receptor small, to match the extremely wide dynamic range of scene intensities.

This work differs from previous work on continuous-time analog photoreceptor circuits in its critical analysis of the limiting behavior, and its attention to constructing, characterizing, and understanding a receptor with maximum dynamic range and sensitivity. The inspiration for this work was biological, but the impetus was engineering reality. Previous photoreceptor circuits have deficiencies that render their usefulness questionable in anything but exploratory designs. Rather than a demonstration-of-concept, we intend this work to be a serious engineering study.

Some people view the competition as video cameras and frame grabbers. This viewpoint is blind to the inherent differences between continuous-time and sampled receptors, but it is valid in the sense that modern charge-coupled-device (CCD) imagers are very good, and a serious comparison must be made between the raw quality of the transduction process in these two technologies. In a video camera, the goal is to build an *imager* that faithfully reproduces electronically the image intensities, with simultaneous optimization of pixel density, dynamic range, sensitivity, noise, and nonuniformity. This art is highly developed in the form of CCD imagers [15]. Present-day cameras achieve 1000 by 1000 density in a 2 cm^2 die, with a dynamic range of up to 4 decades at a single shutter setting, a noise of less than 100 equivalent electrons per integration time, a maximum signal-to-noise ratio (SNR) of over 60 dB†, and a nonuniformity of less than 1 %. These numbers are amazing, but we must remember that they are the result of concentrated effort of many thousands of man-years, driven by intense market forces. If we can match these characteristics, then we can be confident that we are doing well. The receptors reported in this chapter, which are fabricated in a stock CMOS process, compare favorably with several state-of-the-art CCD performance characteristics.

† A 60 dB SNR means the power ratio is 10^6, i.e., 10 dB means a factor 10 in power. The signal-*amplitude* to noise-*amplitude* ratio is 10^3.

A key component in these receptors is an **adaptive element**. The adaptive element is a resistor-like device that regulates the rate that the receptor learns about the history of the signal. CMOS technology does not allow for the use of real resistors except in very specialized applications, like tilting a global bias voltage across a chip. For circuits meant to do local processing, available resistances are much too small to construct useful time constants on the order of seconds. The only alternative to a resistor is a transistor. Transistors have naturally exponential behavior while working in subthreshold. Instead of bemoaning this fact, we take advantage of it by constructing nonlinear adaptive elements from transistors. We show, however, that the obvious choices for adaptive elements that use transistors have problems that arise from the same interaction of light with silicon that lets us detect the light. Scattered light and diffusing minority carriers cause large systematic offset voltages and asymmetric operation. In this chapter, we report two novel adaptive elements that are inherently insensitive to the effects of light.

People sometimes think that speed is not really an issue for receptor circuits designed to work in artificial vision systems. After all, both human and fly vision (two extremes) cut off at less than 200 Hz, so why worry about microsecond time scales? The bandwidth of the primary transduction process is always proportional to the intensity: The more light, the larger the photocurrent, and the quicker the response. No matter how fast the photoreceptor, at some intensity, the response becomes too slow to be useful. We describe how the active feedback circuit speeds up the receptor, compute the effect of circuit and stimulus parameters on the time response, and show comparative measurements of different types of receptors to find the optimum configuration.

Photoreceptor usefulness is not only limited by time response, but also by noise. To understand the limits on photoreceptor sensitivity, we measured basic transistor noise (Appendix 2c), as well as noise in our photoreceptor circuits. We find that the receptor noise is dominated by the initial stage of transduction. The adaptive feedback circuit adds negligible noise. An application of the simple noise theory developed in Appendix 2c to the photoreceptor circuit leads to beautiful results about noise in logarithmic receptors. We say beautiful because it is rare that such noise results are simple and intuitive.

Much of the technology in these receptors involves the interaction of light with silicon. To round out our study of visual transduction, we measured the spectral response properties of silicon phototransducers (Appendix 2a). We find that the response properties of the photodiodes and phototransistors available in a stock CMOS or BiCMOS process are consistent with a simple theory of carrier diffusion and the known properties of the absorption of light by silicon.

It turns out that the adaptive properties of these receptors depend a lot on the parasitic and unintended interaction of light-generated minority carriers with the parts of the photoreceptor circuit. Appendix 2b is about a set of measurements of the diffusion of minority carriers and the effectiveness of guard structures.

The rest of this chapter is organized as follows. We start with a heuristic explanation of the photoreceptor circuit operation that is sufficient for designers to use these receptors in their own designs. We next give a detailed description of the adaptive elements, and compare the old elements with the new and improved ones. In order to test our theoretical understanding of the photoreceptor operation, we develop a small-signal model of the circuit operation and compare it with measurement. Finally, we discuss theory and measurement of the limits of operation, to answer the questions: What are the lowest usable intensities, and what are the smallest detectable signals? Two appendices go into detail about the interaction of light with silicon, and the final appendix is about general transistor noise.

CONTINUOUS-TIME PHOTORECEPTORS—PREVIOUS WORK

The basis for the work reported here is Mead's logarithmic photoreceptor [9], used in the early Mahowald and Mead silicon retinas [5][11] and in the SeeHear chip [16]. This receptor has three deficiencies that are corrected in the present work. The problems with this receptor are the poor matching between different receptors, the poor match between the low sensitivity of the receptor and the high gain required to sense the small contrasts present in real images, and the slow time-response that limits the dynamic range. Delbrück and Mead built an adaptive receptor that contained essentially the same ideas as in this chapter, though in naive form [2]. Mahowald incorporated essentially the circuit described in this chapter, except for the adaptive element, into a silicon retina [4]. At about the same time, Mann [7] developed several adaptive photoreceptor circuits that are more flexible than the ones described here, except that they use a larger number of components and were never fully tested. None of the above receptors were satisfactorily characterized, in the sense that the simple engineering metrics like usable dynamic range and sensitivity were not obtained.

The imager community is not fully won over by CCD dominance, and periodically we see published reports of new approaches to the use of solid-state detector arrays. We will omit discus-

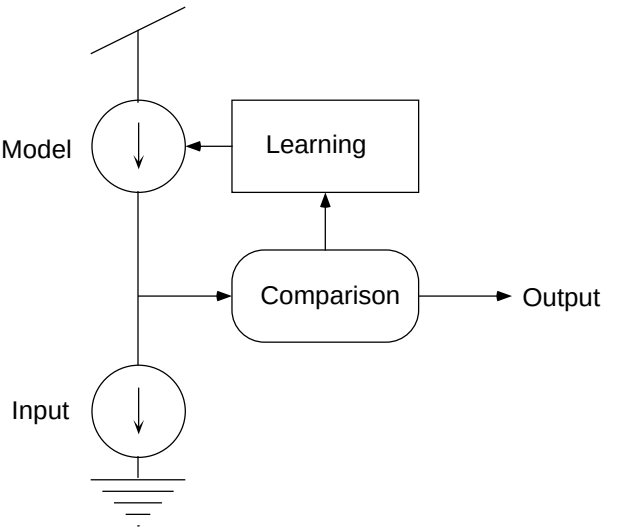


FIGURE 2.1 A photoreceptor in abstract form, as an example of an adaptive information processing device. The photoreceptor compares an input current with a model current, and outputs the amplified comparison. The comparison is also used to learn the model. The combination of comparison, model, and learning allows for a system with high sensitivity and wide dynamic range.

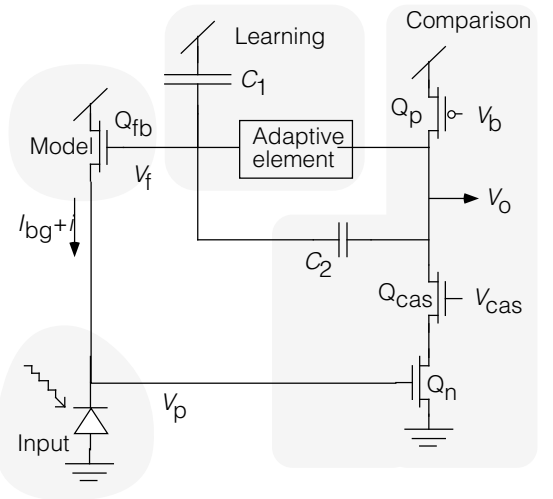
sion of these approaches for three reasons: First, this community is overwhelmingly concerned with imaging, rather than with information processing. Second, these detectors are designed to be sampled in time. Third, these devices usually cannot be built in a standard CMOS process.

CIRCUIT HEURISTICS

There is a direct analogy between the components of Mead's conceptual model of brain information processing ([2] in Chapter 1) and the parts of the adaptive photoreceptor circuit, as shown in Figure 2.1. The scheme is that an *input* intensity is compared with a *model* of the input intensity, and only the result of a *comparison* between input and model is sent on to the next stage of processing.The output of the comparison is used in *learning* the model of input intensity. The advantage of this type of scheme is that only the *unexpected* information fills the dynamic range of the channel.

Figure 2.2 shows the photoreceptor circuit in transistor form, labeled with the elements of the conceptual model. This labeling is useful for understanding the *functional* role of the various components, without getting tangled up immediately in the technical details of the circuit operation. In this section, we provide a heuristic description of the operation of the feedback loop. We will first describe the individual components, and then go through a complete cycle of the loop, in response to a small change in the intensity.

FIGURE 2.2



The adaptive photoreceptor circuit with circuit components labeled with elements of Figure 2.1. Light shining on the receptor generates a photocurrent in the input photodiode linearly proportional to intensity. The rest of the circuit, consisting of Model, Comparison, and Learning, senses the photocurrent with high sensitivity and adjusts the operating point around the historical value of the intensity. The DC gain is low and the transient gain is high. The capacitive divider formed by C_1 and C_2 determines the gain of the amplifier. A resistor-like adaptive element allows the receptor to respond over a large dynamic range, by storing the average intensity on C_1.

The input current comes from a photodiode that generates a photocurrent that is linearly proportional to intensity (see Appendix 2a). The current consists of a steady-state background component I_{bg}, and a varying, or transient, component i. We split the current up in this manner because the goal of the computation is to compute the ratio i/I_{bg}; this ratio is the important quantity to measure for vision, because it corresponds to scene reflectivity rather than scene luminance. The photodiode is simply an extension of the source of Q_{fb}. All parts of the circuit except for the photodiode are covered with metal.

The model of intensity is stored as a charge on C_1, and a feedback transistor Q_{fb} supplies the model photocurrent. The source voltage V_p is directly determined by the gate voltage V_f and by the photocurrent $I_{bg}+i$. Intuitively, the feedback clamps the voltage V_p at whatever it takes for Q_n to sink the bias current supplied by Q_p. For typical intensities, Q_{fb} operates in subthreshold. V_p sits below V_f at whatever voltage it takes to turn on Q_{fb} to supply the photocurrent. Because the current i is exponential in the source voltage and in the gate voltage, the receptor is immediately and naturally a logarithmic detector.

The comparison between input and model is performed by the inverting amplifier consisting of Q_n and Q_p. An additional transistor Q_{cas}, in a cascode configuration, isolates the drain of Q_n from the large output voltage swings, and will be discussed in the next paragraph. The input voltage V_p controls the current sucked from the output node by Q_n. The current pushed into the output node by Q_p is fixed by the bias voltage V_b. The voltage gain $-A$ of the amplifier is determined by the ratio of the transconductance of Q_n to the parallel combination of the drain conductances of Q_n and Q_p, and is typically several hundred. The bias voltage, V_b, determines the cutoff frequency for the receptor, by setting the bias current in the inverting amplifier. We often use this control to filter out flicker from artificial lighting.

The cascode transistor Q_{cas} is a source-follower to shield the drain of Q_n from the large voltage swings of V_o. We hold V_{cas} high enough to hold the source of Q_{cas} high enough to saturate Q_n's drain, but not so high that the source is above V_o. The purpose of Q_{cas} is to nullify the large voltage swings across the gate–drain capacitance of Q_n that load down the input node. These voltage swings can make fF-scale gate-drain capacitance appear to the input node to be on the pF scale, greatly slowing the time–response. This phenomenon is well known and is called the **Miller effect**. In addition, the cascode effect of Q_{cas} multiplies the drain resistance of Q_n by a factor of approximately A, and hence, increases the gain of the amplifier by a factor of 2 or more. Both the reduction in effective input capacitance, and the increased gain, translate into speedup, and hence, increased dynamic range. The addition of this single cascode transistor increases the dynamic range of the receptor by about a decade, as we will see later.

The output V_o is fed back to V_f through the adaptive element and through the capacitive divider formed from C_1 and C_2. On long time scales, the feedback is a short circuit, because charge flows through the resistor-like adaptive element and onto the V_f node until the voltage across the adaptive element is zero. On short time scales, no charge flows through the adaptive element, but changes in V_o are directly coupled to changes in V_f, through the capacitive divider.

Let us follow the effect of a small intensity change around the loop. We denote a small-signal representation by lowercase. In response to a change of intensity i, the output voltage v_o moves enough so that v_f moves enough so that v_p is held nearly clamped. The voltage v_p is clamped because A is large—only a tiny change in v_p is needed to move v_o to wherever it needs to be to change v_f to supply i. (What a mouthful!)

In response to a change i, v_f must move by some amount to hold v_p clamped. On long time scales, the gain of the receptor is low, because the feedback is a short circuit across the adaptive

element, and v_o does not need to move much to hold v_p clamped. On short time scales, no charge flows through the adaptive element. When v_o moves a lot, v_f only moves a little, because the feedback capacitor C_2 is smaller than the storage capacitor C_1. The larger the capacitive divider ratio, the more v_o must move to move v_f by the required amount. The gain of the receptor is set by the capacitive-divider ratio. The larger C_1 is relative to C_2, the larger the gain of the circuit.

It is simple to compute the steady-state and transient gain of the receptor from the arguments just given; the key assumption is that A is large so v_p is held clamped. The gain from v_f to v_p is just κ, the back-gate coefficient. The gain from v_o to v_f is 1 in steady-state, and is $C_2/(C_1+C_2)$ transiently, when no charge flows through the adaptive element. The gain from input i to v_p is just $-V_T$ per e-fold intensity change. We obtain the complete closed-loop gain for both steady-state and transient inputs by computing the v_o needed to hold v_p clamped. Hence, we obtain the linearized steady-state closed-loop gain,

$$\frac{v_o/V_T}{i/I_{bg}} = \frac{1}{\kappa} \tag{1}$$

and the linearized transient closed-loop gain,

$$\frac{v_o/V_T}{i/I_{bg}} = \frac{1}{\kappa}\frac{C_1 + C_2}{C_2} \tag{2}$$

Note that we write the gain in dimensionless form, to display in a simple way the logarithmic, contrast-sensitive response properties. Equation 1 is the small-signal equivalent of the large signal expression for the steady-state output voltage, given here for reference by

$$\frac{V_o}{V_T} = \log I_b + \frac{1}{\kappa}\log\frac{I_{bg} + i}{I_0} \tag{3}$$

where I_b is the amplifier bias current and I_0 is the preexponential in the subthreshold transistor law. The ratio between transient and steady-state gain is simply the capacitive-divider ratio $(C_1+C_2)/C_2$. An important parameter of this system (and any feedback system) is the total loop gain, obtained by following the gain all the way around the loop. The total loop gain is

$$A_{Loop} = A\kappa\frac{C_2}{C_1 + C_2} \tag{4}$$

The adaptation happens when charge is transferred onto or off the storage capacitor. This charge transfer happens through the adaptive element. The adaptive-element is a resistor-like device that has a monotonic I-V relationship. In VLSI, however, true resistors are much too small for adapta-

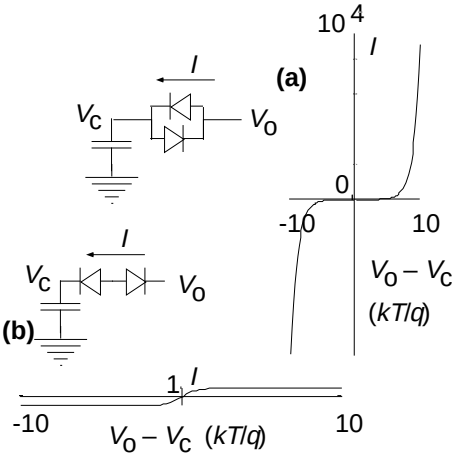


FIGURE 2.3 Two idealized adaptive elements. **(a)** the expansive "hysteretic" element. The current is an exponentially increasing function of the voltage for either polarity. **(b)** the compressive element. The current saturates for voltages higher than a few *kT/q*. In practice, these elements are always unsymmetrical in their response, and there is always a leakage current from the capacitor to a reference voltage other than V_o.

tion on the time scale of seconds,† so we use transistors in our adaptive element. We have developed two novel adaptive elements with dual nonlinearities—expansive and compressive (Figure 2.3). The expansive adaptive element acts functionally like a pair of diodes, in parallel, with opposite polarity. The current increases exponentially with voltage for either sign of voltage, and there is an extremely high-resistance region around the origin. The compressive adaptive element acts functionally like a pair of diodes in series, with opposite polarity. The current saturates at a very small value for either sign of voltage, with a small linear central region. Here we will discuss the functional implications of ideal adaptive element like those shown in Figure 2.3. Starting on page 17, we will describe the actual adaptive elements in detail.

The I–V relationship of the expansive adaptive element means that the effective resistance of the element is huge for small signals, and small for large signals. Hence, the adaptation is slow for small signals and fast for large signals. This behavior is useful, since it means that the receptor can quickly adapt to a large change in conditions—say, moving from shadow into sunlight—while maintaining high sensitivity to small signals.

The I-V relationship of the compressive adaptive element means that the effective resistance is constant for small voltages, and becomes huge for large voltages. For large voltages, the element

† Assume we need an *RC* time constant of a second, and that C = 1 pF (a 50 μm by 50 μm poly-to-poly capacitor). Then we need $R = 10^{12}$ Ω. Poly has a resistance of 20 Ω/square, so we would need 5×10^{10} squares—a 2 μm-wide poly resistor with area 0.6 m^2!

simply acts like a current source and not a resistor. The effect is that large changes in intensity are effectively ignored on short time scales. The receptor is equally sensitive to both large and small intensity changes. The adaptation time is proportional to the size of the signal.

An obvious advantage of using active feedback is that the separate bias current in the output leg of the receptor is capable of driving arbitrarily-large capacitive loads, simply by adjusting the bias current. More fundamentally, however, the feedback configuration, by clamping the v_p node, extends the usable dynamic range of the receptor by speeding it up. The small photocurrents need only charge and discharge the small changes in v_p, rather than the large swings of v_o. We can immediately deduce that the larger the gain A of the amplifier, the more speedup we obtain. Also, for a given amplifier, the more closed-loop gain we design into the receptor, by adjusting the capacitive-divider ratio, the slower the response. We take as a baseline the speed of the input node if v_f is held constant, say, by breaking the loop there. The speedup, over this baseline value, that we obtain by using the active feedback is proportional to the total loop gain A_{Loop}. We explore these points further in the small-signal analysis, starting on page 28. A typical measured speedup, using the active feedback and the cascode, is 1–2 decades.

A demonstration

We can understand the behavior of this receptor with both the expansive and compressive adaptive elements, by examining the measured response to a varying intensity (Figure 2.4). The incident light consists of a small intensity variation sitting on a steady background. The contrast of the signal, relative to the background, is a fixed percentage, independent of the absolute intensity. We vary the overall intensity level by interposing neutral density filters—like sunglasses—with various attenuation factors. All the behaviors just discussed appear in this figure. Because the response is logarithmic, the amplitude of the response to the small contrast variation is almost invariant to the absolute intensity. The receptor that uses the expansive adaptive element adapts very rapidly in response to the large change in intensity caused by a filter change. Because the steady-state gain is much smaller than the transient gain, the adapted response to an intensity change of a decade is almost the same as the response to the 15 % variation. The time constant of the adaptation is much shorter than the period of the input signal. We can see how a receptor like this could be useful in systems that care about the contrast changes in the image, and not the abso-

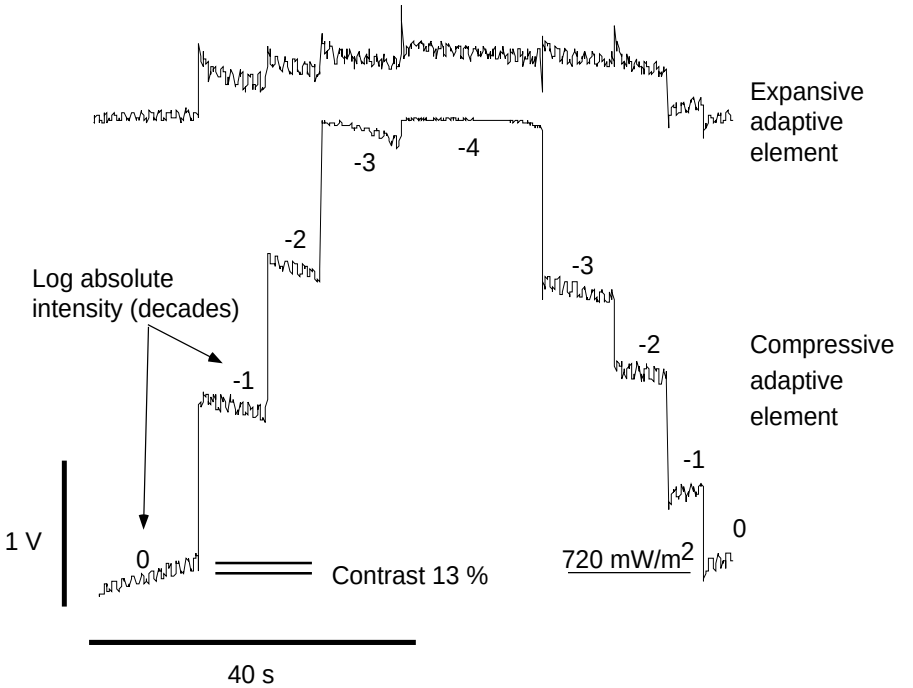


FIGURE 2.4 Response of two adaptive receptors. Stimulus is a square-wave variation in the intensity around a mean value. The numbers by each section are the log intensity of the mean value; 0 log is 720 mW/m^2. The amplitude of the square-wave variation is about 6 % of the mean value. **(a)** is from a receptor with the expansive adaptive element. **(b)** is from a receptor with a compressive adaptive element. Each receptor is primarily sensitive to contrast, rather than absolute intensity, and each has roughly the same gain for small signals, but (b) is much slower to adapt to the large changes of intensity.

lute intensities. In contrast, the receptor that uses the compressive element adapts very slowly to the large intensity changes, but also faithfully reports their size.

THE ADAPTIVE ELEMENTS

The key difficulty in the practical use of these adaptive elements in photoreceptor circuits, when light falls on the surrounding circuitry, is related to the effect of light on silicon. Light creates minority carriers, and minority carriers diffuse home to their majority regions, creating junction currents. Appendix 2b is about measurements we did to show that the diffusion length is tens of

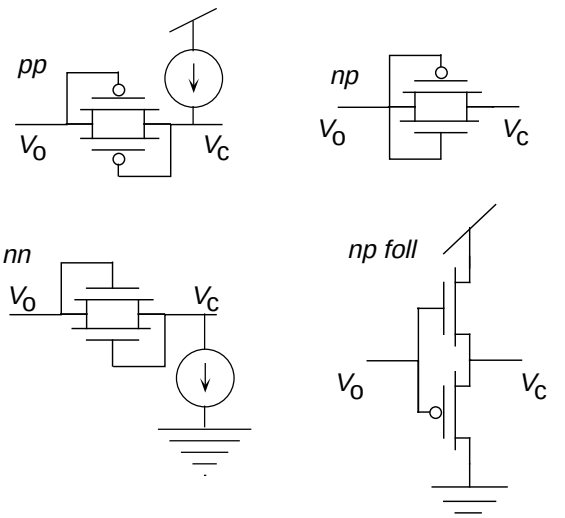


FIGURE 2.5 Expansive adaptive elements [8]. Ideally, each element has the sinh-like *I-V* relationship shown in Figure 2.3a. Because of junction leakage, none are satisfactory. Leakage currents are shown explicitly for the *pp* and *nn* elements; in the other two elements, the bulk leakage dominates.

microns, and that it is not practical to build guard structures to soak up the excess carriers. The adaptive elements, constructed from transistors, are at the same time also photodiodes. In our photoreceptor circuit, the adaptive element is hooked up so that one side is driven and the other side is not. The leakage currents pull the undriven side to the wrong voltage, and unbalance the operating point of the element. Circuits constructed from naively-designed adaptive elements are unbalanced both in steady-state and in response to up- and down-going changes. We invented two adaptive elements that are inherently resistant to these junction leakage effects. Both elements use the idea that the undriven node is isolated from any reference voltage other than the driven side. Before we describe the new adaptive elements, we shall discuss previous solutions and why they are not satisfactory.

Expansive adaptive elements†

The old expansive adaptive elements shown in Figure 2.5 each consist of complementary diode-characteristic elements that turn on for opposite polarities of voltage, and each has an *I–V* relationship, to first order, like $I = \sinh(V)$. The problem with each of these elements is that junction leakage currents cause a large offset voltage and asymmetric operation.

† Adaptive elements with these exponentially-increasing characteristics are often called **hysteretic elements**.

For example, consider the element labeled *nn*, consisting of two native diode-connected transistors. The source-drains of the two diode-connected transistors sit on the bulk substrate. These source-drains are also photodiodes that soak up stray minority carriers generated, say, by light falling on the nearby opening in the metal. Any charge sitting on the source-drain is leaked away towards the bulk potential, which is ground. In order to hold the capacitor at the correct voltage to complete the feedback loop, the output of the photoreceptor must turn on the element to counteract the leakage. The large source voltage combines with the back-gate effect to produce a steady-state offset voltage $V_c - V_o$, across the adaptive element, of at least 1 V, even in the dark.

We hoped to avoid this problem in the *pp* adaptive element. The idea is that the transistors are protected from stray minority carriers that are generated in the bulk substrate. These carriers are absorbed by the well, where they are lost in the sea of majority carriers. Empirically, the minority carriers generated in the well by scattered light are sufficient to unbalance the circuit, resulting in an offset voltage that is nearly as large as in the *nn* element, and a very unsymmetrical response to bright and dark edges when incorporated in an adaptive receptor. Even in the dark, junction leakage currents, combined with the large back-gate effect, leave a remnant offset of about a volt.

We also tried using a combination of native and well transistors, in hopes that the leakage currents would cancel (elements *np* and *np foll* in Figure 2.5). This balancing act fails, because the substrate leakage dominates the well leakage by orders of magnitude.

A new expansive adaptive element

These problems led us to invent the new element, shown in Figure 2.6, that is inherently insensitive to the effect of junction leakage currents. It consists of a single well transistor sitting in its own isolated well. The capacitor, one source-drain of the transistor, and the transistor gate are attached together to form the isolated node. The amplifier output, the other source-drain, and the well itself are attached together to form the driven node.

The operation of the element depends on what direction the current flows. When the output voltage is higher than the capacitor voltage, the current is conducted by the surface-channel, field-effect transistor. The source of the transistor is raised above the gate and drain, and the current flows through the surface channel (Figure 2.6(b)). The effect of raising the source and the well together is effectively the same as lowering the gate, because there is no other reference voltage that matters. Hence the current e-folds every $kT/q\kappa$ volts.

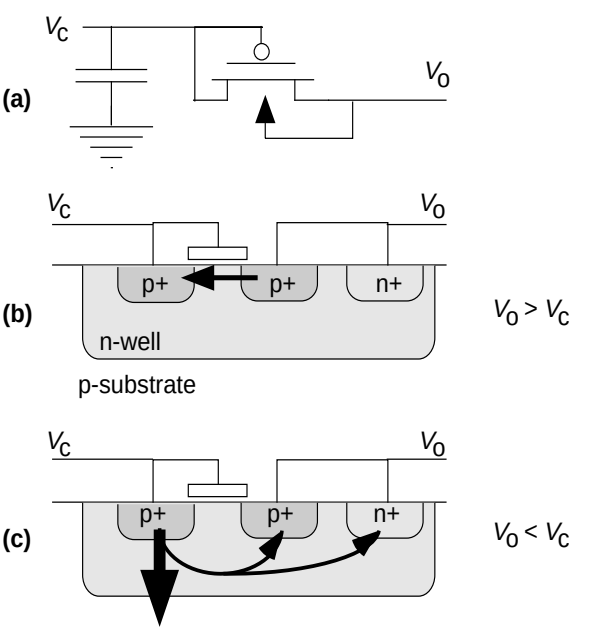


FIGURE 2.6 A new expansive adaptive element **(a)**, shown in schematic form along with the capacitor that stores the adaptation state.

(b) The mode of conduction when the output voltage is higher than the capacitor voltage: The structure acts as a diode-connected FET.

(c) The opposite case: The p+/n junction acts as a true diode, and the device as a whole acts as a lateral bipolar transistor.

When the output voltage is lower than the capacitor voltage, the current is conducted by a bipolar bulk mechanism. The junction between the well and the source-drain diffusion attached to the capacitor is forward-biased (Figure 2.6(c)). Many of the carriers that are emitted from the capacitor source-drain diffusion are collected by the bulk substrate and not by the other source-drain diffusion. In other words, the bipolar current gain means that it takes only a little current sunk by the output node to forward-bias the junction. The current e-folds every kT/q volts.

The low offset functioning of this element lies in two details. First, the capacitor has no reference voltage other than the receptor output voltage. There is no power supply rail to leak to. The only available reference voltage is the voltage of the well—which sits at the output voltage of the receptor. Second, the transistor bulk—the well—is at the source voltage, so there is no back-gate effect and the off-state conductance in the field-effect surface-channel transistor is large enough to overcome the junction potential between capacitor and well that is induced by scattered light. In other words, if there were no field-effect transistor, a junction potential between well and capacitor would form to cancel the remnant junction current caused by thermally or optically generated carriers. The field-effect transistor leaks enough—its threshold is low enough—that this junction potential is negligibly small.

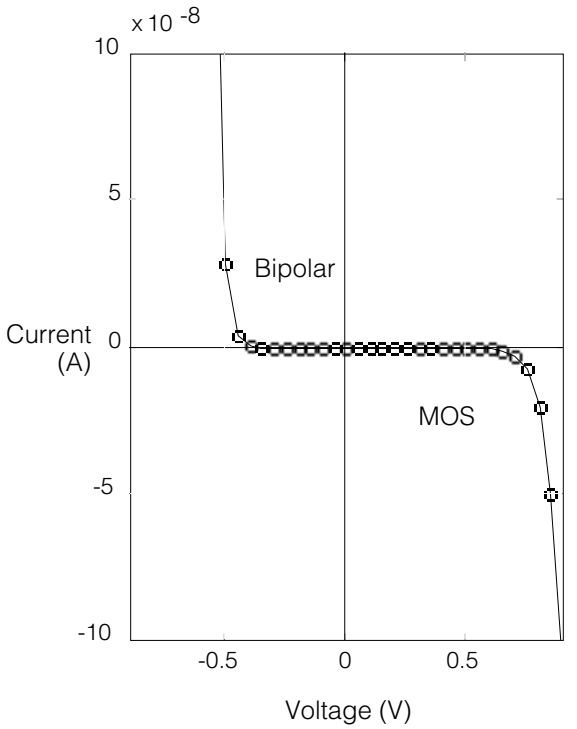


FIGURE 2.7 Measured Current–Voltage relationship for the new expansive adaptive element shown in Figure 2.6. The Bipolar mode conduction is stronger, leading to a quantitative difference in the voltage at which the current explodes. At any scale of current, the curves have the same appearance; the voltage scale changes logarithmically with the current scale. We took this data from a p-well chip.

We can see in Figure 2.7 how the exponential characteristic of the expansive adaptive element leads to an extremely flat response near the origin of the I–V curve. The bipolar mode conduction explodes exponentially at a smaller voltage, and explodes at a faster rate, but otherwise is similar to the MOS mode conduction.

Figure 2.8 shows the measured I–V relationships in more detail. We can see from this data how the bipolar mode current e-folds about every 28 mV, close the predicted value $\frac{kT}{q} \approx 25.4$ mV . The MOS mode e-folds about every 48 mV, in agreement with a value $\kappa = 0.53$ for the back-gate coefficient that is reasonable given the small gate–bulk voltage in the measurement. Figure 2.8 also shows that light falling on an opening in the metal near the element has an immeasurable effect on the current flowing in the gate terminal of the device, while it has a huge effect on the current flowing into the well node of the device.

Figure 2.9 compares the measured static DC offset voltage in the old adaptive elements in Figure 2.5 and in the new element in Figure 2.6. We can see from these measurements that the offsets are minimal in the new element, compared with the previous ones.

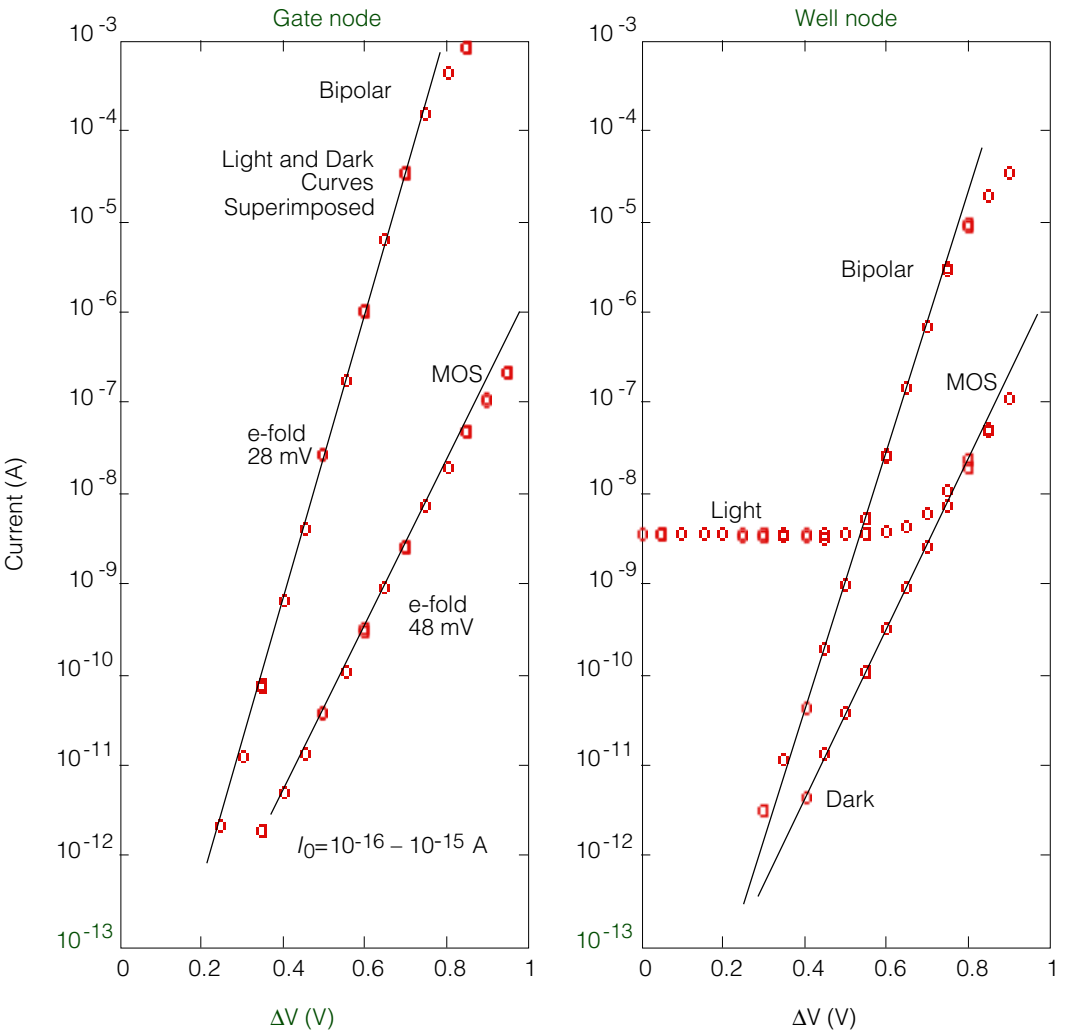


FIGURE 2.8 Current–Voltage relationships for the new adaptive element. There plots show the current flowing into each terminal of the adaptive element shown in Figure 2.6 in each of the modes—Bipolar and MOS—and under darkness and bright (1 W/m^2) lighting conditions, as a function of the voltage across the adaptive element. This data is from a p-well device. The device is illuminated with a red LED. The opening in the metal is 11 by 18 μm^2 and is about 15 μm distant from the device. The MOS transistor is 6 μm wide by 4 μm long. The well is 32 by 17 μm^2 (slightly large than necessary). The common-mode voltage is 2.5 V relative to the substrate.

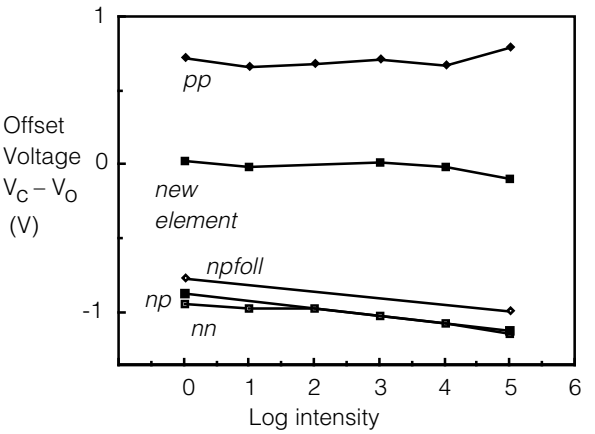


FIGURE 2.9 Expansive adaptive element offset voltage as a function of log incident irradiance. Labels on plots refer to elements shown in Figure 2.5, except for the new expansive element shown in Figure 2.6. Highest irradiance is 5.6 W/m^2. Red LED (λ=635 nm) is used to illuminate devices. Devices are shielded with a surrounding 10–μm-wide native source-drain diffusion guard bar tied to ground, and everything is covered with metal; the nearest substrate opening is 30 μm. All the elements with native devices are dominated by leakage to the substrate (ground). The *pp* element leaks to the well (V_{dd}). Only the new adaptive element has minimal offset voltage.

Figure 2.10 compares the response to step changes in the intensity of a photoreceptor built using the old *pp* element, and built with the new expansive element. The old receptor responds very unsymmetrically to increases and decreases of intensity, because the adaptive element sits on one side of the I–V profile. The asymmetry is worse under bright light (not shown), indicating the role of light-generated carriers. The new receptor, in contrast, responds much more symmetrically to the same step changes of intensity. The response is still unsymmetric, but the asymmetry is explained by the difference between the bipolar and FET conduction modes in the two directions. In the light-to-dark direction, the current through the expansive element e-folds every kT/q, in the dark-to-light direction, every $kT/q\kappa$. The receptor should adapt more quickly in response to transitions from light to dark, and we can see from Figure 2.10 that it does.

The compressive adaptive element

Misha Mahowald used the compressive adaptive element shown in Figure 2.11 in one of her silicon retinas [4]. Ideally, this element has the compressive I-V relationship shown in Figure 2.3(b). The problem with this element is similar to the problems of the elements in Figure 2.5. There is a diode leakage current from the well to the bulk substrate that is exacerbated

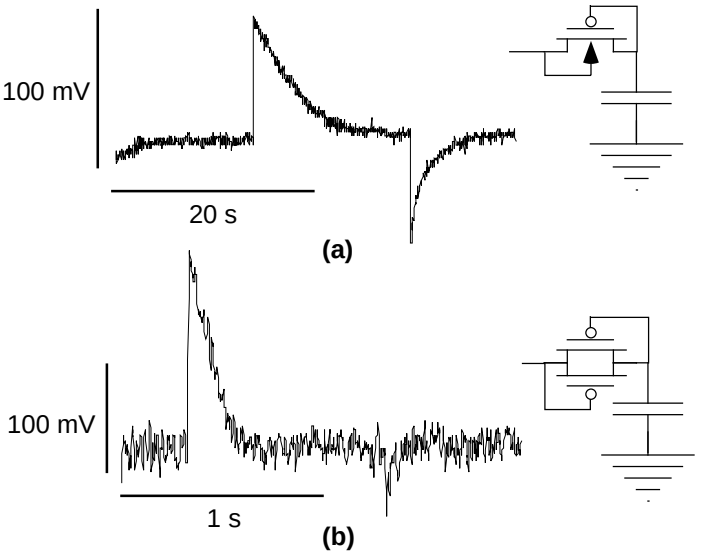


FIGURE 2.10 A dynamic comparison of the operation of the adaptive photoreceptor with **(a)** the new adaptive element versus **(b)** with one of the old elements, in response to step changes in intensity. The receptor with the new element responds much more symmetrically to increases and decreases of intensity, and the time constant of adaptation is much longer (note difference in time scale). We adjusted the stimulus to produce approximately the same change in voltage across the adaptive element for the two cases. The illumination was by red LED with a background irradiance of 60 mW/m^2, about 1/10 the level of direct office fluorescent lighting.

by light-generated minority carriers. This leakage causes a large offset in a feedback circuit, and makes the response of the feedback circuit unsymmetrical in response to increases and decreases in the input. In other words, in one direction of change, adaptation is rapid, and in the other it is slow.

This problem is much less severe in the new compressive element shown in Figure 2.12, which uses the vertical bipolar transistor structure available in a BiCMOS process to construct an isolated node. Again the trick is that the isolated node can only leak to the driven voltage, leading to small offset. The difference in the doping profiles between the two junctions in the bipolar device leads to a saturation current that is about 60 times higher in one direction than the other. The response of a receptor built using this compressive element is shown in Figure 2.4b. We can see from the time scale on this figure that despite the difference in the slew rate, the effective time constant of the

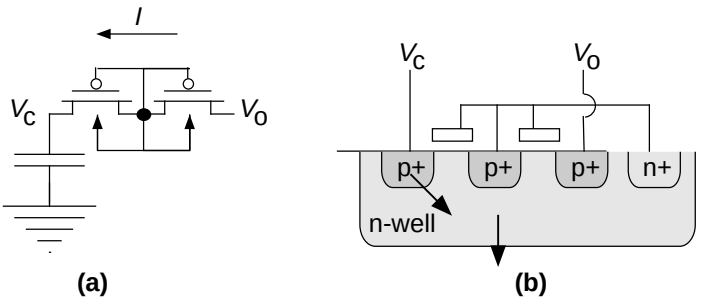


FIGURE 2.11 Old compressive adaptive element. **(a)** shows the circuit schematic. **(b)** shows the physical structure. Arrows show path of current flow when V_O drops below V_C. Diode-leakage current from the well to the substrate causes V_C to drop much faster than reverse current in the transistor by itself allows. The opposite voltage polarity does not suffer this problem because the diode-leakage current is supplied by the driven node V_O.

adaptation is adequate at even the highest intensities for vision tasks on the scale of seconds or faster.

The details of the leakage effects in the compressive element are as follows. In the dark, the saturation current is 500 electrons/second onto the capacitor, and 30,000 electrons/second off the capacitor. With a capacitance of 0.5 pF (the capacitance we get with a 30 mm by 30 μm poly1 to poly2 capacitor), these currents translate to a downward slew rate of a thermal voltage (25 mV) every 3 minutes, and an upward slew rate of a thermal voltage every 3 seconds. This ratio of 60 in saturation current is explained by the different characteristics of the two junctions; the saturation current of the base–emitter junction is about 60 times smaller than the saturation current of the base–collector junction. This ratio of saturation currents is independent of intensity, although the absolute value is proportional to intensity. The leakage current of the base–emitter junction is smaller because the doping of the n++ emitter and the p+ base is larger than the doping of the n-well, leading to smaller minority concentrations, and hence, smaller reverse-saturation current.

EXPERIMENTAL RESULTS ON RECEPTOR GAIN

Figure 2.13 summarizes experimental results on the steady-state and transient gain of the photoreceptor. We can see from this figure the graphical form of the idea that the transient gain is

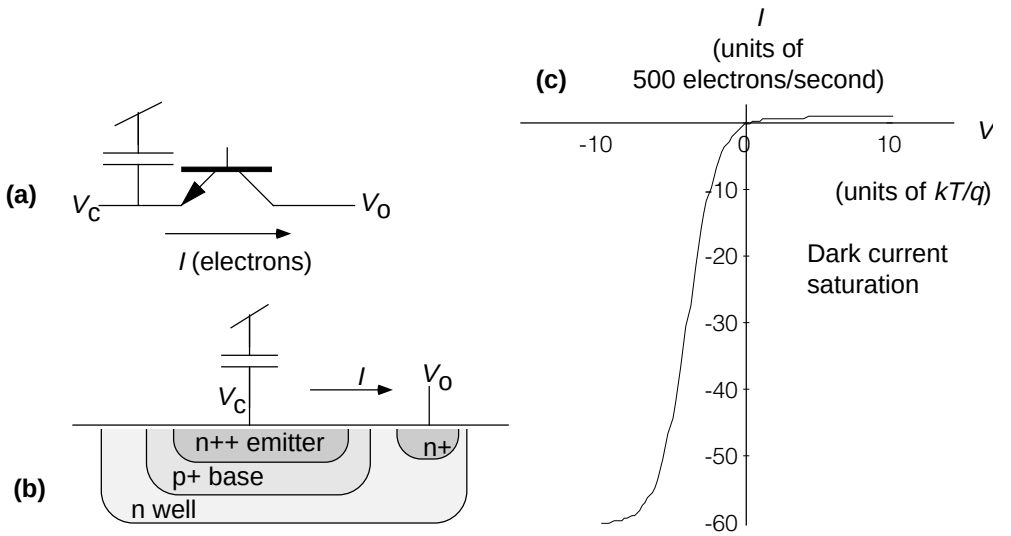


FIGURE 2.12 New compressive adaptive element. **(a)** The schematic form. **(b)** A cross-section through the structure. Structure is a vertical bipolar transistor with floating base. Charge-storage node can only leak to driven node. **(c)** Inferred I-V relationship in the dark.

larger than the steady-state gain. The slopes of the operating curves are predicted by Equations 1 and (2).

