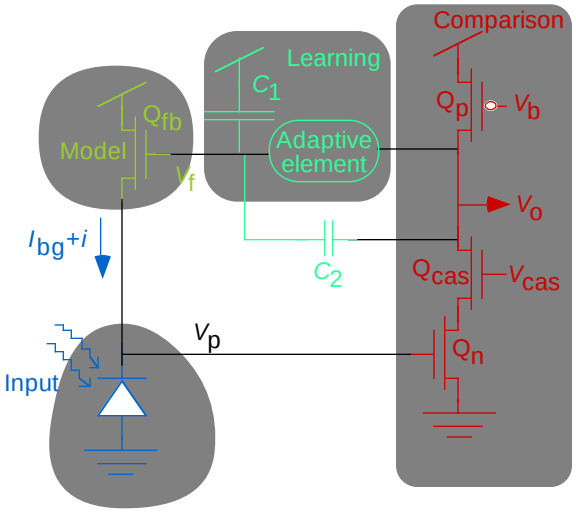
CALIFORNIA INSTITUTE OF TECHNOLOGY
COMPUTATION AND NEURAL SYSTEMS PROGRAM

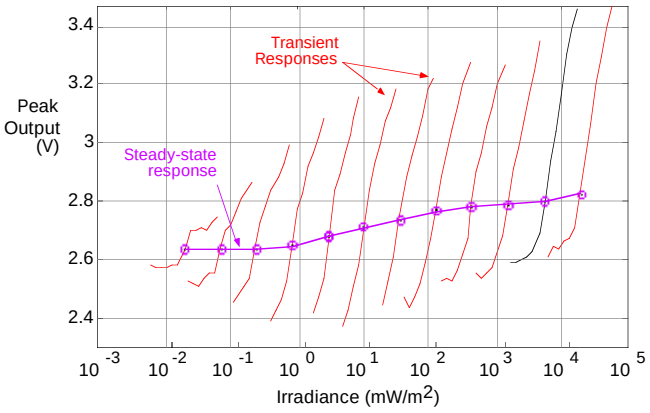
CNS Memo No. 30, April 2, 1996

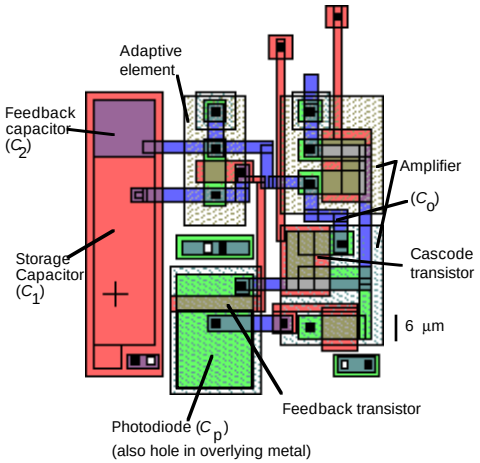
ANALOG VLSI PHOTOTRANSDUCTION

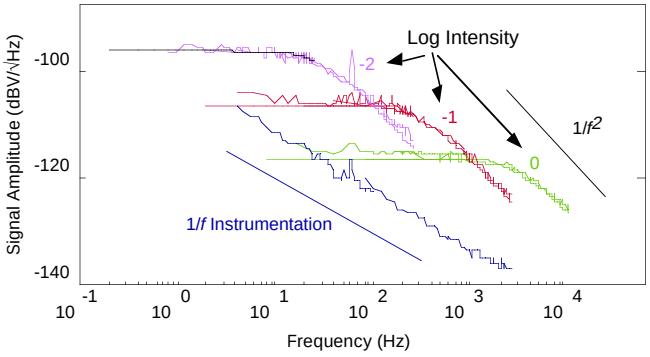
by continuous-time, adaptive, logarithmic photoreceptor circuits

T. Delbrück & C.A. Mead









PASADENA, CALIFORNIA 91125

ABSTRACT

This paper describes an adaptive photoreceptor circuit that can be used in massively parallel analog VLSI silicon chips. The receptor provides a continuous-time output that has low gain for static signals (including circuit mismatches), and high gain for transient signals that are centered around the adaptation point. The response is logarithmic with illumination, which makes the response to a fixed image contrast invariant to absolute light intensity.

The 5-transistor receptor can be fabricated in an area of about 70 by 70 μm^2 in a 2-μm single-poly CMOS technology. It has a dynamic range of 1–2 decades at a single adaptation level, and a total dynamic range of more than 6 decades. Several technical improvements in the circuit yield an additional 1–2 decades dynamic range over previous designs without sacrificing signal quality.

The lower limit of the dynamic range, defined arbitrarily as the illuminance at which the bandwidth of the receptor is 60 Hz, is at approximately 1 lux, which is the border between rod and cone vision and also the limit of current consumer video cameras.

We describe an adaptive element that is resistant to excess minority carrier diffusion. We show measurements of the effectiveness of guard structures, and of the spectral sensitivities of devices that can be built in a BiCMOS process.

The logarithmic transduction process makes the time constant scale inversely with intensity. As a result, the total A.C. RMS receptor noise is constant, independent of intensity. The spectral density of the noise is within a factor of two of photon shot noise and varies inversely with intensity. The connection between shot and thermal noise is beautifully illustrated.

CONTENTS

Over the last few years, people have built a number of neuromorphic analog vision chips that do focal-plane time-domain computation. These chips do local, continuous-time, spatiotemporal processing that takes place before any sampling or long-range communication, for example, motion processing,[2,5,6,9,29] change detection,[7] neuromorphic retinal preprocessing,[10,11,12,13,17,18] stereo image matching,[10,11,14] and synthesis of auditory images from visual scenes[18,22].

This processing requires photoreceptor circuits that transduce from light falling on the chip to an electrical signal. If we want to build analog vision chips that do high-quality focal plane processing, then we need good photoreceptors. It's not enough to just demonstrate a concept; ultimate usefulness will be determined by market forces, which, among other factors, depend a lot on raw performance. The receptor circuits we discuss here have not been used in any commercial product, so they have not yet passed that most crucial test, but by every performance metric we can come up with, including successful fabrication and test of demonstration systems, they match performance criteria met by other phototransduction techniques that *are* used in end-product consumer electronic devices.

We hope that this article will serve several purposes: We want people to have a reference where they can look to see the functioning and practical problems of phototransducers built in a typical CMOS or BiCMOS process. We want to inspire people to build low-power, integrated commercial vision devices for practical purposes. We want to provide a photoreceptor that can be used as a front end transducer in more advanced research on neuromorphic systems.

The transduction process seems mundane, but it is important—GIGO comes to mind. Subsequent computation relies on the information. We don't know of any contemporary (VLSI-era) literature that comprehensively explore the subject. Previous results are lacking in some aspect, either in the circuit itself, or in the understanding of the physics, or in the realistic measurement of limitations on behavior.[6,8,15,19,17]

We'll focus on one highly-evolved adaptive receptor circuit to understand how it operates, what are the limitations on its dynamic range, and what is the physics of the noise behavior. The receptor has new and previously unpublished technical improvements, and we understand the noise properties and illumination limits much better than we did before. We'll also discuss the practical aspects of the interaction of light with silicon: What are the spectral responses of various devices? How far do light-generated minority carriers diffuse and how do they affect circuit operation? How effective are guard bars to protect against them? Finally, we'll talk about biological receptors: How do their functional characteristics inspire the electronic model? How are the mechanisms of gain and adaptation related?

CONTINUOUS-TIME VS. SAMPLED RECEPTORS

The photoreceptors we'll discuss produce a continuous analog output that can be directly coupled to adjacent analog circuits—for example, circuits that compute image motion. This characteristic contrasts with the vast majority of imaging devices used commercially.

CCD imagers, for example, have become dominant in commercial cameras for many reasons—high density, low noise, minimal nonuniformity, high sensitivity, and relatively simple manufacturing process. Their easy availability and reliable operation have led to wide-spread use in machine vision applications.

However, their use has hindered investigation of vision algorithms and architectures that use *time* in an natural and efficient manner, because is it difficult to couple time-domain information from a serial stream of sampled imager outputs to analog circuits. If we want to do time-domain analog visual processing, it makes sense to build analog continuous-time photoreceptor circuits and couple their outputs locally to the circuits that do the computation.

Photoreceptors have not received much commercial attention because the system requirements for analog visual computation and imaging are so different. Sampled imagers are designed to go with serial television displays, and they must faithfully reproduce the visual scene. No computation is necessary (except for gain control) nor particularly desirable, since the output is supposed to look like what we see. Also, CCD cameras need 10^6 pixels, because humans who look at the TV picture want near-foveal resolution everywhere. (There is no way the camera can know where in the scene people are going to look.) The fact that CCD devices naturally produce a serial stream of data is perfectly suited for display and transmission to television monitors.

On the other hand, it is obvious that biological visual systems are massively parallel systems—at least in the preprocessing stages—and not TV cameras. Flying insects, for example, with less than 10^4 pixels, are existence proofs that interesting vision problems well beyond any current technology are doable with 100 times fewer pixels than the cheapest Sony Handycam.

BIOLOGICAL MOTIVATION

How do biological photoreceptors deal with two basic requirements: the simultaneous need for high sensitivity and large dynamic range, and the requirement of rapid re-

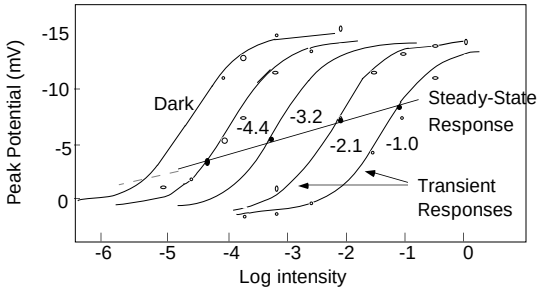


FIGURE 1 Gain adjustment in turtle cones (recorded intracellularly) caused by background illumination. The stimuli are 0.5 s increments or decrements on a steady background (except for the curve for the dark adapted cone which only is for increments). The stimulus spot is 3.2 mm in diameter on the retina. 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) $(cm^2s)^{-1}$. Adapted from Normann and Perlman (1979).[23]

Authors may be contacted by email at tobi@pcmp.caltech.edu. These circuits have been patented as T. Delbruck, C.A. Mead, "Adaptive photoreceptor including adaptive element for long-time-constant continuous adaptation with low offset and inensitivity to light," U.S. Patent 5,376,813.This document (and other related ones) exist in electronic form available via the World Wide Web; the URL for the Physics of Computation Group is *http://www.pcmp.caltech.edu/*

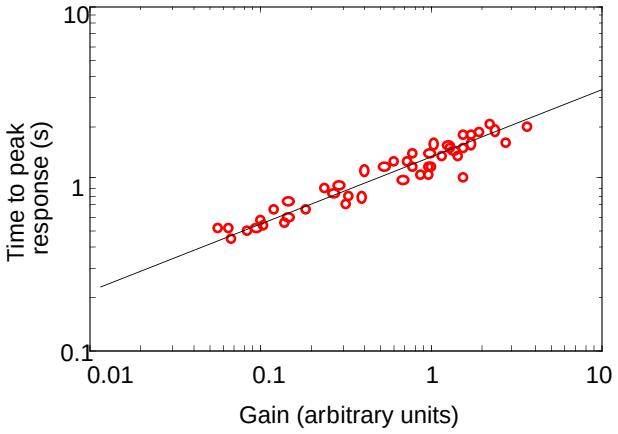


FIGURE 2 Time to peak response in response to a flash of light, vs. gain, in toad rod receptors. Replotted from Baylor et al., (1980).[1]

sponse time, invariant to lighting conditions? The approaches to phototransduction and amplification we use in our silicon photoreceptor are inspired by biology, although the detailed implementation is quite different. We shall only discuss functional *characteristics* here; in the discussion we will discuss biological *mechanisms* in relation to our silicon receptor.

GAIN CONTROL. Figure 1 shows the relationship between input and output in typical cone photoreceptors. The curves show the membrane voltage of the cone in response to a given illumination. The shallow curve shows the voltage in response to *steady* intensity. The steep curves show the response to small *variations* in illumination around a steady value. Two characteristics stand out: The receptor has a larger response to changes in illumination than to steady illumination—it *adapts*, and the slope of the responses is constant on a log-illumination scale, meaning it has a constant response to image *contrast*, independent of illumination level. The adaptation means that the receptor can respond with high gain over a wide operating range without saturating, and perhaps more important, the receptor output reflects actual changes in the illumination and not offsets that can build up due to various biochemical imbalances in a living system. (We don't know if the last supposition is correct, since no one has shown that biological system components have offsets, but it seems likely.)