The steady-state gain is predicted by Equation 1 to be V_T/κ per e-fold intensity change. This translates, at room temperature, to a gain of 70– 80 mV/decade, assuming $\kappa \approx 0.7$– 0.8. From Figure 2.13, we can see that the measured steady-state gain is 30–40 mV/decade. The actual steady-state voltage is difficult to measure precisely, because of the extremely long adaptation time constant. Even so, we believe that the measured result is correct, and that the discrepancy is due to a small, but finite, offset voltage in the adaptive element that is intensity-dependent in a direction that reduces the total gain. When the intensity increases by a decade, the feedback node must increase by 70–80 mV, but since the offset voltage across the adaptive element also increases about 30–40 mV from increased junction potential, the output need only increase 30–40 mV.

In Figure 2.14, we plot the measured small-signal transient gain as a function of background intensity. We can see from this plot that the gain is relatively constant over a range of 4–5 decades.

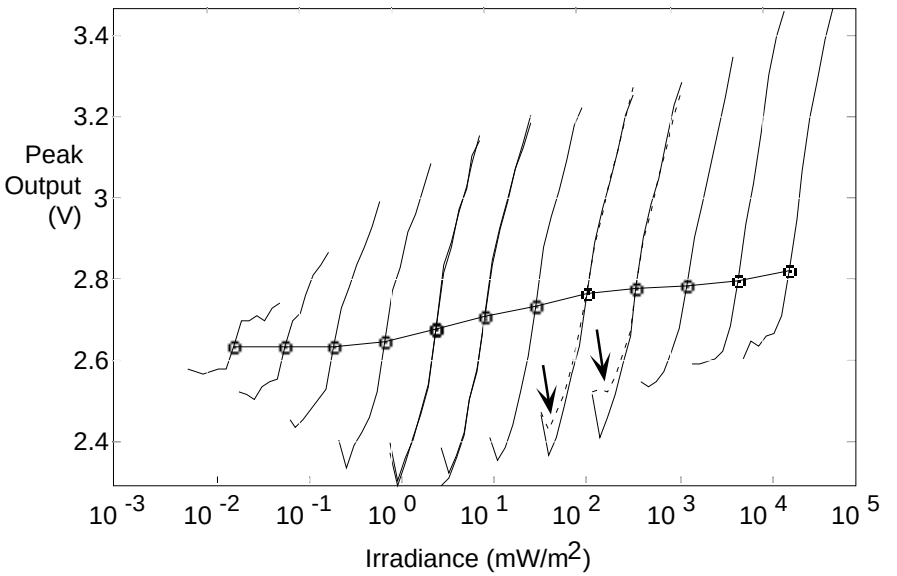


FIGURE 2.13 Step response operating curves. Each s-shaped curve shows the peak value of the response to a step change of intensity, starting at the intensity marked with a circle. The ordinate shows the peak value of the response to the step, and the abscissa shows the absolute incident irradiance. The total range we tested spans 6.5 decades. The receptor was allowed to adapt back to its steady-state value for 5 s before each step stimulus, except for the dashed curves (marked with the arrows). For these curves, the adaptation was for 15 s. Increasing the adaptation time allowed the receptor output to fully restore to the steady-state value. For each of the curves, the first step intensity was to the lowest intensity. These points are higher than immediately higher intensities, because the receptor was allowed to adapt for a longer period of time between adaptation levels than between points on one curve. The steady-state gain measured from the data above is between 30 and 40 mV/decade; the transient gain is plotted in Figure 2.14.

The gain in the constant region is about 1.4 V/decade, or about 18.7 times larger than the steady-state gain. From the layout for this receptor, we measure a capacitive divider ratio of $(C_1 + C_2)/C_1 \approx 17$, so the results are consistent within experimental error. We could not explore the upper limit of the intensity range because our LED's are not bright enough. At the lower limit, the gain decreases, probably from a combination of adaptation time constant competing with response rise-time, and illumination-independent dark current.

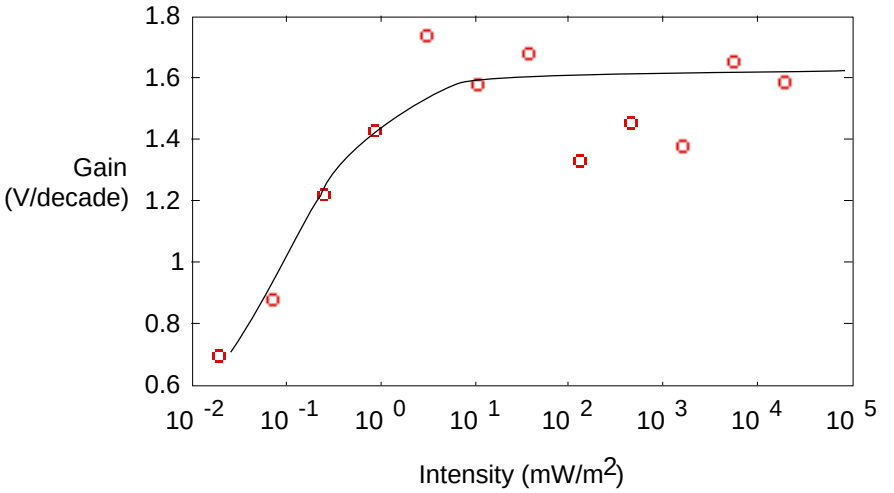


FIGURE 2.14 Adaptive receptor small-signal transient gain, as function of background irradiance. We measured these points from Figure 2.13. The curve is fitted by eye through the data points. The measured gain of about 1.4 V/decade is consistent with the measured capacitive divider ratio $(C_1 + C_2)/C_2 = 17$.

SMALL-SIGNAL TIME-DOMAIN ANALYSIS

In this section, we quantitatively analyze the behavior of the photoreceptor circuit in the small-signal regime. Our analysis formalizes the intuitive analysis we gave earlier, and helps us understand the second-order behavior of the circuit. Figure 2.15 shows the feedback circuit along with definitions of the circuit elements we use in our analysis. Figure 2.16 shows the layout corresponding to the schematic in Figure 2.15. Many of the measurements we show later are taken from this receptor circuit.

To perform the analysis, we make the simplifying assumption that the phototransducer has a lumped capacitance, C_p, to ground. This capacitance is important, because, under low-light conditions, it limits the response time of the photoreceptor. The lumped-capacitance assumption is correct as long as we use a photodiode, rather than a phototransistor, and as long as the response time is not limited by finite minority-carrier diffusion time. Under these conditions, the model shown in Figure 2.15 is a good match to reality. We will discuss the other limiting effects starting on page 38.

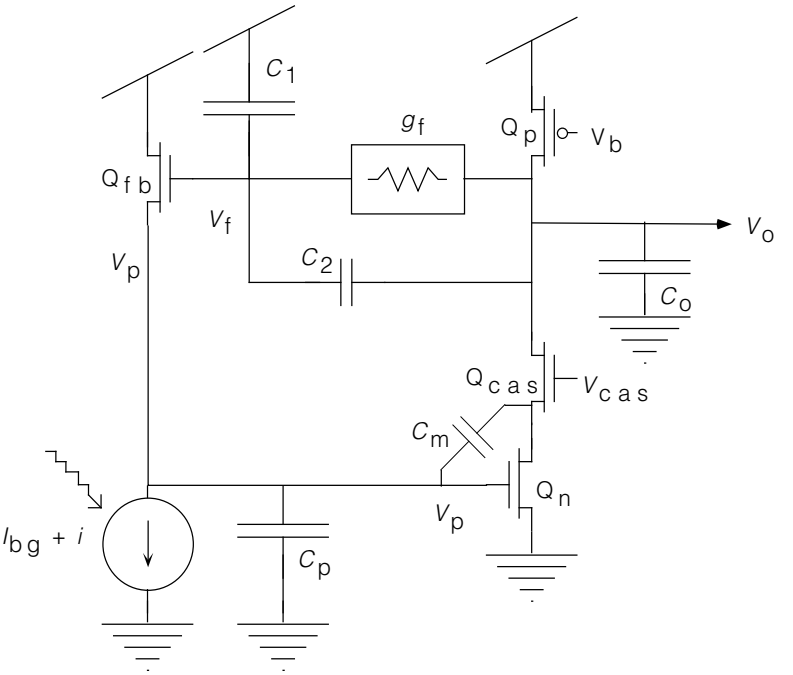


FIGURE 2.15 Photoreceptor circuit used in small signal analysis.

In the following analysis we will assume, for full generality, that Q_{cas} is shorted, rather than acting as a source-follower or cascode. This assumption lets us see quantitatively the effect of C_m. The three equations that govern the behavior of the small signal voltages v_o, v_p, and v_f are

$$\begin{aligned} C_o \dot{v}_o &= -g_n v_p - g_o v_o \\ C_p \dot{v}_p &= -i + \frac{I_{bg}}{V_T}(\kappa v_f - v_p) + C_m(\dot{v}_o - \dot{v}_p) \\ C_1 \dot{v}_f &= C_2(\dot{v}_o - \dot{v}_f) + g_f(v_o - v_f) \end{aligned} \tag{5}$$

The first equation describes the inverting amplifier. The output node, v_o, with capacitance C_o, is discharged by an increase in the input voltage, v_p, acting through the input transconductance g_n of Q_n. The output node is also discharged through its output conductance, g_o. The output capacitance, C_o, consists of an external load in parallel with C_2. The second equation describes the input node. The lumped capacitance to ground is C_p. This capacitance is discharged by the input current i, and is charged by the feedback transistor. We have linearized the Q_{fb} current around its adapted operating point, where, in subthreshold, the source conductance is I_{bg}/V_T and the gate transcon-

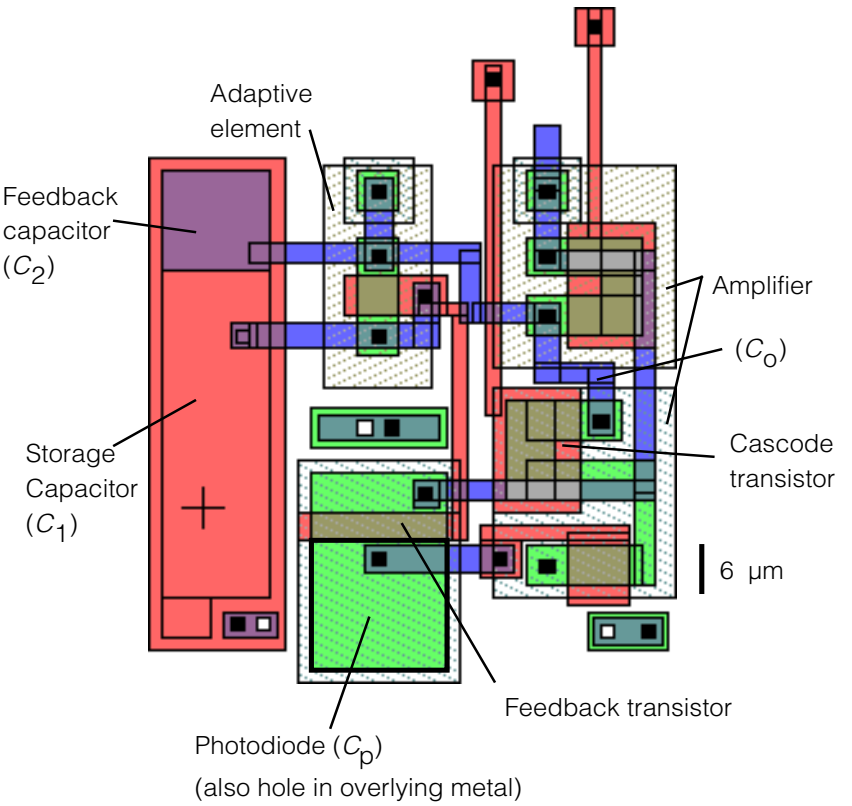


FIGURE 2.16 The layout of a photoreceptor. The photodiode is formed as an extension of the source of the feedback transistor. Second-level metal covers everything but the photodiode. 6 μm scale bar is shown at bottom right; total area is about 80x80 μm^2 in a 2-μm technology.

ductance is $\kappa I_{bg}/V_T$. The last term of the second equation describes the current flowing across C_m. The third equation describes the feedback node. This node is charged through the capacitive divider and through the adaptive element. In this discussion, we assume, somewhat artificially, that the adaptive element has conductance g_f. In practice, this conductance is proportional to intensity, and increases monotonically with the amplitude of the stimulus.

There are four natural time constants that fall out of Equations 5. They are an adaptation time constant τ_f, an output amplifier time constant τ_o, a Miller-capacitance time constant τ_m, and an input-node time constant τ_p, given by

$$\tau_f \equiv \frac{C_1 + C_2}{g_f} \qquad \tau_o \equiv \frac{C_o}{g_o} \qquad \tau_m \equiv \frac{C_m}{I_{bg}/V_T} \qquad \tau_p \equiv \frac{C_p}{I_{bg}/V_T} \tag{6}$$

We can also see how the open-loop amplifier gain arises from the combination of output conductance and input transconductance:

$$A \equiv \frac{g_o}{g_n} \tag{7}$$

One other important quantity is the capacitive divider ratio

$$A' \equiv \frac{C_1 + C_2}{C_2} \tag{8}$$

To solve for the transfer function between input i and output v_o, we rewrite Equations 5 in the s-plane, where each time derivative becomes s. Then it is simple to solve for the ratio of output to input. The result is

$$H(s) = \frac{v_o / V_T}{i / I_{bg}} = \frac{\frac{A'}{\kappa}\tau_f s + 1}{\frac{A'}{\kappa A}(\tau_o s + 1)\,((\tau_p + \tau_m)s + 1)\,(\tau_f s + 1) + \tau_f s + \frac{A'}{\kappa}\tau_m s + 1} \tag{9}$$

The transfer function is dimensionless; it expresses the ratio of output voltage in units of the thermal voltage to input current in units of I_{bg}. We can now see that the reciprocals of the time constants of Equation 6 are not identical to the poles, although they are closely related. The time constants are separated by several orders of magnitude under typical conditions. τ_f determines the adaptation rate, and hence the low frequency cutoff. τ_o determines the absolute high-frequency cutoff for bright light. τ_p and τ_m generally limit the bandwidth of the detector. We note that τ_p and τ_m are inversely proportional to the intensity.

Figure 2.17 shows the measured transfer function for different background intensities. The shape of the measured transfer function is similar to the theoretical one. The high-gain region between low- and high-frequency cutoff spans about 4.5 decades. An interesting behavior is that the low- and high-frequency cutoffs are each proportional to intensity, so that the log bandwidth is a constant, independent of intensity. (We use this fact in the noise analysis starting on page 42.) The linear bandwidth is proportional to intensity. The dependence of the low-frequency cutoff with intensity probably arises from the effect of scattered light on the adaptive element.

We can understand the transfer function by going to various limits of the parameters. The obvious limits obtain the steady-state and transient gains we computed earlier. Let us concentrate

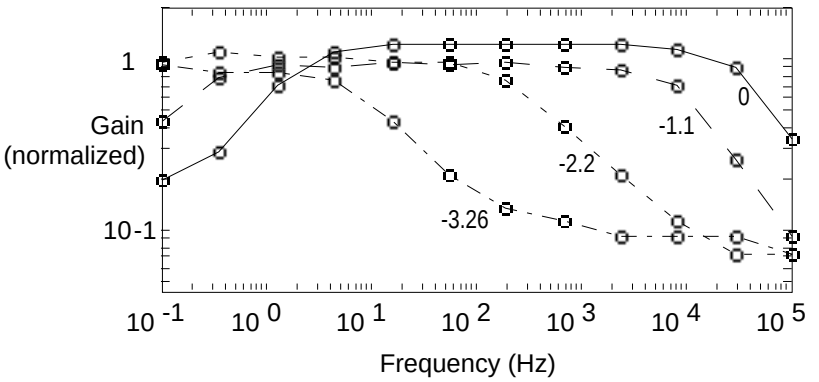


FIGURE 2.17 Measured amplitude transfer functions for the fed-back receptor with a photodiode constructed from a piece of native diffusion. The number by each curve is the log background irradiance, in decades. The highest irradiance is 19 W/m^2, about 10 times direct office-fluorescent lighting. The intensity affects both the high and low frequency cutoffs. The receptors have a constant gain over a range of 4–5 decades of frequency. At the highest background intensity, the gain is larger than at the other intensities, because the feedback transistor comes out of subthreshold, reducing its transconductance. We normalized the curves to the mean gain for the median intensities, about 1.4 V/decade for this receptor. This receptor has no Q_{cas} to nullify Miller capacitance.

our efforts on understanding the limits on the time-response, where the utility of the small-signal analysis becomes apparent. In the absence of adaptation ($\tau_f \to \infty$), H becomes

$$H = \frac{A'/\kappa}{\frac{A'/\kappa}{A}\left[(\tau_p + \tau_m)s + 1\right](\tau_o s + 1) + \frac{A'}{\kappa}\tau_m s + 1} \tag{10}$$

When the amplifier is infinitely fast ($\tau_o \to 0$), the transfer function becomes

$$H = \frac{A'/\kappa}{\frac{A'/\kappa}{A}\langle(\tau_p + (A+1)\tau_m)s + 1\rangle + A'} \tag{11}$$

This first-order system has a single pole on the negative real axis. The first-order time-constant is

$$\tau_{rise} = \frac{(C_p + (A+1)C_m)}{I_{bg}/V_T}\frac{1}{1 + \frac{\kappa A}{A'}} \tag{12}$$

In the limit $A \gg A'/\kappa$, we obtain

$$\tau_{\text{rise}} = \frac{(C_{\text{p}} + (A+1)\,C_{\text{m}})}{I_{\text{bg}}/V_{\text{T}}}\frac{A'}{\kappa A} \tag{13}$$

This expression again tells us that the response time is inversely proportional to the background intensity. We might have guessed this, because we have assumed that the amplifier is infinitely fast. We can also see how the Miller capacitance, C_{m}, is multiplied by A, so no matter how large we make A, we cannot neglect C_{m}. However, $C_{\text{m}} \ll C_{\text{p}}$, so we can neglect C_{m} over a range of A. We can also neglect C_{m} when we activate Q_{cas}. In this case, the time constant is inversely proportional to the amplifier gain A and proportional to the feedback gain A'. In other words, the higher the gain of the amplifier, the faster the receptor, but the more gain we want from the receptor, the slower it will be, for a given feedforward amplifier. These relationships make sense: The better we clamp the input node (with larger A) the faster the circuit. The more gain we want from the circuit (with larger A'), the more the input node must move.

Figure 2.18 shows measurements of the small-signal rise time plotted against the absolute intensity. We can see that the inverse relationship between rise time and intensity, predicted by Equation 12, is true over a wide range of intensities. The inverse relationship is only violated at the very highest intensities, where the photodiode receptors start to be limited by minority carrier diffusion lifetime (see page 38).

The units of intensity

We show the intensity axes on our plots in units of irradiance. **Irradiance** is defined as the energy per unit area per unit time incident on the receptor, and it is measured in $\text{W/m}^2\text{s}$. The amount of *visible* light incident on a surface is called **illumination**, and is related to the irradiance by the spectral visibility function of the eye, which we have shown in Figure 2a.6. Illumination is measured in lux. The connection between the two measures is that 1 W/m^2 equals 680 lux with photons of wavelength 555 nm—the wavelength of peak sensitivity (from [12], highly recommended for anyone dealing with converting between these units). People also use cd/m^2 as illumination units, where cd = candela. 1 cd/m^2 is 4π lux. The amount of energy *emitted* from a surface per unit area per unit time is called **radiance**; the corresponding visible quantity is **luminance**.

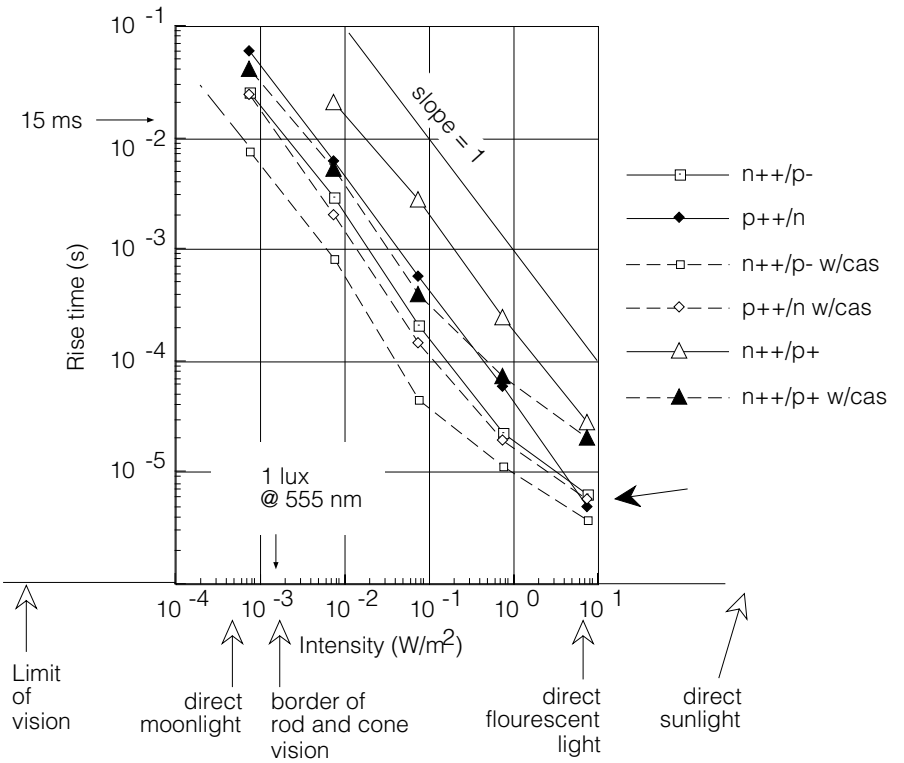


FIGURE 2.18 Response-time measurements for photodiode receptors. Each curve shows the rise time for a small step-intensity change, versus irradiance. The different curves are for different photoreceptors circuits, differing in the phototransducer and the use of the cascode. Keys (refer also to Figure 2a.3 on page 66): *x/y* means junction between *x* and *y*, where *x* and *y* are as follows: *p-* is bulk substrate, *n* is well, *p+* is p-type base layer, *n++* is n-type source-drain diffusion, and *p++* is p-type source-drain diffusion. *w/cas* means cascode is activated. The effect of minority-carrier diffusion lifetime can be seen at the solid arrow (➚).

The illumination of the receptor is different than the illumination of the scene. The two are related through the optics and the reflectivity of the scene. The irradiance of the receptors is the square of the numerical aperture of the optics, times the radiance of the scene. The numerical aperture is approximately the reciprocal of the lens *f*-number. The *f*-number, in turn, is the ratio between the lens diameter and the focal length. Assuming a typical lens with an *f*-number of 2.8, and a Lambertian scene reflectivity of 50 %, we compute that the receptors see an illumination that is about 16 times smaller than the scene illumination. This number may come in handy for estimating the usefulness of these photoreceptor circuits in real world situations.

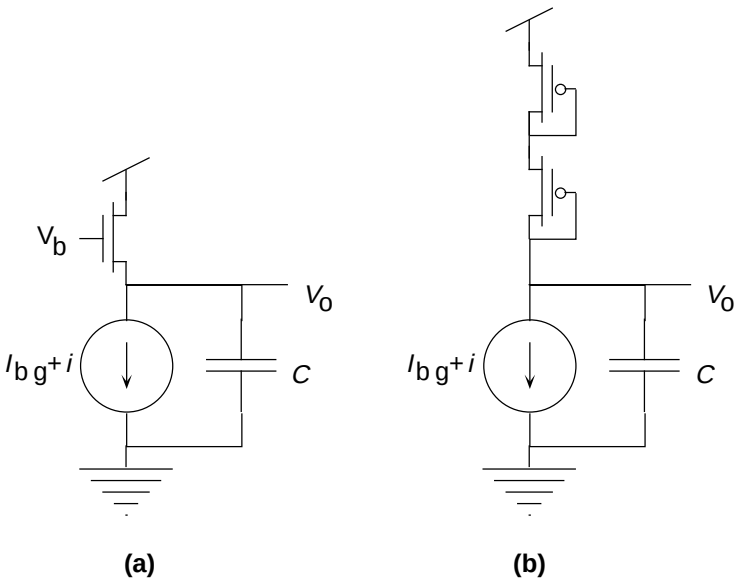


FIGURE 2.19 Two simple logarithmic receptors. **(a)** is a source-follower detector. **(b)** is the simple two-diode receptor used in many early designs.

Gain-bandwidth product

We can see from Equation 13 that the photoreceptor displays a typical trade-off between gain and bandwidth. However, by building a higher total loop gain into our adaptive receptor, we have increased both the gain and the bandwidth over the values for the previous logarithmic receptors. Let us compute the benefit.

Figure 2.19 shows two simple logarithmic receptor circuits that do not utilize active feedback. Part (a) shows the simplest current-mode logarithmic receptor, which consists of a single transistor with exposed source diffusion. Part (b) shows the classic logarithmic photoreceptor used in early silicon retinas [5][8][9]. In both these devices, the output is directly charged and discharged by the photocurrent. The first-order time constant and gain for each of the receptors are shown in Table 2.1

The response time of each of these receptors is again inversely proportional to the background intensity. For both these receptors, the gain is a factor of approximately $\frac{1}{A'}$ times smaller, and the response time is approximately $\frac{A}{A'}$ times larger than for the adaptive receptor with feedback. Hence, the gain–bandwidth product is smaller by a factor of approximately $\frac{1}{A}$ in these simple log-

Receptor	t	A
Source-follower (a)	$\frac{CV_T}{I_{bg}}$	1
Two-diode (b)	$\frac{1+\kappa}{\kappa}\frac{CV_T}{I_{bg}}$	$\frac{1+\kappa}{\kappa^2}$

TABLE 2.1 The first-order time constant τ and gain $A = (v_o/V_T)/(i/I_{bg})$ of the simple logarithmic receptors in Figure 2.19.

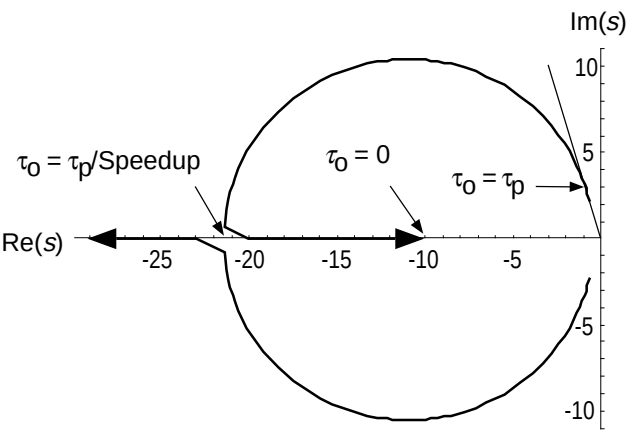


FIGURE 2.20 Root-locus plot for adaptive receptor, showing the poles of the transfer function in Equation 10, parameterized by the output time-constant τ_o of the feedback amplifier. Parameters: $A = 100$, $A'/\kappa = 10$, $\tau_p = 1$, $\tau_m = 0$.

arithmic receptors than in the adaptive receptor. The price we pay for the increased performance of the adaptive receptor is increased power consumption and circuit complexity.

Second-order behavior

In our computation of the expected speedup due to the active feedback clamping of the input node, we assumed that the feedback amplifier is infinitely fast. We can investigate this assumption with a root-locus plot. Figure 2.20 shows the locations of the poles of Equation 10 as the speed of the feedback amplifier is increased from slow to fast. In the infinitely-fast limit, the two poles of the second order system separate along the negative real axis. One pole shoots off to -∞, and the other ends up at the value derived earlier, corresponding to a speedup of A'/κA over the open-loop value. To achieve this speedup, the feedback must be very fast. If it is not, the poles will have

a nonzero imaginary part, and the output will ring. We can derive the condition for a damped, nonringing step response by finding the value of τ_o that makes the imaginary part of the poles equal to zero.

It is easiest to approach this problem from a canonical point of view for second-order systems [8]. A canonical form for the transfer function of a second order system is

$$H(s) = \frac{1}{\tau^2 s^2 + \frac{\tau}{Q}s + 1} \tag{14}$$

where, in the case of an underdamped system, $1/\tau$ is the radius of the circle on which the poles sit, and Q, stated loosely, is the amount of ringing in response to a step input. $Q = 1/2$ means a critically damped system. We can identify τ and Q in the transfer function for the photoreceptor, Equation 10, as follows:

$$\begin{aligned} \tau &= \sqrt{\tau_o \tau_p \frac{A'}{\kappa A}} \\ Q &= \sqrt{\frac{\kappa A}{A'}} \frac{\sqrt{\tau_o \tau_p}}{\tau_o + \tau_p} \end{aligned} \tag{15}$$

From these expressions we can easily solve for the $Q = 1/2$ condition:

$$Q = \frac{1}{2} \qquad \text{when} \qquad \tau_o = \frac{A'}{4\kappa A}\tau_p \tag{16}$$

For a nonringing, critically-damped response, the amplifier-output time constant must be smaller than the input node time constant by a factor proportional to the speedup. This restriction is severe, because the amplifier output is already a factor of A slower than it would be if the amplifier had unity gain. In other words, the amplifier generates high gain by using a small output conductance, and this small output conductance makes the amplifier slow. For a critically-damped response, the transconductance of the input to the feedback amplifier, and hence the bias current, scales as the *square* of the desired speedup.

We can also use Equation 10 to find the condition for maximum Q. The result is

$$Q = \frac{1}{2}\sqrt{\frac{\kappa A}{A'}} \qquad \text{when} \qquad \tau_o = \tau_p \tag{17}$$

If we bias the amplifier so that the amplifier output has a time constant equal to the time constant of the input node, then we obtain the maximum possible amount of ringing. We have labeled these conditions on the root-locus plot in Figure 2.20.

Usually we operate the receptor without regard to these distinctions, and turn the bias current up enough to give a response that is fast enough for the situation at hand, but slow enough to filter out flicker from artificial lighting.

Other limits on the response time

In this section, we discuss the other limits on the response speed of our photoreceptors that we have not included in our small-signal model.

Base capacitance in bipolar phototransistors. Comparison of photoreceptor circuits constructed with bipolar phototransistors and with photodiodes show that our assumption of a lumped capacitance from the input node to ground is incorrect in the case of phototransistors. Figure 2.21 shows that the circuits that we built using phototransistors are not sped up by the active feedback circuits. The missing element in the model is the internal base capacitance in the phototransistor. Even though our feedback circuit clamps the emitter voltage, the base-emitter and base-collector junctions must still charge or discharge in response to a change of intensity.

Minority carrier lifetime. The response time is inversely proportional to intensity over many orders of magnitude intensity (Figure 2.18). This observation rules out a limiting time-constant due to low-pass filtering from minority-carrier lifetime effects, since such effects would not limit responses made slow otherwise by small photocurrents. Also, minority-carrier lifetimes are usually measured in the literature in the microsecond range [18].

To place a limit on minority carrier effects, we compared the speed of two photoreceptors, using an infrared LED (950 nm) that generate many carriers deep in the substrate (Figure 2a.2). Both of the photoreceptors use a photodiode. One of the photodiodes is constructed from source-drain diffusion in the well (volume-limited junction, part n/p++ of Figure 2a.3), and the other is constructed from source-drain diffusion in bulk substrate (Layout is shown in Figure 2.16, and detector is n++/p-, similar to n/p- of Figure 2a.3). At all but the highest intensities (irradiance < 25 W/m^2), the diffusion-limited receptor is faster than the volume-limited receptor. The reason is that the diffusion-limited receptor has higher quantum efficiency (Figure 2a.6), especially for infrared light, and hence a larger photocurrent. At the highest intensity, the volume-limited receptor is slightly faster than the diffusion limited receptor. The reason is that the receptor response time starts to be limited by minority-carrier diffusion lifetime, and the volume-limited receptor,

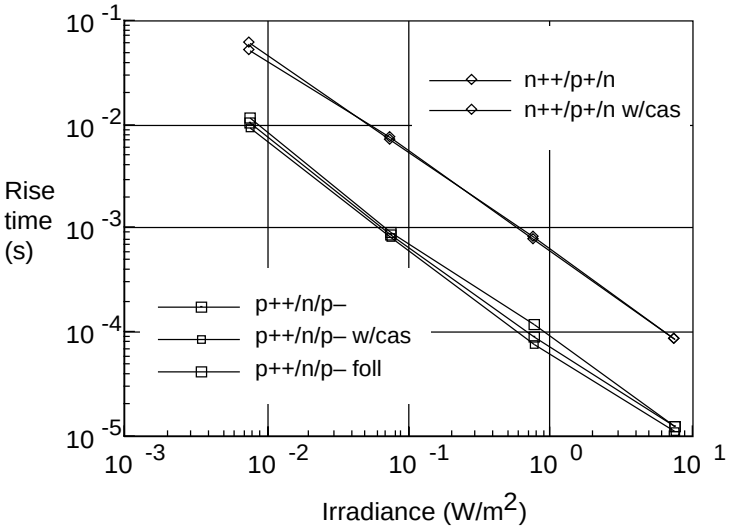


FIGURE 2.21 Effect of feedback circuit and cascode on receptors with bipolar phototransistors. The reference curve is the one labeled *p++/n/p– foll*. This receptor is the simple logarithmic receptor in Figure 2.19(a) with parasitic bipolar phototransistor. We can see that the other *p++/n/p–* curves are nearly identical. The *n++/p+/n* photoreceptor, which uses a vertical NPN phototransistor, is even slower. The speed of both configurations is not affected by the use of a cascode that eliminates Miller capacitance and provides additional gain.

with a smaller collection volume, is faster despite a smaller photocurrent. The crossover occurs at a receptor rise time of a few μs.

We can see the beginning of a similar effect in the rise time measurements shown in Figure 2.18, where an arrow points out the fact that the volume-limited receptor speeds up, relative to the diffusion-limited receptor, at the highest intensities. We can see from this measurement that the diffusion lifetime effects are only important for receptors that must respond on the microsecond time scale. This limitation is hardly important for models of biological vision systems, but possibly critical for specialized sensor applications.

Comparison of response time for different receptors

Figure 2.22 compares the response time of all the test receptors, under a particular set of conditions. The fastest receptor is the one that uses a simple source-diffusion photodiode, in conjunc-

tion with the cascode. This receptor is a factor of almost 100 faster than the slowest receptor, which is constructed from the vertical bipolar transistor. The fastest receptor is separated from the next-fastest by a factor of about 2 in speed.

The cascode, under these particular operating conditions, speeds up the response by a factor of only about 2–3. This small speedup is much less than we expect, given the following argument. The photoreceptor circuit with the layout shown in Figure 2.16, has a Miller capacitance C_m, computed from the Spice Level 3 parameters supplied by MOSIS, of 1.8 fF. The measured amplifier gain A is about 300 for a fast, 1.1 V, above-threshold, bias without the cascode, and a factor of 3–6 higher with the cascode activated. Thus, the effective Miller capacitance gets multiplied up to the substantial value of about 500 fF. Using the same set of MOSIS parameters, we estimate that the parasitic capacitance C_p on the input node of the photoreceptor is about 100 fF. This capacitance consists of the junction capacitance and the gate capacitance of the input to the amplifier. Hence, the cascode should speed up the receptor by a factor of 15 to 30. We had difficulty in making a direct measurement of this speedup in a step response at higher intensities. We believe the reason is that, at higher intensities, the input node is already so fast that we cannot bias the amplifier strongly enough to achieve the full speedup, because of the restriction imposed by Equation 16. However, we shall see in the later discussion of the noise characteristics that a direct measurement of the transfer function shows the full expected speedup (Figure 2.24).

THE ABSOLUTE OPERATIONAL LIMITS

What is the dimmest light these photoreceptor circuits can work with, and how does this value compare with commercial CCD cameras and biological rods and cones? What are the smallest signals that we can reliably detect with our photoreceptor circuits? What circuit and stimulus parameters determine each of these limits? We will discuss these important practical questions in turn.

The absolute illumination limit

As long as we can increase the speed and gain of the feedback amplifier, we can speed up the response by an arbitrary amount. We pay two prices: Increased power consumption, and degradation of the SNR. We can see from Equation 13 that, for a given amplifier, the speed of the receptor is proportional to the ratio $R \equiv I_{bg}/C_p$ of the photocurrent to the capacitance of the photodiode.

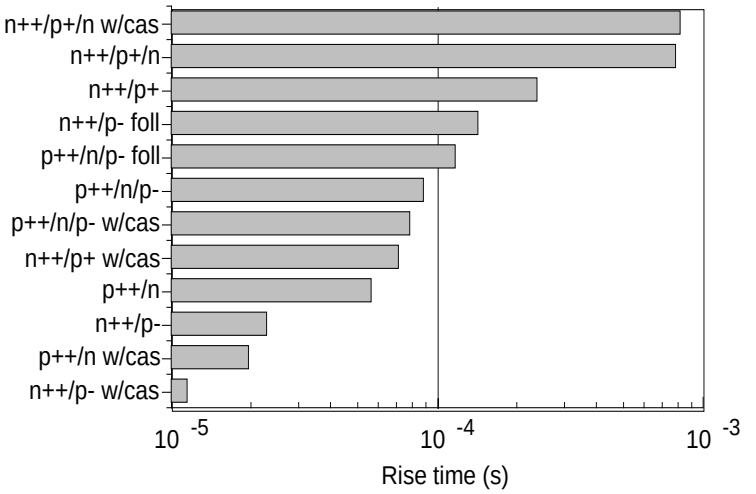


FIGURE 2.22 Rise times for different receptor configurations, in response to a small step-increase of intensity. Background illumination is about 0.1 W/m^2. Keys (refer also to Figure 2a.3 on page 66): x/x means single junction, x/x/x means bipolar transistor, where x's are as follows: p– is bulk substrate, n is well, p+ is p-type base layer, n++ is n-type source-drain diffusion, and p++ is p-type source-drain diffusion. *w/cas* means cascode configuration is active. *foll* means receptor is simple source-follower, like Figure 2.19(a) (but possibly with swapped device type).

The photocurrent is the quantum efficiency of the device times the flux. For a given illumination, R is a fixed process parameter. The ratio could be increased if we had access to PIN diodes, where the p and n regions are separated by an intrinsic region that increases the volume, and hence the quantum efficiency, and simultaneously decreases the capacitance of the device [18]. In the next section, we show that the noise properties degrade with decreased capacitance, so this solution must be balanced against the desire to detect small signals. In the devices available in a stock CMOS process, maximum R is obtained from one of the substrate-junction photodiodes.

A reasonable definition of the limiting intensity is the intensity at which the photoreceptor has a rise time of 15 ms, the same as a single frame of a video camera that scans at 60 Hz, and approximately the same as the cutoff frequency for human vision. We can see from Figure 2.18 that the fastest photoreceptor circuit has a rise time of 15 ms at about 1 mW/m^2 irradiance. This irradiance is equivalent to an illumination of about 1 lux.

CCD detectors are integrating devices; every frame, they dump out all the charge they collected since the last frame. The sensitivity limit is determined by the electron counting noise in the

charge-sensing amplifier. Currently, these amplifiers, in consumer end-product devices, function at about 100 electron noise level, meaning that the RMS noise in the output is equivalent to 100 electrons in the charge bucket. If we assume that the cameras must have at least 4 bits resolution to be acceptable, then the number of electrons collected must be 16 times the noise level, or 1600, each 1/60 s. A typical CCD pixel area is 100 μm^2. The quantum efficiency is about 30 %. From these numbers, we compute that the irradiance is 1.4 mW/m^2, or 1 lux at 555 nm, consistent with the advertised ratings of a few lux.

The borderline between rod and cone vision occurs at an illumination of about 1 lux, which is about bright moonlight level illumination. In Figure 2.18, we have labeled the rod-cone border and the moonlight irradiance.

In summary, the current photoreceptor circuit functions down to about the same intensities as consumer CCD cameras and human cone receptors.

The absolute detection limit

What determines the receptor detection capability? This question is closely related to the noise properties of the receptors. In this section we investigate the noise properties of the receptors and their detection ability, from both empirical and theoretical viewpoints. Electronic noise is mysterious, so we initially take an empirical approach to this discussion. This approach lets us eliminate many uncertainties, and greatly simplifies the later analysis.

Empirical observations of receptor noise

The first set of observations are shown in Figure 2.23. We captured the power spectra from a simple source-follower detector like the one shown in Figure 2.19(a). The measurements show that the noise spectra are essentially white. Flicker noise ($1/f$), surprisingly, is negligible. The total bandwidth is inversely proportional to intensity, as we derived earlier. Also, the total noise power appears to be constant, independent of intensity.

The second set of observations are shown in Figure 2.24, which compares the measured noise spectra of the simple source-follower logarithmic receptor and the fed-back adaptive receptor. In this measurement, we injected a small test signal to examine the SNR degradation by the adaptive feedback circuit. We can see from the SNR in the two signals the unexpected result that the adaptive feedback circuit adds less than 3 dB noise to the original signal. Hence, our analysis can treat

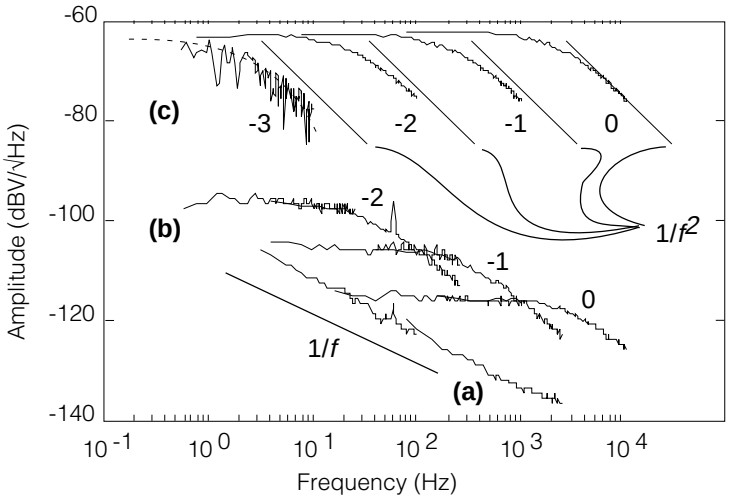


FIGURE 2.23 Noise spectra for source-follower logarithmic receptor, for different intensities. The curve labeled **(a)** is from an isolated follower pad, and shows that the measured 1/*f* instrumentation noise in the follower pad is smaller than measured spectral noise from the detector. The curves labeled **(b)** show the unstimulated noise spectra from the receptor. The number by each curve is the log background irradiance. 0 log irradiance is 1.7 W/m^2. The curves labeled **(c)** show the response of the receptor to small-signal white-noise stimulation. The straight lines have a slope of $1/f^2$, the same as from a first-order low-pass filter. Again, the number by each curve is the log background irradiance. For definition of dBV units, see Figure 2c.1.

the feedback and adaptation as a noiseless amplifier, resulting in an immense simplification. (These data also show the feedback circuit and the cascode widen the bandwidth; the extent of the widening, a factor of 1.5 to 2 decades, is in agreement with earlier predictions based on the parasitic capacitances and gain measurements (see page 40).)

Theory of receptor noise

Our goal in the following analysis is to understand the origin of the noise behavior of the source-follower detector. A full analysis of the noise properties of the receptor circuit could, in principle, turn out to be very complicated, since it would involve placing noise sources in parallel with all source-drains and in series with all gates, where the noise sources may in themselves not have flat spectral properties, and then computing the effective input noise based on all these noise

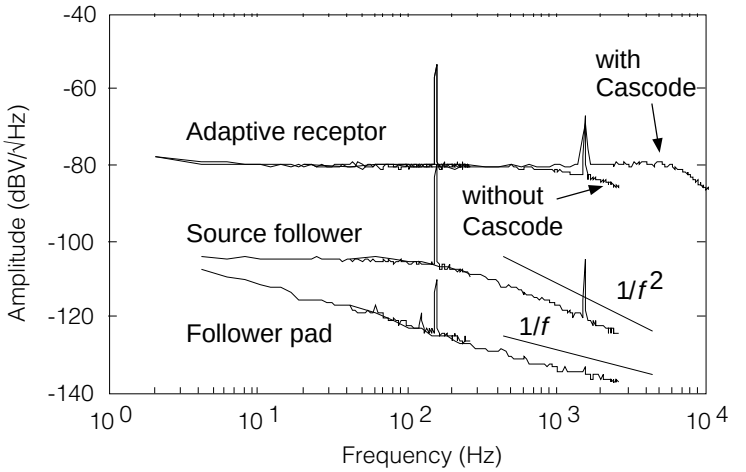


FIGURE 2.24 Comparison of noise spectra for simple source follower detector of Figure 2.19(a) and the adaptive receptor shown in Figure 2.16 and Figure 2.15. The gain of the source follower detector is about 60 mV per decade. The gain of the adaptive receptor is about 1.3 V/decade, or 27 dB more. The plots show the measured power (actually amplitude) spectra for each detector, along with the spectrum of the follower pad instrumentation. The DC irradiance for this measurement is 0.2 W/m^2, and the injected signal is a combination of a 150 Hz sinusoidal signal and a 1.5 kHz signal each with Rayleigh contrast of about 1 %. We can see from the height of the signal spike relative to the surrounding noise that the SNR for each detector are indistinguishable. For definition of dBV units, see Figure 2c.1.

sources. Fortunately, we observe experimentally that nearly all the noise in the circuit is generated in the input node by the photodiode and feedback transistor. This simplification means that we can compute an understandable theoretical expression for the noise in the input leg due to the major noise sources. In Appendix 2c, we derive a general expression for thermal and shot noise, expressed as a variability in the amount of charge on a capacitor that is thermally coupled to a source of charge. To compute the receptor noise, we apply that theory to the combination of current-generating photodiode and source-coupled exponential feedback transistor.

The variation in the output voltage precisely mirrors the fractional variation in the photocurrent and in the number of charges collected during an integration time. In equation form

$$\frac{\Delta v^2}{V_T^2} = \frac{\Delta i^2}{I_{bg}^2} = \frac{\Delta N^2}{N^2} = \frac{\Delta Q^2}{Q^2} \tag{18}$$

where Δx^2 means the mean-square variation of x. Notice once again how the thermal voltage appears as a natural scale under subthreshold operation.

A theoretical expression for shot plus thermal noise is derived in Appendix 2c. The expression is

$$\Delta Q^2 = qC(V + V_T) \tag{19}$$

where ΔQ^2 is the mean-square variation in the charge $Q = CV$ on the capacitance C. V is a little tricky; $\cdot V = \frac{Q}{C} = \frac{qN}{C}$ is the voltage on the capacitor corresponding to the number of charges N that flow in the circuit in one integration time, and not the total voltage on the capacitor. V expresses the shot noise component in units of voltage.

This beautiful expression is easy to apply in the present computation. But what is the charge Q? The total charge Q is the photocurrent integrated over one integration time τ:

$$Q = \tau I_{bg} \tag{20}$$

In our photoreceptors, the bandwidth scales with intensity, so the effective integration time scales reciprocally with intensity. We can compute τ knowing the capacitance of the input node and the conductance of the feedback transistor source:

$$\tau = X\frac{C}{g} = X\frac{CV_T}{I_{bg}} \tag{21}$$

where V_T is the thermal voltage. The receptor integrates the signal using an exponentially-decaying kernel backwards in time, with time constant τ. The factor of X comes from the fact that the integration time is actually longer than the time constant of the low pass filter, so the effective amount of charge is larger. The way to think of this is that the white noise source drives a low pass filter. The output of the low pass filter contains a certain amount of power. How much? That depends on the power density of the white noise source and on the shape of the low pass filter. The **effective noise bandwidth** of the low pass filter is the bandwidth of a flat filter that lets through the same power as the low pass filter. The reciprocal of the effective noise bandwidth is the correct integration time for white noise inputs. It turns out that the correction factor X for a simple, first-order, low pass filter is 4, so that the effective integration time and hence number of charges is really

$$Q = 4CV_T \tag{22}$$

This expression is valid for subthreshold feedback current. The measured photodiode current at the maximum irradiance we tested, 2 W/m^2, is about 120 pA, well within the subthreshold operating region for the feedback transistor. Hence, for all but bright sunlight conditions, the subthreshold approximation is valid. (Note that larger phototransducers can invalidate this assumption.)

We can see how the background intensity cancels, resulting in an effective charge that is a constant, independent of intensity. Combining Equations (18), (19), and (20), along with $Q = CV$, we obtain the following expression for the mean-square fractional variation in the measured current:

$$\frac{\Delta i^2}{I_{bg}^2} = \frac{5}{16}\frac{q}{CV_T} \tag{23}$$

The total noise power is a constant, proportional to the reciprocal of the capacitance. What could be simpler? In the clarity of hindsight, what other result could we possibly have obtained? The quantity CV_T is a given amount of charge; the only way to combine it with the elementary unit of charge and come out with a dimensionless number is with a ratio. The shot contribution comes from an integration time that is inversely proportional to intensity and proportional to capacitance, resulting in constant number of charges. The mean-square deviation always goes as the number of units. The thermal contribution comes from a fixed capacitance, and hence, a fixed thermal fluctuation. The factor of 5/16 comes from a shot contribution that is 4 times larger than the thermal contribution.

Receptor flicker noise

Let us now discuss flicker noise. (In this particular receptor, the white noise is clearly dominant, but this situation is due to the small capacitance on the input node.) We will show that the total flicker noise in the adaptive receptor is also an intensity-independent constant. Flicker noise is characterized by constant noise power per log frequency; any octave of bandwidth contains the same flicker-noise power. In our photoreceptor circuit, both the high and low cutoff frequencies are proportional to intensity (see Figure 2.17), so the total log bandwidth of the adaptive receptor is a constant.

In Appendix 2c, we show measurements of flicker noise as a function of bias current, and that the flicker noise is a fractional current noise at any one frequency is a constant, independent of bias

current in subthreshold operation. The total mean-square fractional variation in the current is given by an integral over the passband of the form

$$\frac{\Delta i^2}{I_{bg}^2} = \int_B \frac{K}{f} df = KB \tag{24}$$

where K is the flicker noise parameter, defined in Appendix 2c as the flicker noise power per e-fold frequency band, and B is the log bandwidth of the receptor, measured in e-folds. Once again, we arrive at a very simple expression for the total noise. Note that we compute the mean-square variation in the input-referred units of $\Delta i / I_{bg}$.

Let us now apply our flicker-noise theory to the actual receptor data, to see if the theory matches the data. We will consider two receptors—the source-follower logarithmic receptor in Figure 2.19(a), and the adaptive receptor in Figure 2.16. From the MOSIS layer-capacitance parameters for this run (N25Y), $C \approx 310$ fF and $C \approx 80$ fF, respectively. The capacitance of the source follower is larger because the input node is instrumented out to a follower pad. We already know by the direct measurements in Figure 2.23 that flicker noise is swamped by white noise in our receptors. Let us see if the measured flicker noise data in Appendix 2c are consistent with this observation.

We shall compute the frequency at which we expect the flicker noise to equal the white noise. From Figure 2.23, the white-noise amplitude at an intensity of -1 log units is about -106 dBV, corresponding to a fractional current-noise power equaling 3.9×10^{-8} /Hz. The flicker noise current is characterized in Appendix 2c to have a subthreshold fractional current-noise power of K/f per unit frequency, where $K < 10^{-7}$ for a 6-µm by 6-µm n-fet, the same-size transistor as is used in the source-follower receptor. Matching up the power, we find that the frequency at which flicker noise should equal the measured white noise is less than 10 Hz. In Figure 2.23, we see the power spectra for the receptor noise start to turn up at frequencies around this value, although the measurements are not conclusive. In any case, the value for K is an estimate that was obtained from a single transistor on a different run, so this entire computation should be taken as an order-of-magnitude argument. As such, it a reasonable explanation of why we do not observe dominant flicker noise.

In the adaptive logarithmic receptor, we can do the same computation. The capacitance is smaller by a factor of about 4 in the adaptive receptor, compared with the source-follower receptor, and hence the theoretical estimates of shot plus thermal noise are larger by this same factor. Hence, in the adaptive receptor, we expect the frequency where the flicker noise would appear to

be lower by a factor of about 4 than in the source-follower receptor. From Figure 2.24, we can see that in the adaptive receptor, flicker noise also does not appear within the measurement bandwidth, which extends down to about 1 Hz.

We can compute the total theoretical estimate for the flicker noise of the adaptive receptor as follows. By inspection of Figure 2.17, the log bandwidth is about 4 decades; in e-fold units $B \approx 9$. Hence the theoretical RMS flicker noise is

$$\frac{\Delta i}{I_{bg}} = \sqrt{KB} < 0.001 \tag{25}$$

which is smaller by a factor of at least 3 in amplitude, and 9 in power, than the theoretical shot plus thermal noise that we will calculate next. Based on the flicker noise estimates just computed, and on the measured dominance of white noise, we will ignore flicker noise from now on.

Receptor noise comparison with theory

Let us now consider the total RMS noise. We will start with the source-follower receptor. Using the capacitance $C = 310$ fF given earlier, we compute a theoretical RMS fractional variation, from Equation 23, of

$$\frac{\Delta i}{I_{bg}} = 0.0025 \quad \text{(Source follower, theory)} \tag{26}$$

Hence the theoretical RMS variation in the output due to thermal and shot effects is a constant 0.25 %, independent of intensity. The data in Figure 2.23 show the noise spectra for different intensities for the source-follower receptor. From these data, we can compute the total RMS noise in the receptor output. Consider the noise spectrum taken at -1 log units of background intensity. The 3 dB frequency is about 300 Hz, and the noise level is about -106 dBV. These numbers translate to a total RMS noise

$$\frac{\Delta i}{I_{bg}} = 0.0034 \quad \text{(Source follower, experiment)} \tag{27}$$

Our receptor does pretty well! Our theory predicts a noise of 0.25 %, and we measure only 0.34 %. The order of magnitude is correct, but our theoretical noise estimate is too low by a factor of about 0.74 (a factor of 0.55 in power). Still, this estimate is not bad. It could be that the capacitance is actually larger than we think. A smaller capacitance would increase the theoretical noise

estimate. It could also be that other noise sources, not accounted for in our analysis, are contributing the missing noise power.

We can do the same computation of total RMS noise for the adaptive receptor. The result is the theoretical prediction that the total RMS noise should be

$$\frac{\Delta i}{I_{\text{bg}}} = 0.005 \quad \text{(Adaptive receptor, theory)} \tag{28}$$

From Figure 2.26, we see that the total measured RMS noise is about 8 mV. The gain of the receptor is about 1.6 V/decade, which is the same as 0.695 V/e-fold, so the fractional variation is

$$\frac{\Delta i}{I_{\text{bg}}} = \frac{8 \text{ mV}}{0.695 \text{ V}} = 0.0115 \quad \text{(Adaptive receptor, experiment)} \tag{29}$$

This time, the measured noise is larger by a factor of 2.3 in amplitude, and 5.3 in power, than the theoretical estimate. We cannot account for this rather small discrepancy.

In summary, our theoretical estimates of the absolute values for receptor noise properties are close to within a factor of 3 of the measured values.

Effect of limited bandwidth on receptor noise

Let us now consider the effect of limiting the bandwidth of the detection, by using our control of the amplifier bias current. So far, we have shown that the total noise power is an intensity-independent constant, and that the power is spread over a bandwidth that is inversely proportional to intensity. This effect can be seen in Figure 2.25, which shows the noise spectra from the adaptive receptor for different intensities. When we artificially limit the receptor bandwidth, we reduce the total noise proportionally. This effect can be seen in the measured total RMS noise voltage versus intensity shown in Figure 2.26. For low intensities, the total RMS noise is a constant. For higher intensities, the noise power decreases as the reciprocal of the intensity. For detection, the implication is that, if we want to detect signals only below some cutoff frequency, then the brighter the light, the better we do. However, detection capability only goes as the square-root of the intensity, and the theory only holds as long as the signal is within the passband of the receptor. When the intensity becomes small enough, the signal is filtered away, and we can no longer detect it.

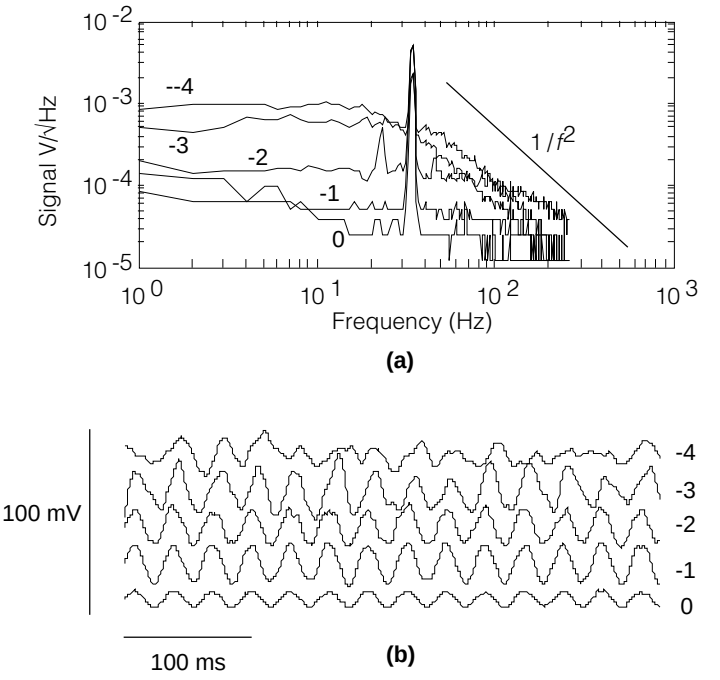


FIGURE 2.25 The spectrum of the output and measured response of the adaptive receptor for different background intensities. V_b is tuned to limit the passband to about 100 Hz. The signal is a small intensity variation around a large mean value. The log intensity of the mean value is shown by each curve. The Rayleigh contrast of the signal is about 3 %; 0 log intensity is about 2 W/m^2 irradiance. **(a)** shows the power spectra; the signal appears as a spike. A line shows the slope corresponding to a first-order low-pass filter. **(b)** shows representative traces corresponding to the spectra in (a).

The Minimum Detectable Signal and the Signal to Noise Ratio

Our definition of the minimum detectable signal (MDS) is a signal that results in an SNR of 1. In other words, the RMS signal amplitude is the same as the total RMS noise amplitude. Intuitively, this signal results in a barely detectable variation in the output. By this definition of the MDS, a plot of total RMS noise, like the one shown in Figure 2.26, is also a plot of the MDS. Hence, Equation 29 predicts that the MDS is about 1.25 %. We measured the MDS as a function of background intensity; the results are shown in Figure 2.27. Also shown in this figure are an example of the subjective measure of observability, and the power spectrum consisting of the signal plus the noise. We can see from the measured MDS that our theory is approximately correct. When we use a bias current that results in maximum-bandwidth operation, the MDS is a constant, as long as the signal is within the passband. When we limit the bandwidth, the MDS decreases as the 1/4 power

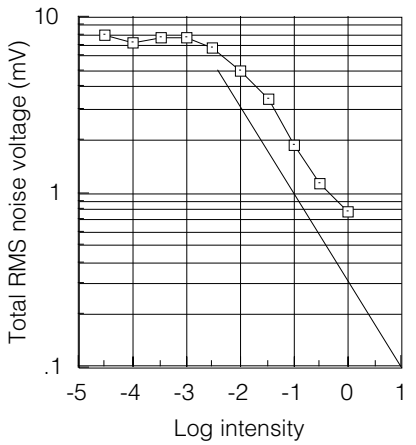


FIGURE 2.26 The total RMS noise of the adaptive receptor, as a function of intensity, when the bias current is used to limit the bandwidth. For intensities lower than the intensity corresponding to the bandwidth limit, the noise power is constant. The line parallel to the data has a slope of –0.5 on this log scale. For intensities higher than the intensity corresponding to the bandwidth-limit, the noise power goes as the reciprocal of the intensity. 0 log intensity is an irradiance of about 1 W/m^2.

of the intensity. We are not sure why the power law is 1/4 and not 1/2, but it could be due to the subjectiveness of the measurement.

Finally, since everyone talks about the SNR, we should compute it for our receptor. Manufacturers have learned that the SNR depends on how it is defined. Following their lead, let us compute an optimistic estimate. We measure the SNR as the ratio of maximum signal power to total noise power. Unlike the case for a linear detector, the maximum signal for a logarithmic detector is defined less precisely; from the operating curves shown in Figure 2.13 it is about a volt, corresponding to about a decade of intensity or 2.3 e-folds. For the least-favorable case of maximum-bandwidth operation, the RMS noise, from Figure 2.26, is about 8 mV, or 0.0125 e-folds. Hence the maximum SNR ratio is an amplitude ratio of about 190, or 45 dB. This figure is respectable, but quite a bit lower than good CCD camera charge-sensing amplifiers, which can achieve 60 dB [15].

Geometrical factors affecting the detection performance

We have seen theoretically that the total noise power scales as the reciprocal of the input capacitance C, and also linearly with the flicker noise parameter K. The implications are straightforward: To build a more sensitive detector, we need only increase C and decrease K. The trivial solution of just sticking a capacitor on the input node is not very satisfactory, because it is equivalent to simply low-pass filtering the signal. If we low-pass low enough, we can make the noise as small as we want. It is more sensible to talk about the noise power per unit bandwidth—what we

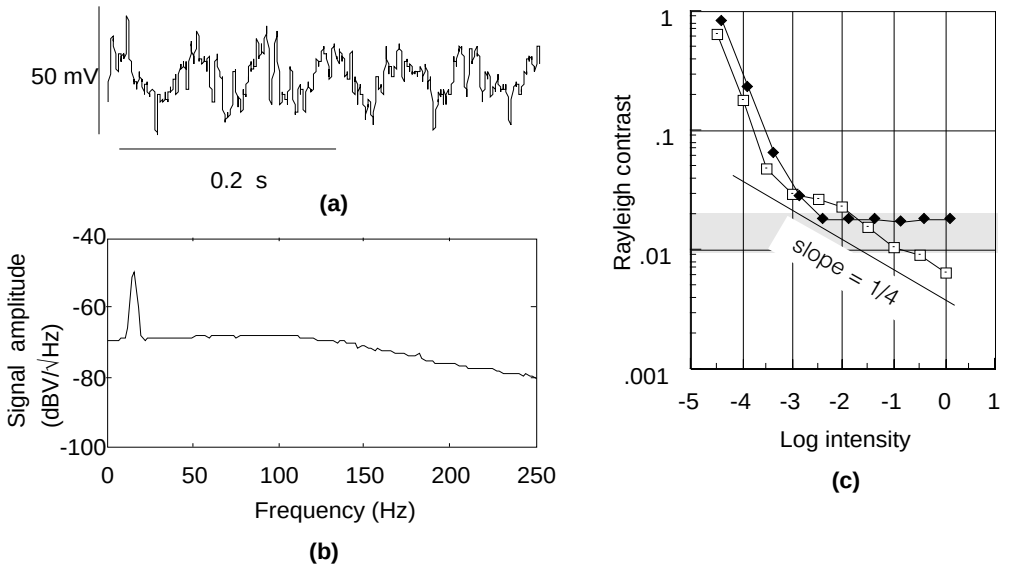


FIGURE 2.27 Measured MDS as a function of intensity. **(a)** shows the criterion for subjective observability. **(b)** shows the power spectrum corresponding to (a). **(c)** is the MDS as a function of log irradiance; 0 log is 2 W/m^2. The Rayleigh contrast is the amplitude of the sinusoidal intensity variation (0.5 of the peak-to-peak variation), divided by the mean intensity. The two plots in (c) are for bandwidth-limited (solid diamonds) and maximum-bandwidth (open squares) operation of the receptor. Shaded area in (c) shows the approximate measured human performance limits [2]

will call the **spectral noise density**. This parameter determines how well we can detect a signal in a small but known frequency band.

If we increase the area of the phototransducer, we scale the photocurrent just as we scale the capacitance. We decrease the total noise power without giving up bandwidth. Hence we decrease the spectral noise density. This observation is a good reason to build large phototransducers.

The flicker-noise parameter scales inversely with transistor size.† When we make the phototransducer large enough, the white noise will eventually be dominated by the flicker noise. At this point, it pays to make the feedback transistor larger. Our data shows that in practice, flicker

† K scales proportional to the reciprocal of the transistor area or the squared transistor length. Whether whether area or squared length is not really determined yet; see reference [2] of Appendix 2c for good measurements of this phenomenon.

noise is negligible even with large photodetectors. We will ignore flicker noise in the following calculations.

We can quantify the preceding discussion as follows. The capacitance on the input node consists of two parts, the photodiode capacitance C_d, and the other parasitic capacitance, C_p. The bandwidth B of the receptor is proportional to the photocurrent, and inversely proportional to the total capacitance:

$$B \propto \frac{I_{bg}/V_T}{C_d + C_p} \tag{30}$$

The total noise power P is inversely proportional to the total capacitance:

$$P \propto \frac{q}{V_T}\left(\frac{1}{C_d + C_p}\right) \tag{31}$$

The spectral noise density S is

$$S = \frac{P}{B} \propto \frac{1}{I_{bg}} \tag{32}$$

This is an interesting result. It says that the SNR is inversely proportional to the photocurrent, regardless of stray capacitance.

To see if our predictions were born out, we designed a chip with 4 receptors of different geometries, as shown in Figure 2.28(a). There are four receptors, arranged in a conceptual 2 by 2 matrix with all combinations of large and small photodetector and large and small feedback transistor. The area of the large detectors is four times the area of the small detectors. The area of the large feedback transistor is also four times the area of the small feedback transistor. We measured the noise in these receptors; the results are shown in Figure 2.29. As before, we injected a test signal to directly measure the SNR at a particular frequency. We can make the following observations from this data.

- Flicker noise is negligible. The noise properties do not depend on the size of the feedback transistor. The noise, for all these receptors, is dominated by the white noise shot and thermal processes.
- The gain of all the adaptive receptors are identical, as we expect.

- The SNR of the receptors with the large photodetector is 5 dB (a factor of 3) larger than those with the small photodetector, very close to the theoretical prediction of 6 dB (a factor of 4) from Equation 32. The discrepancy is probably caused by additive noise from the feedback amplifier. In a separate measurement, we determined that this additional noise is about the same magnitude as the noise from the receptor with the large detector, at this intensity.
- The effects of stray capacitance are substantial. Quadrupling the photodiode area doubles the bandwidth. The stray capacitance is of the same magnitude as the photodiode capacitance.

We can conclude that our theory is correct: the spectral noise density scales inversely with the photodetector area.

Noise performance of photodiodes and phototransistors

Based on the theory just presented, how do we expect the noise of photodiode and phototransistors to compare? We already have discovered that photodiodes have the important advantage that using them, we can speed up the response of the receptor with the active feedback

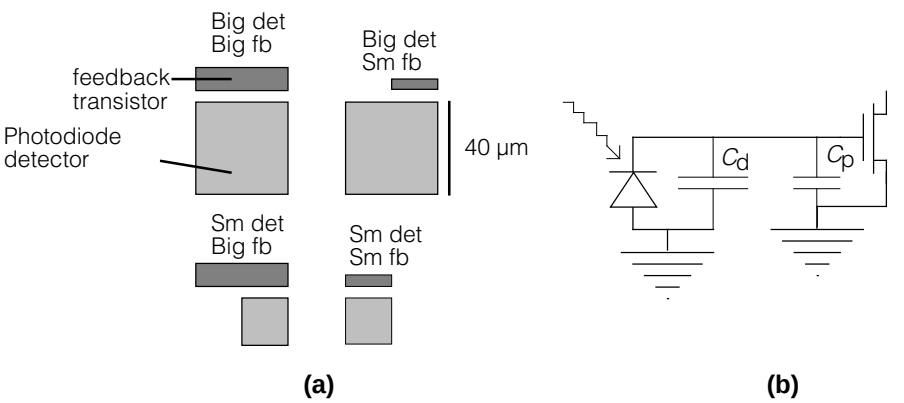


FIGURE 2.28 Receptor geometries used in test of noise scaling properties, along with capacitance model. **(a)** The four receptors have the geometries shown for the photodetector and for the feedback transistor. All other receptor components (capacitors, amplifier transistors, feedback element, etc.) were kept constant. Dimensions are shown by the scale bar; the large diode measures 40 by 40 μm^2. The notation *Big det, Sm fb* means big detector, small feedback transistor, etc. **(b)** Model of the input-stage capacitance, consisting of the photodiode capacitance C_d and the other fixed parasitic capacitance C_p (gate input C_g plus Miller effect C_m). From MOSIS data, $C_{d,big}$ = 360 fF, $C_{d,sm}$ = 94 fF, $C_p = C_g + C_{m=}$=77 fF +150 fF.

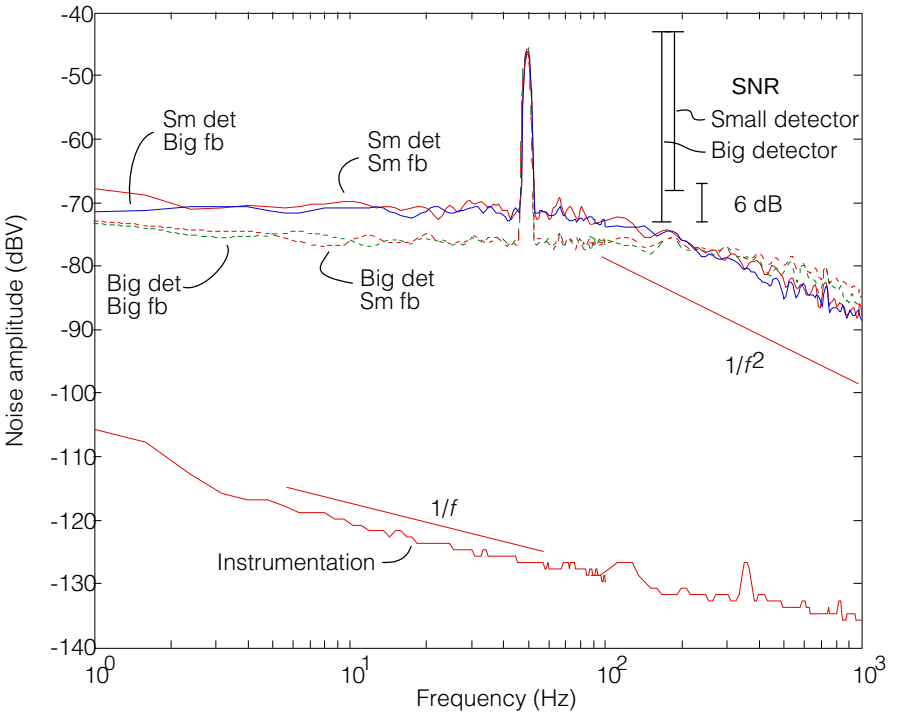


FIGURE 2.29 Power spectra results of noise measurements on the receptor geometries shown in Figure 2.28. The photodiode detector size has the dominant effect on the noise. The feedback transistor effect—if any—is lost in the measurement uncertainty. The straight lines show the power spectra of a first-order lowpass filter and of a flicker noise source. The noise spectra of the receptors are flat and well above the noise from the instrumentation (which comes from the pad driver and other sources.) We injected a test signal to directly measure the SNR. The gain of all of the adaptive receptors is identical, as shown by the test signal power. The use of detectors with quadrupled area improves the SNR by about 5 dB, close to the predicted value of 6 dB.

circuit. It could happen that we give up noise performance by using photodiodes. Perhaps there is something intrinsically less noisy about the gain mechanisms in bipolar transistors, or perhaps the increased current output from a bipolar transistor, compared with the current from a diode, decreases the noise in the feedback circuit. In fact, it turns out that the noise properties of phototransistors are indistinguishable from those of photodiodes, at least in the context in which we use these devices. We can see this result in Figure 2.30

This behavior is easy to understand, once we realize that the noise is dominated by the shot and thermal sources discussed earlier. In the photodiode receptor, the shot and thermal effects

appear as a result of the capacitance of the diode and the conductance of the feedback transistor. In the phototransistor detector, the shot and thermal effects arise not in the feedback transistor, but rather in the phototransistor itself, at the forward-biased, base–emitter junction. It is true that the output current from the phototransistor is much larger than from the photodiode, but the dominant noise source comes from the base–emitter junction, which is driven by the same photocurrent that drives the photodiode.

The noise effect of the adaptive feedback circuit

Earlier we noted the empirical observation that the feedback and adaptation circuit added so little noise to the signal that we could neglect that part of the circuit in the noise analysis. This simplification made the analysis intuitive and simple. Why is the noise in the feedback circuit so small? Clearly, the dominant noise source is the input node of the receptor. In this section, we want to answer the question of why the amplifier contributes so little extra noise, when it is in fact constructed from the same materials and is running at comparable speed to the input node.

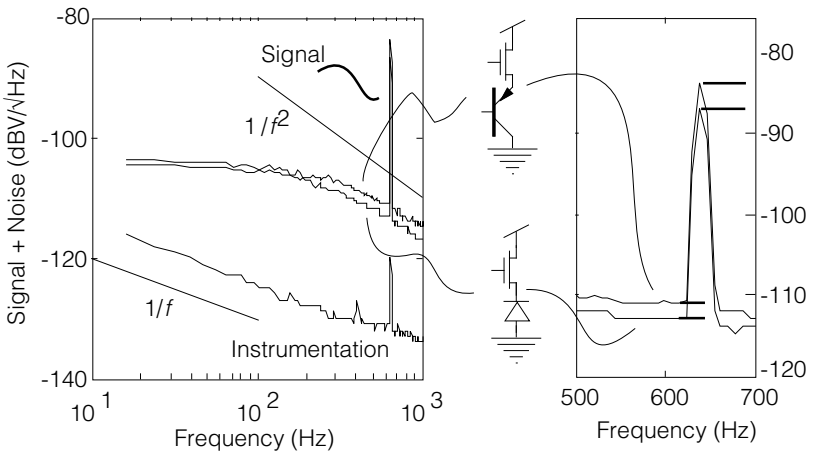


FIGURE 2.30 A direct comparison of phototransistor and photodiode noise properties. The plots show the power spectra of two source-follower detectors, one using a phototransistor, the other using a photodiode. Both detectors are illuminated with the same source, consisting of a steady-state background mixed with a small sinusoidal signal. The two devices have nearly indistinguishable noise properties, as shown in the blowup on the right. The signal-to-noise ratios are identical, and the only difference in the noise spectra results from the slightly faster response of the phototransistor detector. 1/*f* instrumentation noise is negligible, and the noise spectra are white, shaped by the low pass cutoff.

We can see that all of the analysis of the source-follower detector applies straightforwardly to the amplifier. A direct measurement of the noise in an isolated amplifier shows that the noise spectrum is mostly white, suggesting the same shot and thermal sources as in the source-follower detector. Just as in that detector, we know that the noise comes from charge fluctuations in the output node that are seen as voltage fluctuations. The same idea as before applies about an integration time and a consequent characteristic number of charges. As a result, we can immediately write that the charge/current fluctuation on the output node of the amplifier is simply

$$\frac{\Delta i^2}{I_{\text{bg}}^2} = \frac{\Delta Q^2}{Q^2} = \frac{5}{16}\frac{q}{CV_{\text{T}}} \tag{33}$$

This is exactly the same noise magnitude as for the input node, except for the size of the node capacitance—hence we have called it C'. One might now wonder what role does the high voltage gain of the amplifier play—does this voltage gain modify how the noise affects the detection capability? The answer is no: The amplifier noise is a *fractional* variation in the signal, just as it was in the case of the input node of the receptor. If the capacitance is of equal magnitude, then the noise will be too. This jibes neatly with the 3 dB difference in the SNR seen in the measured data in Figure 2.21, between the noise in the input node and the noise in the entire receptor.

BIOLOGICAL PHOTORECEPTORS

Our approaches to phototransduction and amplification were inspired by both biological and engineering principles. Engineers know that dynamic range and sensitivity are of paramount importance in the design of systems that deal with real-world situations. Of course, evolution figured that out long ago, but left us the job of reverse engineering a working system. Let us very briefly discuss the approaches that biological photoreceptors use to deal with two problems, the simultaneous need for high sensitivity and large dynamic range, and the requirement of rapid response time, invariant to lighting conditions. We can see if our electronic receptors deal with these problems in the same way as do the biological receptors.

Gain control

Biological rods and cones use the membrane voltage as a state variable that regulates transmitter release. However, we know that this same state variable is not used to store the history of

the signal—this would be paradoxical. Biological systems have the advantage that they can use a chemical concentration as a state variable, and hence can achieve a wide range of adaptation time scales by using different chemical reaction rates. Cones, and some rods, *do* temporally adapt, although there is not such a dramatic difference between transient and steady-state gain as in the photoreceptor circuits we showed in this chapter. See [3] for measurements of fly photoreceptors that also have curve-shifting temporal adaptation.

The operating curves for turtle cones are shown in Figure 2.31. We can see that in these receptors, the ratio between the transient gain and the steady-state gain is only about 5, compared with the ratios of 10 to 20 that we usually build. However, it is known that the synapse from receptor to bipolar cell adds another gain of 3, resulting in a total gain ratio of about 15. We do not know the significance of the biological value, although it is probably matched somehow to scene statistics. Our engineering choice of 10 to 20 comes from a desire to generate a large signal, a convenient layout, and a desire for some speedup.

The mechanism of photoreceptor adaptation is an active field of study (for a review, see [19]). In a popular working-model, a key ingredient in the adaptation mechanism is in the interaction of two state variables, the concentrations of cyclic GMP and of calcium. The feedforward state-variable is the concentration of cyclic-GMP, which determines the number of open channels in the membrane. The absorption of light by a pigment molecule causes a cascade reaction that chews up cyclic GMP. Cyclic-GMP is resynthesized by an enzyme that is controlled by the feedback state-variable, the calcium concentration. When the channels close, the concentration of calcium is quickly reduced by ion pumps. In terms of our model of information processing, we can think (rather loosely) as the calcium concentration storing the model of the intensity. A complete model of the feedback loop, however, does not exist yet and it is probable that this working model is lacking in many respects.

Time-constant control

In our electronic photoreceptors, the minimum response time is inversely proportional to the intensity: $\tau \propto I^{-1}$. Of course, by using the bias control on the amplifier, we can (and often do) limit the response time to filter out artificial lighting. It is very interesting that biological photoreceptors have a response time that is practically invariant to the light intensity, over a wide range of intensities. Experimentally, the relationship is approximately $\tau \propto I^{-1/4}$ in toad rod outer segments

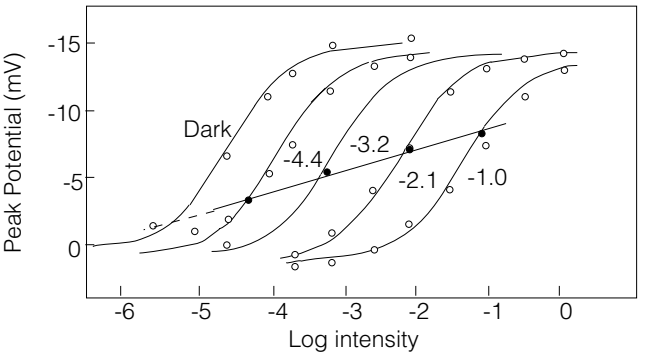


FIGURE 2.31 Gain adjustment in turtle cones (recorded intracellularly) caused by background illumination. The stimuli are 0.5 s increments or decrements on a steady background (except for the curve for the dark adapted cone which only is for increments). The stimulus spot is 3.2 mm in diameter on the retina. Peak responses measured from the dark-adapted resting potential (dotted line) are plotted as a function of test illumination. The thin curves connect the measured points. The thick curve is the steady membrane potential measured at least two minutes after background onset. The average slope of the transient responses is 9.5 mV/decade, and the slope of the steady-state, adapted response is about 1.8 mV/decade. The ratio of transient gain to the steady-state gain is about 5. The total dynamic range is about 15 mV. The illuminations are given as log attenuation from a baseline value. The unattenuated test stimulus (0 log) is 6.4 10^{15} quanta(640 nm)$(cm^2s)^{-1}$ on the retina, equivalent to an irradiance of about 20 W/m^2. (Direct office fluorescent lighting is about 1 W/m^2.) The unattenuated background illumination is 9.1 10^{15} quanta(640 nm) $(cm^2s)^{-1}$. Source: Adapted from [14].

[1]. The rod time-to-peak varies over a factor of about 4 over a factor of 200 in background intensities. A similar relationship holds in cone receptors.

It is easy to make the mistake of claiming knowledge of the survival value of a particular behavior. There are many possible functional reasons for this invariance. For example, it could be that this invariant is valuable for time-domain image processing, or perhaps to control power consumption, or possibly to control noise. Perhaps it is only a by-product of the transduction mechanism. In any case, how does this invariance come about? One thing is for sure: Not the way we do it in our electronic receptors! In the currently popular working model, the enormous gain generated in rods comes from about four amplification stages, each with modest gain (see [13] and [17] for very nice reviews). The gain of each stage is at most a few hundred, but the combination of all stages results in a maximum gain that closes 10^6 ion channels in response to a single photon.

The gain of the rod is inversely proportional to intensity. The mechanism for this gain control probably lies in a modest gain reduction of each stage of the amplification. A small change in the gain of each stage results in a large change in the total gain. Each stage has a fixed gain–bandwidth product: When the gain is reduced, the response time decreases proportionally. The trick is that the total gain goes as the *power* of the number of gain stages, and the total time response goes *linearly* with the number of gain stages.

To understand this argument better, let us assume, artificially, that each stage of amplification has the same gain. (Actually, some stages have gains of a hundred or more, and others have closer to unity gain.) Mathematically, if the total gain is A, then the gain of each of n stages is $\sqrt[n]{A}$. The response time of each stage is proportional to the reciprocal of the gain of each stage, and the total response time is proportional to the number of stages. Hence, the total response time is proportional to $\sqrt[n]{A}$. The total gain is proportional to the reciprocal of the intensity: $A \propto 1/I$. Hence, the response time goes like $\tau \propto I^{-1/n}$. When $n = 4$, we obtain the observed behavior. The argument just given is a simplified form of the Fuortes-Hodgkin model, reviewed in [17].

In biological receptors, the time–constant-control properties arise as a property of the way that gain control is implemented. In our silicon receptors, the gain is generated in a single stage of amplification, and the time constant is dominated by this stage. There does not seem to be much that we can do about it. We considered a cascade of amplification stages based on Darlington-connected bipolar transistors. At first sight, this scheme seems attractive, because bipolar devices are electrically quiet, compact, current gain devices. The problem is that the system is dominated by the slow time-response of the first stage of transduction, and we know of no mechanism to modulate the current gain.

SUMMARY

We have seen how to design compact photoreceptor circuits that have high sensitivity and wide dynamic range. The receptors use adaptive elements that adapt on different time-scales, depending on the size of the signal. We discussed novel adaptive elements that are resistant to junction leakage effects. We found that photodiodes give another decade of dynamic range than phototransistors, without sacrificing other performance characteristics. We showed that using a simple cascode results in yet another decade of range. We analyzed the noise properties, and obtained a full expression for the receptor noise that matches well with measured noise. We dis-

covered the circuit and stimulus parameters that control the detection limits. We saw how the electronic approaches are inspired biologically, but also how they differ in detail.

ACKNOWLEDGEMENTS

Carver Mead realized the need for adaptive circuits and originated much of the work in this chapter. Buster Boahen suggested the cascode configuration. David Van Essen impelled the noise measurements. Rahul Sarpeshkar provided insight on the operation of the new expansive adaptive element.

REFERENCES

1. D.A. Baylor, G. Matthews, and K.-W. Yau, "Two components of electrical dark noise in toad retina rod outer segments," *J. Physiol.*, vol. 309, pp. 591–621, 1980.

2. T. Delbrück and C.A. Mead, "An electronic photoreceptor sensitive to small changes in intensity," in *Advances in Neural Information Processing Systems 1*, D.S. Touretzky, Ed., San Mateo: Morgan Kaufman, pp. 720–727, 1988.

3. M. Dörrscheidt–Käfer, "Die Empfindlichkeit einzelner Photorezeptoren im Komplexauge von *Calliphora erythrocephala* (The sensitivity of single visual receptors in the compound eye of the blowfly," *J. Comp. Physiol.*, vol. 81, pp. 309–340, 1972.

4. M.A. Mahowald, "Silicon Retina with Adaptive Photoreceptor," *Proc. SPIE/SPSE Symposium on Electronic Science and Technology: from Neurons to Chips*, vol. 1473, pp. 52–58, April 1991.

5. M.A. Mahowald and C.A. Mead, "Silicon Retina," in *Analog VLSI and Neural Systems*, by C. Mead, Reading: Addison-Wesley, pp. 257–278, 1989.

6. M.A. Mahowald and T. Delbrück, "Cooperative stereo matching using static and dynamic image features," in *Analog VLSI Implementation of Neural Systems*, C. Mead and M. Ismail, Eds., Boston: Kluwer Academic Pub., pp. 213–238, 1989.

7. J. Mann, "Implementing Early Visual Processing in Analog VLSI: Light Adaptation," *Proc. SPIE/SPSE Symposium on Electronic Science and Technology: from Neurons to Chips*, vol. 1473, pp. 128–136, April 1991.

8. C.A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison--Wesley, 1989.

9. C.A. Mead, "A Sensitive Electronic Photoreceptor," In *1985 Chapel Hill Conference on VLSI*, H. Fuchs, Ed., Rockville: Computer Science Press, pp. 463–471, 1985.

10. C.A. Mead, "Adaptive Retina," in *Analog VLSI Implementation of Neural Systems*, C. Mead and M. Ismail, Eds., Boston: Kluwer Academic Pub., pp. 239–246, 1989.

11. C.A. Mead and M.A. Mahowald, "A Silicon Model of Early Visual Processing," *Neural Networks*, vol. 1, pp. 91–97, 1988.

12. J.R. Meyer-Arendt, “Radiometry and Photometry: Units and conversion factors,” *Applied Optics,* vol. 7, pp. 2081–2084, 1968.

13. K. Nakatani and K.W. Yau, “Sodium-dependent calcium extrusion and sensitivity regulation in retinal cones of the salamander,” *J. of Physiology,* vol. 409, pp. 525-548, 1989.

14. R.A. Norman and I. Perlman, “The effects of background illumination on the photoresponses of red and green cones,” *J. Physiol.,* vol. 286, pp. 509–524, 1979.

15. E. Oda, K. Nagano, T. Tanaka, N. Mutoh, and K. Orihara, “A 1920(H) x 1035(V) pixel high-definition CCD image sensor,” IEEE J. Solid State Circuits, vol. 24, pp. 711–717, 1989.

16. L. Nielson, M. Mahowald, and C. Mead, “SeeHear,” in *Analog VLSI and Neural Systems*, by C. Mead, Reading: Addison-Wesley, chapter 13, pp. 207–227, 1989. (adapted there from 1987 International Association for Pattern Recognition, 5th Scandinavian Conference on Image Analysis.)

17. R. M. Shapley and C. Enroth-Cugell, “Visual adaptation and retinal gain controls,” in *Progress in Retinal Research,* N. Osborne and G. Chader, Eds., New York: Pergamon Press, vol. 3, pp. 263–346, 1984.

18. S.M. Sze, *Physics of Semiconductor Devices, Second Edition,* New York: John Wiley & Sons, chapter 13, 1981.

19. K.W. Yau and D.A. Baylor, “Cyclic GMP-activated conductance of retinal photoreceptor cells,” *Ann. Rev. Neuroscience,* vol. 12, pp. 289–386, 1989.

A P P E N D I X

2a

SPECTRAL SENSITIVITY

In Chapter 2, we saw from our study of adaptive elements that the physical interaction of light with silicon is important to consider whenever we build circuits alongside our phototransducers. The other aspect of this interaction are its beneficial results. We know that silicon makes an excellent phototransducing material. How does it do as a color sensor? In this appendix, we show measurements of the spectral sensitivities and absolute quantum efficiencies of six phototransducers that we can build in a BiCMOS process. We compare our spectral sensitivity measurements quantitatively with theoretical expectations of the spectral sensitivities. Our theory is derived from the known behavior of light absorption in silicon, combined with the diffusion of minority carriers. We do not know of any measurements, in the literature, of the spectral quantum efficiencies of phototransducers that can be built in a typical CMOS or BiCMOS process.

PHOTODETECTOR DEVICES

A semiconductor junction separates photon-generated electron-hole pairs, letting us count them (Figure 2a.1). The junction between n- and p-type silicon creates an electric field that counteracts the diffusion of carriers from the majority region to the minority region. This electric field pulls electrons into the n-type region, and holes into the p-type region. (The way to remember this polarity is by the saying *carriers go home*.) When a photon creates an electron-hole pair somewhere inside the junction, the electron and hole are split apart by the huge electric field ($\approx 10^6$ V/m) and are each swept home. When the pair is generated somewhere in the bulk silicon, the majority carrier is lost in the sea of majority carriers—it is already home. The minority carrier, however, starts

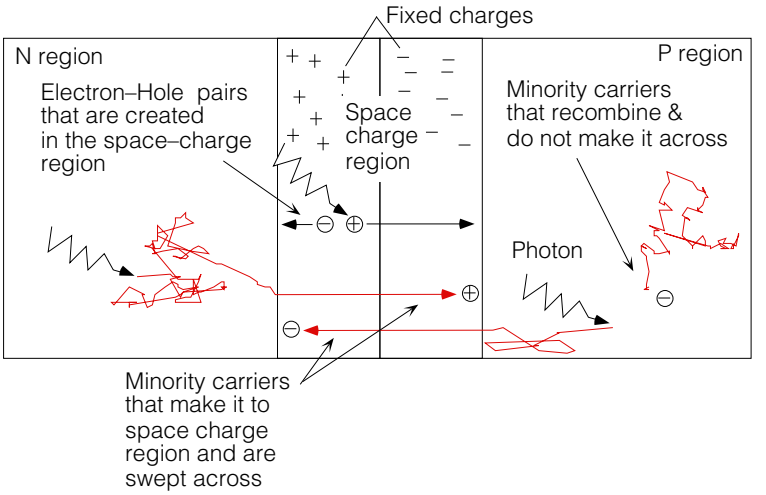


FIGURE 2a.1 Schematic illustration of carrier generation, random walk, and either recombination or collection. When the minority carrier diffuses to the space-charge region, it gets swept across. Otherwise, it recombines with a majority partner and does not contribute to the photocurrent.

diffusing. Two fates can then occur: Either the minority carrier recombines with a majority carrier, in which case it is as though the photon were never absorbed, or the minority carrier diffuses to the junction and is swept across to the other side.

The absorption of light by silicon is strongly wavelength-dependent: Short wavelength photons travel a shorter distance, on the average, before being absorbed.† The absorption length, $L(\lambda)$, as a function of photon wavelength λ, for pure bulk silicon, is shown in Figure 2a.2. For blue light (wavelength 475 nm) L is 0.3 μm, while for red photons (650 nm) L is 3 μm—a ratio of approximately 10 in absorption length over the visible spectrum. This behavior is primarily due to the available density of states, because there are many more available states at higher energies.