TIME-CONSTANT CONTROL. Biological photoreceptors have a bandwidth that is practically invariant to the light intensity, over a wide range of intensities. In toad rods, for example, the bandwidth goes as the fourth root of intensity, as shown in Figure 2. This behavior means that over a factor of 200 in background intensity, the response latency to a flash of light varies over only a factor of about 4. It seems reasonable to speculate that this invariance is useful for dynamic visual processing of moving images, and it suggests that we should at least build a receptor circuit with response speed, if not invariant to illumination, as least as rapid as possible. (We can always lowpass-filter the response afterwards if we don't care about high frequencies, but there is no way to recover the information if the photoreceptor filters it away first.)

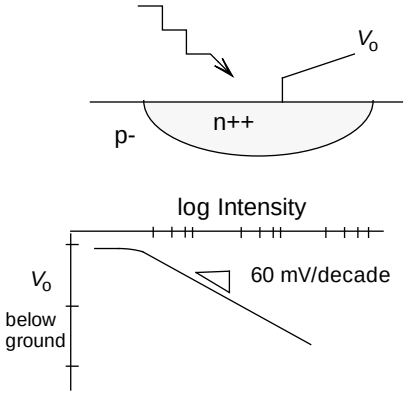


FIGURE 3 A single junction—the simplest logarithmic photoreceptor. V_o sits below the substrate voltage, decreasing logarithmically with intensity.

THE GOAL OF PHOTOTRANSDUCTION

Inspired by biological transduction and common sense, we shall assume that the primary goal of phototransduction is to compute image contrast invariant to absolute illumination. We can think of the total intensity as a sum $I_{bg} + i$ of a **steady state** background component I_{bg}, and a **varying small signal** component i. The contrast of the signal is defined as the ratio i/I_{bg}, and the receptor response should be proportional to this ratio independent of I_{bg}, at least for small ratios. The rationale for this assumption is that objects reflect a fixed fraction of the light that hits them. A receptor responding to a dynamically varying scene, such as would result from ego motion or from a moving object, will produce an output invariant to absolute illumination.†

SIMPLE LOGARITHMIC RECEPTORS

A receptor with logarithmic response to illumination has the right kind of response,

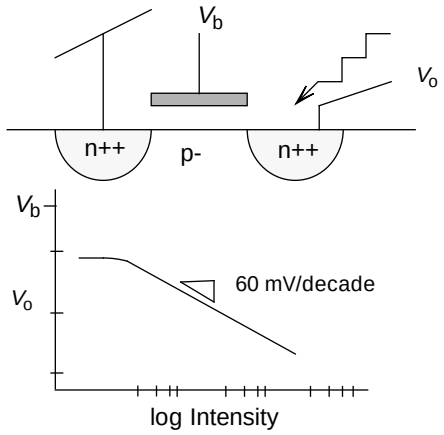


FIGURE 4 The source-follower logarithmic photoreceptor. V_o decreases logarithmically with light intensity, starting with zero intensity at approximately V_b.

because the change in the log intensity is given by

$$d\log I = \frac{dI}{I} = \frac{i}{I_{bg}} \qquad (1)$$

The simplest logarithmic receptor circuit is a single junction formed between the lightly doped substrate and a piece of heavily doped source–drain diffusion (Figure 3). When light shines on the silicon, it makes electron–hole pairs. When electrons freed in the p– substrate diffuse to the junction, they are swept home by the junction's electric field into the n++ region. The same for holes created in the n++ region—these are swept home to the p– region. The result of this photocurrent flowing from n++ to p– is that the n++ region becomes negatively charged with respect to the substrate. This negative voltage sets up a forward current in the junction to compensate for the photocurrent. Since the forward current is exponential in the junction voltage, the voltage on the n++ region is logarithmic in the intensity. However, this signal is not very useful, because it is below the substrate voltage.

The simplest logarithmic receptor that produces a useful output voltage range is shown in Figure 4. It consists of a single MOS transistor, where the source of the transistor forms the photodiode shown in Figure 3. The MOS transistor channel forms the barrier that results in a logarithmic response to intensity. The voltage at the

† This goal of contrast-invariant response makes sense for all but the lowest intensities, where contrast becomes an impractical quantification. When only a few photons hit the detector during an integration time, the fractional variation in the number becomes so large that it makes more sense simply to try to count every photon,[24] a problem that is beyond the scope of this paper.

source decreases as the light intensity increases. The decrease is a thermal voltage, V_T, for each *e*-fold intensity change, or about 60 mV per decade.† The DC operating point is determined by the bias voltage V_b.

This circuit would be a fine logarithmic photoreceptor in an ideal world. The gain of about 60 mV/decade results in a typical range of output voltages of perhaps 20 mV from natural scenes, which would in turn affect later circuits with current variations of factors of two, which is sufficient for computation.

There are two important problems: mismatches and slow response. The differences between supposedly identical receptor outputs are as large as the typical signals variations produced by real scenes. In a system context, we know from practical experience that this circuit is unusable except under demonstration conditions.

The other problem is that under low illumination, the response is too slow. The reason is that the junction capacitance, per unit area, in typical fabrication processes is too large. Making the photodiode larger doesn't help, because the capacitance scales up linearly with photodiode area. This problem will only grow worse as feature size decreases, because substrate doping density and junction capacitance will be increased.

Hence the necessity for adaptation, to deal with the circuit mismatch problem, and active feedback, to deal with the problem of slow response.

ADAPTIVE RECEPTOR CIRCUIT

The adaptive receptor circuit is formed by delayed feedback to the gate of the feedback transistor in the source follower receptor, as shown in Figure 5. Conceptually, the circuit uses an internal model to make a prediction about the input signal. The output comes from a comparison of the input and the prediction. The loop is completed by using learning to refine the model so that predictions are more accurate.[16] The adaptive receptor, with its level adaptation, uses perhaps the simplest type of learning.

The input stage of the adaptive receptor consists of the source-follower receptor shown in Figure 4. The **feedback transistor** Q_{fb}, for typical intensities, operates in subthreshold, so the source voltage V_p is logarithmic in the photocurrent. V_p sits below V_f at whatever voltage it takes to turn on Q_{fb} to supply the photocurrent. Conceptually, the voltage stored on C_1 acts as a model of the input intensity.

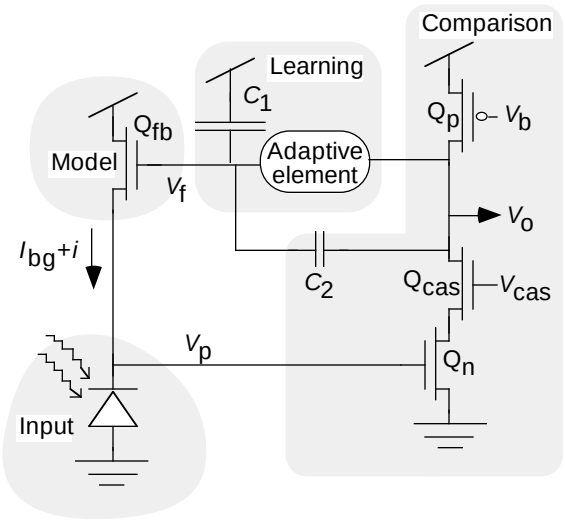


FIGURE 5 Adaptive receptor circuit.

The comparison between input and model is performed by the **inverting amplifier** consisting of Q_n and Q_p. An additional cascode transistor, Q_{cas}, is discussed on page 4. The input voltage V_p controls the current sunk by Q_n. The current sourced by Q_p is fixed by the bias voltage V_b. The voltage gain $-A_{amp}$ of this amplifier is determined by the ratio of the transconductance of Q_n to the output conductance of the amplifier, which in turn is determined by the length of the transistors in the amplifier. For typical layout, the gain is several hundred to several thousand. 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 feedback loop is completed when 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. The adaptation state is stored on C_1, which is shown hooked up to Vdd. (It makes no difference if Vdd or ground is used—all that is required is a fixed potential.††)

THE FEEDBACK LOOP

A small increase *i* of the photocurrent tries to pull down V_p by $(i/I_{bg})V_T$. In response, V_o goes up A_{amp} times as much. The output change is coupled back to the gate of Q_{fb} through the capacitive divider, with a gain of perhaps 0.1. Pulling up the gate of Q_{fb} pulls up on the source, which is where we started. So, instead of pulling down on the source of Q_{fb}, we end up raising the gate. The feedback amplifier and the input fight to control the source voltage of Q_{fb}, but the feedback amplifier wins because it has much higher gain. The input voltage v_p moves enough so that the output voltage v_o moves enough so that v_f moves enough so that v_p is held nearly 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, but changes in v_o are coupled to v_f through the capacitive divider. The transient 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.

Figure 6 illustrates the receptor's adaptive behavior and the invariance of the response to absolute intensity. The traces compare the response of the nonadaptive source-follower receptor with the response of the adaptive receptor. The incident signal is a small intensity variation sitting on a steady background. The small variations represent the type of signal arising from objects in a real scene, while the steady background represents the ambient lighting level. The contrast of the signal is a fixed percentage, independent of the absolute intensity, as would be produced by reflective

† $V_T = kT/q = 25$ mV at room temperature. An e-fold is a factor of e=2.72...The subthreshold transistor drain–source current is

$$I_{ds} = I_0 e^{\kappa V_g}(e^{-V_s} - e^{-V_d}) + (I_{ds}/V_e)V_{ds}$$

where *g* means gate, *s* means source, *d* means drain, and all voltages are in units of V_T measured relative to the bulk. I_0 is the preexponential constant and V_e is the Early voltage characterizing drain conductance . $\kappa \approx 0.7 - 0.9$ is the back-gate coefficient describing the effectiveness of gate voltage changes on channel potential.[18]

†† Mahowald's adaptive retina hooks up C_1 to a reference voltage that is computed by a resistive network, leading to interesting behavior.

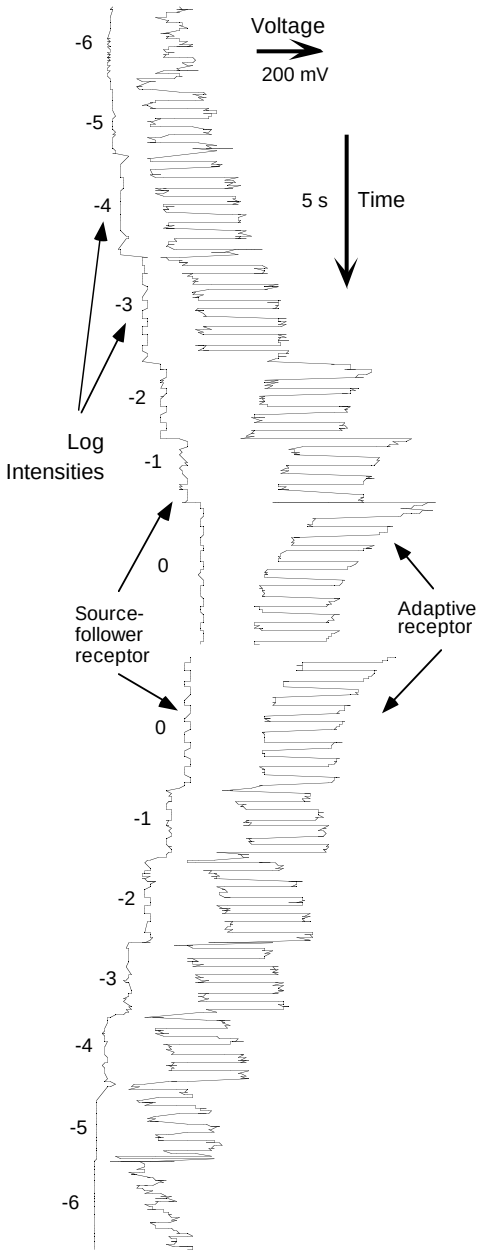


FIGURE 6 Responses of the source-follower and adaptive receptor over 7 decades of background. Stimulus is a square-wave variation in the intensity, centered around a mean value. The numbers by each section are the log intensity of the mean value; 0 log is 2.9 W/m^2, about the level of direct office fluorescent light. The source-follower receptor begins to smooth out the 1 Hz square wave input at the lowest intensities, while the adaptive receptor still responds. The 1.8 Hz square wave stimulus is from a red LED (635 nm). The irradiance varies by a factor of about 2, or about 0.33 decades or 0.8 e-folds.

objects. We varied the overall intensity level by interposing neutral density filters (i.e., sunglasses) with various attenuation factors between the light source and the receptor, while recording the receptor outputs. These changes in the overall intensity level represent changes in the ambient lighting, as would be caused by passing from shadow into sunlight or vice versa.