† This effect is known to degrade resolution in CCD imagers at longer wavelengths. The longer-wavelength photons get absorbed deeper in the substrate, diffuse less precisely to the correct charge bucket, and take enough time to diffuse that they get collected into the wrong charge bucket. Many CCD imagers are built in a shallow p-well to reduce these effects and to better tailor the spectral response to match with human spectral efficiency.

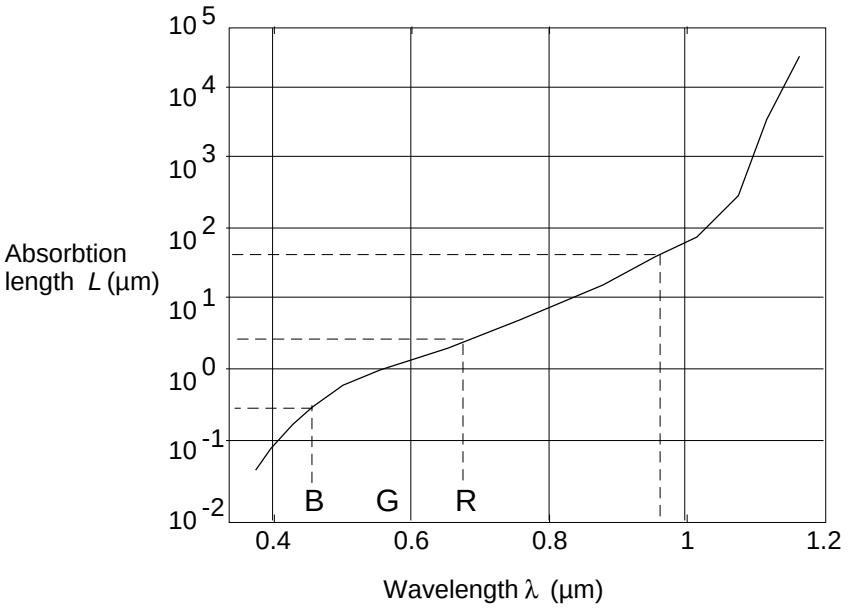


FIGURE 2a.2 The photon absorption length as a function of photon wavelength at 300°K. The absorption length is the distance over which 1/e of the incident photons are absorbed. Approximate wavelength of primary colors (B = blue, G = green, and R = red) from CIE color wheel, along with associated absorption length are shown as dashed lines. Dashed line at about 0.95 μm shows absorption length at peak of spectral response of deep, diffusion-limited junction (see page 68). (*Source:* Adapted from Dash and Newman [3].)

The wavelength-dependent absorption means that photodetectors formed from junctions with different junction depths will have different spectral responses. In a CMOS process, there are a number of different junctions with different doping and depth. A garden-variety CMOS process has two complementary source–drain diffusions and a well diffusion. For concreteness, we consider an n-well process; the well and the source–drain diffusion for the native transistors are n-type, and the substrate and source–drain diffusions for the well transistors are p-type. The complementary process, p-well, results in a set of devices that are exactly complementary to the ones we describe here, and we expect that these devices have similar characteristics. We assign unique names to each device according to the sex and doping strength of the constituent materials. The devices are shown in Figure 2a.3. We order the materials that make up each device so that the pos-

itively-biased side of the device, for reverse bias operation, comes first. We can form four photodetectors in a plain CMOS process:

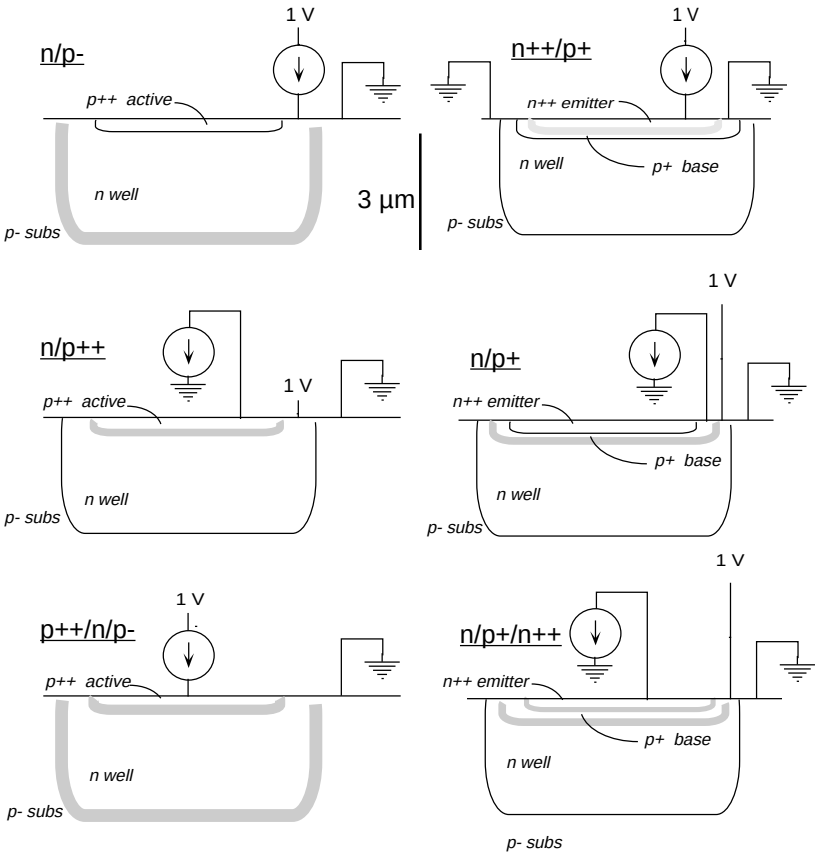


FIGURE 2a.3 The structures and biasing setups used to measure the quantum efficiencies of the devices. The upper four devices are photodiodes and the lower pair are phototransistors. In the text, we refer to the elements by the underlined names in the figure. The active junctions are shaded for each device. A scale bar is shows approximate dimensions (horizontal dimension not to scale).

1. **n++/p-**: A photodiode consisting of the junction between heavily doped n-type source–drain diffusion (n++) and very lightly-doped p-type substrate (p-)
2. **n/p++**: A photodiode consisting of the junction between lightly-doped n-type well (n) and heavily doped p-type source–drain diffusion (p++)

3. **n/p-**: A photodiode consisting of the junction between lightly-doped n-type well (n) and very lightly-doped p-type substrate (p-)

4. **p++/n/p-:** The PNP parasitic phototransistor formed by using heavily-doped p-type source–drain diffusion as emitter (p++), lightly-doped n-type well as the base (n), and very lightly-doped p-type substrate as collector (p-)

A BiCMOS process adds a medium-doped p-type diffusion intermediate in depth between the source–drain diffusion and the well diffusion. This new implant is used to form the p-type base for vertical NPN bipolar transistors. Only n-well BiCMOS processes exist; for technical reasons it is difficult to fabricate good floating vertical PNP bipolar transistors. Also, we note that a BiCMOS process differs from a true bipolar process in that there is no additional low-resistance implant in the collector (the well) to reduce collector resistance. The missing collector contact implant is not a concern for phototransistors, at least in the range in which we are interested, because the generated photocurrents are much too small to generate appreciable ohmic voltage drop. The emitter of the vertical bipolar is the heavily-doped n-type source–drain diffusion, and the collector is the lightly-doped n-type well. In a BiCMOS process, we can form another four photodetectors:

5. **n++/p+:** A photodiode between heavily-doped n-type emitter (n++) and medium-doped p-type base (p+)

6. **n/p+:** A photodiode between lightly-doped n-type well (n) and medium-doped p-type base

7. **n/p+/n++:** The NPN vertical phototransistor made from the vertical bipolar transistor, with lightly-doped n-type well as collector (n), medium-doped p-type base (p+), and heavily-doped n-type source–drain diffusion as emitter (n++). (Note that we bias this device so that the current is measured from the emitter, not from the collector. Circuits that use this device should do likewise.)

8. **p+/n/p-:** The PNP parasitic phototransistor formed from the medium-doped p-type base layer as emitter (p+), lightly-doped n-type well acting as base (n), and very lightly-doped p-type substrate acting as collector (p-). (This device is nearly identical to the p++/n/p- parasitic phototransistor.)

Of these eight devices, we measured the six shown in Figure 2a.3. We did not test device 1 (n++/p-) because we neglected to fabricated it on the same chip in the same configuration, making reliable results difficult to obtain. We expect devices 1 and 3 to behave similarly. We did not test

device 8 (p+/n/p-) because we forgot about it, although almost certainly its behavior is similar to device 4.

A scale bar in Figure 2a.3 shows the approximate vertical dimensions of the junctions, according to the MOSIS fabrication service. The n++ emitter is arsenic-doped, with a junction depth of about 0.3 μm, and a surface concentration of 10^{20} donors/cm^3. The p+ base is boron-doped, with a junction depth of 0.45–0.5 μm, and a surface concentration of 1–2·10^{17} acceptors/cm^3. The n-well is phosphorous-doped with a junction depth of approximately 3 μm, and a surface concentration of 3-4·10^{14} acceptors/cm^3. The p-type substrate has a doping of 3–4·10^{14} acceptors/cm^3 [7].

We distinguish between the deep, **diffusion-limited** detectors like n/p-, where the carrier collection volume is defined mostly by the minority carrier diffusion length, and the shallow, **volume-limited** detectors like n/p++, where the carrier collection volume is mostly defined by junction edges.

EXPERIMENTAL PROCEDURE

The aim of the measurement is to obtain the absolute quantum efficiency $Q(\lambda)$ for each of the devices, as a function of photon wavelength λ, where $Q(\lambda)$ is defined as:

$$Q(\lambda) = \frac{\#\text{ collected charges}}{\#\text{ incident photons at wavelength}\lambda} \quad (1)$$

The phototransistors have built-in current gain; for them we count the collected charges including the gain, so $Q(\lambda)$ can be larger than 1.

We used a prism monochromator, in conjunction with a tungsten incandescent source, to produce a continuously variable, nearly monochromatic, source of light. We had to carefully calibrate several parameters to obtain a reliable measurement. In this section, we will describe the experimental setup and the calibration procedures.

The prism monochromator

We used the setup shown in Figure 2a.4, consisting of a prism monochromator (Gaertner Scientific Corp.), in conjunction with a 400 watt quartz halogen incandescent lamp. We carefully shielded all stray light paths, because any stray source of broadband light severely corrupted the measurements. The quartz crystal prism and lenses in the monochromator pass wavelengths

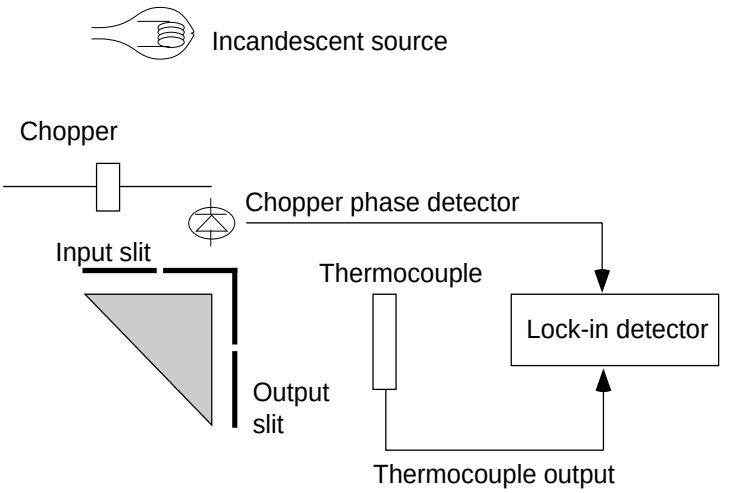


FIGURE 2a.4 Prism monochromator calibration setup. We only used the lock-in synchronous detector for calibrating the source spectrum. For measurement of the device spectral responses, we used a picoammeter and did not chop the source.

down to 200 Å, although our tungsten source generated measurable energy only down to about 350 Å.

Calibrating the source spectrum shape

To interpret any spectral measurement, we need to know the spectrum of the source. We compute the quantum efficiency by dividing the response to the source by the source spectrum. If $R(\lambda)$ is the measured device response, measured in charges/time, to illumination by the source, with known spectrum $S(\lambda)$ measured in units of incident quanta/time, then the quantum efficiency is given by

$$Q(\lambda) = \frac{R(\lambda)}{S(\lambda)} \tag{2}$$

The tungsten lamp spectrum is far from flat. To measure the *shape* of the source *energy* spectrum, we used a thermocouple with a flat spectral response over the range of interest, in conjunction with a chopper and a synchronous lock-in amplifier. Most semiconductor detectors essentially count a fraction of the generated electron-hole pairs in response to absorbed photons, and the fraction is dependent on wavelength. In contrast, the thermocouple measures incident energy directly,

by recording the change in junction potential caused by a change in *temperature* of a semiconductor junction. The junction potential is measured relative to another junction that does not absorb incident energy. The measuring junction is coated with platinum-black, an anodized material that absorbs *uniformly* over a wide spectrum [9]. The electrical signal from the thermocouple, is very small (μV), so it is necessary to use a synchronous detector.

The synchronous detector (Princeton Applied Research), also known as a lock-in amplifier, rectifies the incoming signal from the thermocouple at exactly the frequency and phase of the chopper, and low-pass filters the resulting signal. The synchronizing reference signal comes from a separate detector (see Figure 2a.4). Any signal that is not synchronous with the chopper is averaged away. The chopper blade is driven by a DC motor, and the incandescent lamp is powered with 60 Hz wall voltage. We did not observe any 60 Hz variation in thermocouple output, even at the highest photon energies.

We computed the relative *number* of source quanta per unit wavelength by dividing the source energy density by the photon energy at each wavelength. The photon energy is given by $E = hc/\lambda$, where $h = 6.023 \cdot 10^{-34}$ J·s is Planck's constant and $c = 3.0 \cdot 10^{8}$ m/s is the speed of light. Our calibration is good over 3–4 decades. For photon energies higher than about 3.5 eV, it is necessary to use an arc lamp. We did not go to this trouble, because most interesting photodetector behavior occurs at longer wavelengths. The measured source spectrum shown in Figure 2a.5 is normalized to a maximum value of 1, because we had no convenient means of directly measuring the *absolute* intensity of the source. In a later section we describe the absolute intensity calibration.

We can see from the measured source spectrum in Figure 2a.5 that the incandescent source acts like a blackbody at short wavelengths (number of quanta exponentially decreases with photon energy, for high energies).The exponential characteristic of blackbody radiation is accurate for energies higher than a few *kT*, or approximately 1–2 eV for a 2000 ˚K lamp. At longer wavelengths, the source spectrum deviates from blackbody (not shown in plot). This deviation is in accordance with the literature on tungsten incandescent sources [9].

Source spectrum wavelength and line width calibration

We measured the spectral width of the monochromator output, and calibrated the absolute wavelength, with the aid of a narrow band 543 Å HeN gas laser (bandwidth < 1nm, Melles Griot "Grene"). At the settings of the monochromator input and output slits that we used to measure the

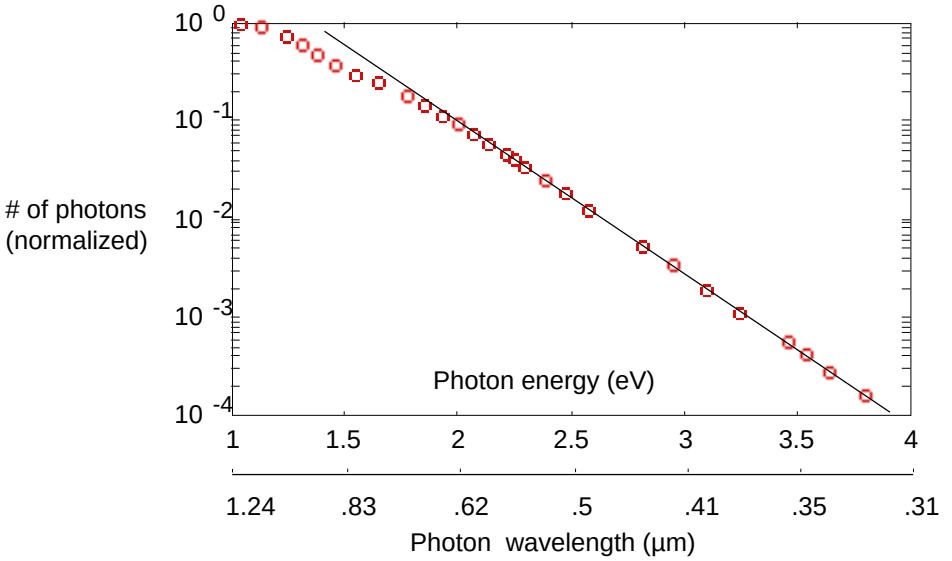


FIGURE 2a.5 The normalized source spectrum $S(\lambda)/S_{max}$ of the tungsten halogen lamp, as a function of photon energy.

responses of the photodetectors, we found that the bandwidth of the monochromator output was less than 10 nm. We also used a miniature 670 nm semiconductor pointing laser to obtain another point on the wavelength calibration. These two wavelength calibration points are shown along with the results of the photodetector spectral response measurements in Figure 2a.6. This two-point absolute wavelength calibration places a reliable axis on the wavelength scale.

Absolute intensity calibration

Using our setup, it was impossible, for mechanical reasons, to accurately measure the actual intensity of the light coming from the monochromator and striking the photodetector. To calibrate the absolute intensity of the source, we used separate measurements of the device photocurrents in response to light emitting diodes (LEDs) of known wavelengths. We calibrated the absolute irradiances of the LEDs with a Tektronix J16 photometer, with a J6512 probe. This probe is rated to have a flat response to within 7% over the range 450 nm to 950 nm wavelength. We used a small pinhole of known dimension to restrict the intensity source falling on the photometer to a uniform patch of intensity, and positioned the LED to maximize the reading. We then normalized the inten-

sity back to the area of the photometer detector, to obtain the true maximum intensity emitted by the LED. We arranged the device under test (the photodiode or phototransistor) so that the distance from the LED was identical to the distance of the LED from the photometer. When we measured the response of the device, we were careful to adjust the location of the LED spot to maximize the response, assuring that the device was illuminated with the same intensity as measured with the photometer. We used only a Yellow (583 nm) and Red (635 nm) LED, because other LEDs did not show results that were consistent with the relative results obtained using the monochromator. The measurements from these other LEDs were probably inconsistent because they were taken with LEDs at the more extreme ends of the spectrum, where the photometer calibration is inaccurate, and the LED spectral line width is larger.

The *absolute* quantum efficiencies measured using the LEDs are shown as the small open circles in the plot of absolute quantum efficiency in Figure 2a.6. When we used the monochromator to measure the relative quantum efficiencies, we were careful to keep the light intensity constant for all the measurements of the devices; to test a different device, we simply switched our probe to a different pin on the test chip, keeping everything else constant. This constancy was only possible because all the devices were fabricated on the same chip. Earlier attempts to obtain consistent results from a number of different chips were a failure. Hence the measurement of relative spectral responses are all reliable, relative to each other. To obtain the absolute quantum efficiencies, we slid the entire group of measured relative efficiency plots up and down (but not sideways!), as a unit, until they best matched up with the absolute measurements from the LEDs.

Device configuration for spectral measurement

When we measured the photocurrent, we held the reverse bias of the device at 1 V, as shown in Figure 2a.3. The reverse bias on the junction changes the junction width, but this effect is only a square-root function of reverse bias. We did not observe any significant change in the spectral response due to our modest (< 3 V) changes in bias voltage. Junctions that we did measure were either left floating or were grounded, as shown in Figure 2a.3.

Spectral response measurement

We first attempted to measure the responses using an off-chip linear current-sense amplifier in combination with the synchronous detector and chopper used to calibrate the source spectrum.

This approach proved to be inaccurate, due to the transient effects of carrier generation, diffusion, and recombination that contaminated the measurement of the steady-state current. Instead, we used a picoammeter (Keithley 617) to directly obtain the photocurrent. The dynamic range was quite similar to that obtained using the synchronous detector. With care, we could obtain useful results over 2 to 3 decades of response.

THE SPECTRAL RESPONSES

Figure 2a.6 is the summary plot of absolute quantum efficiency versus photon wavelength. Figure 2a.7 shows the same data on a linear scale. There are clear differences between the quantum efficiencies of the devices. The figure also shows the photopic (daylight) visibility curve, the band edge for silicon, the monochromator bandwidth, and the wavelengths of the primary colors. We can make several phenomenological observations.

- There is a clear distinction between photodetectors that collect light-generated carriers from a junction-limited volume (n/p+, n/p++, n++/p+, n/p+/n++), and the photodetectors that collect light-generated carriers from a diffusion-limited volume defined by the diffusion length of minority carriers (n/p–, p++/n/p–). The diffusion-limited detectors are more sensitive to longer wavelengths.

- The response spectra are broad band. All the detectors cover far more than the visible spectrum. Also, all of the detectors have responses that are flat within a factor of 2 or 3 within the visible spectrum.

- The shallow junctions (n/p++, n/p+, n++/p+) are most sensitive to a wavelength around 500 nm, while the deep junction (n/p–) sensitivity peaks at about 900 nm, well outside the visible.

- The peak absolute quantum efficiency for the photodiodes varies from a high of about 0.8 for the deep photodiode (n/p–) at near-infrared photon energy, to a low of about 0.3 for the shallow, volume-limited junctions (n/p++, n/p+, n++/p+). These measurements are consistent with reports from the literature (see, for example, [8]).

- All the quantum efficiencies are about identical at short wavelengths.

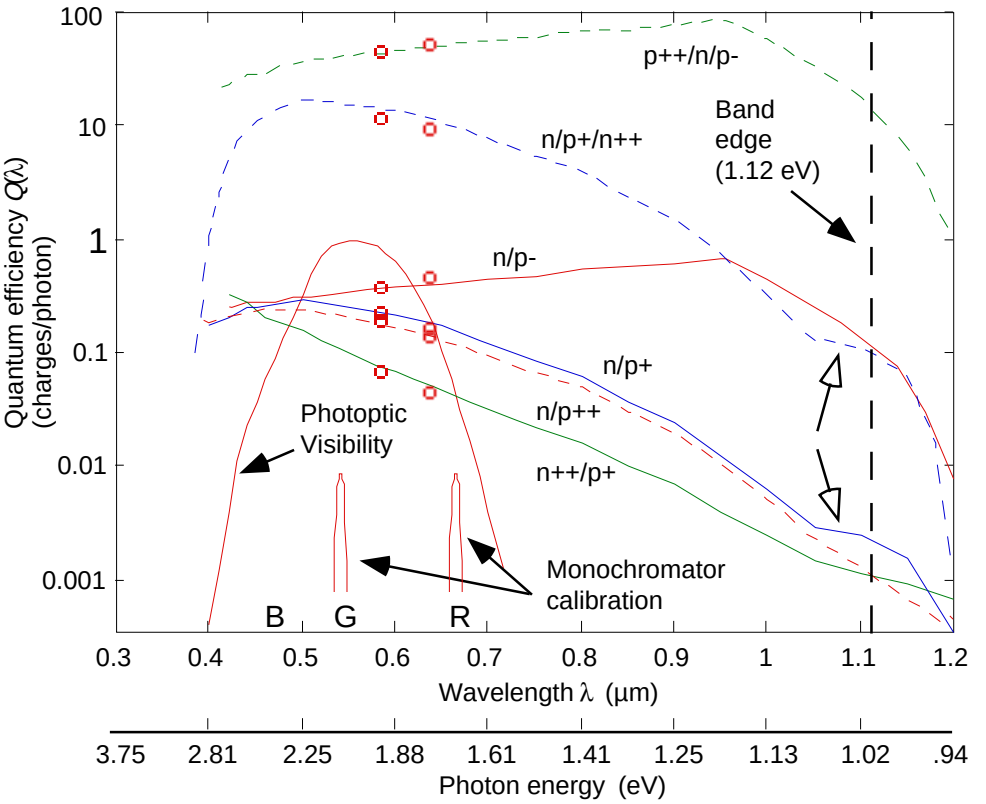


FIGURE 2a.6 Measured spectral quantum efficiencies $Q(\lambda)$ versus photon wavelength and energy. The quantum efficiency is the number of collected charges per incident photon; it is larger than one for the phototransistors because they have built-in current gain. Each curve is labeled with the name of the device as shown in Figure 2a.3. The small circles are the absolute calibration points measured with discrete LEDs (page 71). The photopic visibility curve shows the relative visibility of photons under photopic conditions; this curve is arbitrarily normalized to 1 at its maximum [11]. The primary colors, according to the CIE color chart, are labeled on the wavelength axis. A shelf, marked with a hollow arrow, appears right around the band edge for both the n/p+/n++ vertical bipolar NPN phototransistor and for the n/p+ photodiode that forms the base to collector junction in the NPN bipolar (page 76). The monochromator calibration curves show the spectral line width of the monochromator and the calibration wavelengths (page 70).

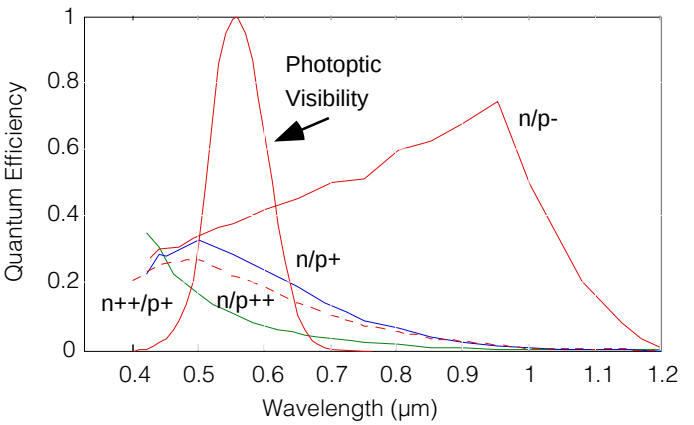


FIGURE 2a.7 Absolute quantum efficiencies plotted on a linear scale. Data are a subset of the data in Figure 2a.6.

- The phototransistor spectral responses are very close to constant multiples of the deeper-junction responses contributing to the base current for the phototransistor. For example, the PNP parasitic bipolar response (p++/n/p–) is very similar in shape to the response of the well to substrate detector (n/p–), but is completely unlike the response of the shallow emitter to base detector (n/p++). This fact is also true for the vertical bipolar phototransistor: The phototransistor response (n/p+/n++) is similar to the base to collector detector (n/p+), but is different than the emitter to base junction (n++/p+). This observation simply means that most of the base current in the phototransistors comes from the deep junction.:

- The current gain is about 100 in the parasitic bipolar phototransistor, and is about 30 in the vertical bipolar phototransistor. This gain is a soft function of the current level, and decreases at both the high and low intensities. Figure 2a.8 shows the separately-measured current gain for the bipolar transistors as a function of emitter current, measured by base-current injection. These measurements are larger by a factor of about 2 than the spectral measurements.

- All of the spectral responses show an absolute cutoff around the band edge. The cutoff is not perfectly sharp, and extends past the actual band edge. The quantum efficiency drops off at a rate of about e-fold per 25 meV around the theoretical band edge. We are confident of this result, because we used a two point absolute calibration of wavelength, and we measured the monochromator bandwidth to be much smaller than the measured cutoff behavior.

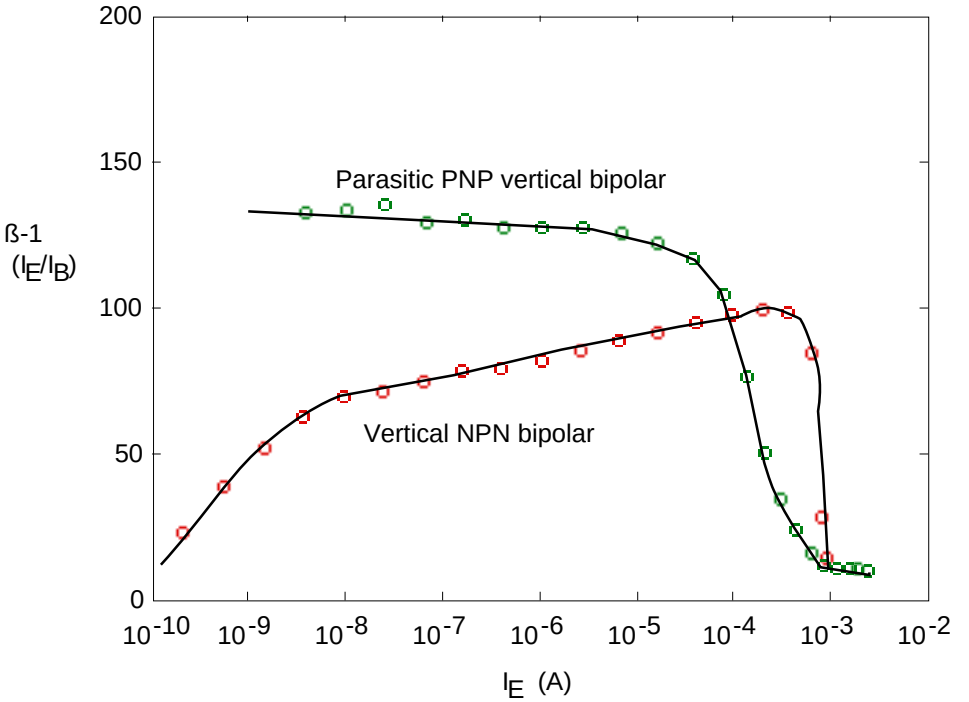


FIGURE 2a.8 Bipolar transistor current gain as a function of emitter current. Ordinate is ratio of emitter current to base current. Data is from n-well, double-poly, 2 µm feature size, BiCMOS MOSIS technology. Collector to emitter voltage was held at 1 V. Transistor dimensions are given in Table 2.1, below.

	Base	Collector	Emitter
Parasitic PNP	20x20	diffusion limited	10x10
Vertical NPN	10x10	20x20	8x8

TABLE 2.1 Bipolar transistor dimensions, in µm.

- There is an interesting “shelf” in the spectral response of the n/p+ (n-well to p–base) detector right around the band edge, that can also be seen in the vertical bipolar response (n/p+/n++). We do not know the origin of the shelf but it could be due to a shallow recombination-generation center unique to the p-base implant that stretches out the spectral response an additional fraction of an eV.

Absolute current level

A frequently asked question is how much current to expect from a given size of photodiode or phototransistor. Since light intensity varies over more than 6 decades under photopic and mesopic conditions, the answer obviously depends on the operating conditions. For reference we will compute a typical situation. Office fluorescent lighting conditions are an irradiance of an exposed surface of about 1 W/m^2, corresponding to an illumination of about 680 lux if the light is at the peak luminance wavelength 555 nm [6]. Under these conditions, each 10 μm by 10 μm photodiode area, with quantum efficiency 0.5, generates a current

$$1\frac{J/s}{m^2}\times\frac{eV}{1.6\times10^{-19}J}\times\frac{\text{quantum}}{2.5eV}\times(10\mu m)^2\times0.5 = 10^8\frac{\text{quanta}}{s} = 25pA \tag{3}$$

UNDERSTANDING THE MEASUREMENTS

We will develop a semi-quantitative model of the spectral responses. Extensive quantitative treatments have appeared in the literature (for example, the Linvill lumped model [4]), but the computations are dependent on unknown physical parameters like the surface recombination velocity. To better explain our measured results, we show the results of theoretical computations based on conventional diffusion theory [5]. Our theoretical treatment is underconstrained by the data, and the theoretical results do not tell us much more than we know from a little qualitative understanding. Still, this theoretical treatment is entertaining from a pedagogical point of view.

The diffusion equation and boundary conditions

Minority carriers are generated by light and by thermal processes, diffuse around, and recombine with majority partners. The concentration of minority carriers is governed by the diffusion equation. In steady-state, the one-dimensional diffusion equation is

$$-D\frac{d}{dx}\frac{dn}{dx}+\frac{n}{\tau} = G \tag{4}$$

where x is space, D is the diffusion coefficient, n is the concentration, τ is the average minority carrier lifetime, and G is the generation rate per unit volume per unit time. If we consider a volume of width Δx and cross-sectional area 1, with carriers generated at rate G per unit volume, gobbled up

at rate n/τ, and diffusing out at rate $D\frac{dn}{dx}$ on each side, then Equation 4 makes sense, because it simply states that the influxes and effluxes of carriers must balance in steady-state.

The simplest solution to Equation 4 is a decaying exponential, with minority-carrier diffusion length space constant L_0 satisfying

$$L_o^2 = D\tau \tag{5}$$

The diffusion equation together with the appropriate boundary conditions lets us solve for the steady-state concentration profile in response to light shining on the device. The photocurrent at a junction is given by the sum of the currents flowing into the space charge region from each side, plus the generated carriers within the space-charge region. We will ignore the space-charge current because the width of the depletion region is small (≈ 1 μm) relative to the diffusion length (> 10 μm). The current flowing into the junction from each side is given by $D\frac{dn}{dx}$. Given the concentration, we can easily compute the photocurrent.

To solve the equation, we need boundary conditions and we need to know the generation rate due to light. The generation rate is proportional to the local flux of light. The light decays exponentially into the device. The measured space constant of the decay is shown in the Dash and Newman absorption data shown in Figure 2a.2. The generation rate, as a function of depth x into the silicon, is given by

$$G(x) = \frac{G_0}{L} e^{-\frac{x}{L}} \tag{6}$$

where G_0 is the flux measured in photons/area/time at the surface and L is the absorption length.

There are two types of boundary conditions.

1. At the edge of a junction under strong reverse bias (like all the junctions we measure) the minority carrier concentration is zero. This condition is not strictly true for the phototransistor base-emitter junction, but is not very inaccurate because the concentration will be much smaller than the equilibrium far from the junction. Also, deep inside the silicon the excess concentration of minority carriers goes to zero. Symbolically, these boundary conditions take the form

$$n|_{\text{junction edge}} = n|_{\text{deep}} = 0 \tag{7}$$

2. At the surface of the wafer, the flux of carriers is determined by the surface recombination rate. The rate of recombination is given by the excess minority concentration times the surface velocity s. Hence we obtain a surface boundary condition of the form

$$D\frac{dn}{dx}\bigg|_{\text{surface}} = sn_{\text{surface}} \tag{8}$$

Given Equations 4, 6, 7, and 8, it is a straightforward although tedious matter to solve for the analytic form of the minority carrier concentration. This computation is well-known and is found in the literature (see, for example [5]), so we will not give the full form of the solution here. In general, the solution is given by a sum of exponentials with space constant L_0 and an exponentially decaying term with space constant $L(\lambda)$. Figure 2a.9a shows a generic p++/n/p– layering, with dimensions approximately the same as the real structure. Part (b) shows the theoretical excess minority carrier concentration for a number of different wavelengths of light, computed from the theory just discussed.

At short wavelengths, the electron-hole pairs are generated near the silicon surface, and the minority carrier has a large chance of recombining at a surface recombination center. This process limits the short-wavelength response. At an intermediate wavelength, the pairs are generated in a depth of the same size scale as the junction. Most of the carriers are collected. At long wavelengths, the pairs are generated deep in the silicon. Only the carriers within a diffusion length of the junction are collected.

We can see from Figure 2a.9 the reason for the difference in the spectral response between the surface, volume-limited device and the deep, diffusion-limited device. In the surface device, the response is mostly determined by the carriers that are generated inside the volume bounded by the surface and the n/p– junction—any carriers generated in the substrate simply do not come into play. We see from Figure 2a.2 that the absorption length increases exponentially with photon wavelength over a wide range of wavelengths. Hence, we expect that for the surface junction detectors, the spectral sensitivity should decrease exponentially with wavelength, with the same decay constant. From the absorption length data in Figure 2a.2, the slope of the exponential is about a factor of 10 absorption length increase per 300 nm wavelength increase. We see from the spectral efficiency data in Figure 2a.6 that the shallow junctions have about this same slope, supporting this reasonable idea.

The deep junction acts differently, because the collection volume is limited largely by the diffusion length of minority carriers. If the light is absorbed much deeper than the diffusion length,

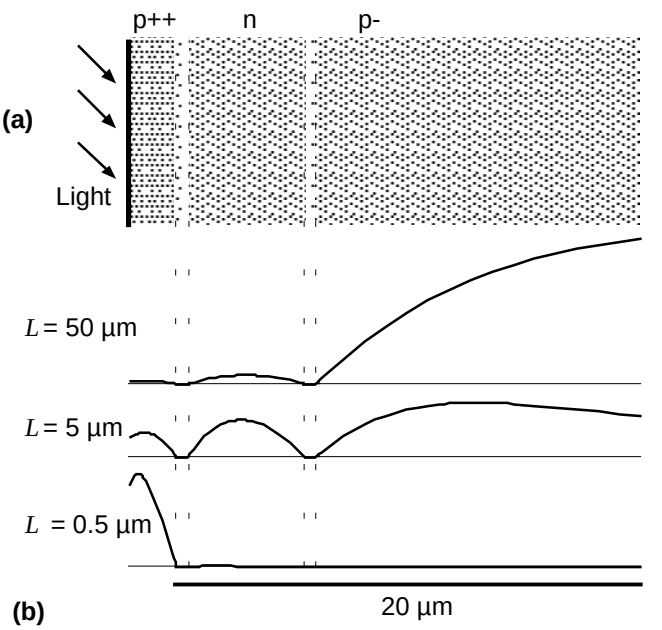


FIGURE 2a.9 Theoretical excess minority carrier concentration profiles. **(a)** shows two junctions, formed from source–drain diffusion, well, and substrate, approximately to scale. This structure is general: Increasing the central region width to infinity makes a single, deep, diffusion-limited device. Counting only the current in the n/p++ junction makes a shallow, volume-limited device. **(b)** shows the calculated minority carrier concentrations, for three different light–absorption-lengths *L*. The junction current is the slope of the concentration at the edge of the junction. For all regions, parameters were as follows: Diffusion length of minority carriers: 10 µm, lifetime: 1 µs, diffusion constant: 10^8 $\mu m^{2/}s$, surface velocity: 10^8 µm/s.

then the carriers will recombine instead of making it across the junction. We expect the response of the deep junction to fall off at the wavelength corresponding roughly to an absorption length corresponding to the diffusion length of minority carriers. In Appendix 2b, we show measurements suggesting that the diffusion length is about 30 µm in the bulk substrate. The wavelength corresponding to an absorption length of 30 µm is about 950 nm, close to the peak spectral efficiency for the n/p– junction. The evidence thus supports our simple model.

Theoretical quantum efficiency curves

Using the same concentration profiles computed for Figure 2a.9, we can compute the theoretical quantum efficiency as a function of wavelength, by combining the absorption data from [3] (shown in Figure 2a.2), with the photocurrents computed from the concentration gradients. The

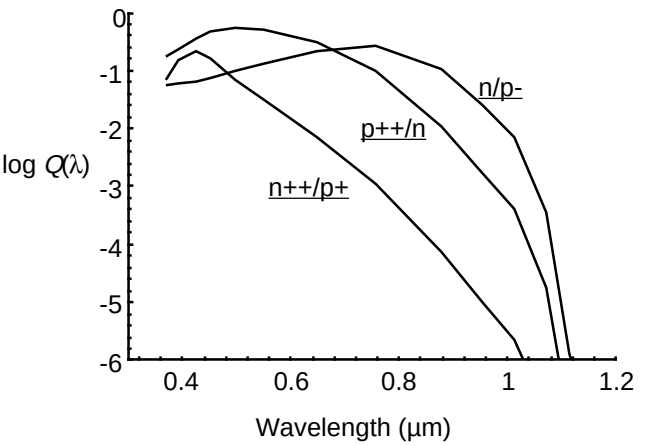


FIGURE 2a.10 Theoretical quantum efficiency curves, based on combination of diffusion theory and Dash and Newman absorption data [3]. These curves are similar to measured data. First two curves are from shallow, volume-limited detectors, last curve is from a deep, diffusion-limited detector.

Parameters are: Junction widths: n++ diffusion, 0.25 μm, p+ base, 0.5 μm, n well, 7 μm. Diffusion length: 10 μm. Lifetime ratios: τ_{p+}/τ_{n++} = 10, τ_{p-}/τ_{p+}=10. Surface velocity=3×10^8 μm/s. A key assumption is that lifetime is inversely proportional to doping. If we assume identical lifetime, then short wavelength behavior is incorrect, if we assume small surface recombination, then quantum efficiency is too large for short wavelengths.

results of the computation are shown in Figure 2a.10, for a particular set of parameters that make the theoretical curves approximate the real curves. We compute three curves, corresponding to three types of devices: A shallow, volume-limited device corresponding to the n++/p+ junction, a deeper, volume-limited device corresponding to the n/p+ and n/p++ junctions, and a deep, diffusion-limited device corresponding to the n/p– junction. For each point on one of the theoretical curves, we look up the absorption length from the Dash and Newman data in Figure 2a.2, and then compute the quantum efficiency for that absorption length. For each curve, we have assumed a set of parameters. All the parameters are identical between the different curves except for the junction depths and the diffusion coefficients.

In our first attempt to fit the measured data, we kept all parameters the same except for the junction depths. The problem with this approach is that it results in quantum efficiencies that do not have the correct short-wavelength behavior. If we assume a small surface-recombination velocity, then we invariably obtain a very high quantum efficiency for all of the junctions at short wavelengths. A larger surface-velocity reduces the quantum efficiency, but has a much larger

effect on the deep junction. The problem is to come up with a model that ends up with roughly identical quantum efficiencies for each of the junctions in the short-wavelength region.

One hypothesis is that the observed behavior is due to absorption in the protective oxide (overglass). This explanation must be ruled out, because the oxide is visibly transparent at 450–500 nm wavelength where the quantum efficiencies area identical. Doug Kerns has anecdotally measured the effect of overglass on the rate of ultraviolet-light (UV) adaptation. He found that overglass actually increased the effect of the approximately 200 nm light (personal communication).

The other hypothesis takes into account the fact that the minority carrier lifetime is not identical for each of the materials, and in fact probably is inversely proportional to doping. When we make this assumption, we obtain the curves in Figure 2a.10. These curves are qualitatively similar to the measured data.

APPLICABILITY TO COLOR MEASUREMENT

Researchers have reported circuits that are designed to measure a scalar color, or a vector of spectral responses, using either sets of different junctions, or a single junction whose junction width is electrically modulated (for examples see [2][10][12]). These approaches seem to require a special process or a large amount of support circuitry, and hence are not practical for analog VLSI, where a key ingredient to successful implementation is the rapid turnaround and low cost of multiproject generic CMOS or BiCMOS processes and the use of compact circuits. To me it seems pointless, at the present time, to look for alternatives to the simple process steps involved in deposition of colored polymer films.

SUMMARY

We have shown measurements of the quantum efficiency for the most commonly-used photodetectors. The measured quantum efficiencies are consistent with our quantitative theory based on known photon absorption characteristics and junction characteristics.

People involved in the analog VLSI vision business make extensive use of standardized bulk VLSI fabrication processes. They sometimes ask the question: Can we make vision sensors that are color sensitive, using the standard processes? The measurements we report in this appendix show

that there are only small variations, over the visible range, of the spectral sensitivities of the various kinds of photosensitive devices available to users of standard CMOS and BiCMOS processes (like those available through MOSIS). Given the rather awkward and inefficient attempts so far reported for using the built-in wavelength filtering by silicon for color discrimination, it will probably prove more practical in the long run to continue using the highly developed overlying color-filter approach found in the majority of modern solid-state commercial video cameras. This should be OK even with purists, because birds used this trick first [1].

ACKNOWLEDGMENTS

I thank Frank Perez for critical reading of this appendix, and for providing several references.

REFERENCES

1. J.K. Bowmaker, "Color vision in birds and the role of oil droplets," *Trends in Neuroscience,* Aug., pp. 169–199, 1980.

2. K.C. Chang, C-Y. Chang, Y.K. Fang, and S.C. Jwo, "The amorphous Si/SiC heterojunction color-sensitive phototransistor," *IEEE Trans. Electron Device Letters,* vol. EDL–8, pp. 64–65, 1987.

3. W.C. Dash and R. Newman, "Intrinsic optical absorption in single-crystal germanium and silicon at 77 ˚K and 300 ˚K," *Physical Review,* vol. 99, pp. 1151–1155, 1955.

4. P.A. Gary and J.G. Linvill, "Modeling of steady-state optical phenomena in transistors and diodes," IEEE Trans. on Electron Devices, vol. ED-15, pp. 267–274, 1968.

5. J.P. McKelvey, *Solid state and semiconductor physics.* Malabar, FL: Robert E. Krieger Pub. Co., chap. 15, 1966.

6. J.R. Meyer-Arendt, "Radiometry and Photometry: Units and conversion factors," *Applied Optics,* vol. 7, pp. 2081–2084, 1968.

7. MOSIS design rules for BiCMOS low-noise analog process. Available from MOSIS (Internet e-mail address: mosis@mosis.edu).

8. S.M. Sze, *Semiconductor devices, physics and technology,* New York: John Wiley and Sons, chap. 2, 1985.

9. Y.S. Touloukian and D.P. DeWitt, *Thermal Radiative Properties: Metallic Elements and Alloys,* vol. 7 of *Thermophysical Properties of Matter.*

10. H-K. Tsai, S-C Lee, and W-L Lin, "An amorphous SiC/Si two-color detector," *IEEE Trans. Electron Device Letters,* vol. EDL–8, pp. 365–367, 1987.

11. J.W.T. Walsh, *Photometry,* 3rd ed., New York: Dover Publications, 1965

12. R.F. Wolffenbuttel, “Color filters integrated with the detector in silicon,” *IEEE Trans. Electron Device Letters*, vol. EDL–8, pp. 13–15, 1987.

A P P E N D I X

2b

MINORITY CARRIER DIFFUSION

We saw in Chapter 2 how light-generated minority carriers affect low-current circuits. In this appendix, we show measurements of the diffusion length of minority carriers, and of the effectiveness of several guard structures that are designed to protect circuitry from stray minority carriers.