The amplitude of the response to the small contrast variation is almost invariant to the absolute intensity, owing to the logarithmic response property. The adaptation makes the receptor have high gain for rapidly varying intensities and low gain for slowly varying intensities. Hence, the response to an intensity change of a decade, after adaptation, is almost the same as the response to the 15 % variation. A receptor like this is clearly useful in systems that care about the contrast changes in the image, and not the absolute intensities.

The receptor adapts very rapidly in response to the decade changes of intensity. This rapid adaptation is due to the use of an adaptive element with an expansive nonlinearity. Large changes in the output adapt rapidly, while small signals around an adaptation point only adapt slowly.

CASCODE. Q_{cas} has two effects:

1. It shields the drain of Q_n from the large voltage swings of V_o. Because the source conductance of Q_{cas} is larger than the drain conductance by a factor of approximately A_{amp}, the drain of Q_n moves only about as much as v_p. Without Q_{cas}, the large voltage swings across the gate–drain capacitance of Q_n load down the input node. They make fF gate-drain capacitance appear to the input node to be on the pF scale, a phenomenon called the **Miller effect**.
2. Q_{cas} also multiplies the drain resistance of Q_n by a factor of approximately A_{amp}, through a cascode action. This increase in drain resistance increases the gain of the amplifier by a factor of about 2.

Both the reduction in effective input capacitance and the increased gain translate into speedup. The additional speed of the receptor makes it usable at lower intensities. The addition of this single cascode transistor increases the dynamic range of the receptor by about a decade.

Figure 7 shows the effect of the cascode on the small-signal time response of the receptor. Under these operating conditions the cascode speeds up the response by a factor of about 6.

PHOTORECEPTOR GAIN

It's simple to compute the steady state and transient gain of the receptor if we assume that A_{amp} is large. When the input current changes an e-fold, the gate of the feedback capacitor must change by V_T/κ to hold v_p clamped. That means in steady state the small-signal gain is given by

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

(linearized *steady-state* closed-loop gain) (2)

and for transient signals, where the output must go through the capacitive divider, the gain is

$$A_{cl} \equiv \frac{v_o/V_T}{i/I_{bg}} = \frac{1}{\kappa}\frac{C_1 + C_2}{C_2}$$

(linearized *transient* closed-loop gain) (3)

where $\kappa \approx 0.7$ is the back-gate coefficient describing the effectiveness of gate voltage changes on channel surface potential. We shall call the gain for transient signals the **closed loop gain** A_{cl} from now on.

Writing the gain in dimensionless form displays the logarithmic, contrast-sensitive response properties. The response to a constant i/I_{bg} is independent of the background current. The ratio between transient and steady-state gain is the capacitive-divider ratio $(C_1+C_2)/C_2$. We generally use a capacitive divider ratio of about 10, which is the ratio of poly-poly to metal-poly capacitance. Figure 8 shows measured transfer curves. The plots illustrate the shifting of the adaptation point to the ambient intensity and that the transient gain is much larger than the steady-state gain.

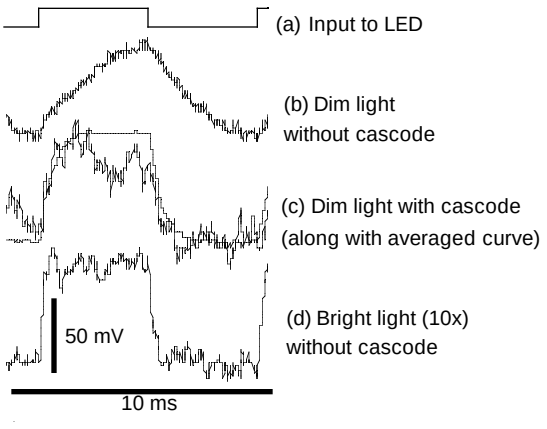


FIGURE 7 Effect of cascode on time-response and noise. **(a)** shows the small-signal input signal. **(b)** shows the response of the adaptive receptor when the cascode is shorted. **(c)** shows the response with the cascode activated, along with a time-averaged response. **(d)** shows the response of the receptor when the light is 10 times brighter. The noise properties are discussed in the text (see page 11).

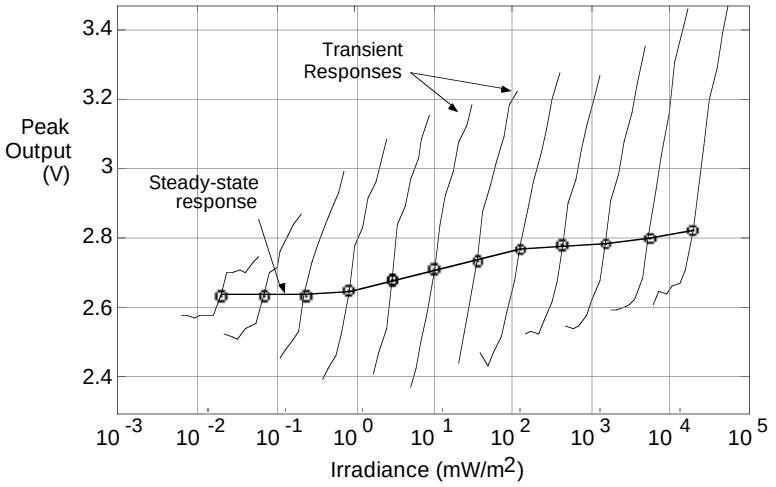


FIGURE 8 Step response operating curves. Each s-shaped curve shows the peak value of the response to a step change of irradiance, starting at the intensity marked with a circle. The ordinate shows the peak value of the response to the step, and the abscissa shows the absolute incident irradiance. The 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. The steady-state gain is between 30 and 40 mV/decade; the transient gain is approximately 1.5 V/decade, decreasing at the lowest intensities due to interaction between rise time and adaptation time constants.

Equation 2 is the small-signal equivalent of the large signal expression for the steady-state output voltage,

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

(large-signal steady-state response) (4)

where V_p is the clamped voltage at the gate of Q_n and I_0 is the preexponential in the subthreshold transistor law.

ADVANTAGES OF ACTIVE FEEDBACK

An active feedback circuit has three advantages over a passive feedback circuit like the one used in many early Mead lab projects: The bias current in the output leg of the receptor is capable of driving arbitrary capacitive loads. The bias control allows us to low-pass filter at a chosen frequency, which lets us filter out flicker from artificial lighting. Most important, the feedback, 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.

SPEEDUP

When we computed the closed loop gain, we assumed that the gain of the feedback amplifier formed from Q_n and Q_p is infinite, which of course is only true in the sense that it is much larger than the closed loop gain. Our assumption meant that the input node is perfectly clamped, which is also not true—the input has to move a little bit to change the output voltage by the required amount. The larger the gain A_{amp} of the feedback amplifier, the less the input node needs to move, and the more speedup we obtain. On the other hand, for a given feedback amplifier, the more closed-loop gain A_{cl} we design in the receptor (by adjusting the capacitive divider ratio), the more the input node must move, and the slower the response. We'll compute the speedup of the receptor relative to the baseline speed of the source-follower receptor. Its time constant is

$$\tau_{in} = \frac{C}{g} = \frac{CV_T}{I_{bg}} \quad \text{(baseline speed)} \quad (5)$$

The adaptive receptor is shown again in Figure 9, this time with relevant capacitors, both explicit and parasitic. In the absence of C_{fb} and C_n, the **speedup** obtained by using the active feedback to clamp the input node is given by

$$A_{loop} = \frac{A_{amp}}{A_{cl}} \quad \text{(speedup)} \quad (6)$$

The speedup is equal to the **total loop gain**, obtained by following the gain all the way around the loop. However, this result is naive, because it ignores the important parasitics C_{fb} and C_n.

MILLER CAPACITANCES

The Miller effect occurs when a capacitor feeds back the output of an inverting, high-gain amplifier back to the input. If the input needs to move a certain amount, it must charge not only its own side of the capacitor, but also but charge the other side of the capacitor which moves A times as much in the opposite direction. Hence a small capacitance C looks like a capacitance $(A+1)C$ to the input. Since $A >> 1$, we usually ignore the 1. The Miller capacitors C_n from the gate to the drain of Q_n and C_{fb} from the gate to the source of Q_{fb} have a substantial effect on the time-response of the receptor, even though the capacitance is only 0.1 fF/µm of channel width. We can compute the time constant of the receptor response including these parasitics by using the fact that the Miller effect increases the relevant gate–drain and gate–source capacitances by a factor of A, and the feedback increases the effective conductance at the source of Q_{fb} by a factor of the total loop gain. (In this analysis we shall imagine that we have left out Q_{cas}.) The time constant including these effects is given by

$$\tau = \frac{C}{g} = \frac{C_p + A_{amp}C_n + \frac{A_{amp}}{\kappa A_{cl}}C_{fb}}{\frac{A_{amp}}{A_{cl}}\frac{I_{bg}}{V_T}} \quad \text{with cascode} \quad (7)$$

$$= \frac{V_T}{I_{bg}}\underbrace{\left(\frac{A_{cl}}{A_{amp}}C_p + A_{cl}C_n + \frac{1}{\kappa}C_{fb}\right)}_{C_{eff}}$$

We can think of this result as implying an effective capacitance C_{eff}, as shown above, at the input node, which can be combined

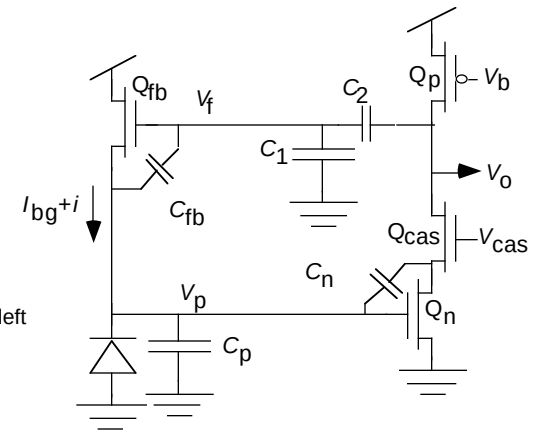


FIGURE 9 Adaptive receptor circuit with important parasitic capacitors. (Adaptive element is left out.)

with an input conductance of I_{bg}/V_T to compute the time constant. This way of thinking of the capacitance has the virtue that it clearly shows how the effective C_{fb} is unaffected by either the closed-loop gain or the amplifier gain. When a cascode is used in the receptor, the C_n term essentially disappears. When the total loop gain is large, the C_p term also essentially disappears, leaving only C_{fb} to ultimately limit the response time. Clearly, the receptor should be designed to minimize C_{fb} by using a narrow Q_{fb}.

We tested this theory on a fully-instrumented receptor by turning the cascode off and on and measuring the resulting speedup. Capacitance and gain values are shown in Table 1.

	True value	Effective value, without cascode	Effective value, with cascode
A_{amp}		1000	3700
A_{cl}		11	11
A_{loop}		91	336
C_p	100 fF	1.1 fF	0.3 fF
C_n	0.54 fF	5.9 fF	0.01 fF
C_{fb}	2.7 fF	3.0 fF	3.0 fF
C_{eff}		10.0 fF	3.3 fF

TABLE 1 Capacitance and gain values for adaptive receptor. Measured κ is 0.89. Typical areal capacitance values are 0.5 fF/μm^2 for photodiode, 0.8 fF/μm^2 for above-threshold gate, 0.1 fF/μm gate-drain or gate-source width.

The ratio of C_{eff} without and with the cascode is 3.0. This value is in excellent agreement with the measured speedup value 3.1, suggesting that we have correctly accounted for all the important parasitic capacitors. Table 1 is worth examining to see how C_{fb} becomes totally dominant once the cascode and high loop gain take the other capacitors out of the picture. With proper layout, C_{fb} should be reducible to 0.6 fF, a factor of 5 smaller than in our example. Hence, we should be able to achieve a speedup of 10 with the cascode.

RISE TIME & BANDWIDTH

Equation 7 shows that the response time of the receptor is inversely proportional to intensity. Figure 10 shows measurements of the small-signal rise time plotted against the absolute intensity for receptors built with different types of photodiodes. The inverse relationship between rise time and intensity is only violated at intensities above 1 W/m^2 with light from a red LED, where the receptor starts to be limited by minority carrier diffusion lifetime to a minimum response time of about 1 μs. This limit is unimportant for most vision applications, but specialized applications requiring very rapid response can use a junction with limited collection volume to speed up the response, at the cost of lower quantum efficiency.

GAIN–BANDWIDTH PRODUCT

A feedback amplifier whose closed-loop gain is determined by the feedback element generally has a fixed gain–bandwidth product. The GB product is an invariant for a particular design that can be used to compare different designs. We can compute the GB product from Equation 7; the result is

$$\text{GBproduct} = \frac{I_{bg}}{V_T} \frac{A_{amp}}{C_p + \frac{A_{amp}}{A_{cl}} C_{fb}} \qquad (8)$$

In the limit where we can ignore C_{fb} (in reality never), the GB product is higher by a factor of A_{amp} in the adaptive receptor than in the source-follower receptor. In practice, we have measured increases of the GB product of 500 to 2000.

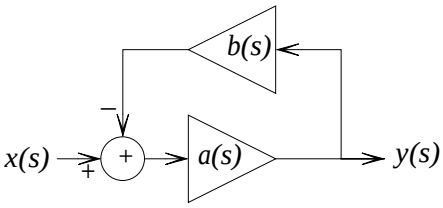


FIGURE 11 Generic feedback model.

PHOTODIODE AREA?

How big to make the diode? Often the choice is dictated by design requirements. We've built receptors with areas ranging from 100 μm^2 (10x10) up to 4000 μm^2 (65x65). The big wins from using a bigger photodiode are that it is faster (as long as you some loop gain to knock out the photodiode capacitance), and it is electrically quieter. The penalty is lower resolution.

SMALL-SIGNAL ANALYSIS

To understand the second-order behavior of the circuit, we'll compute the general transfer function including the time constant of the feedback amplifier. We'll ignore Miller effects in this analysis.

The generic feedback model shown in Figure 11 has the transfer function

$$H(s) = \frac{y(s)}{x(s)} = \frac{a(s)}{1 + a(s)b(s)} \qquad (9)$$

where *a(s)* and *b(s)* are the feedforward and feedback transfer functions in the s-plane.

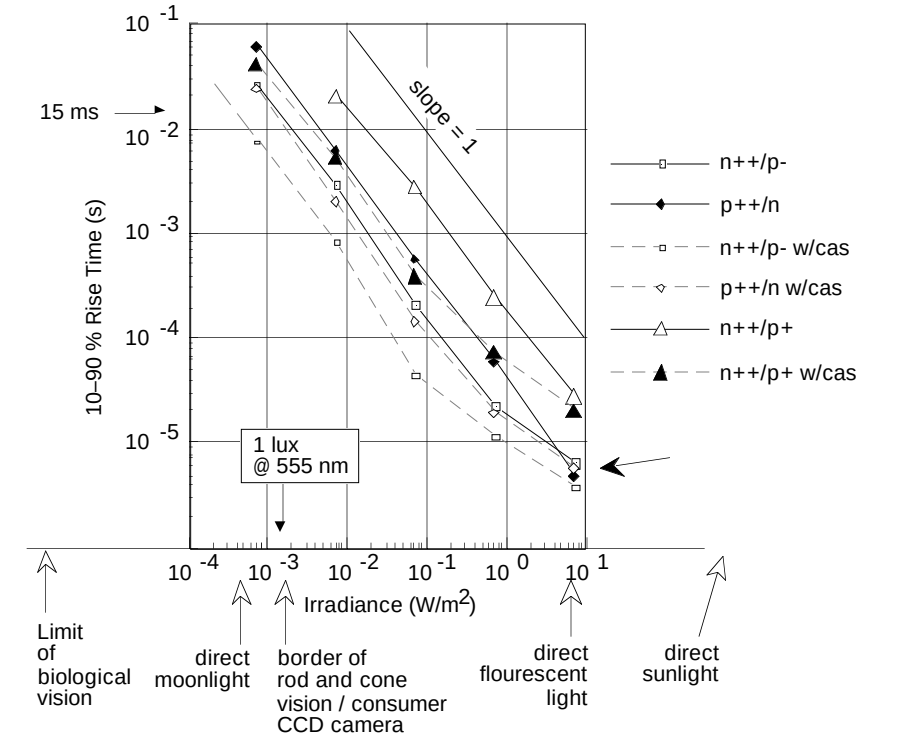


FIGURE 10 Response-time measurements for photodiode receptors. Each curve shows the 10–90 % rise time for a small step-intensity change, versus irradiance by a red LED. The different curves are for different photoreceptors circuits, differing in the phototransducer and the use of the cascode. Keys: *x/y* 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 ( ). This effect is discussed on page 6.

In the photoreceptor circuit, the input and output variables x and y are given by $x(s) = i/I_{bg}$ and $y(s) = v_o/V_T$.

The feedforward gain element $a(s)$ consists of the photodiode, the source of Q_{fb}, and the amplifier consisting of Q_n, Q_{cas}, and Q_p. The transfer function is given by

$$a(s) = \frac{A_{amp}}{(\tau_{in}s + 1)(\tau_{out}s + 1)} \qquad (10)$$

where τ_{out} is the time constant of the output node, set by the capacitance and output conductance in the amplifier, and τ_{in} is the time constant of the input node, set by the photocurrent and input capacitance, and given by Equation 5.

The feedback element $b(s)$ consists of the capacitive divider and the gate of the feedback transistor. In this analysis, we shall assume that no charge transfers through the adaptive element and hence that we are operating the circuit in the high-gain, transient-response mode. $b(s)$ is given by

$$b(s) = \frac{1}{A_{cl}} \qquad (11)$$

Using Equation 9, we obtain the transfer function:

$$H(s) = \qquad (12)$$

$$\frac{A_{cl}}{(A_{cl}/A_{amp})(\tau_{in}s + 1)(\tau_{out}s + 1) + 1}$$

The shape of this transfer function when τ_{out} is small is shown in the measured curves in Figure 12 (except that Equation 12 doesn't include the adaptation). As s goes to zero, $H(s)$ approaches A_{cl} when $A_{amp} >> A_{cl}$. As s goes to infinity, $H(s)$ approaches zero. Assuming τ_{out} is zero results in first order system with equivalent time constant $\tau_{in}(A_{cl}/A_{amp})$.

SECOND-ORDER TEMPORAL BEHAVIOR

In our computation of the expected speed-up due to the active feedback clamping of the input node, we assumed that the feedback amplifier is infinitely fast. The second-order behavior when τ_{out} is not zero can be visualized in the root-locus plot shown in Figure 13, which shows the locations of the poles of Equation 12 as τ_{out} is decreased. 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_{cl}/A_{amp} 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 in response to a step input. We can derive the condition for a damped, nonringing step response by finding the value of τ_{out} that makes the imaginary part of the poles equal to zero.

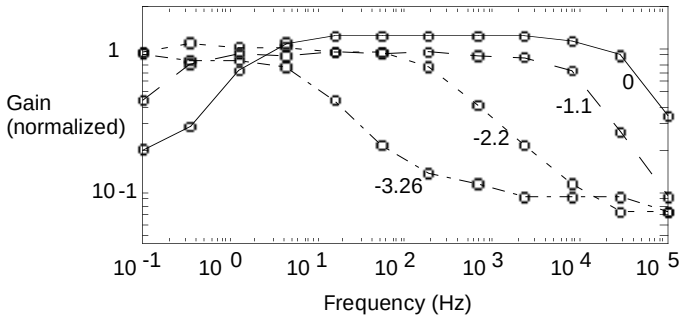


FIGURE 12 Measured amplitude transfer functions for the adaptive receptor using a photodiode constructed from native diffusion. The number by each curve is the log background irradiance, in decades. The highest irradiance is 19 W/m^2, about 10 times direct office-fluorescent lighting. The intensity affects 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. The adaptation rate, given by the low-frequency cutoff, appears to scale with intensity.

It is easiest to approach this problem from a canonical point of view for second-order systems.[18] A canonical form for the transfer function of a second order system is

$$H(s) = \frac{1}{\tau^2 s^2 + \frac{\tau}{Q}s + 1} \qquad (13)$$

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 number of cycles 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 12, as follows:

$$\tau = \sqrt{\frac{\tau_{out}\tau_{in}}{A_{loop}}}$$
$$Q = \sqrt{A_{loop}}\frac{\sqrt{\tau_{out}\tau_{in}}}{\tau_{out} + \tau_{in}} \qquad (14)$$

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

$$Q = \frac{1}{2} \quad \text{when} \quad \tau_{out} = \frac{1}{4A_{loop}}\tau_{in} \qquad (15)$$

For a nonringing, critically-damped response, the amplifier must be faster than the input node by about the total loop gain. This restriction is severe, because the amplifier output is already a factor of A_{amp} 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. Hence, for a critically-damped response, the transconductance of the input to the feedback amplifier, and the bias current, scales as the *square* of the desired speedup.

We can also use Equation 12 to find the condition for maximum Q. The result is

$$Q = \frac{1}{2}\sqrt{A_{loop}} \quad \text{when} \quad \tau_{out} = \tau_{in} \qquad (16)$$

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

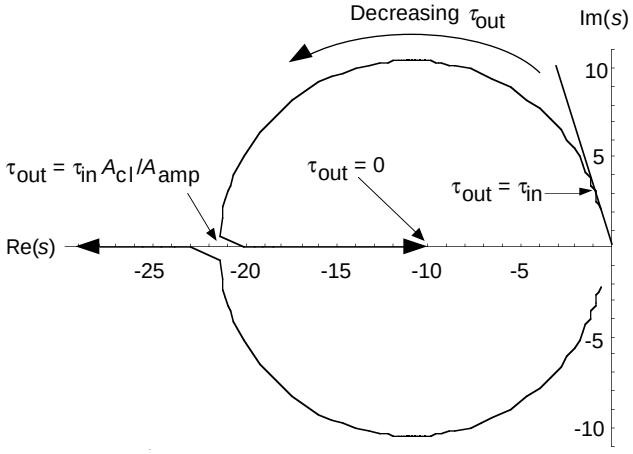


FIGURE 13 Root-locus plot for adaptive receptor, showing the poles of the transfer function in Equation 12, parameterized by the output time constant τ_{out} of the feedback amplifier. Parameters: $A_{amp} = 100$, $A_{cl} = 10$, $\tau_{in} = 1$.

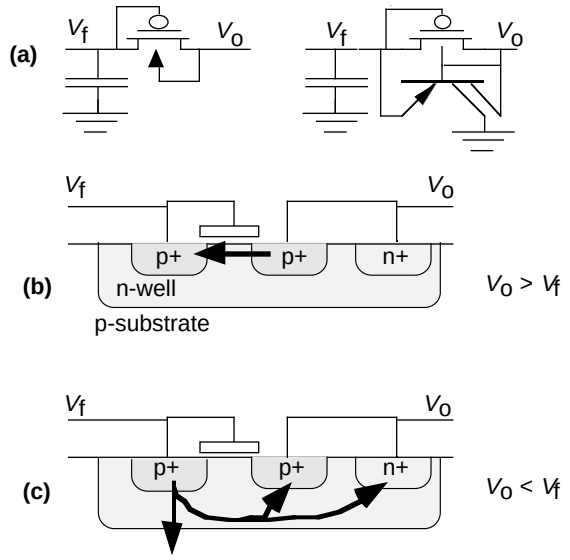


FIGURE 14 Expansive adaptive element **(a)**, shown in two schematic forms, along with the capacitor that stores the adaptation state.

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

(c) The opposite case: The p+/n junction is forward-biased, and the device as a whole acts as a bipolar transistor with two collectors.