The diffusion equation (Equation 4 on page 77) governs the steady-state distribution of minority carriers. The most important parameter in that equation is the diffusion length, given by

$$L = \sqrt{D\tau} \tag{1}$$

The diffusion equation is a linear equation; hence the solutions are exponentials with space constant equal to the diffusion length. In the presence of a space variant generation rate, other solutions corresponding to the spatial distribution of the generation term may also appear.

The situation considered in this appendix is in detail much more complex than a simple one-dimensional system, since potentially the full three-dimensional structure of the silicon may be involved. Empirically, we find that the decay of minority carriers is close to a plain exponential in the distance away from the source of the generated carriers. Hence we here define the space constant to be the measured space constant of decay of measured minority carriers away from the site of the source of the carriers.

We built a simple structure designed specifically to measure this empirical lateral diffusion length, and to test for the effectiveness of various guard structures in blocking the diffusion of light-generated minority carriers. This structure is shown in Figure 2b.1. It consists of a central parasitic bipolar phototransistor surrounded on three sides by guard bars with different widths. The fourth side is left bare. The parasitic bipolar transistor in the center of the structure collects

and amplifies the local concentration of excess minority carriers in the bulk, i.e., we used the bipolar transistor as a probe for minority carriers. We expect this measured current to be roughly proportional to the excess minority carrier density. This chip is built in a 2 μm p-well technology. We measured the emitter current using a Keithley 617 picoammeter, with a 2 V collector-emitter bias on the NPN transistor.

The guard bars consist of two rectangles of ordinary p-type source–drain diffusion, and a rectangle of p-well diffusion. All guard bars were grounded to the bulk potential. The dimensions of all the structures are shown in Figure 2b.1.

We imaged a small spot of light onto the test structure at various locations away from the sensing transistor and measured the induced current. We imaged the spot onto the structure through a small pinhole and the 100x objective lens on the microscope. We moved the chip under the spot using a two-dimensional motorized positioning system. By viewing the spot through the normal optics of the microscope, we computed the size of the spot to be 12 μm by comparison with the known size of layout features.

The results of the measurements on the structure in Figure 2b.1 are shown in Figure 2b.2. The three parts of this figure show the measured current as a function of the distance of the test spot from the center of the sensing transistor. The different curves in parts (a) and (b) of Figure 2b.2 show the measured current as the test spot is moved away from the transistor and across the different guard structures.

We can see that the exponential-decay approximation is not very good for short distances. For distances greater than approximately 70 μm, the measured e-fold distance is about 30 μm and is fairly constant. For distances less than this, the space constant is shorter, probably because of geometrical effects due to the finite size of the test spot or three-dimensional effects of the interaction of surface recombination with the pure exponential decay. In the case of no guard bar, the measured current is reduced by a factor of 10 in a distance of about 40 μm.

To test for the possibility that this relatively long diffusion length of 30 μm is due to scattering of light, and not minority carrier diffusion, we compared the diffusion length measured with a red and green LED. From Figure 2a.2 (on page 65), we see that the ratio of absorption length for the red compared with the green LED is a factor of 2–3, so if light scattering is a significant effect in the measurement of diffusion length, we expect a difference between the diffusion lengths measured with the two LEDs. In fact, there is no measured difference. The decay with distance is nearly

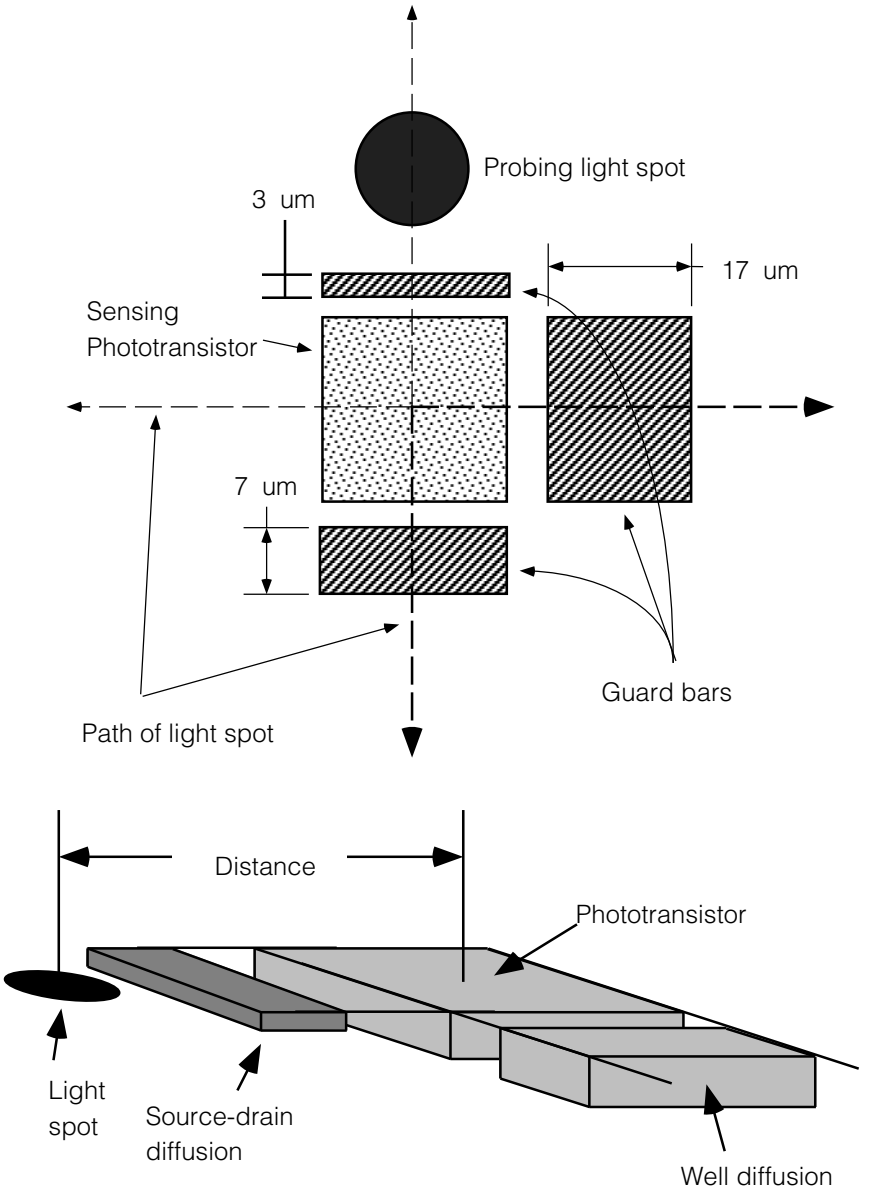


FIGURE 2b.1 Layout used to test for light-generated minority carrier diffusion and for the efficacy of various guard bars in blocking minority carriers. The large square in the middle of the figure is a parasitic bipolar transistor that probes for minority carriers.We shined a probing light spot onto the structure at various distances and directions away from the phototransistor and measured the current generated in the phototransistor. The guard structures consist of p-type diffusion in the n-type substrate. The 3 and 7 micron-wide structures are of-drain diffusion, and the 17 micron-wide structure is a p-well.

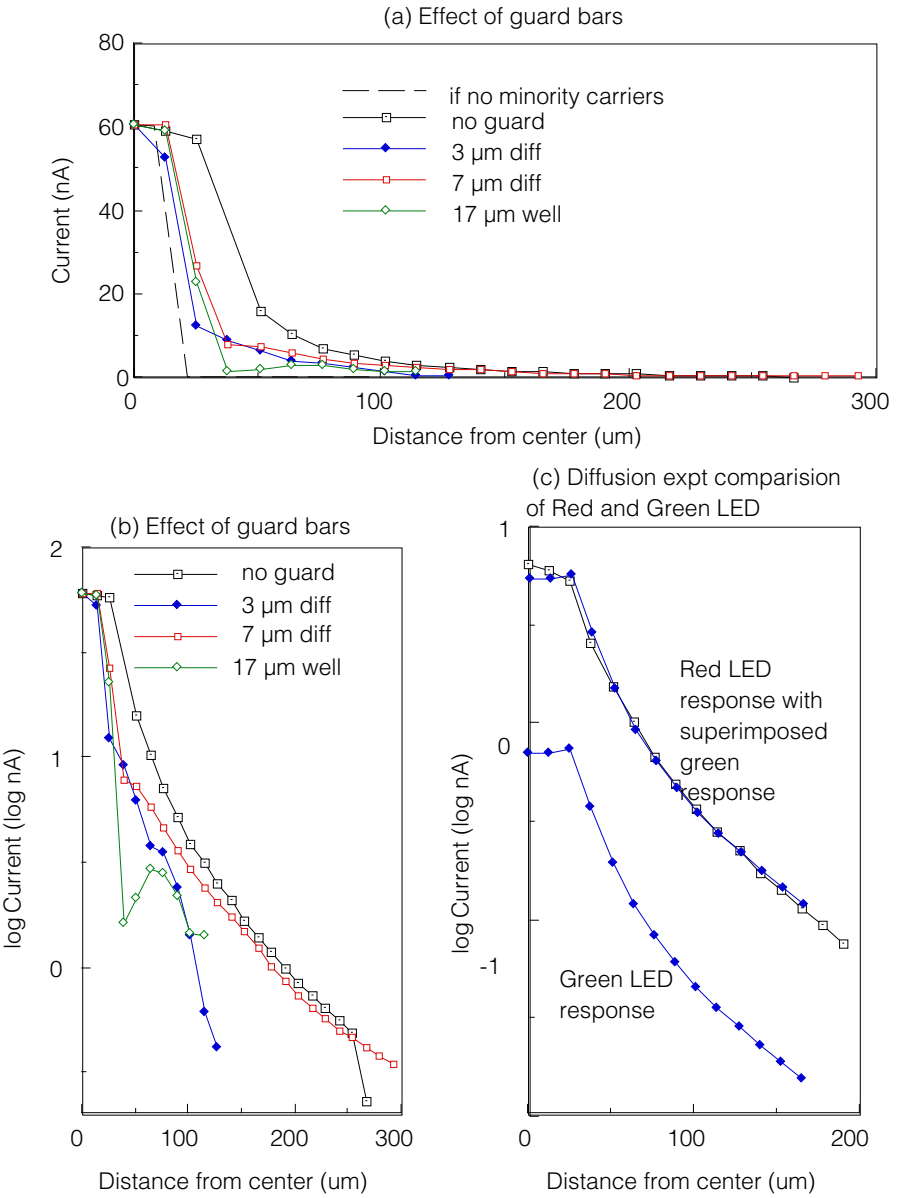


FIGURE 2b.2 Results of measurements on diffusion test structure shown in Figure 2b.1. **(a)** shows the measured photocurrent due to minority carrier diffusion as a function of the distance of the test spot from the sensing phototransistor. The different curves show the current for movement of the spot in different directions out from the middle of the phototransistor. **(b)** shows the same results on a log(current) scale. **(c)** shows a comparison between illumination with red and green light. The absorption for red light is much less than for green light, so if the results in (a) and (b) were due to scattered light and not minority carrier diffusion, we expect a difference in the measured diffusion length that we do not observe.

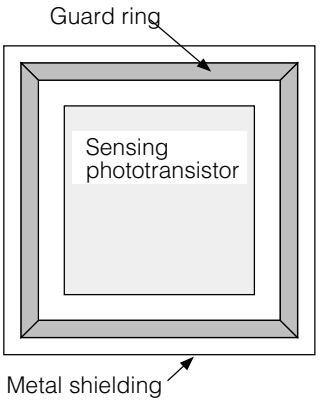


FIGURE 2b.3 Structure used to test for effect of guard bar reverse-bias voltage. Guard ring is 3 μm wide, and reverse-bias voltage relative to the bulk is controllable. Entire structure is covered with metal out to the edge shown.

indistinguishable, as shown in Figure 2b.2c. We could anticipate this result by observing, from Figure 2a.2, that the absorption length for red and green light centers around only 1 μm.

The guard bars were only moderately effective in reducing the minority carrier density. The widest guard bar, a 17 μm well, reduces the minority carrier density by up to a factor of 10, particularly when the test spot shines directly onto the guard bar. The other two guard bars, which consisted of source-drain diffusion, are less effective, reducing the carrier density by a factor of 2–3. Interestingly, the 3 μm guard bar seems to be more effective than the 7 μm guard bar. We do not know the reason for this result, but we are able to repeat it.

EFFECT OF GUARD BIAS

Using a different structure, we tested for the effect of changing the reverse-bias voltage of the guard structure. In this test structure, the sensing phototransistor was completely covered with metal, and was surround by a guard bar consisting of a 3 μm-wide ring of source-drain diffusion. This test structure is shown in Figure 2b.3. We can see from the results in Figure 2b.4 that the effect of reverse bias are minor. Changing the reverse bias from 0 V to 5 V only decreases the measured photocurrent by a factor of about 2.

SUMMARY

We have seen a direct demonstration of the diffusion of minority carriers. We measured a diffusion length of about 30 μm at distances of greater than 70 μm from the source. A minimum-size silicon retina pixel has about the same dimension, which means that minority carriers can have a

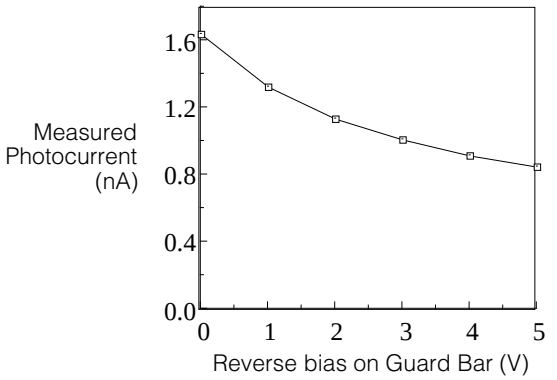


FIGURE 2b.4 Measured effect of changing guard-ring reverse bias voltage on structure in Figure 2b.3.

large effect on circuitry surrounding an opening in the overlying metal. The use of guard bars can reduce the number of minority carriers, but not very much. Even a well—the deepest diffusion we have available—that is 17 μm wide can only reduce the minority carrier concentration by a maximum factor of approximately 10. The reverse bias on the guard structure has only a minor effect—approximately a factor of 2 between a bias of ground and V_{dd}. Hence our measurements support the need for structures such as the new adaptive elements described in Chapter 2, that are resistant to the effects of minority carriers.

A P P E N D I X

2c

TRANSISTOR NOISE

Ultimately, the efficacy of our synthetic systems in dealing with real-world input will be determined in part by their raw detection capabilities. Those detection capabilities will in turn depend on intrinsic circuit noise. In this appendix, we show measurements of transistor noise. We give a very phenomenological treatment, because our aim is to know the scale and behavior of real transistor noise, not to reproduce an already vast literature. We regard this treatment as a current reference for noise parameters for a commonly-used fabrication process. With this knowledge, designers will have a greater capability of designing circuits that are electrically quiet.

This appendix is organized as follows. We start with a measurement of noise in isolated transistors. We discuss the characteristics of this flicker noise. We then discuss the physical origin of flicker noise in the briefest possible terms. We discuss how the noise is affected by bias-current level, and measure it. We conclude this appendix with a derivation of a physically-intuitive model of the other noise sources—shot and thermal. This model, and the measurement of flicker noise, are used in Chapter 2.

MEASURED TRANSISTOR NOISE

Figure 2c.1 shows noise signals from single n- and p-channel MOSFETs, measured with the setup in Figure 2c.2. The time recordings of the transistor current in part (a) of the figure show that the current is bursty, consisting of long-lasting steps and a rapidly-fluctuating background. This type of noise signal is called **flicker noise.** The power per unit frequency in the noise signal goes like the reciprocal of the frequency, leading to the name **1/*f* noise** Figure 2c.1b).

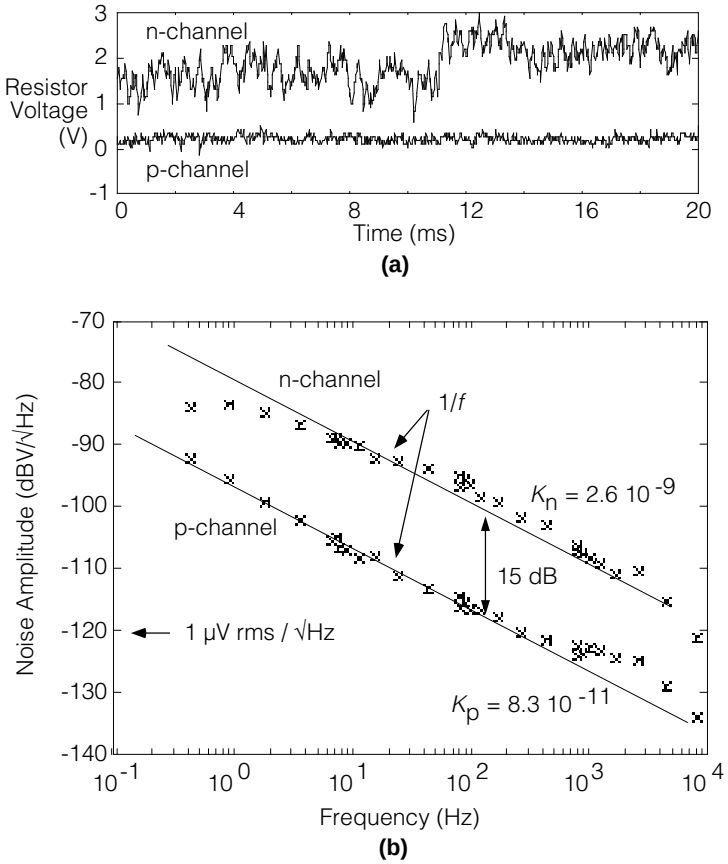


FIGURE 2c.1

Transistor noise properties. The data were taken from 6x6 μm^2 n- and p-fets from a 2–μm p-well, double poly process available through MOSIS. **(a)** Time recordings of noise voltage developed across a 1.1 MΩ resistor with a 2 μA bias current. **(b)** Power spectra. The dBV unit means the power ratio, in dB, of the signal relative to a 1 V^2 signal; 10 dB is a factor of 10 in power, and 20 dB is a factor of 10 in amplitude.

In MOSFETs, flicker noise is usually the dominant source of noise for low frequency operation, outweighing white spectral sources like thermal or shot noise, so we will focus our discussion on it. (However, the results in Chapter 2 suggest there are clearly important exceptions.) A good way to think of $1/f$ noise is that it is a *noise source with constant power per log frequency.* For example, there is as much power in the 1 Hz frequency band from 1 Hz to 2 Hz as there is in the 1 kHz band from 1000 Hz to 2000 Hz. The reason for this behavior is that, if we integrate $1/f$, we get log$f$. Hence the noise power is dependent only on the bandwidth measured in log units, whether those units be e-folds, octaves, or decades.

p-channel transistors are often quieter than are n-channel devices. For the measurements in Figure 2c.1, the p-fet is quieter than the n-fet by a ratio of approximately 15 dB, or an amplitude ratio of about 6. This is not always true; Eric Vittoz has told us that in the SACMOS process, n- and

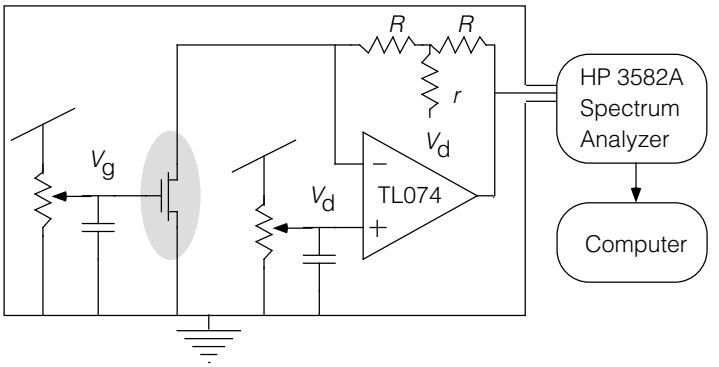


FIGURE 2c.2 The setup for transistor noise measurements. Only the shaded transistor is on-chip. The transistor and the current-sensing amplifier are battery powered and are enclosed in a completely shielded box, and all connectors are shielded to cover any open loops. These precautions are necessary, since even a single small exposed connector severely corrupts the measurement, due to the electromagnetically noisy character of our laboratory. The bias voltages are supplied by 1 kΩ potentiometers, and are low-pass filtered with 100 μF capacitors. The J-fet input TL074 opamp has equivalent input noise of 14 nV/√Hz = -157 dBV/√Hz. This input noise corresponds to a thermal noise resistance of about 10 kΩ – completely negligible compared to the range of feedback resistors we used. In measuring currents smaller than 100 nA we used the "T" feedback-resistor configuration to make the effective feedback resistance up to 100 MΩ. For larger currents we used a single feedback resistor. In all the results we checked that the thermal noise due to the feedback resistor was negligible.

p-channel devices are about equally noisy. Other researchers have reported p-fets that are 100 times quieter than equivalent n-fets [2].

It makes sense to measure the noise in the dimensionless units $\Delta i/I$, where Δi is the noise current and I is the steady-state bias current. (In other words, the mean-square variation in i is Δi^2.) This fractional noise current is a sensible parameter for subthreshold transistor operation. In subthreshold, the current is exponential in the terminal voltages, and hence $\Delta i/I$ is equivalent to a terminal noise voltage. The flicker noise power in a small bandwidth Δf is

$$\frac{\Delta i^2}{I^2} = \frac{K}{f}\Delta f\,, \tag{1}$$

n-fet	p-fet
2.6×10^{-9}	8.3×10^{-11}

TABLE 2.1 The constant K that describes the $1/f$ fractional current noise power per e-fold bandwidth in a 6 μm by 6 μm transistor operating at a bias current of 2 μA.

K is the single *dimensionless* constant that describes the noise, and is the noise power in one e-fold bandwidth. Note that the noise power is $2K$ in one e^2-fold bandwidth. From the data in Figure 2c.1, we compute that the constant K has the values shown in Table 2.1 for the n- and p-fet. In general, K is a function of the process, the transistor type and geometry, and the operating point. We can apply Equation 1 to find the variability of the current in a given frequency band. It is well-known that K is usually inversely proportional to the transistor area (see [2] for good measurements).

As a digression, let us see if the time recordings in Figure 2c.1 make sense, by computing the expected peak-to-peak (P-P) fluctuation in the current using our phenomenological theory, and comparing it with the time recordings. From Equation 1, we can compute the mean-square fractional-current noise power, for any given log bandwidth, by integrating the power over the ratio of frequencies in the band. For a bandwidth ratio of e, the mean square fluctuation is just K. The digital oscilloscope traces contain N=200 points each. We were careful to match the input signal bandwidth to the digital sampling rate to avoid aliasing. Hence, the oscilloscope traces span the frequency range 1/2 (the Nyquist frequency) to $1/N$, in sampling units—a log bandwidth of $\log N/2$. Statistically, 99% of the fluctuations of a signal will occur within 5 times the root-mean-square (RMS) fluctuation. Combining these observations, we expect to see fractional current fluctuations for the n-fet on the order of

$$\left[\frac{\Delta i}{I}\right]_{\text{P-P}} \approx 5\sqrt{K\log\frac{N}{2}} \approx 0.06\ \%\ , \tag{2}$$

in close agreement with the time recording. The data for the p-fet agrees in a similar way. Any signal that produces a current fluctuation smaller than the P-P noise variation of the current would be indistinguishable from a spurious noise signal, in the absence of special coding techniques.

It is obvious from the measurements in Figure 2c.1 that transistor noise is overwhelmingly flicker, under the tested experimental conditions. In Chapter 2, however, we saw a counterexample, where shot and thermal noise clearly dominate flicker noise. In this appendix, we show mea-

surements only of flicker noise, with the caveat that flicker noise only dominates under conditions when the effective number of charges is large enough. The photoreceptor circuit of Chapter 2 is a notable counterexample, typical of many subthreshold circuits, where the white noise sources dominate.

THE SHORT STORY ON FLICKER NOISE

An intuitive, but physically correct explanation for flicker noise goes as follows.† Flicker noise is caused by traps inside the gate oxide. The traps capture and emit carriers from the channel, via quantum-mechanical tunneling. When a trap is occupied by a charge, the channel current is affected. The traps are all within a few angstroms of the channel–oxide junction, so, on the average, the effect of each trapped charge is the same. The less mobile charge is in the channel, the larger the fractional effect of a charge on the mobile concentration of carriers. The time constant of a trap is exponential in its distance from the channel. A single trap causes a bump in the noise-power spectrum, measured as power per log frequency, at the average transition frequency of the trap. If the traps are uniformly distributed in the volume of the oxide, then the power spectrum of the flicker noise has uniform power per log frequency, due to the combination of the time constant distribution, and the characteristics of a single trap noise spectrum. The noise spectrum of flicker noise hence arises from the quantum-mechanical nature of the trapping and detrapping mechanism. More complete accounts can be found in the literature [1][7][8].

Flicker noise as a function of bias current

It is interesting to understand how the bias-current level affects flicker noise. The physics of the problem is simple and intuitive if approached in the right way. The trapped charges in the oxide affect the current in the channel in very different ways in the weak- and strong-inversion operating regimes. In weak inversion, the effect of a trapped charge is to modulate the channel surface potential, on the average, by a fixed *voltage*. Boltzman statistical mechanics means that this voltage modulates the amount of mobile charge in the channel by a constant *fractional* amount. In

† At one point, a fuller explanation of the physics underlying flicker noise, including discussion of the many physical assumptions, had ballooned to over 50 pages! This discussion is beyond the scope of this thesis, was never cleaned up sufficiently, and was long on ideas but short on data, so I decided not to include it.

strong inversion, the effect of a trapped charge is simply a removal of the charge from conduction through the channel. Hence, in strong inversion, the trapping affects channel charge by a constant *amount* of charge. In the intermediate moderate-inversion region, we expect a smooth transition between the two behaviors.

The implications for the current flowing through the transistor are as follows. In weak inversion, the constant fractional variation in the mobile charge means that the fractional current noise, $\Delta i / I$, is a constant, independent of the bias current I. In strong inversion, the constant variation in the amount of mobile charge means that the current noise Δi is a constant, and hence the fractional current noise goes as $1/I$.

We can see these relationships in the measured fractional current noise shown in Figure 2c.4, which shows the noise power as a function of bias current. Figure 2c.4 shows separate measurements of drain current versus gate voltage, to determine the weak, moderate, and strong inversion regions of operation.

The amount of flicker-noise power in the subthreshold region is an interesting parameter that is useful for understanding and calculating the effects of flicker noise in real circuits. The data in Figure 2c.4 come from a square-geometry, 6-μm by 6-μm n-fet. The amount of fractional-current flicker-noise power is K per e-fold bandwidth. From Figure 2c.4, the subthreshold value of K is between 10^{-7} and 10^{-8}. Intuitively, one e-fold bandwidth has a variability of between 10^{-4} and 3×10^{-3}, or between 0.01 % and 0.03%. In a p-fet of the same size, K is about 36 times smaller, or between 10^{-9} and 10^{-8}. Since we tested only a single chip, we do not know how variable K is between processes, runs, or even chips.

The measurement of flicker noise versus current confirms results from the literature [4]. Reimbold derived the above theory much more formally, using electrostatic theory. His derivation is correct, given the unstated assumption that the channel can be modeled by a uniform sheet charge. The effect of a trapped charge on the channel current depends critically on the *local* surface potential of the channel under the trapped charge. Random fluctuations in the fixed ion concentration modulate the surface potential by more than a thermal voltage (computations not shown here). A single trapped charge can modulate the surface potential by a thermal voltage for a distance of tens of angstroms. Hence, the subthreshold transistor channel is energetically a hilly terrain, and the functional pathways for carriers are defined by the percolation of carriers through the terrain. The assumption of uniform sheet charge accounts for the observed results, perhaps by

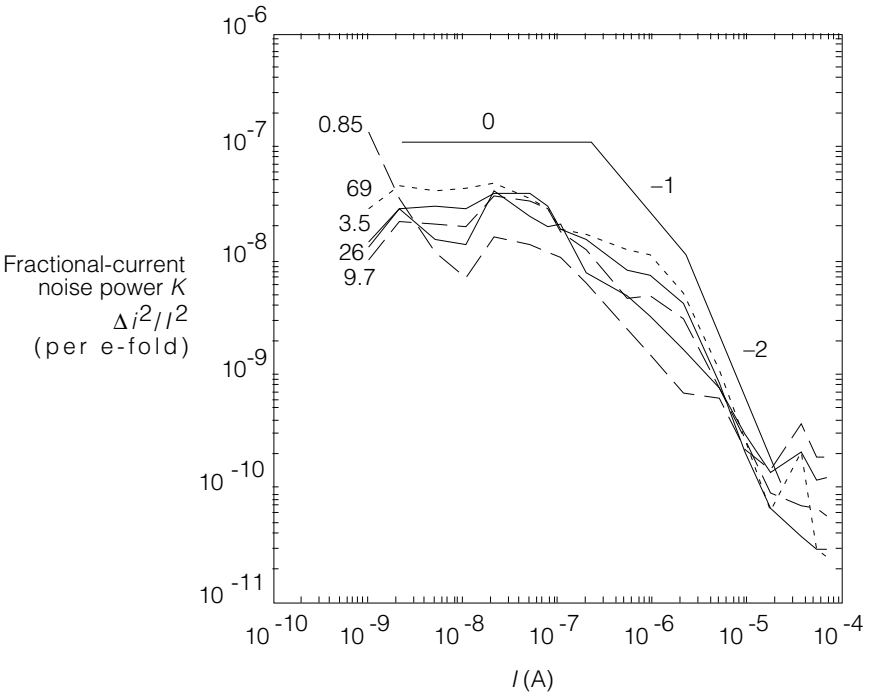


FIGURE 2c.3 Fractional-current noise power as a function of drain current, for a 6 μm by 6 μm n-channel transistor. Each curve shows the noise power in approximately one e-fold bandwidth. The curves are labeled by the center frequency in Hz of the frequency band. The line segments show slopes of 0, -1, and -2 on a logarithmic scale. The drain–source voltage is about 100 mV, so the current is not in saturation. Below threshold, the fractional noise power is approximately a constant, independent of the current. Above threshold, the noise power drops as the square of the current, i.e., the noise amplitude drops inversely with the current.

some kind of averaging, but future studies should take the real statistics of the surface-potential terrain into account.

A CHARGE-DOMAIN VIEW OF NOISE

Carver Mead, inspired by the work of Al Rose [5], figured out a simple and beautiful way to think about shot and thermal noise. We describe it here, so that we can use it to understand the noise properties of the photoreceptor circuits. The idea is simply based on the trick that if we think about *charge*, or equivalently, the *number* of charges, then noise theory becomes very intuitive.

From this viewpoint, each of the white-noise sources—shot and thermal—cause a fluctuation in the *charge* sitting on an electrical node. This electrical node has a capacitance C that determines

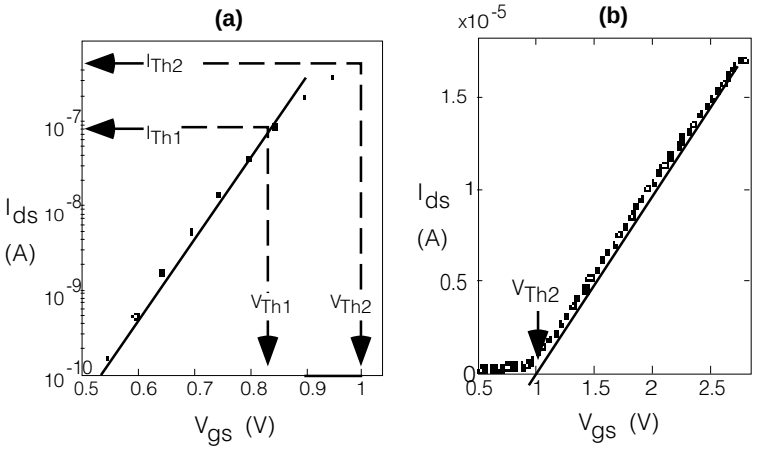


FIGURE 2c.4 Measurements of threshold operating point for the transistor in Figure 2c.3. **(a)** Transistor current measured on logarithmic scale as function of gate-source voltage. V_{Th1} shows lower estimate of threshold voltage, based on deviation from linearity. **(a)** Transistor current on linear scale. Higher estimate of threshold voltage V_{Th2} is given by intercept of linear fit to current.

the voltage going with a given amount of charge. Consider thermal effects first. If C is coupled to a thermal bath, then the basic equipartition law of statistical mechanics says that each independent degree of freedom of the capacitor has average energy $kT/2$. There is only a single degree of freedom, measured equivalently by capacitor voltage V or charge Q. Hence,

$$\frac{C\Delta V^2}{2} = \frac{\Delta Q^2}{2C} = \frac{kT}{2} \tag{3}$$

where Δx^2 means the mean-square fluctuation of x away from its mean value. Now consider shot noise. Suppose there is some electrical event or signal that places N charges on C, on the average. If each charge is an independent event, by counting statistics, the variability in this number is $\sqrt{N}$. We can express this fact in terms of the charge on C as

$$\Delta N^2 = \frac{\Delta Q^2}{q^2} = N = \frac{Q}{q} \tag{4}$$

where q is one charge unit, i.e., one electron, and Q is the amount of charge. Combining Equation 3 and Equation 4 and rewriting, we obtain the final result

$$\Delta Q^2 = qC\left(V + \frac{kT}{q}\right) \tag{5}$$

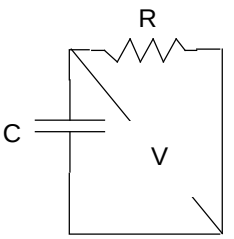


FIGURE 2c.5 RC circuit used in analysis of thermal noise. *V* is the thermal noise voltage computed in the text.

where *V* means the voltage change on *C* corresponding to the mean number *N* of charges placed on or removed from *C* during the time of observation. This intuitive expression is easy to remember, because it makes sense (at least from a physicist's perspective).

Equation 5 expresses the mean-square variation in the charge. It does *not* say what is the power spectrum of that fluctuation, only what is the *integral* of the power spectrum over all frequencies. However, it is well known that thermal and shot noise each have a white power spectrum out to well beyond normal circuit operating frequencies. Thermal noise is limited by quantum mechanics to frequencies lower than the photon frequency corresponding to the thermal energy, or about 10^{13} Hz at room temperature. Shot noise is limited to the cutoff frequency corresponding to the impulse response to a single charge in the device. In real circuits, the power spectrum is limited by an effective *RC* time constant, as we saw in Chapter 2.

Note that truly floating capacitors, like floating-gate structures, do not obey the law just derived, because they are not thermally or otherwise coupled to a source of charge. (Technically, I suppose, the law does apply, as long as we let the RC time be a million years or so.)

Let us conclude this section by deriving the well-known but nonintuitive Johnson-noise expression for thermal noise in a resistor, using Equation 5. Figure 2c.5 shows the RC circuit. The mean-square charge fluctuation on *C* is given by Equation 5. This mean-square fluctuation is spread over a bandwidth *B* given by

$$B = \frac{1}{4}\frac{1}{RC} \tag{6}$$

where the factor of 1/4 is the usual equivalent-bandwidth correction for a first-order low-pass filter. The charge fluctuations on *C* produce currents in *R*, and these currents in turn produce voltage fluctuations across *R*. It is easiest to go the other route. The RC circuit forms a complete loop, so

the voltage fluctuations on C are the same as on R. Hence, from $Q = CV$, Equation 5, and Equation 6, the total mean-square voltage fluctuation on R is given by

$$\Delta v^2 = \frac{\Delta Q^2}{C^2} = \frac{kT}{C} = 4kTRB \tag{7}$$

which is the well-known result.

SUMMARY

We showed measurements of transistor noise. We found that flicker noise is dominant in isolated transistors. We also reproduced the known result that flicker noise is a constant in subthreshold, and decreases inversely proportional to current above-threshold. We gave a very brief description of the source of flicker noise, leaving the full description for interested readers to find in the literature. We derived an intuitive expression for the variability of the charge on a capacitor due to shot and thermal noise.

REFERENCES

1. A. Ambrozy, *Electronic Noise*, New York: McGraw-Hill International, 1982.

2. Z.Y. Chang and W.M.C. Sansen, *Low-noise, wide-band amplifiers in bipolar and CMOS technologies*, Boston: Kluwer Academic Publishers, 1991.

3. P.R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits, Second Edition*, New York: John Wiley and Sons, chapter 11, 1984.

4. G. Reimbold, "Modified 1/f trapping noise theory and experiments in MOS transistors biased from weak to strong inversion – influence of interface states," *IEEE Trans. on Electron Devices*, vol. ED-31, pp. 1190–1198, 1984.

5. A. Rose, *Vision, Human and Electronic*, New York: Plenum Press, 1973.

6. C. Schutte and P. Rademeyer, "Subthreshold 1/*f* noise measurements in MOS transistors aimed at optimizing focal plane array signal processing," *Analog Integrated Circuits and Signal Processing*, vol. 2, pp. 171–177, 1992.

7. W. Shockley and W.T. Read, Jr., "Statistics of the recombinations of holes and electrons," *Phys. Rev.*, vol. 87, pp. 835–842, 1952

8. C.T. Sah, "Theory of low-frequency generation noise in junction-gate field-effect transistors," *Proc. of the IEEE*, pp. 795–814, July, 1984.

CHAPTER

3

BUMP CIRCUITS†

In this chapter, we describe two small analog circuits that compute generalized measures of the similarity of voltage inputs. The similarity outputs from the circuits, given as currents, become large when the input voltages are close to each other. One of the circuits computes only this similarity output. The other circuit computes the similarity output as well as a dissimilarity measure; each of its dissimilarity outputs becomes large only when the corresponding input is sufficiently larger than the other input. The dissimilarity outputs can be summed together or left separate; when left separate, they resemble generalized rectifier outputs. The output characteristics of these circuits may be useful in the construction of classifier networks based on the idea of radial basis functions. Using the same circuits, we also describe a transconductance amplifier with increased linear range compared with the usual 5-transistor simple transconductance amplifier. Investigation of these circuits uncovered an interesting across-channel-transistor effect that makes transistors behave weaker than their geometrical layout. We discuss data suggesting that this effect is due to fringing fields across the channel, perpendicular to the flow of current.

---

† This work has been published in part as T. Delbrück, ""Bump" circuits for computing similarity and dissimilarity of analog voltages," in *Proc. of International Joint Conference on Neural Networks*, vol. 1, pp. I-475--479, 1991. These circuits have been patented as U.S. Patent 5,099,156.

One of the fundamental nonlinear operations is measurement of the similarity of quantities. In neural-network theory, classes of networks based on units that measure the similarity between input vectors and stored vectors are called **radial-basis-function** (**RBF**) networks, in contrast to conventional neural networks whose units bisect the space by hyperplanes. RBF networks have been found to be good for classification tasks where the data is clustered in the input space — like character recognition [4]. They are not so good for problems that have an overlaid slow variation in some parameter, because it takes many RBFs to represent this variation.

In the brain, there are both radial and threshold units. Most receptive fields are radial basis functions, whose input vector can consist of spatial location, spatial frequency, direction of movement, color, etc. What distinguishes these receptive fields is their localization in parameter space. In contrast, threshold neurons show a monotonic response to some stimulus parameter, like photoreceptors, or motoneurons.

This discussion suggests that a useful thing would be a circuit primitive that computes a measure of the similarity, or distance, between inputs. In this chapter, we discuss a class of such circuits—**bump** circuits — and at the end show a number of example systems that use the bump circuits as simple RBF's.

A SIMPLE CURRENT-CORRELATING CIRCUIT

Carver Mead recognized that in subthreshold operation, the circuit in Figure 3.1 computes an interesting generalized measure of the correlation of the two input currents I_1 and I_2. We will refer to this circuit as the **simple current-correlator**. Intuitively, the series-connected transistors act as a sort of analog logical AND combination. If either of the gate voltages on these series-connected transistors is low, then the output current is shut off; conversely, if both of the input voltages are high, then the output current is large. In the intermediate regions, the circuit computes a kind of product of the input currents.

This configuration of series connected transistors has been exploited for a long time because it computes a fundamental nonlinear interaction. For example, RF modulators sometimes use this **split-gate** configuration to multiply, or mix, a carrier signal and a modulation signal. Pairs of series-connected transistors are available commercially.

We will analyze the simple current-correlator in the subthreshold operating region, where the transistor current is exponential in the terminal voltages. To compute the mathematical form of the

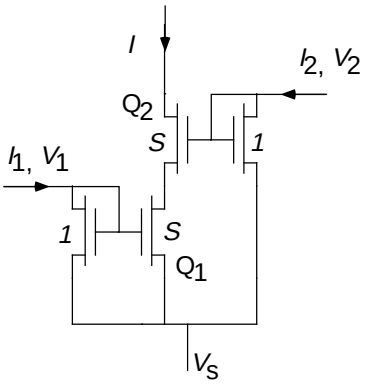


FIGURE 3.1 figThe simple current-correlator. *S* is the strength ratio between the transistors in the middle leg to the transistors in the outer legs. *I* is only large when both I_1 and I_2 are large.

response of the simple current correlator, we use the transistor law for subthreshold operation [13]:

$$I_{ds} = I_0 \frac{W}{L} e^{\kappa V_g} \left(e^{-V_s} - e^{-V_d} \right) \tag{1}$$

where I_{ds} is the current from drain to source, W/L is the effective strength of the transistor, V_g is the gate voltage, V_s is the source voltage, and V_d is the drain voltage. All these voltages are in units of kT/q, the thermal voltage, and are measured relative to the bulk potential. The factor $\kappa \approx 0.7$ accounts for the back-gate, or body, effect. All pre-exponential parameters have been absorbed into $I_0 W/L$.

For the rest of this chapter, we will refer to the effective W/L ratio for a transistor as the *strength* of the transistor. For circuit configurations like the simple current-correlator, we will refer to the strength *ratio* between the strength of the transistors in the middle leg to the strength of the transistors in the outer legs. This parameter is given by

$$S = \frac{(W/L)_{\text{middle}}}{(W/L)_{\text{outer}}} \quad . \tag{2}$$

The quantity S is an important circuit parameter for the simple current-correlator as well as the later circuits.

To compute the output current I, we assume that the top transistor Q_2 in Figure 3.1 is saturated, and that the currents through Q_1 and Q_2 are identical. Using (1), we obtain, after a little algebra,

$$\begin{aligned} I &= Se^{-V_s}\frac{e^{V_1}e^{V_2}}{e^{V_1}+e^{V_2}} \\ &= S\frac{I_1 I_2}{I_1 + I_2} \end{aligned} \quad . \tag{3}$$

As long as the transistors are operating in subthreshold, the circuit operation is symmetric in the two input currents, despite the apparent asymmetry in the stacking order. Above threshold, the function is more complicated and is no longer symmetric in the input currents.

This simple current-correlator circuit computes a *self-normalized* correlation. The output current is proportional to the product of the two input currents, divided by the sum of the inputs. All of the other circuits in this chapter rely on the simple current-correlator.

We can extend the simple current-correlator to more than a pair of inputs. The output current for n input currents (a stack of n series-connected transistors) is

$$\frac{1}{I_{\text{out}}} = \sum_{k=1}^{n} \frac{1}{I_k} \quad . \tag{4}$$

The n-input current correlator computes the parallel combination of the n input currents. The maximum number of inputs is large, because the only requirement for correct circuit operation is that the top transistor in the correlator be saturated. However, the output current scales as $1/n$.

SIMPLE BUMP CIRCUIT

The first bump circuit, which we will refer to as the **simple bump circuit**, is shown in Figure 3.2a. The input is the differential voltage $\Delta V = V_1 - V_2$; the output, plotted versus ΔV in Figure 3.2b, is the current I_{out}. We can see that I_{out} becomes large only when the inputs are close together.

Intuitively, the simple bump circuit operates as follows. The currents I_1 and I_2 through the two legs of the differential pair are comparable only when the differential input ΔV is near zero. When ΔV is larger than a few units of kT/q, the current in one of the two legs shuts off. The transistors Q_1-

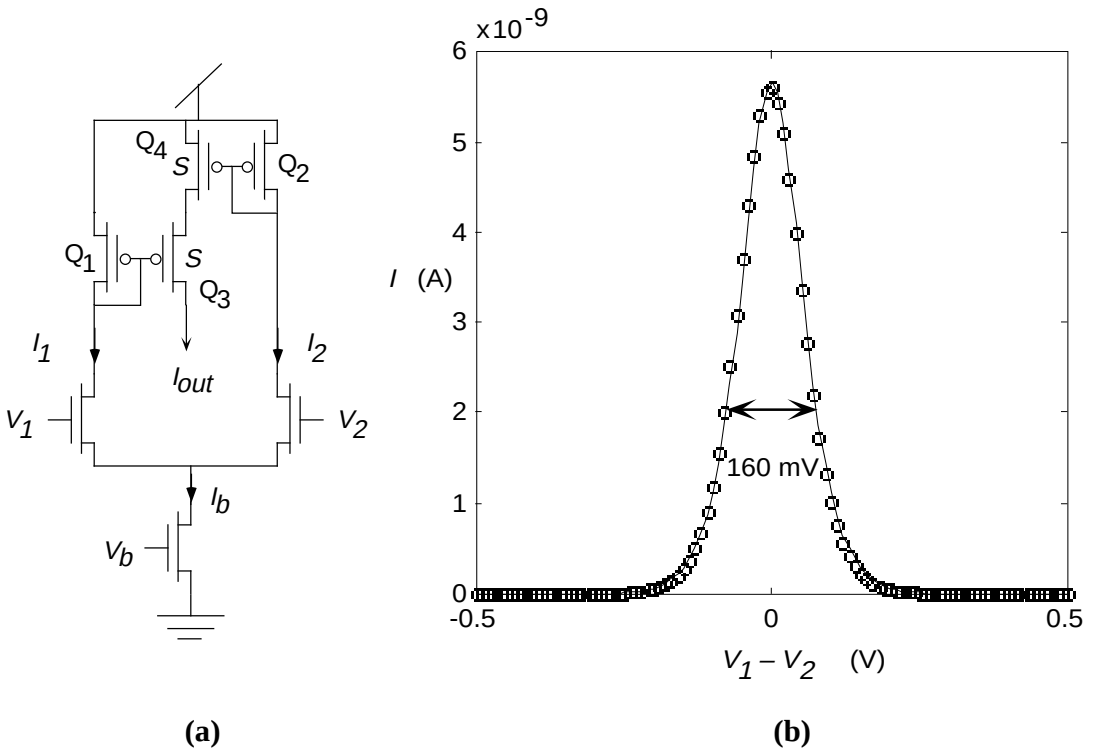


FIGURE 3.2 The simple bump circuit and response. **(a)** Circuit. The similarity output is I_{out} . The transistors $Q_{1\text{-}4}$ are the simple current-correlator. **(b)** The output current from the circuit in (a) as a function of the differential input voltage. The solid curve is a fit of the form in Equation 6. The double arrow shows the width of the bump, measured at *1/e* below the maximum.