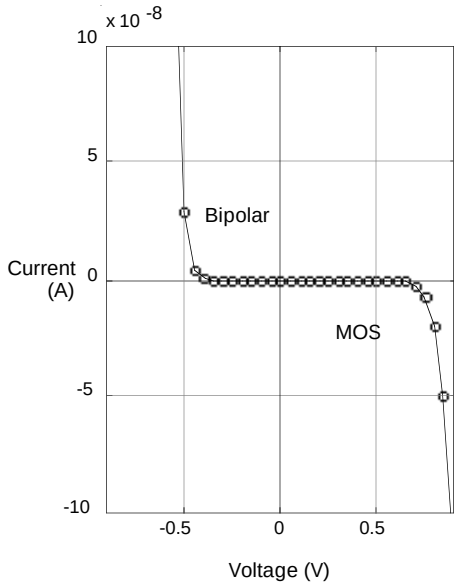


FIGURE 15 Measured current–voltage relationship for the new expansive adaptive element shown in Figure 14. The bipolar mode conduction e-folds every 28 mV, compared with 48 mV for the MOS mode, leading to a quantitative difference in the voltage at which the current rapidly increases. At any scale of current, the curves have the same appearance; the voltage scale changes logarithmically with the current scale. This data was taken from a p-well chip.

the maximum possible amount of ringing. This ringing is not very severe, because the maximum Q of the circuit is generally less than 5. We have labeled these conditions on the root-locus plot in Figure 13.

Usually we 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. A nice feature of this mode of operation is that the speedup is only effective at low intensities, while at higher intensities, the low pass filtering reduces the noise.

ADAPTIVE ELEMENT

Adaptation occurs 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. For analog VLSI circuits, however, true ohmic resistors available in a plain CMOS process are much too small for adaptation on the time scale of seconds.† Instead, we use transistors in our adaptive element—a sacrifice with unanticipated benefits. We have developed two novel adaptive elements with dual nonlinearities—expansive and compressive.[6] Here, we shall discuss only the expansive element, shown in Figure 14. The expansive element acts 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, as shown in Figure 15.

The I–V relationship of the expansive 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, because 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 and slowly varying signals. The adaptation time rate is proportional (in some sense) to the signal amplitude.

For voltage polarity $V_o > V_f$ across the element, the MOS transistor is turned on and the bipolar transistor is turned off. The driven side (V_o) in the MOS case acts as the source of the transistor, but because the back gate (the well) is driven at the same time, the current e-folds every V_T/κ.

For opposite polarity, the bipolar is turned on. The driven side forward-biases the p++/n emitter–base junction. The bipolar transistor has two collectors: the driven side and the substrate. The current e-folds every V_T volts. (The back gate of the MOS transistor is also turned on, leading to a current that e-folds every $V_T/(1-\kappa)$, but this small component is invisible relative to the large bipolar current.)

These characteristics may be seen in the data shown in Figure 16. This data was taken with and without light shining on a nearby hole in the metal covering of the adaptive element, to illustrate that the currents in the capacitor node (V_f) are unaffected by minority carriers generated in the substrate.

The I–V relationship for the element is given by

$$\begin{aligned} I &= I_m - I_b + I_{par} \\ \Delta V &\equiv V_o - V_f \\ I_m &= I_{0,m}[e^{\kappa \Delta V} - e^{(1-\kappa)(-\Delta V)}] \\ I_b &= I_{0,b}[e^{-\Delta V} - 1] \end{aligned} \quad (17)$$

where I is the current flowing onto the capacitor. I consists of three components, the MOS transistor current I_m, the bipolar transistor current I_b, and the parasitic photocurrent in the emitter–base junction I_{par}. The

† 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. Polysilicon 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! Some DRAM processes have an extremely high-resistance undoped polysilicon with ohmic properties, but it is unstable, with large variations, and very temperature dependent.

parasitic photocurrent flows out of the element. Voltages are in units of V_T. The preexponential constants $I_{0,m}$ and $I_{0,b}$ are for the MOS and bipolar transistors. The current-gain factor β has been included in $I_{0,b}$. Figure 17 shows Equation 17 plotted near zero differential voltage.

ADAPTATION RATE

The conductance of the adaptive element at the adapted condition ($I = 0$) determines the time constant of adaptation for small signals at the output. When $I_{par} = 0$ we can see by inspection of Equation 17 that the conductance is

$$g = I_{0,m} + I_{0,b} \quad (18)$$

The detailed measurements in Figure 16 indicate that $I_{0,m}$ and $I_{0,b}$ are both approximately equal to 0.1–1 fA. The parasitic photocurrent I_{par} in the emitter–base junction is small but nonetheless important, because it determines the imbalance across the element that must be maintained in an adapted condition of zero net current. Of course, I_{par} is never zero, due to scattered light and dark light (thermal photons). In any case, the conductance near the adapted state is determined by the condition $I = 0$. I_{par} is a positive current flowing onto the capacitor; hence it shifts the I-V curve upwards, resulting in a conductance at $I = 0$ that is approximately proportional to I_{par}, when I_{par} is somewhat larger than $I_{0,m} + I_{0,b}$.

The rate of adaptation is proportional to intensity, judging from the transfer curves shown in Figure 12. This behavior is to be expected if scattered light creates a parasitic photocurrent I_{par}. However, the more careful measurement shown in Figure 18 indicates that only under intensities above approximately 500 mW/m^2 is the adaptation rate affected by light intensity; at low intensities, the adaptation rate is constant, suggesting a conductance of about 4 fA/V_T in the adaptive element. This conductance is consistent with the $I_{0,m}$ and $I_{0,b}$ from inspection of Figure 16.

OTHER ADAPTIVE ELEMENTS

The MOS–bipolar adaptive element is inherently resistant to the deleterious effect of diffusing minority carriers at the capacitor node. Earlier attempts to construct adaptive elements purely from MOS transistors, such as the one shown in Figure 19, suffer from collection of minority carriers by the parasitic photodiodes formed by source–drain diffusions. There are offsets in these other elements of about a volt in an adapted condition (where no net current flows onto the capacitor). The large offset voltages arise from the huge back-gate voltage from the bulk that tries to turn off the transistor. The new adaptive

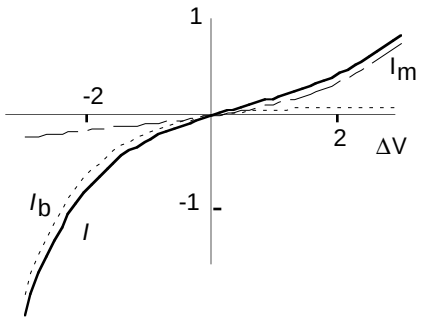


FIGURE 17 Theoretical plots of currents in Equation 17, assuming κ=0.7, $I_{0,m}$ and $I_{0,b}$ both equal 1, and I_{par} is zero.

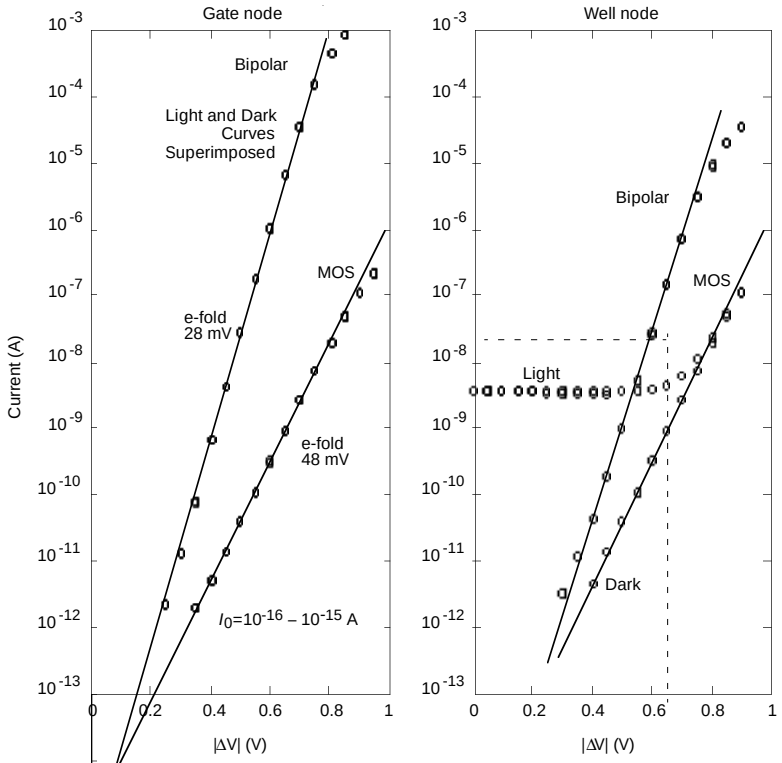


FIGURE 16 Detailed I–V characteristics of the bipolar–MOS adaptive element, taken in the dark and under illumination. At the gate node (left plot), the I-V characteristics are unaffected by light because the emitter/drain diffusion is shielded from minority carriers by the well. At the well node (right plot), there is a large parasitic photocurrent from the well to ground. Current gain for the bipolar mode is visible as a gate node emitter current that is approximately 100 times larger than the well node collector current.

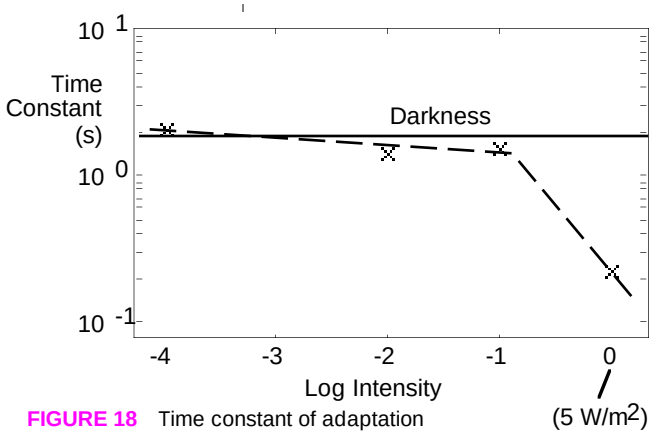


FIGURE 18 Time constant of adaptation node *vs.* intensity, using red LED.

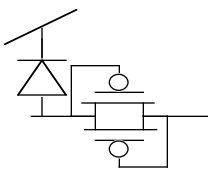


FIGURE 19 A bad choice for an adaptive element. Parasitic photodiode pulls adaptation node towards substrate voltage.

element has offsets of less than 100 mV in an adapted state, owing mostly to the zero back-gate voltage.

PHOTODIODE VS. PHOTOTRANSISTOR

Previous logarithmic photoreceptor designs from the Mead lab all used the parasitic vertical bipolar transistor, shown in Figure 20, instead of a photodiode. We used the bipolar transistor because bipolars have much less $1/f$ noise than MOS surface channel devices, and because the larger output current is more capable of driving capacitive loads. Later on, when we started to use active feedback circuits, we continued using bipolars for these same reasons.

Once we started to characterize the noise, we immediately discovered that $1/f$ noise is negligible compared with shot and thermal noise. In fact, there is a big *disadvantage* to using bipolar phototransistors in the adaptive receptor. The active feedback speeds up the response by clamping the input node. When we use a phototransistor instead of a photodiode, the feedback clamps the emitter node, but the base is left floating. Indeed, the base must not be clamped if the bipolar transistor is to function with its normal current gain mechanism. The current available to charge and discharge the base of the transistor is approximately the same photocurrent available from the photodiode. As a result, we obtain no speedup. The dynamic range is at least 1–2 decades smaller with a phototransistor than with a photodiode. In the discussion of receptor noise, we shall see that the noise properties of photodiode receptors and phototransistor receptors are indistinguishable (page 15). In the context of an active feedback circuit, it makes no sense to use bipolar phototransistors.

RECEPTOR LAYOUT

Figure 21 shows the layout corresponding to the schematic in Figure 5. The photodiode can be constructed from any of the *pn* junctions that are part of a CMOS process, but the one that we use in practice is the junction between native source–drain diffusion and substrate. In an n-well process, the junction is between the p– substrate and the n++ diffusion. We use this junction because the quantum efficiency is high and the capacitance per unit area is relatively small, resulting in a fast response. This junction also has the advantage that it may be constructed simply as an extension of the source of Q_{fb}. All parts of the circuit except for the photodiode are covered with metal. A nearby substrate contact sinks the photocurrent to ground. The transistors in the feedback amplifier are long, to maximize the gain and hence the speedup. The adaptive element is made from an isolated well with a single MOS transistor. The capacitive divider is formed from the two levels of polysilicon plus a metal plate, but a MOS capacitor may be used instead. The total area in this conservative layout is about 80x80 μm^2 in a technology with a 2-μm feature size.

The capacitive-divider ratio determines the gain of the receptor for transient signals. The capacitance of the adaptive element itself is usually not negligible compared with the explicit feedback capacitor C_2 and must be taken into account. It consists mostly of the active-well junction. A typical value is 15 fF, equivalent to a square metal–poly capacitor with an edge length of about 17 μm.

THE ILLUMINATION LIMIT (SPEED)

From an evolutionary perspective, it is clear that animals that can see in the dark occupy an important niche. The same is true for our silicon receptors. If they can only be used in bright sunlight, then they are not very useful.

The source follower receptor has a response speed that is determined primarily by the ratio of quantum efficiency to capacitance per unit area in the photodiode. The larger we make the photodiode, the larger the photocurrent, but the larger the capacitance. Since we are pretty much stuck with a fixed technology, we regard this speed as a constant of the problem. †

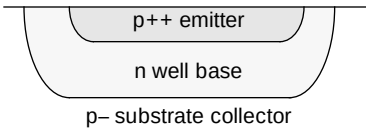


FIGURE 20 Parasitic vertical bipolar transistor in an n-well process.

In the devices available in an ordinary CMOS process, the maximum receptor speed is obtained from one of the substrate-junction photodiodes, because these have the largest quantum efficiency and the lowest capacitance. We have chosen to use the active–substrate diode rather than the well–substrate diode in all of our designs because the layout is much more compact. We expect that the well–substrate diode will have a similar capacitance and quantum efficiency.

† The speed could be increased if we had access to PIN diodes [28], 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—but we know of no CMOS process with these characteristics.
In a typical 2-μm n-well process, the active-substrate capacitance is around 120 aF/μm^2+200 aF/μm. In the more heavily doped 1.2-μm process, these values jump to 500 aF/μm^2 + 400 aF/μm.

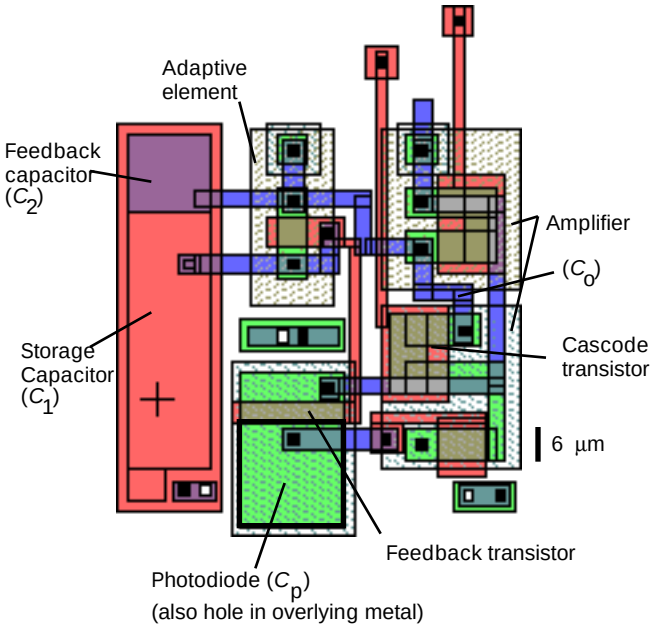


FIGURE 21 Photoreceptor layout. This layout is nonoptimal because the feedback transistor is wider than it needs to be, leading to excessive C_{fb} (see page 5).

A reasonable definition of the lower limiting intensity for operation of the receptor is the intensity at which the photoreceptor has a rise time of 15 ms—corresponding to the time for a single field of a video camera that scans at 60 Hz, and approximately the same as the cutoff frequency for human vision under photopic† conditions. We can see from Figure 10 that the fastest photoreceptor circuit that we have tested has a rise time of 15 ms at about 1 mW/m^2 irradiance—equivalent to an illuminance of about 1 lux††, or approximately the lighting of the full moon. This receptor is built with a feedback amplifier with a gain of several hundred, a closed loop gain of about 10, and a photodiode with an area of 20 µm by 20 µm.

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. Current commercial amplifiers, in consumer end-product devices, function at a noise level of about 100 electrons, 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 illuminance of about 1 lux, which is approximately the level of bright moonlight. In Figure 10, 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.

ILLUMINATION LIMIT: HIGH END

MOS transistors Q_n and Q_{cas} that form the bottom of the amplifier circuit in the adaptive receptor contain parasitic photodiodes from their drains and sources to the substrate. Ordinarily, these parasitic photodiodes are irrelevant, but under intense illumination the current to ground that they produce may exceed the bias current supplied by Q_p, pulling the output node to ground. This situation may be ameliorated by shielding the native-type transistors in the amplifier from light using a metal wiring layer, and by surrounding them with a guard bar made from native diffusion that is preferably tied to V_{dd}. (It doesn't make much difference if you tie the guard bar to ground or V_{dd}.) Guard structures are discussed further starting on page 15.

† **Photopic** means cones vision, **mesopic** means cones and rods are both used, and **scotopic** means rod vision.

†† At 555 nm wavelength (the peak of human sensitivity under photopic conditions),
1 lux = 1.4 mW/m^2. For white light (uniformly distributed over visible),
1 lux=4 mW/m^2=10^4 photons/µm^{2}s. [21]

††† Whatever these ratings mean—supposedly each manufacturer makes up their own definition, but none of them tell you what it is!

THE DETECTION LIMIT (NOISE)

What is the smallest signal that can be detected? How is this value affected by light intensity, and by detector area? How does the performance compare with commercial CCD cameras and biological rods and cones? What is the physical basis for the receptor noise?

We'll investigate the noise properties of the receptors and their detection ability, empirically and theoretically. We'll start empirically, with measurements of noise properties. These measurements show that the underlying behavior is very simple. The simplicity of this behavior motivates an equally simple theory that intuitively and quantitatively describes how noise works in logarithmic photoreceptors.

EMPIRICAL OBSERVATIONS

The observations shown in Figure 23 were measured from the simple source-follower detector shown in Figure 22. We captured the noise power spectra using two types of stimuli, a steady illumination from an LED and a white-noise source. We used the white noise source as a direct measurement of the receptor transfer function. We did each measurement at several levels of intensity, separated by decades. It is clear that all characteristics are well above the instrumentation noise.

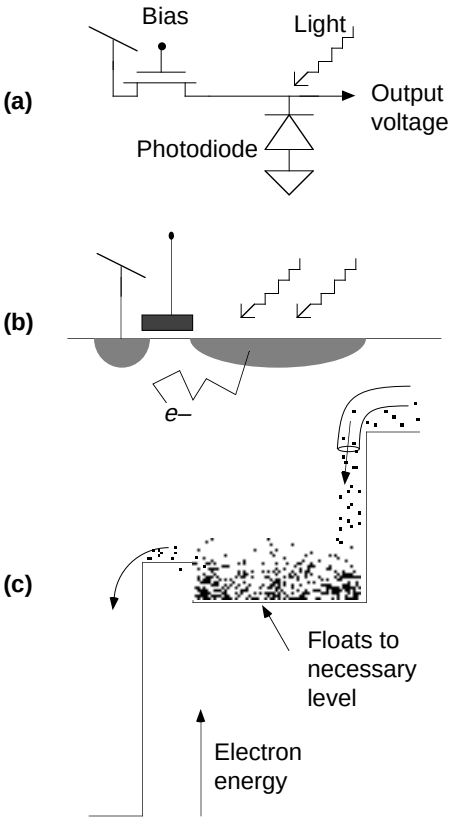


FIGURE 22 The simple logarithmic photoreceptor used in the study of receptor noise. This circuit forms the input stage to the adaptive receptor. **(a)** shows the schematic form. The bias voltage sets a reference for the source voltage, which is the output. **(b)** shows that the compact receptor consists of a single MOS transistor whose source forms the photodiode. **(c)** shows the electron energy diagram. The source voltage floats to whatever level is required to spill the photocurrent over the channel and into the drain.

The important observation is that *the total noise power, integrated over the entire passband of the receptor, is a constant independent of intensity.* The lower the intensity, the smaller the bandwidth of the receptor, but the larger the noise level within the passband. The responses of the logarithmic receptor to white noise stimulation show that the shape of the noise spectrum is the same as the shape of the receptor transfer function, at each intensity level. But while the transfer function measurement shows that the receptor contrast gain is constant, independent of intensity, the underlying noise behaves differently, becoming larger, the lower the intensity.

We know that the noise arises from within the receptor, and is not an artifact of the measurement (for instance, from noise in the LED light source). The reason is that the amplitude of the noise depends on the level of intensity. If the noise arose from the steady light source, then reducing the intensity by interposing neutral density filters would result in a set of curves like the top set of curves in Figure 23 showing the response to a white noise source. In other words, if the noise arose in the supposedly steady source, then the bottom curves would duplicate the top set of curves but be shifted down by a constant amount. Since the behavior is clearly quite different, we are certain that the noise arises in the detector itself.

Another important observation is that flicker noise (1/*f*) in the receptor is negligi-

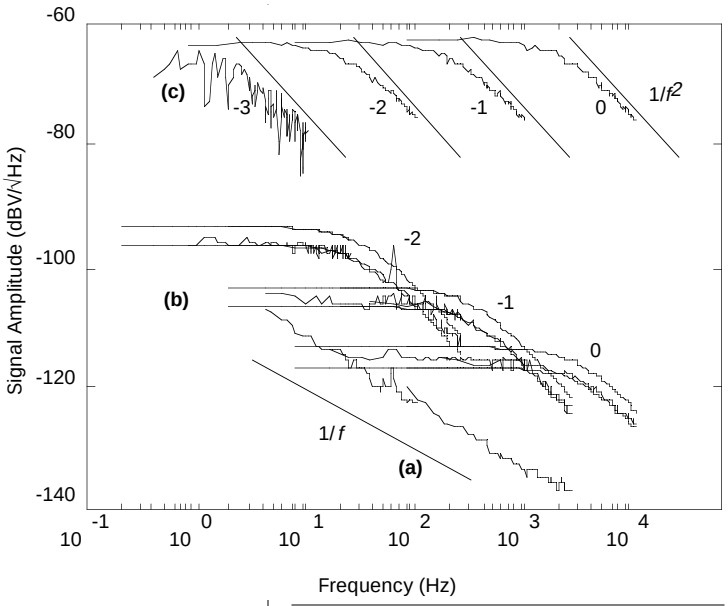


FIGURE 23 Noise spectra for the source-follower logarithmic receptor shown in Figure 22, at different intensities. The curve labeled **(a)** is from an isolated follower pad, and shows that the measured 1/*f* instrumentation noise in the follower pad and spectrum analyzer is smaller than measured spectral noise from the detector. The curves labeled **(b)** show the intrinsic noise spectra from the receptor at a given level of steady background intensity. The number by each curve is the log background irradiance. 0 log irradiance is 1.7 W/m^2. The lower of each smooth curve is the theoretical fit based on the theory given in the text; the upper curve is twice the theoretical value. The curves labeled **(c)** show the response of the receptor to small-signal white-noise stimulation from an LED. The stimulus for each curve has the same contrast, i.e., it is formed by interposing neutral density filters between a white noise source and the receptor. The straight lines have a slope of $1/f^2$, the same as from a first-order low-pass filter. The number by each curve is the log background irradiance. Definition of dBV units is the signal power, in dB, relative to a 1 V signal. Parameters used in the fits: node capacitance C=341.2 fF, temperature T=300°K, and time constant τ chosen to make cutoff frequency correct at the brightest intensity. τ is scaled inversely with intensity for the other curves. The capacitance consists of a 20x20 µm^2 photodiode with areal capacitance of 0.122 fF/µm^2 and edge capacitance 0.451 fF/µm (total 85 fF), a 6x6 µm^2 gate with areal capacitance 0.828 fF/µm^2 (oxide thickness 417 Å, 29 fF), and a metal wire with total capacitance 92 fF.

ble, although it is clearly dominant in the instrumentation. One often gets the impression from the literature that flicker noise dominates MOS transistor operation, but here it clearly does not.

The second set of observations (Figure 24) compare the noise spectra of the simple source-follower receptor and the adaptive receptor. We injected a small test signal to examine the SNR degradation by the adaptive feedback circuit. We can make the following observations:

1. The adaptive feedback circuit amplifies both signal and noise, but degrades the SNR by less than 3 dB.
2. 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 given in the discussion of the adaptive receptor, that were based on the parasitic capacitances and gain measurements.