$_4$ form the simple current correlator shown in Figure 3.1. Thus, if ΔV is large, I_{out} is zero. If $\Delta V = 0$, I_{out} takes on its maximum value.

To analyze the simple bump circuit, we assume that there is a bias current I_{b} set by the bias voltage V_{b}, that transistors Q_3 and Q_4 are S times stronger than are Q_1 and Q_2, and that the output transistor Q_3 is saturated.

To compute the mathematical form of the response for the simple bump circuit, we use the fact that each differential tail current I_1 or I_2 is a Fermi function of ΔV — for example,

$$I_1 = \frac{I_\mathrm{b}}{1 + e^{-\kappa \Delta V}} \ . \tag{5}$$

Using this result in Equation 3 and simplifying, we obtain

$$I_\mathrm{out} = I_\mathrm{b} \frac{S}{4} \mathrm{sech}^2\left(\frac{\kappa \Delta V}{2}\right) = \frac{I_b}{\frac{4}{S}\cosh^2 \frac{\kappa \Delta V}{2}} \quad , \tag{6}$$

which forms a bell-shaped curve centered on $\Delta V = 0$, with maximum height $SI_\mathrm{b}/4$. Equation 6 represents a particular measure of the similarity of the two signals V_1 and V_2. It happens that $\mathrm{sech}^2 x$ is the derivative of $\tanh x$, the transfer characteristic of a transconductance amplifier operating in subthreshold. Experimental results from this circuit are shown in Figure 3.2b, along with a fit of the form Equation 6.

Results for different bias currents are shown in Figure 3.3, where we can see that the form of the response is invariant under subthreshold biasing condition. Above threshold, the width of the bump grows and becomes somewhat unsymmetrical. The direction of the above-threshold asymmetry is such that the maximum output current occurs when I_1 is slightly larger than I_2. By duplicating the correlating transistors with swapped connections, we could symmetrize the response above threshold.

We can compute the width, in voltage units, of the simple bump circuit response by finding the differential input voltage at which the output decreases to some fraction of its maximum value, say *1/e* of its maximum. The full *1/e* width of the simple bump circuit, measured around the origin, is

$$\Delta V_{1/\mathrm{e}} = \frac{4.34}{\kappa} \tag{7}$$

in units of kT/q, or approximately 160 mV, assuming $\kappa = 0.7$ and room temperature. We note here the important point that the width of the simple bump response is independent of the S transistor strength ratio.

By including four additional transistors in the simple bump circuit in Figure 3.2a, we can add a wide-range transconductance output [13] that computes $I_1 - I_2$. The new circuit takes a differen-

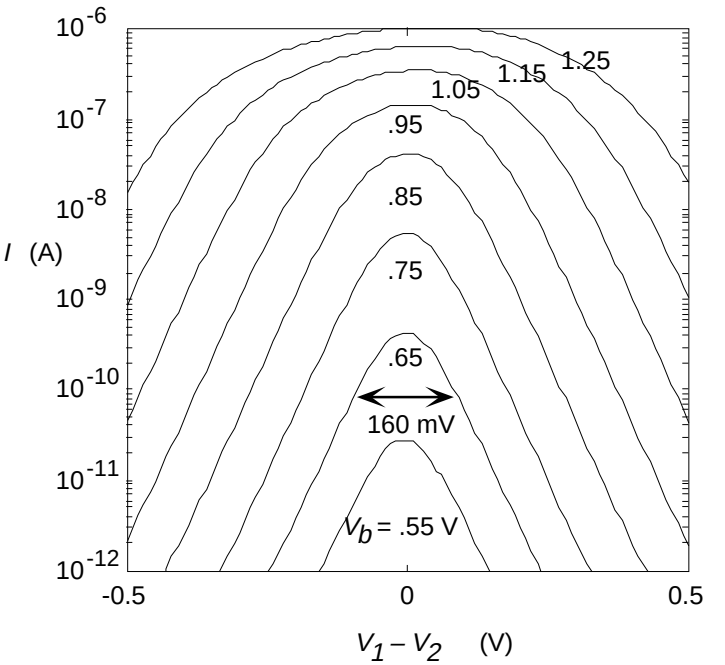


FIGURE 3.3 Simple bump circuit output as function of differential input, for various bias voltages. In these measurements, the input to V_1 was held constant at approximately 2V, and V_2 was swept around V_1. The double arrow shows the width of the bump, measured at *1/e* below the maximum.

tial voltage and produce both a transconductance and a bump output. Because the addition is obvious, we do not show it here.

Mass Sivilotti noticed that it is easy to build a simple bump circuit with more than a pair of inputs, in analogy with the multiinput current correlator, whose behavior is described by Equation 4 [14].

BUMP-ANTIBUMP CIRCUIT

Figure 3.4a shows the **bump-antibump** circuit. This circuit is more flexible than the simple bump circuit because there are three outputs: I_1, I_2, and I_{mid}, shown in Figure 3.4b. Output I_{mid} is the bump output. Outputs I_1 and I_2 behave like rectifier outputs, becoming large only when the corresponding input is sufficiently larger than the other input. If I_1 and I_2 are combined, they form the antibump output—the complement of the bump output.

Intuitively, we can understand the operation of this circuit as follows. The three currents must sum to the bias current I_{b}; hence, the voltage V_{c} follows the higher of V_1 or V_2. The series-con-

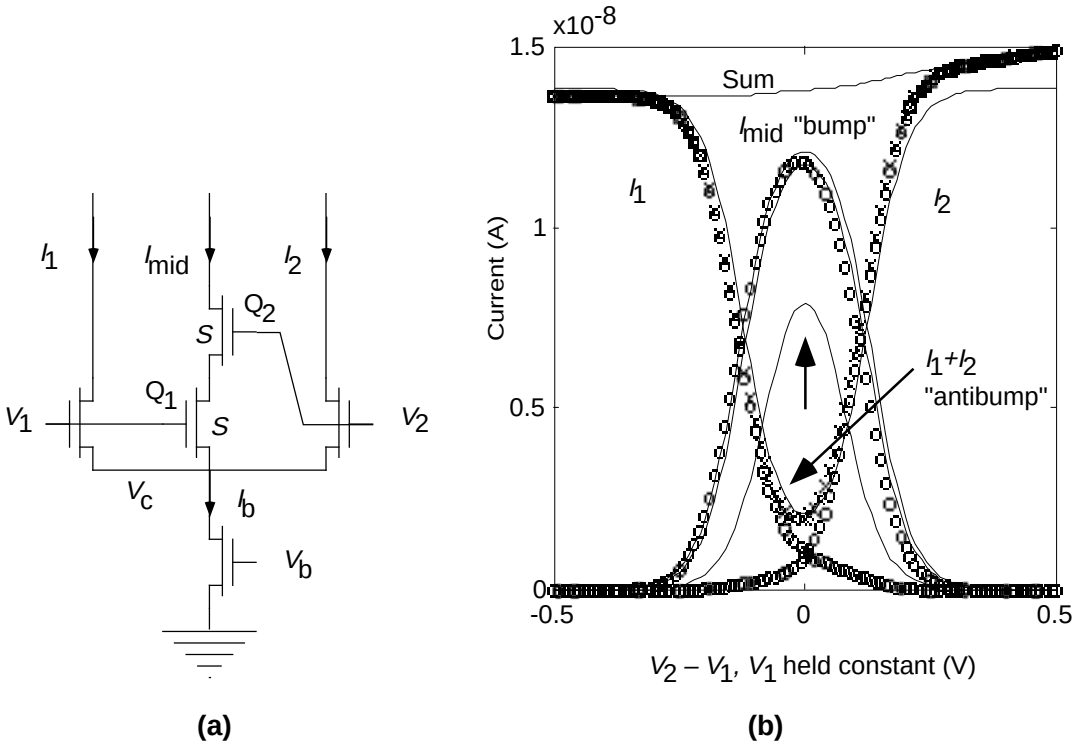


FIGURE 3.4 **(a)** The bump-antibump circuit. **(b)** Output characteristics of bump-antibump circuit. The plots show data points along with theoretical fits of the form given in the text. The curve pointed to by the arrow shows the fit that would result from using the S ratio 5.33 derived from the drawn layout geometry, before any process correction. The two theoretical curves shown for I_{mid} are the result of computing the fit using the best numerical fit to the entire curve (S=22.4), or using the ratio of maximum to minimum current in I_1+I_2 (S=28). The two numerically fit curves are virtually indistinguishable and clearly different that the theoretical curve derived from the layout geometry. The MOSIS fabrication service supplied width and length reduction parameters for this run of the order of 0.5μm. Using these parameters, we compute an S ratio of 6.6, still far short of the observed behavior. The curve labeled Sum is $I_1+I_2+I_{mid}$. The slope on the Sum curve is due to the drain conductance of the bias transistor.

nected transistors Q_1 and Q_2 form the core of the same analog current correlator that is used in the simple current-correlator and in the simple bump circuit. When $\Delta V = 0$, there is current through all three legs of the circuit. When $|\Delta V|$ increases, the common-node voltage V_c starts to follow the higher of V_1 or V_2. This action shuts off I_{mid}, because one of the transistors Q_1 or Q_2 shuts off—the one whose gate is connected to the lower of V_1 or V_2. Both V_1 and V_2 can rise together and I_{mid} does not increase, because the common-node voltage V_c rises along with V_1 and V_2.

Using the basic transistor law (Equation 1) the input-output relation for the simple current-correlator (Equation 5) and Kirchoff's current law applied to the common node,

$$I_b = I_1 + I_2 + I_{mid}, \tag{8}$$

we can compute the current I_{mid}:

$$I_{mid} = \frac{I_b}{1 + \frac{4}{S}\cosh^2\frac{\kappa\Delta V}{2}} \quad . \tag{9}$$

It is interesting that this expression is identical to the input-output relation for the simple bump circuit (Equation 6), except for the 1 added to the denominator.

We can now observe the effect of S, the transistor strength ratio, on the circuit behavior. The width of the bump, measured in input voltage units, depends on this ratio. S controls the fraction of the bias current I_b that is supplied by I_{mid} when $\Delta V = 0$. By examining the denominator of Equation 9, we can see that the width of the bump scales approximately as $\log S$, when $S >> 1$. Using the same definition for the width of the response as we used for the simple bump circuit, we obtain

$$\Delta V_{1/e} \approx \frac{2}{\kappa}\log S \tag{10}$$

in the limit of large S. The units for ΔV in this expression are, as usual, kT/q. For $S = 8.4$, the width of the output is the same as the width of the simple bump circuit output, 160 mV.

Figure 3.4(b) shows measured representative operating curves for this circuit. We can see that the theoretical form for I_{mid} fits the data quite well, with one important exception. This exception is a discrepancy between the measured and expected value for S, the ratio of transistor strengths between the middle and the outer legs of the circuit. This effect is also seen in bump circuits fabricated on a different chip, as shown in Figure 3.5. The bump-antibump circuits acts as though the S

ratio is much larger than it has any right to be, given the drawn layout of the circuit. This effect is fortuitous because it means that the designer who wants to use these bump circuits need not use inconvenient and bulky layout in order to achieve a large width and size for the bump response, which has generally been the desired profile for most designs to date. Starting on page 117, we examine this discrepancy in detail. From a practical point of view, designers can examine the measured data in Figure 3.5, which shows operating curves from eight bump-antibump circuits with different S ratios and layout sizes, to determine the correct layout to use for a particular application.

Bump-antibump circuit bias behavior

Figure 3.6 shows the response of the bump output of the bump-antibump circuit for different bias currents. Below threshold, the width of the bump is a constant, independent of the bias current. Above threshold, the bump first widens, and then eventually becomes asymmetric with respect to ΔV. The direction of the asymmetry is the same as for the simple bump circuit; the maximum current appears when the Q_2 gate voltage is slightly higher than the Q_1 gate voltage. To symmetrize the response for the above threshold operating region, we can add two more transistors to the correlation portion of the circuit with interchanged gate connections. This addition is obvious so we will not show it here.

BUMP TRANSCONDUCTANCE AMPLIFIER

By adding a current mirror to the bump-antibump circuit we can produce an output consisting of the difference current $I_{out} = I_1 - I_2$ (Figure 3.7). We will refer here to the resulting circuit as the **bump amplifier.** The output from a number of these bump amplifiers is shown in Figure 3.8, where they are compared with the theoretical output from a simple transconductance amplifier [13]. We can see that there is a flattened region in the center of the response curve that is due to the current flowing through the center leg of the circuit. The circuit in Figure 3.7 differs from a transconductance amplifier only in the addition of the two correlating transistors.

The rationale for the bump amplifier came from our observations that voltage offsets and current mismatches often dominate the behavior of system-level chips, swamping out the desired functionality. We designed the bump amplifier to produce a transconductance element that would

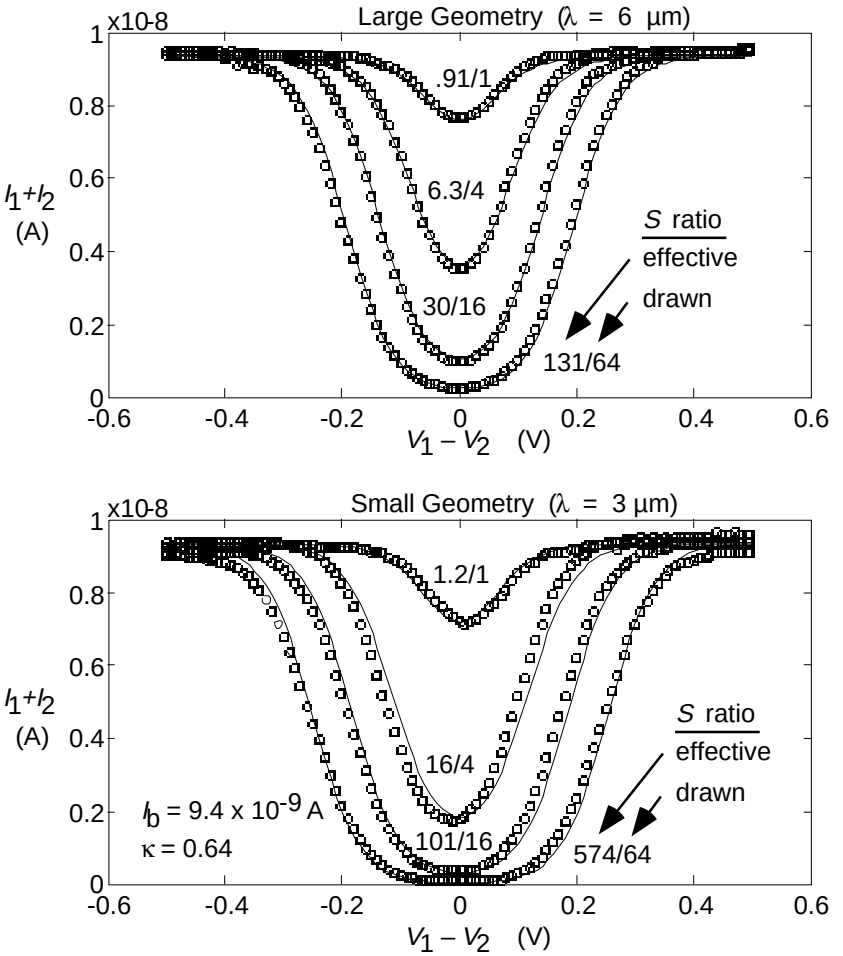


FIGURE 3.5 Antibump outputs from 8 bump-antibump circuits with different geometries. The solid curves show the theoretical fits derived from Equation 9. The numbers beside each curve are the actual and expected *S* ratios for the circuit. The different bump circuits in each set of graphs all had transistors of the same width in the outer legs, and transistors of the same length in the middle leg. For top set of curves, minimum transistor dimension was 6 μm, correlating transistors had widths 6, 12, 24, and 48 μm, and outer transistors had lengths 6, 12, 24, and 48 μm. For bottom set of curves, all transistor dimensions were halved. The discrepancy between measured and expected *S* values are larger for the circuits with smaller dimensions. We fit these curves by minimizing the total squared error, using a single common I_b and κ,.

ignore small voltage offsets of the input voltage, while retaining a monotonic saturating output characteristic for larger inputs.

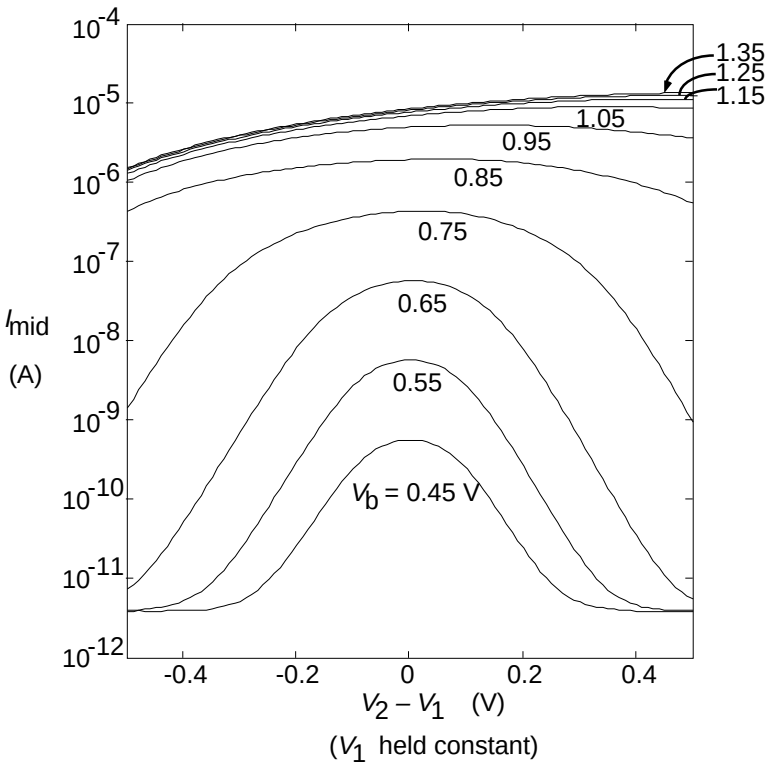


FIGURE 3.6 Effect of bias level on bump-antibump bump output. The curves plot the output current I_{mid} as a function of the differential input voltage $V_2 - V_1$. In this measurement, V_1 was held constant and V_2 was swept around V_1. The different curves are for different bias voltages. Up to a bias voltage of 0.55 V, the curves are simply shifted on a semi-logarithmic axis, indicating that the circuit is operating in the subthreshold region. The bias transistor on this bump circuit is very strong, so the behavior goes above-threshold for a relatively small bias voltage. When the circuit goes above threshold, the bump widens out and eventually becomes asymmetrical.

The response of the bump amplifier can be easily computed in the same manner as used for the bump-antibump circuit. The result of this computation is

$$I = \frac{I_b \tanh\frac{\kappa\Delta V}{2}}{1 + \frac{S}{4}\operatorname{sech}^2\frac{\kappa\Delta V}{2}} \quad . \tag{11}$$

For $S = 0$ (middle leg absent), this expression reduces to the familiar form for a transconductance amplifier:

$$I = I_b \tanh\frac{\kappa\Delta V}{2}. \tag{12}$$

Figure 3.8 shows the measured responses of bump amplifiers with different values of S, along with fits of the form Equation 11. As for the bump-antibump circuit, the fitted value for S is much larger than the drawn values.

Dick Lyon has observed that the bump amplifier may be used as a transconductance amplifier with an increased linear input-range, compared with the simple transconductance amplifier (personal communication). This device may be useful in filter circuits in which one amplifier must sat-

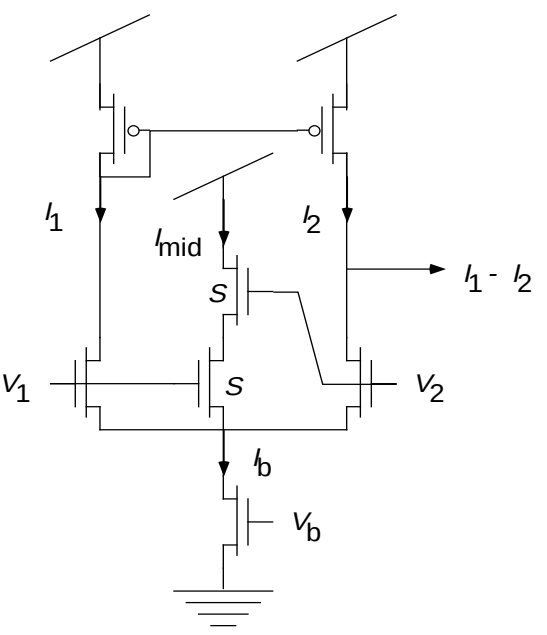


FIGURE 3.7 Bump amplifier.

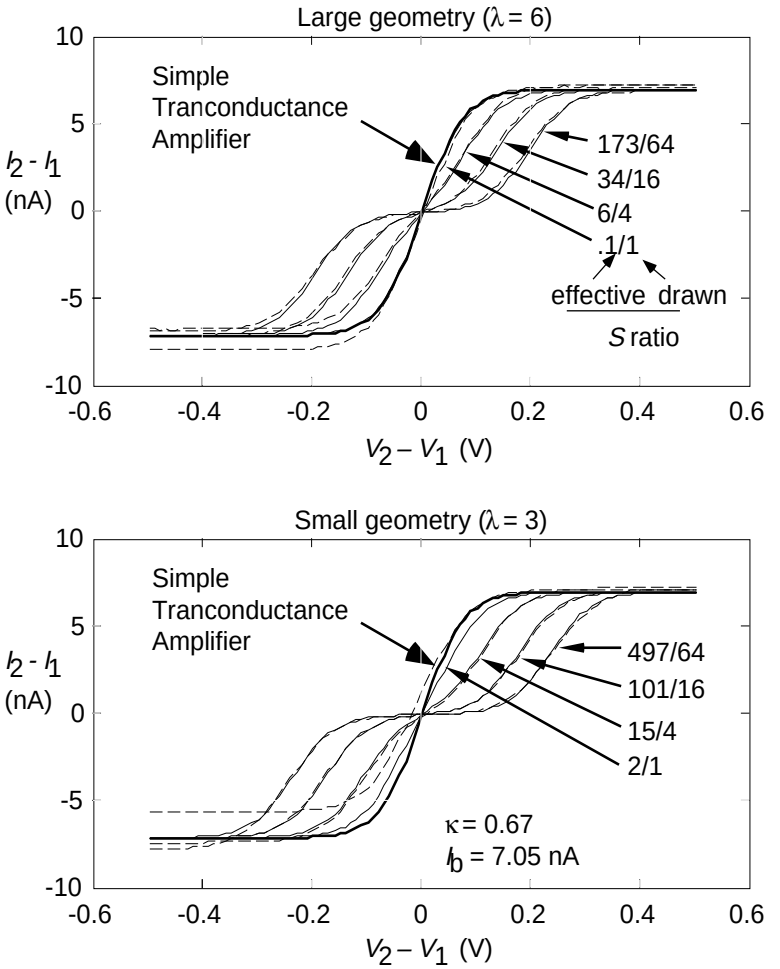


FIGURE 3.8 Responses of bump amplifiers of various *S* ratios. Dotted curves are data, solid curves show fits of the form Equation 11, and thick curves show theoretical simple transconductance amplifier response. Numbers next to curves show fitted and intended *S* values for curves. For top set of curves, minimum transistor dimension was 6 μm, correlating transistors had widths 6, 12, 24, and 48 μm, and outer transistors had lengths 6, 12, 24, and 48 μm. For bottom set of curves, all transistor dimensions were halved. Effects of offsets are visible on some of the curves, and these offsets occur primarily on the side of the response where the output current is mirrored.

urate at a larger voltage than another. It may also be useful in circuits that must decrease power consumption in the middle of the operating range. By choosing a certain value of S, we can make the response of the bump amplifier maximally linear. To find the desired intermediate value of S, we can minimize the "acceleration" of I at $\Delta V = 0$, by setting the third derivative of I with respect to ΔV to zero, and solving for S. It turns out that the desired value of S is 2, independent of κ. By examining Figure 3.8, we can see that small variations in S do not greatly affect the linearity of the response. For example, the curve with an S ratio of 6 is difficult to distinguish from a straight line passing through the origin.

The input range of this modified amplifier is larger than that of a simple transconductance amplifier. We can think of the expanded range as arising from an adaptive bias current. For small input, the effective differential-pair bias current is small, because part of the total bias current I_b is supplied by I_{mid}, and hence is stolen from the differential pair. For larger input, I_{mid} becomes smaller, and the effective differential-pair bias current becomes larger, increasing the linear input range. We can compute the increase in the linear range by doing a Taylor expansion of Equation 11, centered around $\Delta V = 0$, using the value $S = 2$ computed earlier for the optimally linear case:

$$\frac{I}{I_b} = \frac{2x}{3} - \frac{8x^5}{135} + o(x^6), \text{ where } \quad x = \frac{\kappa \Delta V}{2} \quad . \tag{13}$$

A Taylor expansion of the simple transconductance amplifier response is

$$\frac{I}{I_b} = x - \frac{x^3}{3} + \frac{2x^5}{15} + o(x^6), \text{ where } \quad x = \frac{\kappa \Delta V}{2} \quad . \tag{14}$$

From these expansions, we can see that the bump amplifier has a transconductance that is 2/3 that of the simple transconductance amplifier for the same saturation current. Also, the term of order ΔV^3 has vanished, leaving only a residual of order ΔV^5. Figure 3.9 shows a theoretical plot comparing the bump amplifier and the simple transconductance amplifier responses.

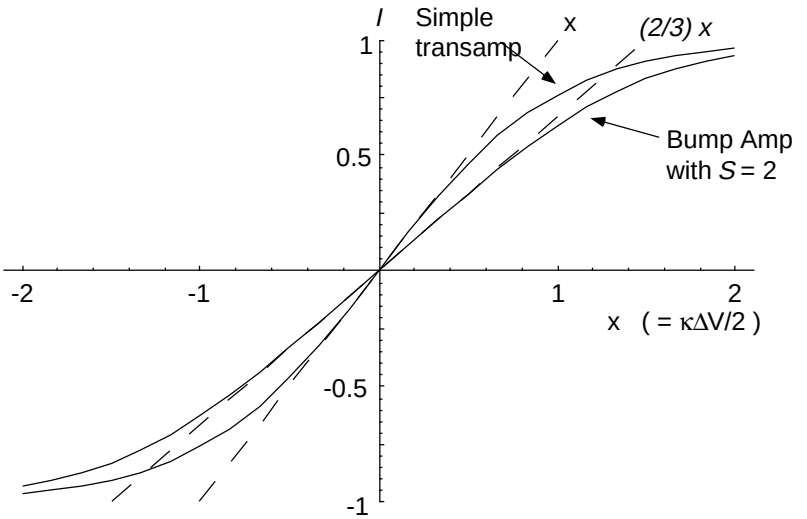


FIGURE 3.9
Theoretical comparison of simple transconductance response with maximally linear bump amplifier response. Solid curves show the simple transconductance amplifier and bump amplifier responses. Dashed curves show line through origin with same slopes as curves. The slope of the bump amplifier response is 2/3 that of the simple transconductance amplifier, and the deviation from linear behavior occurs at a larger voltage.

APPLICATIONS OF BUMP CIRCUITS

Table 3.1 lists examples of applications of bump and antibump circuits in systems. In several cases, we can think of part of the system behavior as a specialized RBF network. The systems in Table 3.1 fall into three categories:

1. Systems that use antibump circuits to compute the distance away from a reference signal — generally in a generalized computation of the power in a signal.
2. Systems that use bump circuits to classify signals into categories.
3. Systems that require a modified transfer characteristic.

None of these example networks incorporate long-term learning. This development awaits invention of efficient storage and learning circuits.

Name	Author(s)	Description
Motion chip.	Delbrück (this thesis) [3]	Antibump circuit used to measure energy in delay line activity, by computing distance of delay line voltage from reference level.
Stereopsis	Mahowald [12]	Bump circuit used in conversion of value encoding (voltage) to place encoding (localized current).
Bump fuse	Liu & Harris [10]	Antibump circuit used to measure voltage difference to break resistive fuse at a controllable threshold.
Spike classifier	Kerns & Watts [5]	Bump circuit used to classify spike height.
Focus	Tobi Delbrück [3]	Antibump circuit used to measure energy in spatial pattern, by comparing image to smoothed reference image.
Herroult-Jutten network	Cohen and Andreou[2]	Bump amplifier configuration used in Gilbert multiplier to approximate a cubic term.
Cochlea	Lyon et al. [11]	Bump amplifier used to stabilize filter circuit.

TABLE 3.1 Applications of bump circuits.

Image Compression	Tawel [15]	Simple bump circuits used to classify pixel blocks for vector quantization.
Gradient Descent	Kirk, et al.[7][8]	4-input bump circuit used as target function for on-chip gradient descent implementation.

ELECTROSTATICS OF THE SUBTHRESHOLD TRANSISTOR CHANNEL

Analysis of the bump-antibump response curves for different S ratios tells us that the circuits show a larger S ratio than we expect from the drawn layout (Figure 3.5). Either short wide transistors are stronger, or long narrow transistors are weaker, than we expect from the drawn layout. This effect is large, and therefore of practical importance.

The transistor strength discrepancy is interesting from the device-physics perspective, because it points out that a transistor is really three-dimensional. We think mostly about the physics along the channel of a transistor (e.g. the Early effect), or the physics vertically through the channel (e.g. the body effect). Only rarely do we consider the physics *across* the channel of the transistor, although this physics is important in determining the strength of the transistor for subthreshold operation.

Electrostatics across and along the channel

Figure 3.10 shows a conceptual† model of the three-dimensional electrostatics around a MOS transistor channel. Part (a) of the figure shows a mesh plot of the surface potential in and around a transistor channel. The source-drain regions are at the lowest potential. The bulk covered with thick oxide is at the highest surface potential. The channel is at an intermediate potential. Part (b) of the figure conceptually shows what happens to the potential across the channel of the transistor, perpendicular to the flow of current, for different transistor widths. The fringing field causes a bowl-shaped potential across the channel. In subthreshold, the concentration of carriers is exponential in the potential, and hence the effective width of the channel is smaller than the drawn width. This effect is larger for channels with smaller width, both because the fractional effect of a fixed width is larger, and because the bottom of the surface potential starts to lift off its wide-channel minimum. Dave Kewley (personal communication, [6]) has designed devices that rely on this effect to modulate the flow of current, through the explicit use of side gates on the transistor.

For digital circuits, the resulting small shift of threshold voltage is not important. Because the threshold voltage is only affected slightly, the literature has not emphasized this across-channel effect (but see [1][9][16]). For subthreshold operation, however, the threshold voltage has an exponential effect on I_0, the pre-exponential factor for the subthreshold transistor law, and hence can have a large effect there.

We distinguish the across-channel effect just discussed from the usual dimensional effects. It is well-known that the physical *length* of a transistor channel is smaller than the drawn layout, because lateral diffusion of the source-drain implants causes the implant to extend under the gate region. The typical distance is about a quarter micron in the 2 μm process used in these chips. The length is further reduced by the finite width of the depletion regions surrounding the source and drain. Modulation of the drain depletion region by drain voltage results in the Early effect. It is also well-known that the physical *width* of a transistor is smaller than the drawn layout, because the process of field oxide growth eats under the masking nitride, causing **bird beaks**. These bird beaks make the oxide thicker, in regions where the drawn geometry indicates thin oxide. The thickened oxide effectively makes the channel narrower. The bird beaks are typically on the order

† By conceptual, we mean that the potentials are what we expect, based on the geometry of the transistor and experience with similar structures. We have not explicitly computed these potentials starting from electrostatic theory.

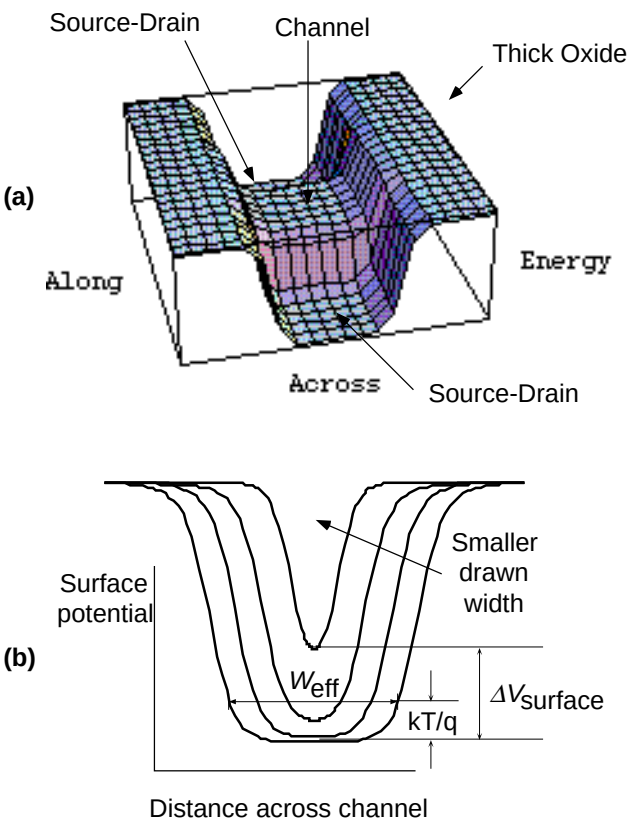


FIGURE 3.10 Conceptual plots of surface potential in and around transistor channel.

(a) Three-dimensional transistor channel. Axes labels show orientation with respect to channel.

(b) Potential across channel, for different drawn channel widths. For wide channels, the effective width is smaller than the drawn width. For narrow channels, the energy barrier is raised over the wide-channel value.

of an eighth to a quarter micron wide. These two effects are schematically illustrated in Figure 3.11.

We shall present three pieces of evidence regarding the transistor strength discrepancy that indicate an effect that is dominated by an across-channel electrostatic field, and that is not simply modeled by the dimensional reductions just discussed.

1. Data from the bump-antibump circuits tells us the scale of the effect is much larger than can be accounted for by the usual dimensional reductions. This data, however, does not disambiguate along-channel and across-channel effects.

2. Data about threshold voltages for various sizes of transistors tells us that the effect is mainly seen in narrow transistors, rather than in short transistors.

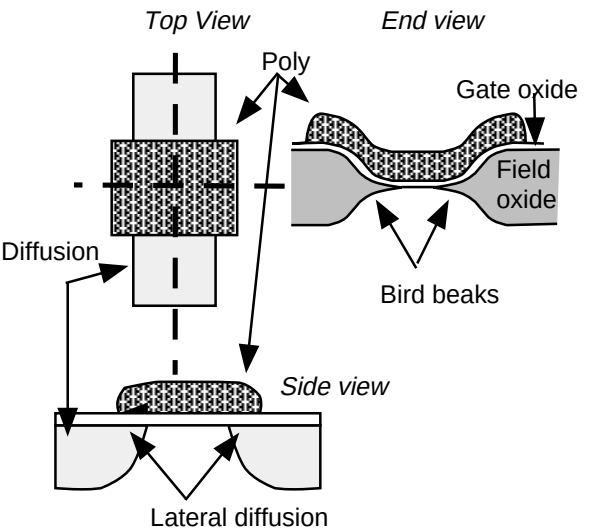


FIGURE 3.11 Lateral diffusion and oxide encroachment effects in transistor fabrication. We show three views of a transistor: from the top, a cross-section along the channel, and a cross-section across the channel.

3. Data that we measured from various sizes of transistors confirms the dominance of the effect in narrow transistors, and shows that the effect depends on the transistor operating regime. In other words, the transistor size, relative to other transistors, is a function of whether the transistor operates in weak, moderate, or strong inversion.

To support our conclusions, we will discuss these three items.

Bump circuit data

Figure 3.12 shows the relation between the intended S ratio and the actual value measured by fitting the antibump response curves in Figure 3.5. To understand these data, we will model the effect of the width and length reductions on the strength of the transistor. Hence, we will *pretend* that the effects are only due to a dimensional reduction in the size of the transistor channel. This analysis will be sufficient for phenomenological understanding, and later we will discuss the physics underlying the behavior, and the validity of the assumption that a simple dimensional reduction can account for the observed behavior.

We can make a simple model for the effect of width and length reductions on S. The effect of *length* reductions are significant only on the *short* transistors in the circuit, namely, the transistors in the middle leg. Similarly, the effect of *width* reductions are significant only in the *narrow* transistors in the circuit, namely, the transistors in the outer legs of the circuit. Hence, our simple model

includes only those geometrical reductions shown in Figure 3.13. In the limit of large S, the effective S is related to the drawn S by

$$S_{\text{eff}} = \frac{S_{\text{drawn}}}{1 - \frac{\Delta W}{W} - \frac{\Delta L}{L}} \quad . \tag{15}$$

Equation 15 does *not* include the effect of the width and length reductions on the long dimensions of the transistors, because these effects are only significant for small S, and including these effects makes the expression much uglier. If we *include* these effects, we obtain the fitted curves in

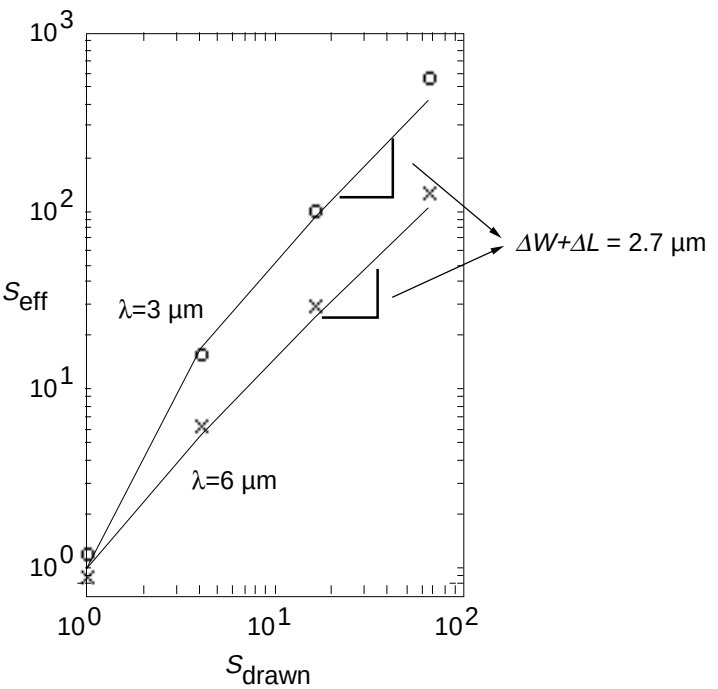


FIGURE 3.12 The relation between S intended by the layout geometry and the measured S. These data were derived from the curves in Figure 3.5. The top data are from the bump circuits with minimum dimension 3 μm, while the bottom data are from the circuits with minimum dimension 6 μm. The solid lines represent a fit to these data assuming that the discrepancy is due to a width reduction ΔW and a length reduction ΔL in all the transistors. The result of the fit show that $\Delta W+\Delta L$ is approximately 2.7μm.

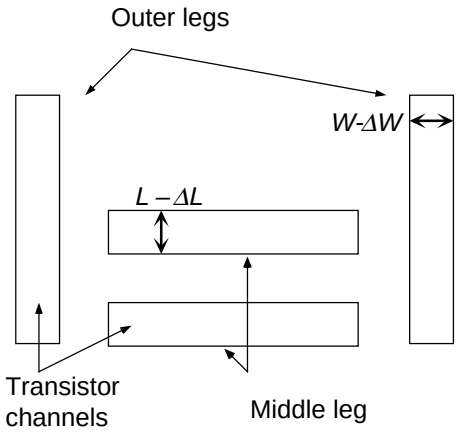


FIGURE 3.13 Geometrical model of length and width reductions in bump circuits. Rectangles show transistor channels.

Figure 3.12, which are accurate fits even for small S. The slopes of these fitted curves, along with the drawn dimensions of the transistors, lets us conclude that

$$\Delta W + \Delta L \approx 2.7\mu\text{m} \tag{16}$$

The data indicate an effect that is much larger than we would expect from the half micron scale for the size of the bird beaks and lateral diffusion. In four of the bump circuits, the small transistor dimensions are only 3 μm, so Equation 16 is a substantial effect. The circuits that gave us the data in Figure 3.12 confound possible length and width variations, because we varied both length and width equally in the layout. As a result, we cannot say what are the relative influences of width and length effects.

MOSIS test parameters

MOSIS, the fabrication service, supplies parameter test results with every processing run. From these tests, we can obtain more clues as to the nature of the strength effect. The tests include measurements of the threshold voltages for different sizes of transistors. The threshold voltage is a direct measure of the channel energy barrier. There are many different flavors of threshold voltage, but all simply measure the gate voltage that makes the mobile carrier concentration "equivalent" to the space-charge density. The exact meaning of equivalent defines the particular variant of threshold voltage. A higher threshold voltage directly translates into a higher channel potential. Each $kT/q\kappa$ in threshold voltage means that at a given subthreshold gate voltage, the carrier concentration is e-fold smaller. The preexponential constant I_0 is simply a measure of the channel car-

rier concentration at the arbitrary gate and source voltage of zero volts. Hence, a higher threshold voltage translates into an exponentially smaller I_0.

The threshold voltage measurements are shown in Table 3.2 for the run that produced the chip that resulted in the data in Figure 3.5 and Figure 3.12. We can see that the threshold voltage for the narrowest transistor (3 µm wide by 2 µm long) is 94 mV higher than the threshold voltage for a wide transistor of the same length (18 µm wide by 2 µm long). A larger threshold voltage translates to a weaker transistor, for subthreshold operation. A threshold shift of 94 mV means that the current, at a given subthreshold gate bias voltage, is about $e^{\kappa(94\text{mV})/V_T} \approx 15$ times smaller. Hence, the narrow transistor is about 15 times weaker than the layout would predict given identical threshold voltages. If we wanted to account for the threshold voltage shift by assuming a geometrical reduction in the width of the transistor, we would need to assume a width reduction of 2.8 µm, because then the effective width would be 0.2 µm, 15 times smaller than the drawn width of 3 µm.

In contrast, the threshold voltage for the long wide transistor (50 µm by 50 µm) is only 26 mV larger than the threshold voltage for the short wide transistor (18 µm wide by 2 µm long). This translates to a length reduction of about 1 µm, using the same reasoning as before. In summary, this data shows that most of the effect of subthreshold geometrical correction can be accounted for by an effective reduction in the width of the transistor.

W/L	V_{Th}
3/2	.905
18/2	.811
50/50	.837

TABLE 3.2 Threshold voltage behavior vs. transistor size, from MOSIS parameters.

Our own transistor measurements

We investigated this question directly, using a test chip supplied by Bhusan Gupta. The test chip has a number of different sizes of transistor. We measured the drain currents as a function of gate voltage, holding the source voltage at the substrate potential.

The results of these measurements are shown in Figure 3.14. The transistor strengths are roughly proportional to the *W/L* ratio. However, there is a systematic bias-level effect that is not explained by a constant geometrical factor. The effect is strongest in the subthreshold operating region and becomes smaller, approaching a constant level, for above-threshold operation. Com-

paring the effect for the 4 wide by 20 long transistor with the 20 wide by 20 long transistor, we see that the 4 wide transistor is about 0.5 times a strong as drawn geometry would predict, for subthreshold operation. For above-threshold operation, the effect reduces to a factor of about 0.8. The width reduction needed to explain this effect makes a corresponding variation from 2 µm in subthreshold to about 0.8 µm above-threshold. In contrast, by comparing the 30 wide by 6 long transistor with the 20 wide by 20 long transistor, we see again that the length effect is smaller and much more constant.