The main conclusion of this measurement is that the degradation of the SNR by the adaptive feedback circuit is small enough that our analysis can treat the feedback and adaptation as a noiseless amplifier. If we can understand what determines the noise in the input stage of the adaptive receptor circuit, then understanding the noise behavior of the complete adaptive receptor circuit is trivial.

THEORY OF LOGARITHMIC RECEPTOR NOISE

In a logarithmic detector, the natural input units are fractions of the baseline signal, and the natural output units are fractions of the e-folding parameter. In the source-follower receptor, the gain of the receptor is simply $V_T = kT/q$ per e-fold change in the intensity, and hence the total dimensionless noise power P is given by

$$P \equiv \frac{\Delta v^2}{V_T^2} = \frac{\Delta i^2}{I_{bg}^2} = \frac{\Delta Q^2/C^2}{V_T^2} \qquad (19)$$

where Δx^2 means the mean-square variation of x. The reason we write the noise in the form of Equation 19 is that we shall derive an expression for the mean square variation in the *charge* sitting on the output node of the source-follower receptor. We will use that expression in Equation 19 to obtain the receptor noise.

There are two equivalent methods to compute the charge fluctuation. The first way uses the principle of equipartition from statistical mechanics. The second way analyzes the statistics of the individual charges.

USING EQUIPARTITION TO COMPUTE THE NOISE POWER.

The principle of equipartition says that the average energy stored in each independent degree of freedom of a system in thermal equilibrium is $kT/2$. A degree of freedom is a parameter that appears quadratically in the energy—for instance, each component of the velocity of a free particle. Similarly, the charge on a capacitor is a degree of freedom, because the energy stored on the capacitor is given by $Q^2/2C$. Using equipartition, we can write the following relation between the fluctuations in Q and the temperature:

$$\frac{\Delta Q^2}{2C} = \frac{kT}{2} \qquad (20)$$

or

$$\Delta Q^2 = kTC \qquad (21)$$

Substituting this result in Equation 19, we obtain the total dimensionless noise power for the source-follower receptor:

$$P = \frac{kTC/C^2}{V_T^2} = \frac{qV_T\,C/C^2}{V_T^2} = \frac{q}{CV_T} \qquad (22)$$

The simplicity of this answer is quite remarkable. It says that the dimensionless noise power—namely, the noise expressed in input units—is the ratio of the unit charge to the "thermal charge," CV_T. In hindsight, what else could it have been?

For a typical input capacitance of 100 fF, Equation 22 says that the total noise power at room temperature is about 10^{-4}, equivalent to an RMS variation of about 1 %.

TOTAL NOISE IN ADAPTIVE RECEPTOR.

The feedback circuit in the adaptive receptor adds minimal noise, but it does extend the bandwidth. The total receptor noise is hence increased to

$$P = \frac{q}{C_{eff}V_T} \qquad (23)$$

where C_{eff}, given in Equation 7 (page 5), is the effective input capacitance assuming a source conductance I_{bg}/V_T at the source of Q_{fb}. For a typical speedup of about 30 (Table 1, page 6) the total noise is increased by the same factor, leading to an RMS variation of about 5 %.

USING SHOT NOISE STATISTICS TO COMPUTE THE NOISE POWER.

The equipartition computation may appear to be magic. Relying solely on this principle may give one an uncomfortable feeling that something has been left out. That something is the intuition and understanding of the origin of the noise, and why it takes the interesting form given by Equation 22.

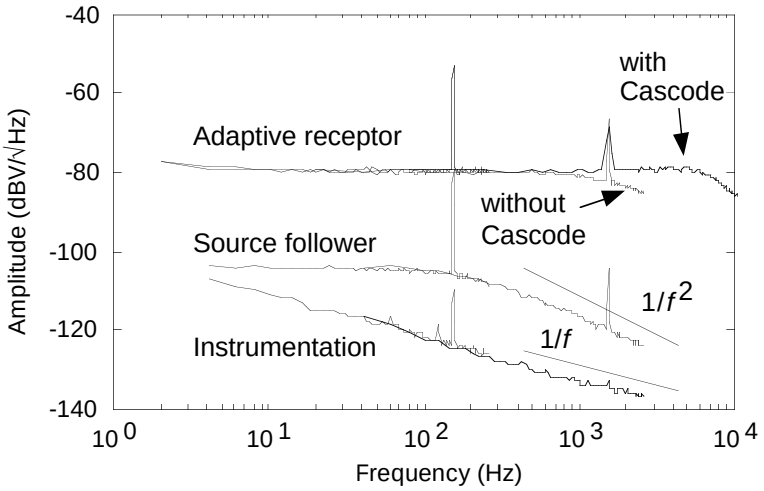


FIGURE 24 Comparison of noise spectra for simple source follower detector of Figure 22 and the adaptive receptor shown in Figure 5, with the layout shown in Figure 21. The gain of the source follower receptor is about 60 mV per decade. The gain of the adaptive receptor is about 1.3 V/decade, or 27 dB more than the gain of the source follower receptor. The plots show the measured power 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 nearly indistinguishable. For definition of dBV units, see Figure 23.

The macroscopic flow of current through the receptor circuit consists of the microscopic movement of discrete charges. As shown in Figure 25(a), single charges cause step changes in the voltage on the capacitor. The charges are collected by the photodiode and leave via the channel of the feedback transistor. The time at which a charge appears, and the time that it stays on the capacitor, are both random. A given charge spends a random amount of time sitting on the output node. These times are distributed according to a Poisson distribution.

Each charge is independent of all the others, which means that if we can compute the statistics of the average charge, then we can easily obtain the statistics of the current as a whole. We shall first compute the average noise *energy* in the step change of charge caused by a *single* charge. Using the independence of the charge events, we shall then compute the noise *power* in a current that consists of a *flow* of single charges.

The amplitude of the step change is the unit charge q of a single electron. The noise energy contained in the step is the integral over time of the squared deviation of the charge from the mean value. The mean value is zero, since this single charge has finite duration. During the presence of the charge, the value is simply q, the size of the charge. Hence, the mean value of the noise energy is given by $q^2\tau$, where τ is the mean length of time that the charge is present. Hence, the noise energy in a single event is given by

$$E \equiv q^2\tau \qquad (24)$$

The current consists of a random stream of single events. All the events are independent, so we can compute the mean square charge fluctuation ΔQ^2 (the noise *power*) in the signal by taking the noise *energy* in a typical event given by Equation 22 and multiplying by the average number of events per unit time, given by I_{bg}/q:

$$\Delta Q^2 = E\frac{I_{bg}}{q} = q^2\tau\frac{I_{bg}}{q} = qI_{bg}\tau \qquad (25)$$

We can think of the preceding computation in two ways: Either as an average of a single particle over many lifetimes, or as an average of many particles over a single lifetime. The two viewpoints lead to identical results; the former is perhaps more rigorous in terms of obtaining the correct multiplier, while the latter offers a better intuition of what process physically governs the noise behavior—an average over a time window. Formally, the equivalence of the two approaches arises from the ergodicity of the system.

A crucial fact is that the average lifetime τ of a charge is the same as the time constant of the node—a fact that is obvious but hard to prove. The reason it is true is that the time constant describes the time scale of the response of the node to a disturbance. If the disturbance consists of a sudden excess amount of charge on the node, the system returns to its equilibrium level after some time—the time being the average lifetime of the individual charges that make up the disturbance. The decay from the excess is exponential, and the e-folding time is once again τ. Hence, in Equation 25 we can replace τ with the time constant of the subthreshold circuit, CV_T/I_{bg}, to obtain

$$\Delta Q^2 = qI_{bg}\frac{CV_T}{I_{bg}} = qCV_T = qC\frac{kT}{q} = kTC \qquad (26)$$

which is the same result as Equation 21, obtained using equipartition. Notice that here we have suddenly slipped thermal behavior into a discussion that was strictly statistical. Boltzman statistics determines the time scale of the integration, which in turn determines the total noise.

SHOT VS.THERMAL NOISE? When we initially attempted to compute the noise in the receptor, our theory consistently differed from the measured noise power by a factor of about two—our theory always predicted a noise power *twice* what we measured. This distressing situation was not resolved until Rahul Sarpeshkar pointed out that shot and thermal noise are alternative aspects of a single underlying phenomenon. We had been under the impression that they

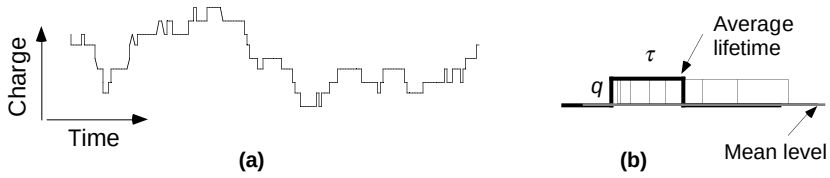


FIGURE 25 Shot noise computations. **(a)** The flow of current in the source-follower receptor consists of unit charges that appear and disappear, acting independently. **(b)** A single charge appears at a random moment, lasts a random amount of time, and disappears. The distribution of times that the impulse lasts is described by a Poisson process. The *mean* charge level is zero. The average energy in the impulse is $q^2\tau$.

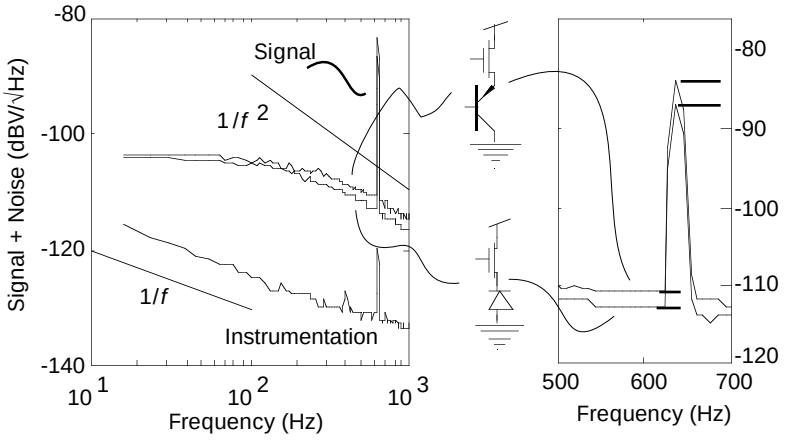


FIGURE 26 Photodiode and phototransistor noise properties are indistinguishable, and bandwidths are identical with passive feedback. The stimulus is a steady background with a small AC component. The test signal has the same SNR for each device.

were separate phenomena, with identical magnitudes. We computed the shot noise, and then insisted that the thermal noise was something additional—which is wrong.

Perhaps the easiest way to see how shot and thermal noise are related is to consider the flow of individual charges that make up a current. Looking at the current, we observe an average arrival rate I_{bg}/q of elementary charges. The charges are uncorrelated. Given an average arrival rate, and the fact that all the charges are independent, in some sense the statistics of the flow are as random as they can possibly be. How could more noise be added? Only by introducing additional correlations. An example is 1/f noise, where modulatory fluctuations in the level of current contribute additional noise power. Temperature has no effect on these statistical fluctuations; it serves only to set the bandwidth—the averaging time—at which the system operates.

EFFECT OF TEMPERATURE. It may seem odd that the total noise power given by Equation 22 is inversely proportional to the temperature. Usually, we think that reducing the temperature *reduces* the noise. This odd behavior is due to a bandwidth effect. When the temperature is reduced, the thermal voltage scale is also reduced. At a given level of current, determined by the light intensity, the source conductance of the feedback transistor is increased, and the time constant is reduced. Hence, reducing the temperature increases the bandwidth. The circuit integrates over a smaller time scale, the current is observed with finer granularity, the individual charges are more closely resolved, and hence, the total dimensionless noise power is increased. We must recall that we are talking about the dimensionless noise, given by Equation 19, that measures noise in input-referred units, and not the noise voltage. We can see from Equation 19 and Equation 22 that the mean-square output noise *voltage* is given by

$$\Delta v^2 = \frac{qV_T}{C} \tag{27}$$

which is proportional to temperature, but the output noise voltage is not a relevant quantity by itself because it must be compared with the natural voltage scale given by V_T.

NOISE SPECTRAL DENSITY. The amount of noise power per unit bandwidth—the **spectral density** of the noise—is interesting, because it says how detectable a signal is if we know that it occurs within some particular frequency band.