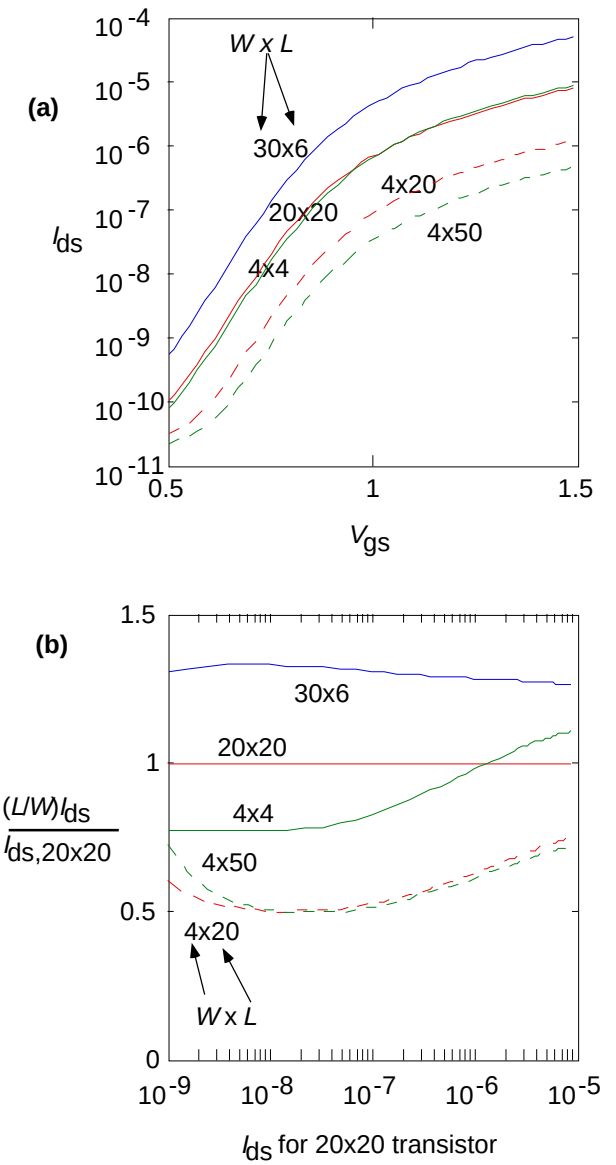


FIGURE 3.14 Effect of transistor sizing on transistor strength. **(a)** shows the drain current as a function of the gate-source voltage, for various transistor sizes. **(b)** shows the ratio of the transistor current, normalized by the *W/L* ratio for the transistor, to the current in the 20 x 20 transistor. These curves are plotted vs. the current in the 20 x 20 transistor. If the transistor current scaled exactly like the transistor *W/L* ratio, then all the curves in (b) would be identical, i.e., unity. For all these data, the source of the transistor was grounded, i.e., held at the bulk potential.

The overall size of the effect that we see in these measurements is comparable to the measured bump circuit effects and the MOSIS parameters. For example, the largest effect we see in our measured transistor data comes from a 4 μm-wide transistor, where the strength is half what the drawn *W/L* ratio predicts. This discrepancy can be accounted for by an effective width reduction of 2 μm.

We can summarize the transistor strength effects in terms of a model where the subthreshold channel potential has a bowl-like shape across the channel. The radius of curvature of the edges of the bowl, in subthreshold, is on the order of a micron by a thermal voltage. In other words, about a micron from the channel edge the channel potential is about a thermal voltage higher than the middle of the channel. For wide channels, the effect appears as an effective width reduction. For narrow channels, the effect appears as an increase in the threshold voltage, or a decrease in I_0.

SUMMARY

We have seen how to build simple circuits consisting of less than 7 transistors that compute powerful similarity and dissimilarity measures. The simple current correlator correlates analog currents, producing a self-normalized output current. The simple bump circuit computes the similarity between voltage inputs, producing a current output. The bump-antibump circuit computes a bump output current—the similarity output—and antibump output currents—the dissimilarity outputs. In addition, for the bump-antibump circuit, the drawn layout controls the width of the bump. The same transistor configuration lets us make amplifiers with a wider input range. A discrepancy between measured and expected transistor strength forced us to look at the physics underlying subthreshold behavior, and we investigated an important effect that makes transistors act much weaker than their drawn layout, in subthreshold operation.

ACKNOWLEDGEMENTS

John Harris had the idea of calling these circuits “bump” circuits. Dave Kewley, Dick Lyon and Carver Mead contributed significantly to this chapter, in terms of a critical analysis of the transistor strength discrepancy. Dick Lyon analyzed the bump amplifier configurations and realized they would be useful in filter circuits. Doug Kerns provided contextual information about commercial split-gate mixer circuits. Dave Kewley showed me his own measurements of transistor

width and length. Dick Lyon and Dave Kewley provided critical comments on earlier versions of this chapter. Bhusan Gupta supplied a test chip.

REFERENCES

1. L.A. Akers and J.J. Sanchez, "Threshold voltage models of short, narrow and small geometry MOSFETS: a review," *Solid State Electronics*, vol. 25, pp. 621–641, 1982.

2. 1 M.H. Cohen and A.G. Andreou, "MOS circuit for nonlinear Hebbian learning," *Electronics Letters*, vol 28, pp. 809–810, 1992.

3. T. Delbrück, "Analog VLSI predictive visual motion processing," *IEEE Trans. on Neural Networks*, in press, 1993.

4. J. Hertz, A. Krogh, R.G. Palmer, *Introduction to the theory of Neural Computation*, Boston, MA: Addison Wesley, 1991.

5. D.A. Kerns and L. Watts, personal communication, 1992.

6. D. Kewley, personal communication, 1992.

7. D. Kirk, *Accurate and precise computation using analog VLSI, with applications to computer graphics and neural networks*, Ph.D. Thesis, California Institute of Technology, Caltech-CS-TR-93-??, 1993.

8. D. B. Kirk, D. Kerns, K. Fleischer, A.H. Barr, "Analog VLSI implementation of multi-dimensional gradient descent," *Neural Information Processing Systems*, vol. 5, 1993 (in press).

9. E.H. Li, K.M. Hong, Y.C. Chen, K.Y. Chan, "The narrow-channel effect in MOSFETS with semi-recessed oxide structures," *IEEE Trans. on Electron Devices*, vol. 37, pp. 692–701, 1990.

10. S.C. Liu and J.G. Harris, "Edge-Enhancing Resistive Fuse," SPIE Technical Symposia on Optical Engineering and Photonics in Aerospace Sensing, Orlando, FL, vol. 1473, pp. 185-193., 1991.

11. D. Lyon, personal communication, 1992.

12. M.A. Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function*, Ph.D. Thesis, California Inst. of Tech., Dept. of Computation and Neural Systems, Pasadena CA, 91125, 1992.

13. C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison--Wesley, 1989.

14. M. Sivilotti, *Wiring Considerations in Analog VLSI Systems, with Application to Field-Programmable Networks*, Ph.D. Thesis, California Institute of Technology, Dept. of Computer Science, Pasadena, CA, June 1991.

15. R. Tawel, "An analog processor for image compression," *IEEE J. Neural Networks, Special Hardware Issue* (in press).

16. P.P. Wang, "Device characteristics of short-channel and narrow-width MOSFETS," *IEEE Trans. on Electron Devices*, vol. ED-25, pp. 779–786, 1978.

2

A SYSTEM EXAMPLE

CHAPTER

4

SILICON RETINA WITH VELOCITY-TUNED PIXELS

In this chapter, we discuss a two-dimensional silicon retina that computes a complete set of local direction-selective outputs.† The chip motion computation uses unidirectional delay lines as tuned filters for moving edges. Photoreceptors detect local changes in image intensity, and the outputs from these photoreceptors are coupled into the delay line, where they propagate with a particular speed in one direction. If the velocity of the moving edges matches that of the delay line, then the signal on the delay line is reinforced. The output of each pixel is the power in the delay line signal, computed within each pixel. This power computation provides the essential nonlinearity for velocity-selectivity. The delay line architecture differs from the usual pairwise correlation models in that motion information is aggregated over an extended spatiotemporal range. As a result, the detectors are sensitive to motion over a wide range of spatial frequencies.

I have designed and tested functional one- and two-dimensional silicon retinas with direction-selective, velocity-tuned pixels. A velocity-selective detector requires only a single delay element

† A brief abstract reporting preliminary results appeared as T. Delbrück, "A silicon network for motion discrimination that uses spatiotemporal interpolation," *Soc. of Neuroscience Abstracts*, no. 143.8, p. 344, 1991.

and nonlinearity for each tuned velocity, and is sensitive to both light and dark contrasts. The use of adaptive photoreceptors and compact circuits makes for a well-conditioned input signal and small circuit offsets, resulting in robust operation. All circuits work in subthreshold, resulting in low power consumption. Pixels with three hexagonal directions of motion selectivity are approximately (225 μm)^2 area in a 2-μm CMOS technology, and consume less than 5 μW of power.

ANALOG HARDWARE MOTION COMPUTATION

In this introduction, I will place the new chip in the context of previous approaches to analog-hardware visual motion computation. Essentially, there are two approaches: Those that use spatiotemporal image *gradients*, and those that use spatiotemporal image *correlation.* Hardware systems that utilize each approach have been previously built. The chip reported here falls into the second category.

Tanner and Mead built the first analog two-dimensional optical-flow chip [35][36][30]. The Tanner chip computes a *single* motion vector for the *global* two-dimensional optical flow, using a gradient-based scheme (for references, see [28]). The motion vector is computed cooperatively by a network of circuits, each circuit using local spatiotemporal derivative information. The circuits reach a consensus on the global image motion vector, using the overall constraint that the total derivative of the image intensity is zero. This constraint allows the chip to solve the aperture problem, which arises because a *local* measurement cannot detect motion *along* an edge [28].

The two-dimensional direction-selective chip built by Benson and Delbrück [5] performs a different and much simpler computation that is based on biological studies. The pixel outputs respond selectively to local rightward motion, and the motion computation is feedforward, rather than cooperative. The architecture is based on the idea of null-inhibition direction selectivity, developed by Barlow and Levick [4] to explain the direction selective responses found in rabbit retinas. In a null-inhibition model, the pixel output is inhibited by edges coming from the null-direction. In the preferred direction, the pixel output in response to an edge is not inhibited, because the inhibition arrives after the excitation.

I distinguish the *value*-encoded global velocity-vector output of the Tanner chip, from the *place*-encoded, direction-selective, velocity-tuned outputs of the Benson–Delbrück chip. The computations are qualitatively different, because one encodes the pattern velocity as a vector with magnitude and direction, and the other encodes the motion of features by the location of active

outputs. The Tanner chip does a higher-level computation, integrating spatiotemporal image information from the entire moving image in a cooperative calculation, whereas the Benson–Delbrück chip simply responds selectively to rightward versus leftward motion.

The operation of the Tanner chip is not robust, perhaps because of the demands of the mathematically sophisticated motion algorithm it implements. High-contrast images, with sharp edges, are required if the chip is to produce a direction-selective output, and the precision of the velocity measurement is poor. The reason for this nonideal performance is the combination of inherent susceptibility of gradient-based schemes to temporal and spatial noise, low-sensitivity receptors, and large circuits with poor offset characteristics. The strongest signals are derived from places in the image with sharp gradients and high contrasts, but gradient-based schemes require images with smoothly varying contrast if they are to obtain reliable estimates of spatial and temporal derivatives. The performance of the chip is compromised severely by these conflicting requirements.

The operation of the Benson–Delbrück chip is robust in comparison with that of the Tanner chip, because the Benson–Delbrück chip, unlike the Tanner chip, responds correctly to real input scenes of moderate contrast. There are several reasons for this performance difference. The algorithm is much simpler, the circuits are more compact and hence less offset prone, and a well-conditioned input is generated by the use of adaptive photoreceptors. However, this comparison is arguably pointless, given the major differences in computation performed by the two chips.

The null-inhibition architecture has the advantage of a large degree of direction-selectivity and hence, inherently robust operation. The disadvantage is that it is difficult to produce a response-versus-speed tuning that is other than a simple lowpass determined by the pixel spacing, so construction of a set of detectors tuned to different speeds is not straightforward. Moreover, this architecture does not deal with the aperture problem, and, in fact, knows nothing of global motion constraints.

The chip described here has place-encoded outputs that encode information about components of local edge motion along the detector directions, like the Benson–Delbrück chip. It is also like the Benson–Delbrück chip in that the motion computation does not deal with solving the aperture problem. It is like the Tanner chip in that the computation integrates information over an extended spatiotemporal region. Our chip is the first functional analog device that computes a full set of local direction-selective outputs in two dimensions. It is also the first reported chip that uses correlation-based *analog* computation to produce direction-selective, velocity-tuned outputs. (Ref-

erence [22] reports a correlation-based one-dimensional motion circuit that correlates digital pulses, produced by image intensity changes, in analog time.)

I omit discussion of other analog motion chips that have been fabricated only as one-dimensional circuits [22][23]. For the most part these architectures cannot be fabricated practically in a two-dimensional architecture, because they are not parsimonious with wire. I also omit discussion of analog change-detecting circuits [8][13], digital motion circuits [2], and discrete-component correlation-detector systems [16][32].

The remainder of this chapter is organized as follows. We start with a heuristic description of correlation motion detectors, and of the delay line motion architecture in one and two dimensions. We then describe a mathematical model of the circuit operation in the time and frequency domains, and investigate the model to understand the limiting behavior, and the differences between the new model and previous pairwise correlation detectors. Starting on page 148, we describe the circuits used in the chip. Starting on page 150, we discuss qualitative and semiquantitative measurements on the two-dimensional motion chip, and quantitative measurement on the one-dimensional chip. The chapter concludes with a short discussion of future directions of this work and a summary of the results.

CORRELATION-BASED MOTION DETECTORS

The simplest correlation-based motion detector is shown in Figure 4.1(a). A pair of spatially separated, feature-detecting cells feeds into a nonlinearity that detects coincidence of the two inputs. One of the inputs to the nonlinearity is delayed. When the motion is matched to the detector spatial separation and delay, the output becomes large. The nonlinearity, although often formulated as a multiplicative operation, can be any operation that detects *coincidence* of the inputs. Hence, the correlation detector may linearly combine the spatiotemporally separated inputs, and then use a threshold element, or any other expansive (faster-than-linear) nonlinearity.

The correlation detector was conceived by Hassenstein and Reichardt during experiments on a beetle visual system [18]. Their model, and later modifications worked out through experiments on the fly visual system, are more complicated than the model in Figure 4.1(a) [33]. The major elaboration is the addition of a complementary direction selectivity. The output from the Hassenstein-Reichardt model is the difference between two outputs tuned to complementary directions. This differencing operation removes the common-mode signal.

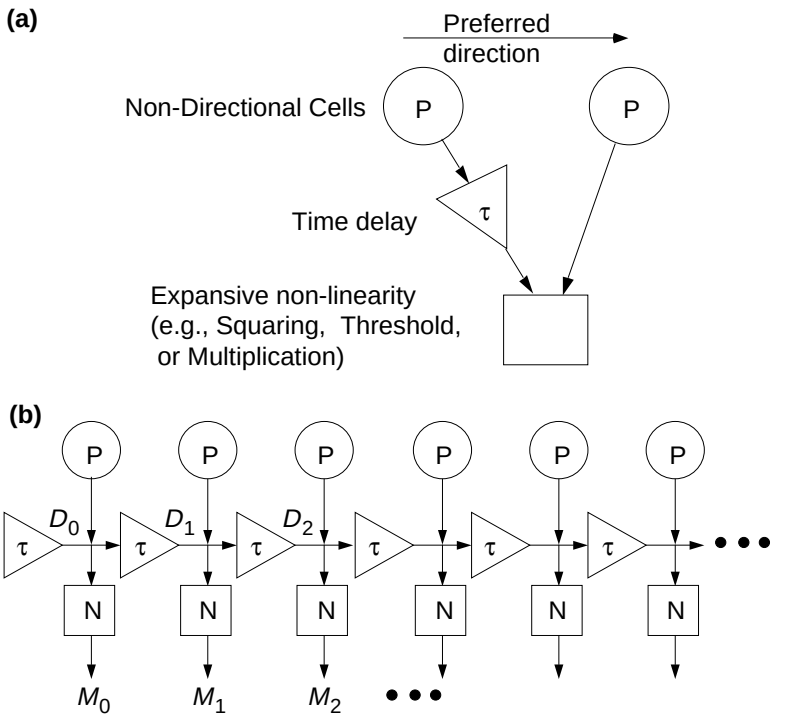


FIGURE 4.1 Correlation-based motion-detectors. **(a)** Pairwise correlation detector. Nondirectional feature-detecting cells feed into an output nonlinear element that measures coincidence between the spatiotemporally separated inputs.The exact nature of the nonlinearity is not important, as long as it is expansive. **(b)** Extended correlation detector. Nondirectional cells (P) couple activity into a unidirectional delay line with delay τ per stage. The delay line signal (D_n) is large when the image motion is matched to the delay line speed. The expansive output nonlinearities (N) compute a velocity selective signal M_n that is monotonic in the magnitude of the delay line signal.

Extension of the simple correlation detector

My motion architecture extends the pairwise correlation model in Figure 4.1(a), by using a unidirectional *delay line* as a tuned filter for a particular velocity, as shown in Figure 4.1(b). The input to the delay line comes from photoreceptors that respond to local intensity change. The photoreceptor outputs are coupled capacitively into the delay line. The delay line is composed of low-pass filters. An edge passing over a photoreceptor creates a traveling signal in the delay line that spreads and decays with time. If the velocity of the edge is matched to that of the delay line, the

successive inputs to the delay line pile up in synchrony. If the input velocity is not matched to the delay line velocity, the successive inputs arrive in asynchrony and do not pile up.

It is well known that the time-averaged output of a linear system cannot be direction or speed dependent. In a linear system, we can alter the time order of the input without changing the time-averaged output, because the response to a sum of inputs is the same as the sum of the responses to each input individually. Hence, the *average* signal on our delay line is not velocity selective. The *amplitude* of the signal clearly *is* velocity selective. Measuring amplitude is a nonlinear operation. We can use any nonlinearity that measures some metric of the amplitude of the delay line signal, and the result will be direction and speed selective. On my chip, I use a bump circuit, which computes an approximate squaring function of the delay line signal. Hence, the nonlinearity approximately measures the filter output power.

The nonlinear operation is an even function of its input, so the delay line activities caused by edges of light and dark contrasts are both detected. A single nonlinearity, and a single delay line, are sufficient to compute a direction-selective, velocity-tuned response for either sign of image contrast.

The delay line architecture is robust against noise, because a particular output integrates information from an extended spatiotemporal range of the moving image. This feature is shared with the human visual system [31].

Two-dimensional architecture

In the two-dimensional chip, I use a hexagonal architecture that encodes any possible direction of motion using three directions (Figure 4.2). Each pixel computes the response to three principal directions of motion. Since the three directions are nonorthogonal, any direction can be represented by weighted combinations of the three *nonnegative* outputs. This basis set may be familiar from the representation of color space by a set of three primary colors. In an orthogonal architecture, we would need four principal directions, each with nonnegative output, to represent an arbitrary direction of motion, or, equivalently, two bidirectional outputs.

Each pixel contains a single photoreceptor, three delay elements, and three output nonlinearities. The three outputs are scanned out and are displayed on a monitor as three different colors, through use of a scanning frame, as described in [29]. The Reichardt detector uses differencing of complementary directions to reduce common-mode responses. In the two-dimensional motion

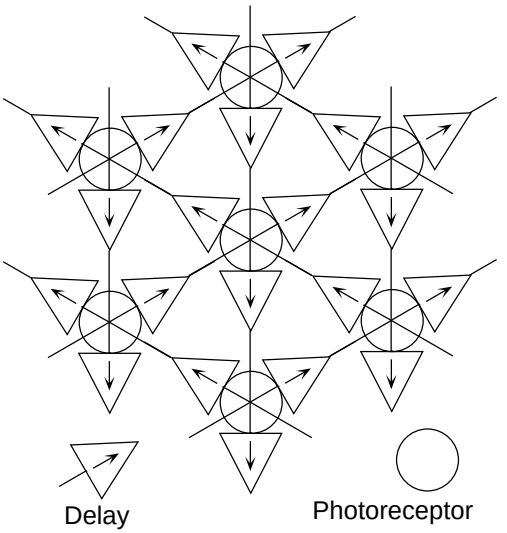


FIGURE 4.2 Two-dimensional architecture of the motion chip. Pixels are arranged in a hexagonal array. Each pixel contains a single photoreceptor, three delay elements, and three output nonlinearities. The delay lines run in the three principal directions shown by the arrows in the delay elements.

chip, I use an analogous scheme to increase the contrast between the three channels of output that is computed serially by the video driver circuits. This linear differencing operation does not do any additional information processing, so I will not discuss it further.

The aperture problem arises in any two-dimensional motion computation. A local edge-based measurement can measure only the component of edge motion orthogonal to an edge, and cannot measure motion along the edge. Another and more general way of stating this fact is that local measurements *constrain* the estimate of the actual pattern movement. An array of detectors measures a set of motion vectors, each member of which is *consistent* with the movement of the pattern as a whole. The different detectors, however, may disagree in their assumptions about the component of motion *along* their local edges. The global motion vector is the vector that most closely agrees with the set of constraints imposed by the local measurements. The Tanner chip ingeniously solves the constraint problem in a cooperative manner.

The delay line motion architecture does no global computation to reconcile local measurements, so the best that it can do, in principle, is the computation of the local motion vectors. In the present architecture, these quantities are computed such that they are consistent with the observed motion, but there is no built-in assumption that the measured motion is only orthogonal to the edge. The nonintuitive feature is that the motion that is consistent with the observed set of outputs may not be orthogonal to the edges. The reason is that a detector is sensitive to only spatiotemporal correlation along its preferred direction. An edge moving at a speed S in a direction oblique to

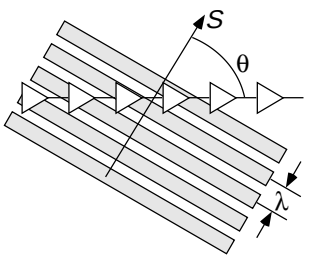


FIGURE 4.3 Apparent motion and wavelength effects. A grating pattern with wavelength λ, moving at speed S and angle θ over motion detector with preferred direction to the right causes an apparent motion along the array at the increased speed *S*/cosθ. The apparent wavelength seen by the array is also increased to λ/cosθ. The "apparent" motion can equally well be caused by true motion to the right at velocity *S*/cosθ of edges that are oriented at angle θ; there is no way, in principle, that the local measurement can distinguish the two.

the preferred direction, with angle θ, causes apparent motion along the preferred direction with speed $S/\cos\theta$ (Figure 4.3).

THEORETICAL ANALYSIS OF MOTION CIRCUIT

To understand the properties of our model, and to derive results that can be verified experimentally, we shall analyze the one-dimensional motion circuit in both the time and frequency domains. The time-domain analysis is more intuitive. The frequency-domain analysis is more useful for comparison with measurement, and for understanding the response of the system to variations in stimulus parameters, such as velocity and spatial frequency. The nonlinearity is a simple feedforward operation done on the delay line signal, so our analysis will concentrate on the linear delay line. Some of the symbols used in this section are shown in Figure 4.1

Time-domain analysis

We shall compute the delay line response to a single moving bright edge. To do this calculation, we compute the transfer function from photoreceptor to delay line, and the transfer function of an *n*-stage delay line. Combining these two transfer functions, we obtain the complete transfer function between the input and the delay line output *n* stages later. Using the transfer function, we compute the step response. The response to a moving edge is a sum over the temporally shifted step responses produced by the edge at successive receptors.

The photoreceptor outputs are coupled capacitively into the follower-integrator delay line section. Hence, the transfer function from receptor to delay line is the single high-pass filter

$$H_{\text{in}}(s) = \frac{\tau s}{\tau s + 1} \tag{1}$$

where s is the Laplace transform variable and τ is the time constant of each first-order section [30]. (It may be helpful to refer to the circuit diagram in Figure 4.9).The delay line—a chain of n first-order, buffered, low-pass filters—has the transfer function

$$H_n(s) = \frac{1}{(\tau s + 1)^n} \tag{2}$$

We obtain the transfer function from photoreceptor input 0 to delay line output n by multiplying H_{n} and H_{in}. By computing the inverse Laplace transform, we find that the time response at the nth delay line tap to a bright intensity step at the first photoreceptor at time $t = 0$ is

$$r_n(t) = \frac{1}{n!}\left(\frac{t}{\tau}\right)^n e^{-t/\tau} \tag{3}$$

for $t > 0$. For $n = 0$, this function is a step-increase from 0 to 1, followed by a decaying exponential. For $n = 1$, this function is an initially-linear increase from 0, followed by the same decaying exponential. In general, (3) represents a wave packet that travels along the delay line, one stage per time τ. Asymptotic analysis shows that as n grows large, the response approaches a Gaussian shape with constant area. The width of the Gaussian broadens as $\sqrt{n}$, and the height decreases as $1/\sqrt{n}$. Note that the stimulus is a *step*-input at the photoreceptor that is high-pass filtered by the capacitive coupling to the delay line.

A motion stimulus causes a succession of inputs to the delay line, where the time delay ΔT between excitation of successive stages is

$$\Delta T = \frac{L}{S} \tag{4}$$

where L is the spatial separation between detectors, and S is the velocity of the moving pattern. Since the delay line is a linear system, the response to a moving bright edge is the superposition of the responses to the edge passing over all the contributing inputs. We define $R_{\text{n}}(t)$ to be the delay

line response at tap n, in response to a moving bright edge that passes over the first input at time $t = 0$. It is easy to compute $R_n(t)$:

$$\begin{aligned} R_0(t) &= r_0(t) \\ R_1(t) &= r_1(t) + r_0(t-\Delta T) \\ R_2(t) &= r_2(t) + r_1(t-\Delta T) + r_0(t-2\Delta T) \\ &\cdots \\ R_n(t) &= \sum_{k=0}^{n} r_k(t-(n-k)\Delta T) \end{aligned} \tag{5}$$

We have assumed that either the delay line or the motion starts at tap 0. Figure 4.4(a, c), show a graphical representation of the *linear* delay line response to a bright bar moving over the delay line at the optimum velocity and at one half of the optimum velocity. The bright edge causes positive activity, and the dark edge causes negative activity. Both signs of edge-contrast information travel along the delay line with a fixed speed and direction, though the activity has opposite polarity for the two edges. When the edge motion is matched to the speed of the delay line, the activity on the delay line piles up maximally. We can see how, in this linear system, the *integrated* response is unaffected by the stimulus velocity, even though the activity has larger peak magnitude when the motion is matched to the delay line speed. The longer the motion continues, the larger the peak activity. Saturating nonlinearities eventually limit the response of the delay line.

Antibump output nonlinearity

From **Chapter 3**, we know that the antibump output-nonlinearity-circuit used in the motion chips computes the function

$$N(x) = \frac{I_b}{1 + \frac{w}{4}\operatorname{sech}^2 x} \tag{6}$$

where I_b is the bias current, x is the input voltage, in units of approximately $\frac{2kT}{q\kappa} \approx 70$ mV , and $w \approx 20$ is a geometrical layout parameter.† The output is a current. The shape of this function is

† The parameter w is the same as S in Chapter 3; the name is changed here to avoid confusion with the S used here to mean the speed of the moving pattern.

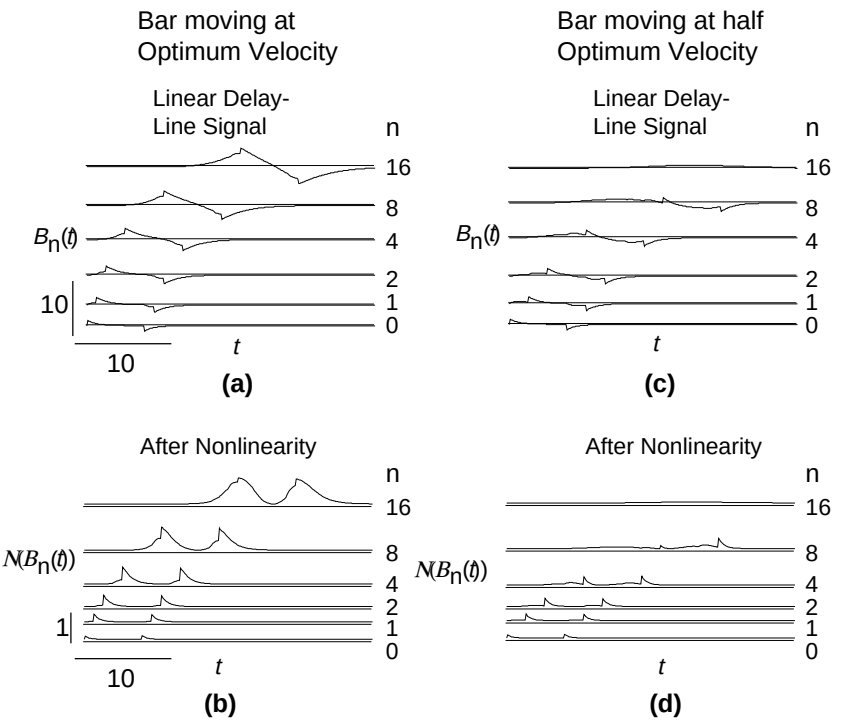


FIGURE 4.4 Theoretical response of delay line to moving bar for two velocities. A bright bar of width 6 pixels, traveling to the right over the delay line, passes over the first tap at time $t = 0$. The mathematical form of the delay line response to the moving bar at the nth tap is $B_n(t) = R_n(t) - R_n\left(t - \frac{6}{S}\right)$, where R is the response to a single bright edge, computed in the text as (5), and S is the speed in pixels/time. Each trace shows the response at a particular tap as a function of time. Tap number is shown to right of trace. The linear delay line signals $B_n(t)$ are shown in **(a)** and **(c)**, and the result of passing the linear delay line signal through the antibump nonlinearity (Figure 4.5), $N[B_n(t)]$, is shown in **(b)** and **(d)**. In (a) and (b), the edge travels at the natural speed of the delay line. In (c) and (d), the edge travels at one-half the speed of the delay line. When the bar motion is matched to the delay line speed, the successive inputs pile up, in response either to the bright or the dark edge. The nonlinearity rectifies $B_n(t)$, so both edge contrasts appear as a positive output. When the motion is not matched, the successive inputs arrive in asynchrony with the delay line signal, and the resulting output is smaller.

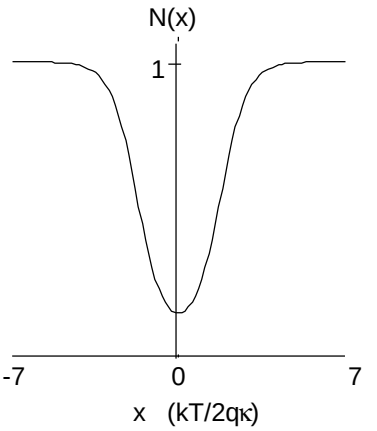


FIGURE 4.5 The antibump output nonlinearity (6). The input is a differential voltage x shown in natural input units, and the output is a current, shown here normalized to 1.

shown in Figure 4.5. The output is parabolic for small input, and saturates at the adjustable bias current. The width of the valley region in subthreshold is determined by layout geometry. The valley width can be increased by using an above-threshold bias current.

The output from the motion circuit is the nonlinearity (6) applied to the signal on the delay line. In response to the moving edge signal $R_n(t)$ on the delay line, the velocity-tuned outputs are $N(R_n(t))$. Figure 4.4(c, d) show the response to the same moving bright bar after the nonlinearity is applied. We can see how the nonlinearity makes both the peak and integrated responses strongly velocity selective, and also how the even characteristic of the nonlinearity results in equivalent detection of both bright and dark edges.

Frequency-domain analysis

In the frequency domain, the stimulus is a one-dimensional sinusoidal variation in contrast, with wavelength λ and velocity S (Figure 4.6). If the velocity is constant, then each input to the motion array is a simple sinusoidal function of time with the same temporal frequency,

$$\omega = \frac{2\pi S}{\lambda} \tag{7}$$

at each input. The signal at a particular tap of the delay line is a sum over all the contributing inputs. Each input is first phase shifted by the spatial separation, and then phase shifted again by the filtering in the low-pass delay line stages. The delay line signal is a pure sinusoidal function of

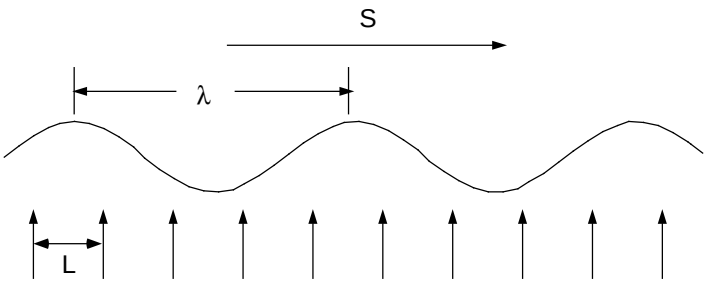


FIGURE 4.6 Sinusoidal grating pattern moving over motion array. The grating moves at speed *S*. Pixel locations are shown as arrows. The pattern has wavelength λ, and the pixel spacing is *L*. The temporal frequency at each pixel is $\omega = 2\pi S/\lambda$. The pattern phase shift between pixels is $\phi = 2\pi L/\lambda$.

time, with the same temporal frequency at each tap, multiplied by the transfer function for that tap.

The pattern phase shift of the input signal between each pixel, due to the pixel spacing and the sinusoidal spatial variation of image contrast, is

$$\varphi = \frac{2\pi L}{\lambda} \tag{8}$$

Each stage of the delay line is separated by a low-pass filter, with transfer function

$$\frac{1}{i\omega\tau + 1} \tag{9}$$

Each input is a single high-pass filter, with transfer function

$$\frac{i\omega\tau}{i\omega\tau + 1} \tag{10}$$

We obtain the complete transfer function for tap n by summing the inputs from taps 0 through n, resulting in

$$\begin{aligned} H_n(\omega) &= \frac{i\omega\tau}{i\omega\tau+1}\left(1+\frac{e^{i\varphi}}{i\omega\tau+1}+\frac{e^{i2\varphi}}{(i\omega\tau+1)^2}+\dots\right) \\ &= \frac{i\omega\tau}{i\omega\tau+1}\sum_{k=0}^{n}\frac{e^{ik\varphi}}{(i\omega\tau+1)^k} \end{aligned} \tag{11}$$

where we have factored out the high-pass filter common to each stage. The summation in (11) is a geometric series, so we can write the transfer function explicitly for an arbitrary number of stages:

$$H_n(\omega) = \frac{i\omega\tau}{i\omega\tau+1}\frac{1-\left(\frac{e^{i\varphi}}{i\omega\tau+1}\right)^{n+1}}{1-\frac{e^{i\varphi}}{i\omega\tau+1}} \tag{12}$$

This expression represents a phasor that rotates about the complex origin at a frequency ω The physically measurable part of the expression is the projection of (12) onto any axis in the complex plane, and is much more complicated than is (12).

The average signal on the delay line in response to a sinusoidal input pattern is zero, regardless of the velocity. The length of the phasor, however, is strongly direction and velocity selective. To obtain a response that is tuned for velocity requires a nonlinear operation that measures the length—on some metric—of the phasor. The delay line signal is fed into the antibump nonlinearity, which computes a squaring operation for small input signals. In explicit mathematical form, the time-dependent velocity-tuned output is

$$O_{\mathrm{n}}(t,\omega) = N\left(\left|H_{\mathrm{n}}(\omega)e^{i\omega t}\right|\right), \tag{13}$$

where N is the antibump nonlinearity in (6). Hence, the operation in (13) is proportional to the squared absolute transfer function, for low-contrast stimuli.

We can learn about the behavior of the system by studying the amplitude of the linear transfer function (12)—that is, the length of the phasor. We shall examine the case of a seven-input system. In Figure 4.7, I have plotted the phasor at the seventh tap, $H_6(\omega)$, in response to the motion of a sinusoidal pattern in the preferred and null directions. The speed of the motion, in each case, is near the optimum speed for the preferred direction. Each phasor is shown as a number of adjoin-

ing line segments; each of these components is the contribution of one pixel input to the complete phasor.

When the grating moves in the preferred direction at the optimum speed, the phasor is as long as it can be, given an input pattern of a given amplitude and wavelength, because the components all point in nearly the same direction. They all point in the same direction because, for each input, the total pattern-phase-shift and the total low-pass-filtering-phase-shift match. There is also an overall rotation of the entire phasor, due to the high-pass filtering at each input to the delay line.

In contrast, when the pattern moves in the null direction, the phasors sum destructively, and the resultant phasor curls up on itself. The complete phasor is short, no matter how many inputs are combined, because the components always approximately cancel one another. The length of the phasor, as a function of the speed of the motion, has wiggles, corresponding to the periodic behavior of the destructive cancellation. These wiggles can be seen in the experimental data shown in Figure 4.15 and Figure 4.17.

Many of the correlation-based direction-selective models in the literature correspond to a two-input ($n = 1$) version of this system. Hence, it is interesting to compare the case of a two-input system with an system that is effectively infinitely long. We shall first derive the mathematical behavior for each system, and then contrast the two systems.

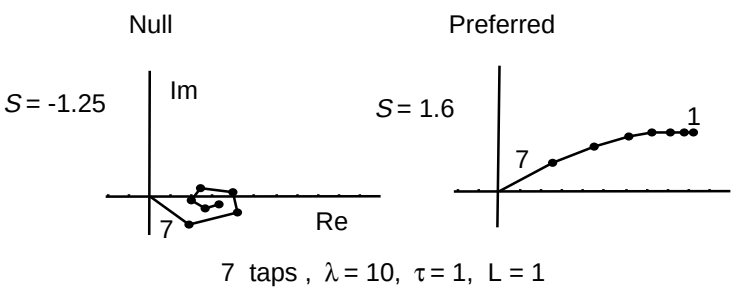


FIGURE 4.7 Phasor summation properties for a seven-stage system. $H_6(2\pi S/\lambda)$ is plotted on the complex plane. The symbols are defined in Figure 4.6. The contributions of individual taps are shown as component line segments; taps 1 and 7 are labeled. The optimum velocity is near $L/\tau = 1$.

Long system

The explicit form of the transfer function for an infinite number of stages is given by (12) with $n \to \infty$:

$$H_{\infty}(\omega) = \frac{1}{1 + \dfrac{1 - e^{i\varphi}}{i\omega\tau}} \tag{14}$$

(Note that this transfer function includes the high-pass input to the delay-line.) By examining the behavior of the n-dependent term in the numerator of (12), we can determine that this limiting form is accurate as long as the following condition holds:

$$n > \left(\frac{\lambda}{L}\right)^2 \tag{15}$$

This restriction seems severe, but examination of the measured tuning curves shows that, in practice, there are only scale differences in the responses after the first few taps. We shall examine the behavior of (14) to understand how it is affected by spatial wavelength.

Since the numerator of (14) is 1, we can find the maximum value of the response by minimizing the absolute value of the denominator. This simple computation reveals that the maximum value of $\left|H_{\infty}(\omega)\right|$ occurs at

$$\omega_{\max} = \frac{2}{\tau}\frac{(1 - \cos\varphi)}{\sin\varphi} \tag{16}$$

corresponding to the speed

$$S_{\max} = \frac{\omega_{\max}\lambda}{2\pi} = \frac{\lambda}{\pi\tau}\frac{(1 - \cos\varphi)}{\sin\varphi} \ . \tag{17}$$

For long wavelengths, $\varphi = 2\pi L/\lambda$ is small, and we find the peak tuning occurs at the frequency

$$\omega_{\max} = \frac{\varphi}{\tau} \tag{18}$$

corresponding to the speed

$$S_{\max} = \frac{L}{\tau} \tag{19}$$

which is reassuring, since this optimum speed corresponds with the intuitive idea that optimum motion is motion matched to the delay-line speed. The optimum speed is invariant with wave-

length, for wavelengths longer than a few pixel spacings. This result is also surprising, because correlation models do not generally display this invariance.

We can see how this behavior comes about by examining the limiting case of a long wavelength. As we *increase* the wavelength λ, we *decrease* the pattern phase shift φ between pixels. (The pattern phase shift is the phase shift of the input, between adjacent pixels, just due to the wavelength of the pattern). The response maximum does not move, because the decrease in φ is compensated by a *decrease* in low-pass phase lag due to decreased frequency ω in the low-pass filter.

The first-order, low-pass phase lag cannot become greater than 90°. Eventually, as we decrease wavelength, the pattern phase shift decreases faster than does the low-pass phase lag. To compensate, the pattern must move faster to obtain an increased phase lag from the low-pass filter. Hence, the shorter the wavelength, the higher the optimum speed. In fact, when $\lambda = 2L$, the optimum velocity approaches infinity. This limit corresponds to the Nyquist sampling criterion—the spatial frequency pattern is sampled twice per cycle by the array. The trend toward higher optimum speed with shorter wavelength is visible in the experimentally measured velocity tuning curves in Figure 4.17.

At the optimum velocity, in the limit $\lambda \gg L$, the squared magnitude of the transfer function takes the value

$$\left|H_\infty(\omega)\right|^2_{S_{\max}} = \left(\frac{\lambda}{\pi L}\right)^2 \tag{20}$$

Hence, the squared magnitude of the response at the optimum speed goes as the square of the wavelength, independent of the optimum speed. At low speeds, $\omega << 1/\tau$, and the squared magnitude of the transfer function is given by

$$\left|H_\infty(\omega)\right|^2_{S \ll L/\tau} = \left(\frac{S}{L/\tau}\right)^2 \tag{21}$$

Hence, for low speeds, the squared magnitude of the response is proportional to the square of the speed and invariant to wavelength. (The apparently conflicting results in (19), (20) and (21) are actually not conflicting, because the assumption that leads to (21) breaks down at $S_{\max}$.)

Two-input system

As stated earlier, the two-input system is similar to existing correlation-based direction-selective detectors. Let us now consider the behavior of a two-input system in the same way we just did for the effectively-infinite delay line system. For two inputs, (12) becomes

$$H_1(\omega) = \frac{i\omega\tau}{i\omega\tau + 1}\left(1 + \frac{e^{i\varphi}}{i\omega\tau + 1}\right) \tag{22}$$

We will form a particular instantiation of the pairwise Reichardt-type correlation detector by taking the difference between the squared magnitude of two pairwise detectors tuned to opposite directions, resulting in the following expression.

$$\left|H_1(\omega)\right|^2 - \left|H_1(-\omega)\right|^2 = 4\left(\frac{\omega\tau}{\omega^2\tau^2 + 1}\right)^2 \omega\tau\sin\varphi \tag{23}$$

We can think of this detector as the difference between two time-averaged pairwise detectors, like the one shown in Figure 4.1(a), with opposite direction selectivities. The common-mode terms have canceled. The tuning is separable into a temporal and a spatial product. Hence, we can immediately deduce that the optimum response occurs at a particular temporal frequency, and not at a particular velocity. The pairwise detector is not a velocity-selective filter, but rather is a temporal-frequency–selective filter whose response is modulated by the geometrical interference term $\sin\varphi$, familiar from the fly literature (see, for example, [7]). We can find the optimum temporal frequency ω_{max} for the two-input system in the same way that we found S_{max} for the infinite delay line. Independent of wavelength, the maximum of (23) occurs at the temporal frequency

$$\omega_{max} = \sqrt{3}/\tau \tag{24}$$

For low speeds, the temporal frequency is small compared with $1/\tau$, and we compute the response (23) as approximately

$$\left|H_1(\omega)\right|^2 - \left|H_1(-\omega)\right|^2 = 4(\omega\tau)^3\sin\varphi\,. \tag{25}$$

Hence, the response to slow motion depends on the wavelength through the $\sin\varphi$ term, and is cubic in the speed.

Figure 4.8 compares the theoretical transfer function amplitudes for the two-input system and the long system cases. Note that in this figure, we compare $\left|H_\infty\right|$ with $\left|H_1\right|$, rather than with the Reichardt detector in (23). In the two-input system, the optimum speed is proportional to wavelength (in accordance with Equation 24), but the amplitude of the response at the optimum speed

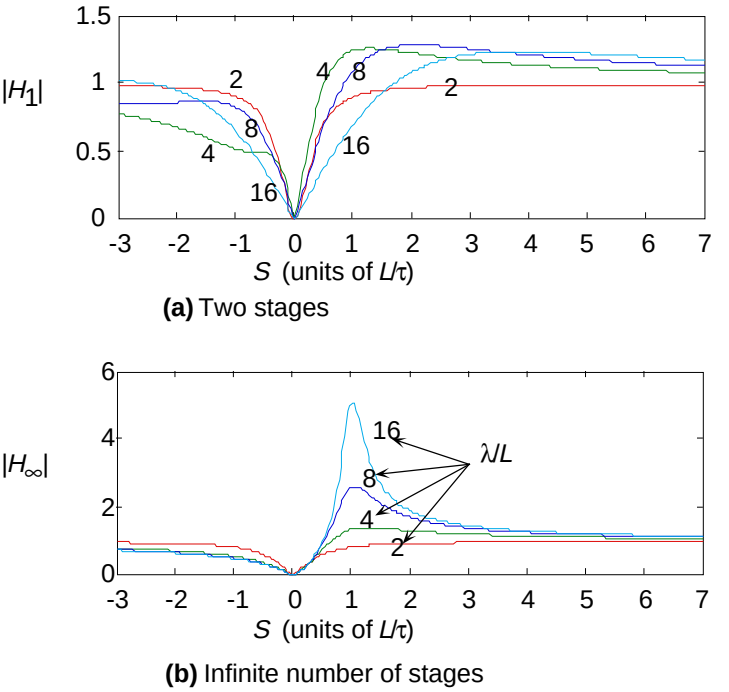


FIGURE 4.8

Theoretical plots of the magnitude of the transfer function for **(a)** two-stage system, compared with **(b)** infinite delay line, for various wavelength stimuli, shown next to curves. Parameters: $L/\tau = 1$.

is relatively invariant to wavelength. In the long system, the optimum speed and the low-speed response are invariant to wavelength, but the amplitude of the response at the optimum speed is proportional to wavelength. We can also see these characteristics in the measured tuning curves shown in Figure 4.17. Table 4.1 lists these theoretical results. In summary, the delay-line architecture differs from the pairwise detectors in that it is sensitive to stimuli with low as well as high spatial frequencies. The pairwise detector, because of the locality of the computation, cannot be very selective for motion of patterns with long wavelengths. The delay-line detector, in contrast, aggregates information over a spatial range corresponding to an arbitrarily-long wavelength, and hence can retain direction and speed selectivity even for long-wavelength patterns.

	Pairwise detector: $\lvert H_1(\omega)\rvert^2 - \lvert H_1(-\omega)\rvert^2$	Delay-line detector: $\lvert H_\infty\rvert^2$
Location of optimum response	$\omega = \frac{\sqrt{3}}{\tau}$	$S = \frac{L}{\tau}$ (for $\lambda \gg L$)

TABLE 4.1 Comparison between pairwise Reichardt-type detector and delay-line detector. Symbols are defined in Figure 4.6.

	Pairwise detector: $\lvert H_1(\omega)\rvert^2 - \lvert H_1(-\omega)\rvert^2$	Delay-line detector: $\lvert H_\infty\rvert^2$
Low-speed response	$(\omega\tau)^3 \sin\varphi$	$\left(\frac{S}{L/\tau}\right)^2$ (for $\lambda \gg L$)
Response at optimum speed	$\propto \sin\varphi$	$\propto \left(\frac{\lambda}{L}\right)^2$ (for $\lambda \gg L$)

TABLE 4.1 Comparison between pairwise Reichardt-type detector and delay-line detector. Symbols are defined in Figure 4.6.

THE MOTION CIRCUIT

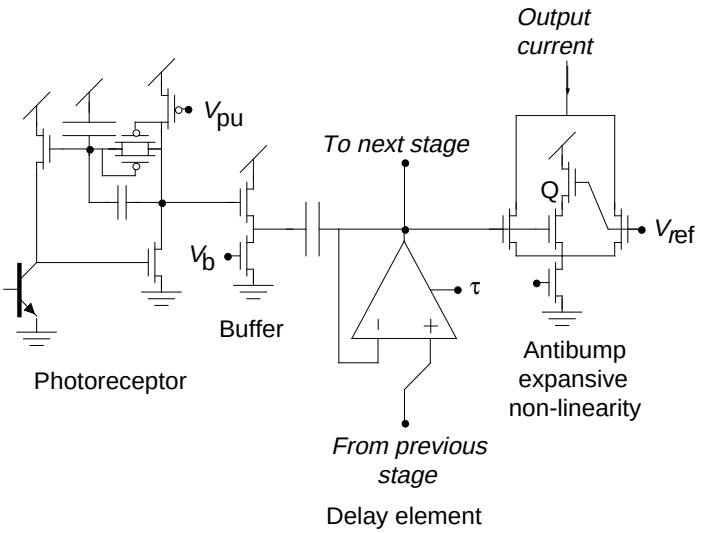
The circuit for a single stage of the motion circuit is shown in Figure 4.9. The output from an adaptive, high-gain, logarithmic photoreceptor circuit is coupled capacitively into a delay line stage. The delay line consists of a line of buffered, first-order, low-pass filters, implemented on the chips with follower integrators. The output nonlinearity is implemented with an antibump circuit.

The adaptive photoreceptor is essential for the robust operation of the circuit, because it provides well-conditioned input to the delay line over a wide range of lighting conditions. The receptor circuit has been reported previously [10][13][27] and is described in detail in Chapter 2.† In brief outline, the operation of the photoreceptor circuit goes as follows. The input leg of the receptor, consisting of the phototransistor and the source-follower feedback transistor, forms a signal voltage that is logarithmic in the intensity. The feedback circuit amplifies the logarithmic signal, and centers the operating point of the output signal around the history of the signal. The time-averaged output signal is stored on the feedback capacitor and serves as the reference around which the response is computed. The pair of diode-connected transistors act as a resistor-like adaptive element with monotonic *I-V* characteristic. The use of adaptation makes for a receptor with simultaneous high sensitivity and wide dynamic range.

The output from the adaptive photoreceptor connects to the input of a source-follower amplifier that buffers the output of the receptor. Different directions of selectivity use separate buffers that share the common photoreceptor input. The source follower isolates different directions in the

† Note that the circuit shown in Figure 4.9 does not use the latest-and-greatest receptor. For accuracy, I have shown the receptor actually used on the motion chip, but the results of Chapter 2 should be used as a reference for building new circuits using these photoreceptors.

FIGURE 4.9



Circuitry for a single stage of the motion circuit. An adaptive photoreceptor, with a contrast-sensitive gain of about 1 V/decade, feeds into a source-follower buffer, whose output couples capacitively into the follower integrator. The velocity-selective output current is computed by the antibump expansive-nonlinearity circuit. When the delay line signal is pushed away from its equilibrium point (V_{ref}) by activity on the delay line, the output current increases, because the center leg of the antibump circuit is turned off. For additional directions of selectivity, only the Buffer, Delay element, and Antibump nonlinearity must be duplicated—all directions share the same photoreceptor input.

case of a motion network with more than a single direction of selectivity. A separate source-follower is used for each direction. The source-follower output can follow a decreasing input signal only as fast as the bias current can discharge the output. It can follow increasing signals at an arbitrary rate, determined by the input voltage. By biasing the follower strongly (i.e., turning up V_b), we ensure that the source-follower output can follow accurately both increasing and decreasing outputs from the photoreceptor.

The follower-integrator uses a simple five-transistor transconductance amplifier [30]. These simple transconductance amplifiers suffer from systematic offset of several mV, due to Early-effect drain conductance in the differential pair and in the current mirror. Over a 25-stage delay line, the offset can be more than 100 mV. Since we compute the output of the pixel relative to a fixed reference voltage V_{ref} that is the input to the first stage of the delay line, this offset is important. We set V_{ref} near V_{dd}, so that the differential pair and the current mirror are both balanced. The speed of

the delay line, and hence of the velocity tuning, is set by the bias τ of the transconductance amplifier.

The voltage on the delay line connects to the input of an antibump circuit [11]. The other input to the antibump circuit connects to the reference voltage V_{ref}. If any activity on the delay line pushes the voltage away from V_{ref}, either in the positive (bright-edge) or in the negative (dark-edge) direction, the output current becomes larger, thus computing the power-like measurement of the linear delay line signal. Details of the antibump circuit operation are given in Chapter 3. In brief, the antibump circuit works because the center leg of the circuit, consisting of the series-connected transistors marked with Q in Figure 4.9, turns off when the differential input voltage is sufficiently large, forcing the bias current to flow through the outer legs of the circuit.

The delay line architecture for this circuit was inspired by an earlier architecture for sound synthesis of visual images [30]. The SeeHear chip sees an image, through use of on-chip photodetectors, and converts the image into an auditory equivalent of the visual image. This SeeHear chip used the same scheme of photodetector signals coupling into linear delay lines. The delay lines serve to synthesize the interaurel delay cues that drive the auditory-localization system in the brain.

The layout for one pixel from the two-dimensional motion chip is shown in Figure 4.10. Most of the pixel area is covered with wire and capacitance, rather than with transistor. The amount of interpixel wiring is minimal compared with the amount of wire used for routing bias signals into the pixel or routing outputs from the pixel. The architecture is efficient, because these input and output signals are essential and occupy most of the pixel area.

EXPERIMENTAL RESULTS

In this section, I present experimental results from a two-dimensional motion chip with the hexagonal architecture, and from a one-dimensional motion chip with two opposing delay lines.

Results from two-dimensional motion circuit

Figure 4.11 shows the real-time scanned output from the two-dimensional motion circuit in response to two moving patterns. Part (a) shows the response to a drifting square-wave grating. The buildup of activity away from the edge of the array in part (a) of the figure is evident as an increasing saturation of the color of the output. The longer the motion is visible to the array, the

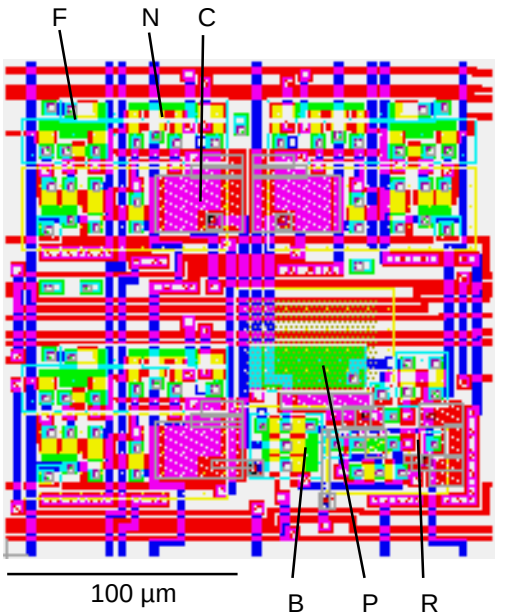


FIGURE 4.10 Layout of pixel of two-dimensional motion circuit with hexagonal architecture. Pixel has one photoreceptor, three follower integrators, and three antibump nonlinearities. Lines indicate circuit components. Keys: F = follower, N = antibump nonlinearity, C = capacitor, B = buffer, P = phototransistor, R = photoreceptor. Scale bar shows dimensions. Pixel area is 224 by 225 µm. This pixel is arranged in the hexagonal architecture shown in Figure 4.2.

larger the response. Part (b) shows the response of the chip to a rotating spiral pattern that produces the illusion of expansive optical flow. The chip computes a pseudolocal flow field. Motion that is not in a principal direction appears as a combination of the colors corresponding to the principal directions.

This chip consists of a 26 by 26 hexagonally arranged array of pixels. The power consumption of the core of the chip (all the analog computation done by the pixels) is 1.5 mW (dark) to 8 mW (light), depending on the brightness of the illumination, and hence on the raw pixel photocurrent. Hence, the power consumption of the analog computation, independent of photocurrent, is 1.5 mW, or 2.2 µW/pixel. About 80 % of this power is supplied to the antibump circuit, which I run at above-threshold bias to widen the transfer characteristic. The rest of the chip—consisting of the scanning frame, the clock driver, and the on-chip video amplifiers [29]—uses another 26 mW. The size of each pixel is 224 by 225 µm; the entire chip fits on a 6.8-by 6.9-mm die and is fabricated in a 2-µm, double poly, *n*-well process available through MOSIS, the DARPA fabrication service [9].

Figure 4.12 shows the directional tuning for the pixels in the two-dimensional chip. These plots show the average angular tuning for each of the directions represented by a pixel, for three different movement speeds. The responses are plotted in polar coordinates, and the distance from

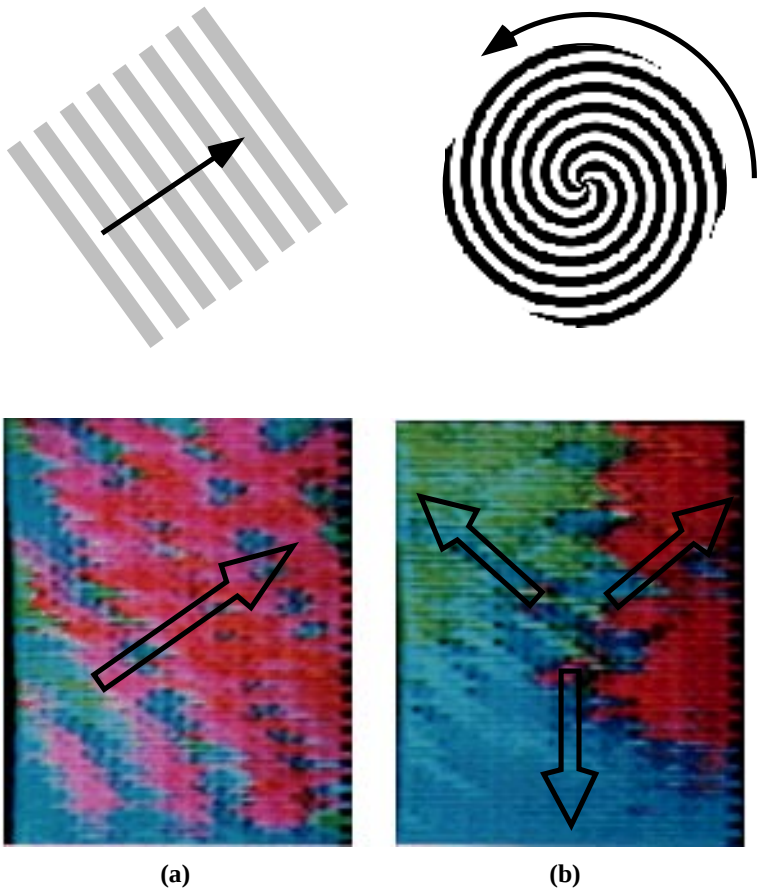


FIGURE 4.11 Color photographs showing the video output of the two-dimensional motion chip to motion **(a)** of a square-wave grating pattern in a single principal direction (arrow), and **(b)** of a rotating spiral pattern that produces illusory expansive optical flow (arrows). The outputs of the chip encode motion downward as blue, motion upward and to the right as red, and motion upward and to the left as green. Encoding of intermediate directions is evident as mixing of the primary colors. The dynamic nature of the output of the chip makes direct photography difficult, so we produced these pictures by video taping the output from the color monitor, and then photographing the freeze-framed screen of the video-tape monitor.

the origin is the magnitude of the average response. For speeds at or higher than the optimum velocity, the relative values of three outputs unambiguously determine the direction of the normal component of edge velocity. However, for speeds lower than the optimum, motion of a grating in a direction *oblique* to the detector orientation excites a detector more than motion at the same speed *along* the detector, resulting in responses with two peaks. This characteristic is an aspect of the aperture effect, since the delay-line detectors correlate image information along only the detector orientation.

Quantitative results from a one-dimensional motion circuit

I collected quantitative data primarily from a one-dimensional version of the motion circuit with two directions of selectivity. Figure 4.13 shows the response from two taps of the one-dimensional circuit in response to a moving sinusoidal pattern. The buildup of direction selectivity is evident in the difference between the response of the early and late taps. A systematic asymmetry is evident in the lack of frequency doubling, except for the largest response. I am not certain of the origin of the asymmetry, although I suspect either the source-follower buffer or the photoreceptor.

Figure 4.14 shows velocity-tuning curves for a number of taps from the same one-dimensional motion circuit, in response to the same moving pattern. The velocity tuning becomes sharper and more direction selective for later taps, although it appears that the tuning begins to approach a limiting form after about tap 15. The tuning curves shown in Figure 4.14 are measurements of the peak-to-peak output voltage from the circuit. The theoretical form of this measurement can be computed from (13), plus the transfer function of the logarithmic current-sense amplifier used in the scanning frame [29]. In Figure 4.15, I show fits to the data in Figure 4.14. Only the tap number was varied among the theoretical curves; all other parameters (wavelength, spatial frequency, time constant, antibump circuit parameters, logarithmic current-sense amplifier parameters) were kept constant.The quality of the fits is remarkable, considering that the input to the system was from a grating pattern printed on a piece of paper.

A number of nonlinearities have been ignored in the analysis. The effect of saturating nonlinearities can be seen in Figure 4.16, which compares the velocity tuning for low- and high-contrast grating patterns. The high-contrast grating pattern saturates the differential inputs in the follower integrators or in the antibump nonlinearity, causing a shift of the peak response toward a lower velocity, and a widening of the peak.

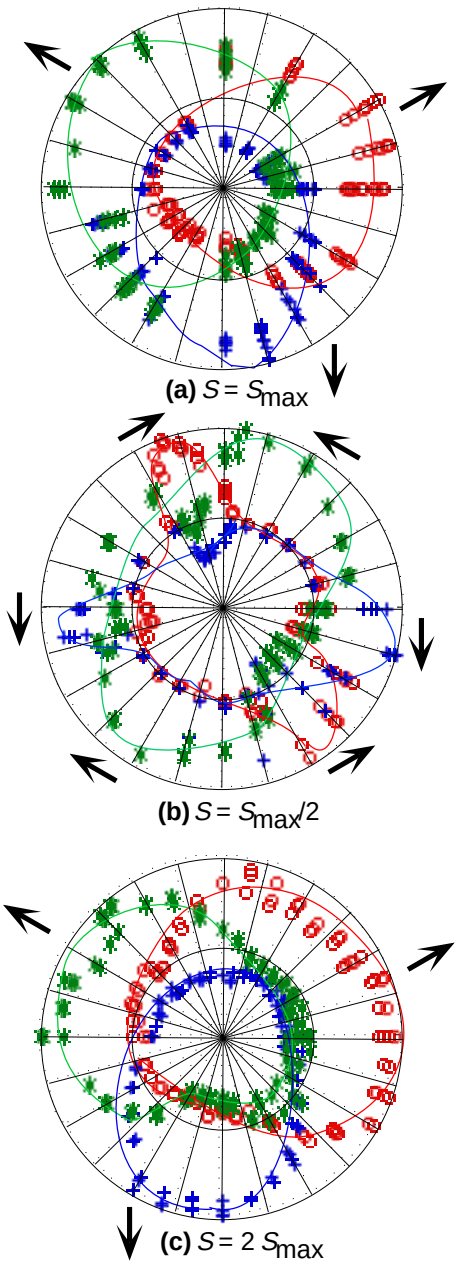


FIGURE 4.12 Directional-tuning of pixel in two-dimensional chip, for different stimulus speeds. Stimulus is sinusoidal grating pattern with wavelength approximately 10 pixel spacings. Each polar plot shows the response, measured as average video output voltage, of the three outputs from a pixel near the center of the chip. Each curve is labeled with an arrow showing the orientation of the delay line. I normalized each polar curve to the same maximum to adjust for monitor brightness corrections in the video circuits. The scatter in the points shows the noise in the 2 second averages. The solid curves are Bezier curves fitted by eye to the data points.

The stimuli in (a), (b), and (c) differ only in grating speed. In **(a)** the grating moves at the optimum speed, in **(b)**, at half optimum speed, and in **(c)**, at twice optimum speed. In (b), the apparent motion effect from Figure 4.3 is evident as doubly peaked responses that are roughly perpendicular to orientation of delay line. In (c), the response does not display the same doubly-peaked effect, because oblique motion is faster than orthogonal motion, so oblique motion excites pixel less than does true motion.

Figure 4.17 shows the velocity tuning of several taps of the network for different wavelength stimuli. For the late taps, the peak response occurs at a constant speed, relatively invariant to wavelength. For tap 1, the optimum velocity is a strong function of wavelength. For all the taps, the primary effect of varying the wavelength of the stimulus is a simple scaling of the amplitude of the response with wavelength. The speed of the optimum response shifts toward higher speeds as the wavelength is decreased. All these effects are consistent with the theoretical analysis presented earlier. The scaling of tap 1 response amplitude with wavelength, however, is inconsistent. I think that this result is due to a defocused image, which decreases the effective modulation depth of the image more for shorter wavelengths, and hence results in a smaller response for shorter wavelength.

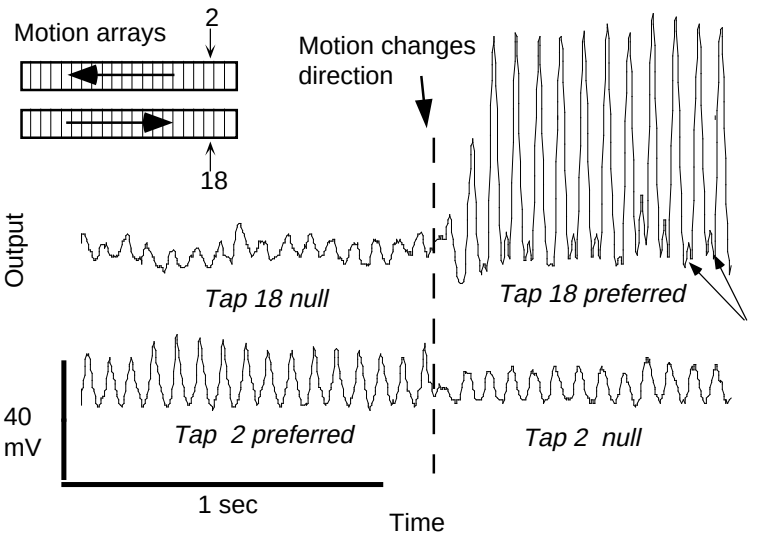
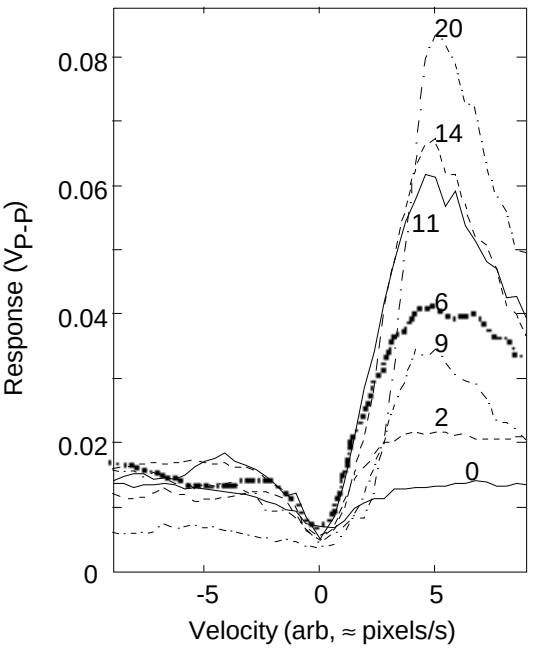


FIGURE 4.13 Measured response from two taps of a one-dimensional motion circuit with both directions of sensitivity (see upper left). The stimulus is a moving low-contrast grating pattern. The motion network had a total of 21 taps. I simultaneously recorded tap 2 in one direction and tap 18 in the other direction (numbering starts at zero, as in the text), and I plot the output voltage from the logarithmic current-sense amplifier. The grating initially moved in the tap 2 preferred direction and in the tap 18 null direction; it then changed direction. After a slight latency, the response from tap 18 became much larger than that of tap 2. The lack of frequency doubling, except in the largest signals (arrows), arises from a systematic offset of unknown origin.

FIGURE 4.14 The velocity tuning for a number of taps of a one-dimensional motion circuit. Plotted are the peak-to-peak outputs from the logarithmic current-sense amplifier fed from the antibump nonlinearity at each tap. The tap number is shown next to each curve. The gain of the logarithmic sense amplifier is approximately 100 mV per decade of current. Circuit offsets are visible as an inversion of the expected order between taps 6 and 9.



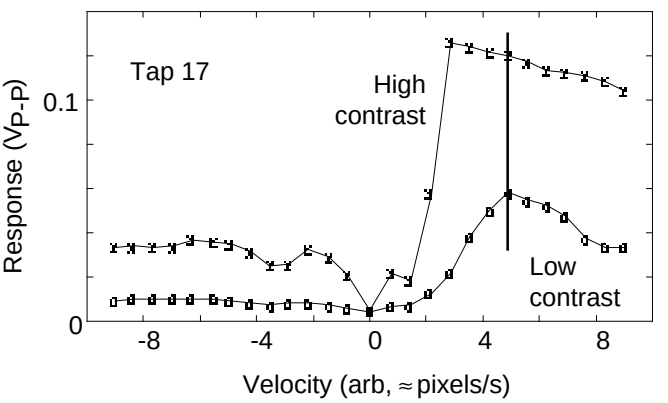


FIGURE 4.16 Effect of saturating nonlinearities on motion-circuit response. The data in the top curve were collected with the use of a high-contrast (≈ 50 %) grating; those in the lower curve were collected with the use of a low-contrast (≈ 10 %) grating.

DISCUSSION

The inspiration for the spatiotemporal integration properties of the chip's motion architecture came from psychophysical observations of human motion perception [31]. I am aware that cortex is far more complex than my simple model. The important thing to point out, however, is that the robust functioning of the chip is at least partially due to the spatiotemporal aggregation, while I know of at least three other attempts to build functional analog correlation detectors that have only marginal performance. The spatiotemporal aggregation in the chip also has the consequence that it allows the computation to see motion over a wide range of spatial frequencies, as opposed to the short spatial scales that a pairwise detector is capable of discriminating.

A given combination of the three pixel outputs can correspond to a range of possible pattern velocities of the actual scene. This characteristic is an expression of the aperture effect, and is true for *any* local motion detector. The three spatiotemporal correlations computed in each pixel are

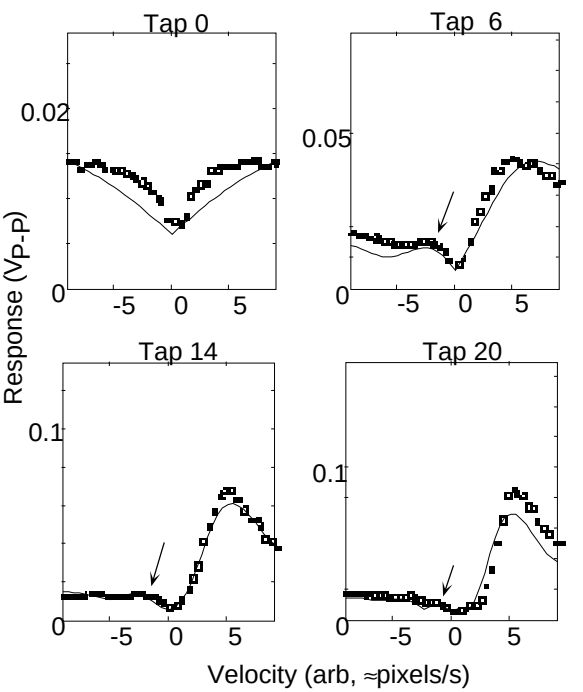


FIGURE 4.15 Velocity tuning with theoretical fits. Each plot shows measured velocity-tuning data as squares, with theoretical fits as solid curves. Response plotted is the peak-to-peak output voltage from a logarithmic current-sense amplifier that senses the antibump output current. Theoretical curves are derived from (14), plus the transfer function for the sense amplifier. The tap number is shown at the top of each curve and corresponds to the numbering used in the theory. Note the vertical scale—the response grows with tap number, as in Figure 4.14. Arrows point out wiggles in null direction response that are fitted by theory. All parameters in the model are identical for each theoretical curve, except for the tap number.

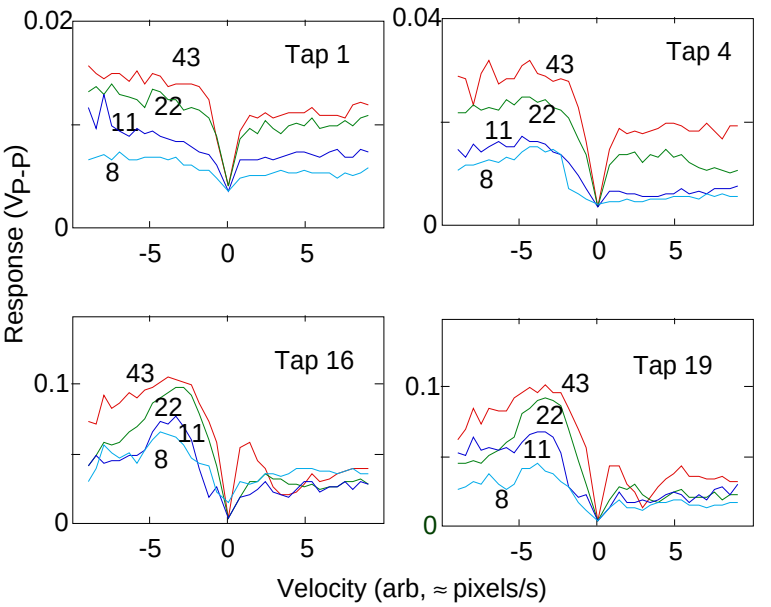


FIGURE 4.17 Effect of changing the stimulus wavelength. Each plot shows the measured output of a particular tap of the one-dimensional motion chip as a function of the stimulus velocity, in response to a moving sinusoidal grating pattern. In each graph, four plots are shown, one for each wavelength of the stimulus. The numbers on each plot show the wavelength, in pixels, of the grating.

valid computations in the sense that the information they report is consistent with a possible motion of a one-dimensional pattern. It can turn out that the output is not consistent with a motion orthogonal to the edge, as seen in Figure 4.12(b). However, there is absolutely no reason why motion must be orthogonal to an edge. At first sight, the apparent-motion effect suggests why cortical direction-selective cells are orientation tuned to edges orthogonal to the preferred direction, and why fly motion-sensitive cells are sensitive to the temporal frequency more than to image speed [14][20]. The argument goes that, in cortex, the orientation tuning removes the ambiguity about edge motion by ensuring that the computation measures only those motion components that are orthogonal to the edge. The same argument may be made in flies [19][39]. However, I believe that it is not settled whether these arguments are computationally valid or only opportunistically based on particular aspects of experimental observations.

The receptive field of the delay line motion detector resembles a simplified Adelson-Bergen spatiotemporal energy model [1]. Different spatial locations in the receptive field have different time delays to get to the output of the detector. The spatial receptive field is a long, skinny, rectangle starting at the location of the pixel and extending back along the delay line. The antibump output nonlinearity is similar to the squaring nonlinearity used in the Adelson-Bergen detectors. My detectors share receptive field with each other, through the common shared signal on the delay line.

One approach to performing full computation of the image motion is to plaster the input space with tuning curves. In cortex, there are spatial- and temporal-frequency tuned neurons with many different tuning curves [3]. The ensemble of broadly tuned, imprecise responses, that covers the whole space of inputs, can be combined to form a precise estimate of optical flow [17][21][24][37]. In comparison with some models of cortical motion computation, the number of velocity tunings in my motion chip is tiny—only three curves are used to cover the entire space of inputs. Models of cortex often plaster the space with dozens or hundreds of different tuning curves. Can we build systems with even tens of tuning curves? Not on one chip! This approach would require a number of motion chips because of the large number of required tuning curves. The problem of integrating multiple massively parallel analog chips is now being addressed [26].

It would be good to automatically adjust the tuning of the motion detector to match the input image, in analogy with other adaptive sensory processing like the automatic gain control used in the photoreceptors. This strategy would use the available dynamic range more fully, and the adaptation state itself would give information about the image motion. Biological studies suggest that animals with cortex do not adjust the tuning curves, while insects do. (However, people have not looked very hard in cortex for short–time-scale adaptive behaviors.) The fly visual system has only a few tens of visual output neurons. These direction-selective, velocity-tuned neurons, which integrate information from the entire visual field, adjust their tuning curves in response to the scene velocity [6][25][34][38]. The adjustment takes the form of a variable time constant that is shorter for higher image velocities or temporal frequencies (which it is—velocity [34] or frequency [6]—is under contention). This scheme is attractive from a chip designer's point of view, because it potentially allows a single velocity-tuned unit to cover a large dynamic range and still retain high sensitivity, in a manner analogous to the adaptive photoreceptor circuits used in the pixels.

SUMMARY

I have described the first functional two-dimensional analog VLSI implementation of a motion detector based on the 35-year-old Hassenstein–Reichardt correlation detector. The delay line extension to the usual pairwise correlation model has the novel functional property that it integrates information over an extended spatiotemporal region. The functionality of this chip is part of an accumulating body of evidence that we will eventually build complete, functional, neuromorphic visual systems.

ACKNOWLEDGMENTS

I must thank anonymous reviewers of the IEEE Neural Network Hardware Issue article for constructive comments, Carver Mead, Martin Egelhaaf, Shih-Chii Liu, Buster Boahen, Misha Mahowald, Humbert Suarez, Rahul Sarpeshkar, and Rod Goodman for valuable discussion, and Lyn Dupré for style editing. I also thank the MOSIS fabrication service for making this type of exploratory chip design possible. This work was funded by the Office of Naval Research under grant number NAV N00014-89-J-1675 and by the California Competitive Technologies Program. The inspiration to build a correlating detector array came from Werner Reichardt's work on the fly visual system, and from Nicola Franceschini's work on recording responses to optical stimulation of single pairs of ommatidial receptors. The idea of extending the correlation model to more than a pairwise correlation was inspired by a talk given by Ken Nakayama at the Society of Neuroscience Annual meeting in Phoenix in 1989.

REFERENCES

1. E.H. Adelson and J.R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am.*, vol. 2, pp. 284–299, 1985.

2. P. Ang, P.A. Ruetz, and D.Auld, "Video compression makes big gains," *IEEE Spectrum*, Oct 1991, pp. 16–19, 1991.

3. C. Baker, "Spatial- and temporal-frequency selectivity as a basis for velocity preference in cat striate cortex neurons," *Visual Neuroscience*, vol. 4, pp. 101–113, 1990.

4. H.B. Barlow and W.R. Levick, "The mechanism of directionally selective units in the rabbit's retina," *J. Physiol.*, vol. 178, pp. 477–504, 1965.

5. R. G. Benson and T. Delbrück, "Direction -selective silicon retina that uses null inhibition," in *Advances in Neural Information Processing Systems 4*, D.S. Touretzky, Ed., San Mateo, CA: Morgan Kaufmann, pp. 756–763, 1991.

6. A. Borst and M. Egelhaaf, "Temporal modulation of luminance adapts time constant of fly movement detectors," *Biol. Cybernetics,* vol. 56, pp. 209–215, 1987.

7. E. Buchner, "Elementary movement detectors in an insect visual system," *Biol. Cybernetics,* vol. 24, pp. 85–101, 1976.

8. C.P. Chong, C.A.T. Salama, and K.C. Smith, "Image-motion detection using analog VLSI," *Journal of Solid State Circuits*, vol. 27, pp. 93–96, 1992.

9. D. Cohen and G. Lewicki, "MOSIS—the ARPA silicon broker," in *Proceedings from the Second Caltech Conference on VLSI,* California Inst. of Tech., Pasadena CA, pp. 29-44, 1981.

10. T. Delbrück, "An electronic photoreceptor sensitive to small changes in intensity," in *Advances in Neural Information Processing Systems 1*, D.S. Touretzky, Ed., San Mateo, CA: Morgan Kaufmann, pp. 720–727, 1988.

11. T. Delbrück, "`Bump' circuits for computing similarity and dissimilarity of analog voltages," in *Proc. of International Joint Conference on Neural Networks*, vol. 1, pp. I-475–479, 1991.

12. T. Delbrück, *Investigations of Analog VLSI Visual Transduction and Motion Processing,* Ph.D. Thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena CA 91125, 1993.

13. T. Delbrück and C.A. Mead, "Time-derivative adaptive silicon photoreceptor array," in *Proc. SPIE,* vol. 1541, pp. 92–99, 1991.

14. H. Eckert, "Functional properties of the H1–neurone in the third optic ganglion of the blowfly, *Phaenicia*," *J. Comp. Physiol.*, vol. 135, pp. 29–39, 1980.

15. M. Egelhaaf and W. Reichardt, "Dynamic response properties of movement detectors: theoretical analysis and electrophysiological investigation in the visual system of the fly, *Biol. Cybernetics*, vol. 56, pp. 69–87, 1987.

16. N. Franceschini, J. Pichon, C. Blanes, "Real-time visuomotor control: from flies to robots," in *IEEE Fifth Intl. Conf. on Advanced Robotics, June 1991, Pisa, Italy,* pp. 91–95, 1991.

17. N.M. Grzywacz and A.L. Yuille, "A model for the estimate of local image velocity by cells in the visual cortex," *Proc. Royal Soc. Lond.*, vol. B 239, pp. 129–161, 1990.

18. B. Hassenstein and W. Reichardt, "Systemtheoretische analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers *Chlorophanus*," *Z. Naturforsch.*, vol. 11b, pp. 513–524, 1956.

19. J.H. van Hateran, "Directional tuning curves, elementary movement detectors, and the estimation of the direction of visual movement," *Vision Res.*, vol. 30, pp. 603–614, 1990.

20. K. Hausen, “Monokulare und binoculare Bewegungsauswertung in der Lobula plate der Fliege (Monocular and binocular computation of motion in the lobula plate of the fly),” in *Verh. Dtsch. Zool. Ges.*, Stuttgart: Gustav Fischer Verlag, pp. 49–70, 1981.

21. D.J. Heeger, “Optical flow using spatiotemporal features,” *Int. J. Computer Vision*, vol. 1, pp. 279-302, 1988.

22. T. Horiuchi, J. Lazzaro, A. Moore and C. Koch, “A delay line based motion detection chip,” in *Advances in Neural Information Processing Systems 3*, R. Lippman, J. Moody, D. Touretzky, Eds., San Mateo, CA: Morgan Kaufmann, pp. 406–412, 1991.

23. T. Horiuchi, W. Bair, B. Bishofberger, J. Lazzaro, J. and C. Koch, “Computing motion using analog VLSI Chips: an experimental comparison among different approaches,” *Int. J. Comp. Vision*, vol. 8, pp. 203–216, 1992.

24. S.R. Lehky and T.J. Sejnowski, “Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity,” *J. Neurosc.*, vol. 10, pp. 2281–2299, 1990.

25. T. Maddess, “Afterimage-like effects in the motion-sensitive neuron H1,” *Proc. R. Soc. Lond.*, vol. B-228, pp. 433–459, 1986.

26. M.A. Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function*, Ph.D. Thesis, California Inst. of Tech., Dept. of Computation and Neural Systems, Pasadena CA, 1992.

27. M.A. Mahowald, “Silicon retina with adaptive photoreceptors,” in *Proc. SPIE/SPSE Symposium on Electronic Science and Technology: from Neurons to Chips*, vol. 1473, April 1991.

28. D. Marr, *Vision*, New York: W.H. Freeman & Co., 1982.

29. C.A. Mead and T. Delbrück, “Scanners for visualizing activity of analog VLSI circuitry,” *Analog Integrated Circuits and Signal Processing*, vol. 1, pp. 93–106, 1991.

30. C.A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison–Wesley, 1989.

31. K. Nakayama and G.H. Silverman, “Temporal and spatial characteristics of the upper displacement limit for motion in random dots,” *Vision Res.*, vol. 24, pp. 293–299, 1984.

32. J.M. Pichon, C. Blanes, and N. Franceschini, “Visual guidance of a mobile robot equipped with a network of self-motion sensors,” in *Proc. SPIE*, vol. 1195, pp. 44–53, 1989.

33. W. Reichardt, “Evaluation of optical motion information by movement detectors,” *J. Comp. Physiol.*, vol. 161, pp. 533–547, 1987.

34. R.R. de Ruyter van Steveninick, W.H. Zaagman, and H.A.K. Mastebroek, “Adaptation of transient responses of a movement-sensitive neuron in the visual system of the blowfly *Calliphora erythrocephala*,” *Biol. Cybernetics*, vol. 54, pp. 223–236, 1986.

35. J.E. Tanner and C. Mead, “An Integrated Analog Optical Motion Sensor,” in *VLSI Signal Processing, II*, S.Y. Kung, Ed., New York: IEEE Press, pp. 59–76, 1986.

36. J.E. Tanner, *Integrated Optical Motion Detection.* Ph.D. Thesis, California Inst. of Tech., Dept. of Computer Science, Pasadena CA, 1986.

37. P.A. Viola, S.G. Lisberger and T.J. Sejnowski, "Recurrent eye tracking network using a distributed representation of image motion," in *Advances in Neural Information Processing Systems 4*, pp. 380–387, 1992.

38. W.H. Zaagman, H.A.K. Mastebroek, and R.R. Van Steveninck, "Adaptive strategies in fly vision: on their image-processing qualities," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC–13, pp. 900–906, 1983.

39. J.M. Zanker, "On the directional sensitivity of motion detectors," *Biol. Cybernetics*, vol. 62, pp. 177-183, 1990.

2

POSTSCRIPT

CHAPTER

5

LESSONS

What are the lessons from this thesis? There are general lessons and specific lessons. The general lessons are quite simple and obvious, but often overlooked. They concern the construction of *functional* analog VLSI systems. In no particular order, I call them **GIGO**, **KISS**, and **AGGSIG**. I will come back to them in a moment. The specific lessons are the tricks of the trade, "the real stuff." I mean of course things like the bump circuits, the physical structures like the adaptive elements, the noise analysis, the fact that photodiodes are good and phototransistors not useful, and so forth. These tricks are already listed in the introductions and summaries of each chapter. Let me discuss here only the general lessons.

Engineers recognize the first acronym above as *Garbage In, Garbage Out*. The motion chip would not, could not, work without the well conditioned input from the carefully designed photoreceptors. Yet all too often, we see people designing systems that assume that the input is going to be perfect. We can never do this, of course, but the better the input, the easier the subsequent computation. We must strive to generate the best possible input for our computational system.

Takeo Kanade coined the second acronym. It stands for *Keep it Simple, Keep it Stupid*. The design of the motion chip is very simple and quite minimal. The number of components is small, and hence the offsets, the pixel area, and the amount of wire are minimized. The design of circuits with many components is easy—what is hard is eliminating components without eliminating

functionality. These statements are truisms, but designers, especially beginners, must apply a litmus test of this sort or they will be sorely disappointed at the results of their efforts.

The third acronym comes from the working title of a chapter in Carver Mead's book, *Analog VLSI and Neural Systems*. It stands for *Aggregation of Signals*. Mead, inspired by biological neural structures, suggests that it is natural to build circuits that aggregate signals, and that such circuits can be used to do efficient computation. For example, the resistive network in the outer plexiform layer of the retina, used in Mahowald's silicon retinas, computes a smoothed version of the input image that is used as a grey-level reference. Another example is found in the moment-computation chips built by Steve DeWeerth, which use follower aggregation to compute the location of the center of mass of a pattern. Yet another example is the motion chip built by John Tanner, which uses global wires to aggregate information about local motion over the entire field. The example in this thesis is the motion chip, where the delay line aggregates local information about image derivatives over space and time to make a robust estimate, with wide dynamic range, of motion energy. These functional examples show by example that this AGGSIG trick is good, and I am sure it will be used extensively in the future.

MISCELLANEOUS DETAILS

This thesis was typeset on a Macintosh using Framemaker (version 3.0). Figures were drawn mostly with Canvas 3.0, or modified from graphs generated from Matlab 3.5 and CricketGraph. The design and format of this thesis is emulated from Carver Mead's book *Analog VLSI and Neural Systems*, largely designed by Calvin Jackson The body font is Palatino 10 point, and the figure caption font is Helvetica Light 9 point.

I collected all the data in the thesis using a Macintosh running Matlab 3.5, driving a National Instruments GPIB controller card, via driver code supplied by National Instruments and module interface code supplied with Matlab. The integration of National Instruments drivers and Matlab was done by Dan Naar of Apple Computer. The controller card controlled various GPIB instruments, including a Keithley 617 electrometer, a Keithley 230 voltage source, a Hewlett Packard spectrum analyzer, and a Tektronics 2430A digital oscilloscope.

The chips are fabricated through the MOSIS facility, which is run by DARPA as a multiproject, shared-cost fabrication service. Chip designs are submitted in CIF (Caltech Intermediate Format) via e-mail to MOSIS (e-mail address: mosis@mosis.edu). Typical chip cost is $500 for TinyChip Projects (2 mm by 2 mm square in 2 µm feature-size technology), and $5000 for 7 mm by 7 mm projects. The chips are designed using Tanner Research's layout tool L-Edit (Tanner Research, support@tanner.com, (818) 792–3000). Design connectivity is verified versus schematically entered layout, using the Tanner Research's *lvs* program or Mass Sivilotti's *netcomp* program. The schematics are generated using *analog*, a digital/analog simulation/schematic entry program written by John Lazzaro and Dave Gillespie. *analog* (and other Caltech tools) are freeware available via anonymous ftp from the internet host *hobiecat.pcmp.caltech.edu.*

PUBLISHED WORK

For reference, here is a complete bibliography of my published work.

1. T. Delbrück, (1993). Silicon retina with correlation-based, velocity-tuned pixels, *IEEE Transactions on Neural Networks*, (in press).

2. T. Delbrück, (1991), A silicon network for motion discrimination that uses spatio-temporal interpolation, *Society of Neuroscience Abstracts* 143.8 (page 344).

3. T. Delbrück and C.A. Mead, (1991), Silicon adaptive photoreceptor array that computes temporal intensity derivatives, in T.S. Jay Jayadev (ed.) *Proc. SPIE, Infrared Sensors: Detectors, Electronics, and Signal Processing.*, vol. 1541, pp 92–99.

4. C.A. Mead and T. Delbrück, (1991), Scanners for visualizing activity of analog VLSI circuitry. *Analog Integrated Circuits and Signal Processing*, vol. 1., pp. 93-106. (Extended version as Caltech Computation and Neural Systems Memo Number 11.)

5. R. G. Benson and T. Delbrück, (1991). Direction-selective silicon retina that uses null inhibition, in D.S. Touretzky, Ed. *Advances in Neural Information Processing Systems 4*. pp. 756–763.

6. T. Delbrück, (1991), "Bump"Circuits for Computing Similarity and Dissimilarity of Analog Voltages. Proceedings of International Joint Conference on Neural Networks, July 8-12, 1991, Seattle Washington, pp. I-475–479. (Extended version as Caltech Computation and Neural Systems Memo Number 10.)

7. J.M. Fox, D.C. Van Essen, and T. Delbrück, (1991), Modulation of classical receptive field responses by moving texture backgrounds in monkey striate cortex: spatial and temporal interactions, in *Analysis and Modeling of Neural Systems 1,* Frank H. Eeckman, (ed.) Kluwer Academic Publishers.

8. J.M. Fox, T. Delbrück, J.L. Gallant, C.H. Anderson, and D.C. Van Essen, (1990), Modulation of classical receptive field responses by moving texture backgrounds in monkey striate cortex: spatial and temporal interactions, *Society of Neuroscience Abstracts* 523.5 (page 1270).

9. T. Delbrück, (1989), A chip that focuses an image on itself, in C. Mead and M. Ismail (eds.) *Analog VLSI implementation of neural systems*, Kluwer Academic Publishers:Boston, pp 170–188.

10. M. Mahowald and T. Delbrück, (1989), Cooperative stereo matching using static and dynamic image features, in C. Mead and M. Ismail (eds.) *Analog VLSI Implementation of Neural Systems*, Kluwer Academic Publishers: Boston, pp. 213–238.

11. T. Delbrück and C.A. Mead, (1988), An electronic photoreceptor sensitive to small changes in intensity. in D.S. Touretsky (ed.) *Advances in Neural Information Processing Systems 1*, Morgan Kaufman: San Mateo, pp. 720–727.

12. M. Mahowald and T. Delbrück, (1988), An analog VLSI implementation of the Marr-Poggio stereo correspondence algorithm, *Abstracts of the first annual INNS meeting*, Boston 1988, Neural Networks, vol. 1, supplement 1, p. 392.

INDEX