The noise fluctuations are spread over the spectrum of a first-order lowpass filter. The power spectrum (the power per unit frequency) is given by

$$S(f) = \frac{dP}{df} = \frac{4\tau P}{1 + (2\pi f\tau)^2} \tag{28}$$

The funny normalization constant $4\tau P$ appears so that the integral over all frequencies is the total power P:

$$\int_0^\infty S(f)df = P \tag{29}$$

Within the passband of the lowpass filter, the spectral density of the shot noise is

$$S = 4\tau P = 4\left(\frac{CV_T}{I_{bg}}\right)\left(\frac{q}{CV_T}\right) \tag{30}$$

or

$$S = \frac{4q}{I_{bg}} \tag{31}$$

We can attribute half of this noise power to each terminal connecting to the capacitor, to obtain a well-known result given in all the standard texts:

$$\frac{d\Delta I^2}{df} = 2qI \tag{32}$$

Equation 32 gives the shot noise power on a node arising from a single source of current onto that node. However, there are subtleties involved in the indiscriminate use of this expression. Sarpeshkar analyzed the noise properties of MOS transistors for both saturated and nonsaturated operating conditions in subthreshold operation, and showed that Equation 32 applies to a MOS transistor only when the transistor is saturated.[26]

THE ESSENCE. From Equation 31, we see that S has units of time: The spectral density of noise is proportional to the time per unit charge. The less time between charges, the smaller the noise. If we observe the signal over the time scale corresponding to the transit of a single charge, then it makes sense that the noise is approximately unity. At a current of 1 nA, the spectral noise density from Equation 31 is approximately 1 ns. The noise power is unity if we observe the signal over a time scale of 1 ns, and it decreases linearly with the time over which the signal is averaged.

MEASUREMENT AND THEORY

We can compare the measured spectra of the source-follower receptor with the theoretical spectrum. The three fit parameters are the temperature T, the capacitance C, and the integration time constant τ—which scales inversely with intensity. After accounting for all node capacitance using parameters supplied by MOSIS, assuming room temperature operation, and choosing the time constant to match the cutoff frequency of the measured spectrum for the brightest intensity, we can see that the fit is very good. The lower of the two theoretical curves is the predicted result; the upper curve is twice the prediction. Another measured receptor with a different capacitance shows comparable results, although we don't show the data here because the results are so similar. Direct measurements of the current noise spectrum in a single transistor confirm these results with a single parameter fit.[26]

ASSUMPTIONS OF DIFFICULTY. A common phenomenon for people who study noise is to start with a pessimistic attitude that assumes that noise, inherently a random phenomenon, is not quantifiable. The statistics of noise are just as quantifiable as the statistics of the steady-state flow of current in a MOS transistor. In the case of steady-state current, we study statistics of diffusion—first-order statistics, while in the case of noise we study statistical fluctuations—second-order statistics. It is satisfying to be able to quantify so precisely a phenomenon that initially seems intrinsically random and unpredictable.

For more insight and detailed discussion about electronic noise and photon counting,

Sarpeshkar *et al.* [26] and Rose's two books[24,25] are helpful.

PHOTODIODE *VS.* PHOTOTRANSISTOR: NOISE BEHAVIOR

Earlier we saw that phototransistors have the disadvantage that the base cannot be voltage–clamped to speed up the response. Figure 1 shows that the noise properties of photodiodes and phototransistors are indistinguishable. The reason is obvious, given that the analysis just given may be directly applied to the base-emitter junction of the phototransistor.

NOISE ADVANTAGE OF CONTINUOUS-TIME RECEPTORS OVER SAMPLED DETECTORS

Continuous-time logarithmic photoreceptors have a natural advantage in noise performance over sampled detector systems. The advantage comes from the scaling of bandwidth with intensity, so that circuit, in a sense, adapts its integration time to how much signal is available. The less light there is, the longer the integration time. The total noise is held constant.

In a CCD detector, the bandwidth is limited by the sampling rate, because the pixels cannot store state across samples. This limited integration time means that the number of integrated photons goes as the intensity. The total noise power goes inversely with the intensity—it gets worse as the intensity decreases instead of staying constant. There is nothing magical about it: the CCD detector holds its bandwidth fixed while the log receptor holds the number of integrated photons fixed.

If we look at the spectral density of the noise power, we see that there is no difference: in both types of detectors the spectrum of the noise comes from the quantal nature of light and charge; in both, the noise spectral density goes inversely with the intensity.

One additional complication is that a CCD detector invariably uses a charge sensing amplifier to convert the photocharge to a voltage—the signal that is actually used. The noise in the charge-sensing amplifier is generally equivalent to a fixed number—on the order of a hundred—of electrons at the input. As the intensity decreases, the number of integrated photocharges becomes comparable and then smaller than this fixed amplifier noise. For very low intensities, the mean-square charge fluctuation in the output signal is fixed instead of decreasing linearly with intensity. Here is a specific example that will make this idea clearer: Suppose the amplifier noise—the total RMS fluctuation—is equivalent to 100 photoelectrons. When the integrated number of photoelectrons is 100, the fluctuation in the signal itself is 10—a 10% variation. In reality, however, this small fluctuation in the signal is totally swamped by the amplifier noise. The total noise is equivalent to a fractional variation of 100% of the signal!

Under dim lighting conditions, shot noise dominates any detector system. The limited integration time of a CCD detector means that shot noise is a real limitation to the usable limit of operation for video imagers, where the integration time is at most 1/30 s. At moonlight illumination levels, the number of charges collected per integration time is less than 1000. The shot fluctuation, 30, means that at best—with a noiseless amplifier—the noise level is 3%, limiting the SNR to 30 dB.

MINORITY CARRIER DIFFUSION & GUARD STRUCTURES

Turning now to an different topic, we'll discuss the diffusion of minority carriers and guard structures. We built the simple structure shown in Figure 27 specifically to measure lateral diffusion length of minority carriers, and to test for the effectiveness of various guard structures in blocking their diffusion. The structure consists of a central parasitic bipolar phototransistor surrounded on three sides by guard bars with different widths. The fourth side is left bare. The bipolar transistor acts as a probe for minority carriers by collecting and amplifying the local concentration of excess minority carriers in the bulk. This chip was 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.

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.

We imaged a 12 µm spot of light onto the test structure at various locations away from the sensing transistor using 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.

The results of the measurements, Figure 28, show the measured current as a function of the distance of the test spot from the center of the sensing transistor. Each curve corresponds to moving across a different guard structure.

For distances greater than approximately 70 µm, the decay of carrier concentration is exponential and the measured e-fold distance is about 30 µm. An exponential-decay approximation is not good for short distances. The space constant is shorter, the closer the spot to the sensor, 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. With 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. 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. There is no difference. The decay with distance is nearly

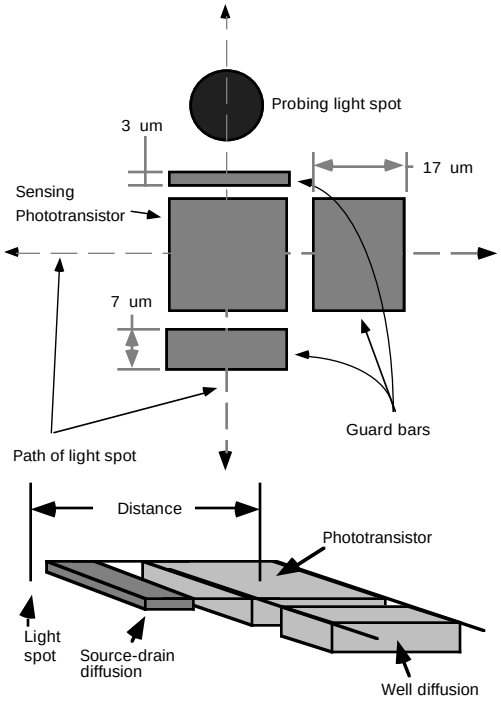


FIGURE 27 Layout used to test for light-generated minority carrier diffusion and for the efficacy of various guard bars in blocking minority carriers. The large square in the middle of the figure is a parasitic bipolar transistor that probes for minority carriers.We shined a probing light spot onto the 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 source-drain diffusion, and the 17 micron-wide structure is a p-well.

indistinguishable, as shown in Figure 28c. This result is not very surprising because the absorption length for red and green light centers around 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 consist 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 don't know the reason for this result, but we can repeat it.

SUMMARY. 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 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 by very much. Even a 17 µm-wide well—the deepest available diffusion—can only reduce the minority carrier concentration by a maximum factor of about 10. Circuits must be inherently resistant to the effect of excess minority carriers—like the adaptive element discussed earlier—if they are to function correctly with light shining on the chip.

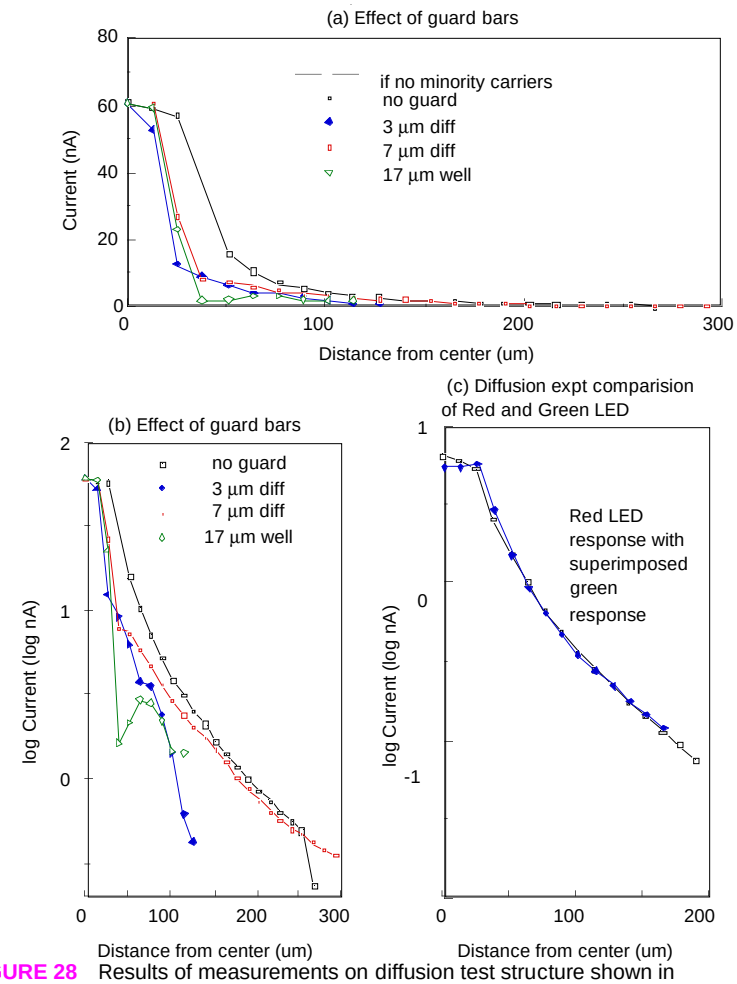


FIGURE 28 Results of measurements on diffusion test structure shown in Figure 27. **(a)** shows the measured photocurrent due to minority carrier diffusion as a function of the distance of the test spot from the sensing phototransistor. The different curves show the current for movement of the spot in 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.

SPECTRAL SENSITIVITY

The absorption of light by silicon is 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 bulk silicon, is shown in Figure 29. 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.

The wavelength-dependent absorption means that photodetectors formed from junctions with different junction depths have different spectral responses. In a CMOS process, there are a number of different junctions with different doping and depth. An ordinary CMOS process has two complementary source–drain active diffusions and a well diffusion. For concreteness, we consider an n-well process; the well and the active diffusion for the native transistors are n-type, and the substrate and active diffusions for the well transistors are p-type.†† We can form four photodetectors in a plain CMOS process, the three photodiodes plus a single parasitic PNP bipolar transistor, whose emitter is active, base is well, and collector is substrate.

Figure 30 shows a set of typical carrier concentrations in the parasitic vertical bipolar for different photon absorption lengths. We computed the curves from the diffusion equation. The carrier concentration goes to zero at the edge of a junction. At the surface, the concentration is reduced by surface recombination. The junction current is proportional to the slope of the carrier concentration at the edge of the junction. Long-wavelength light creates most carriers deep in the substrate, so the current is mostly found in the deepest junction. In contrast, short-wavelength light causes a current mostly in the shallowest junction.

A BiCMOS process adds a medium-doped p-type diffusion intermediate in depth between the active diffusion and the well diffusion. This new implant is used to form the p-type base for vertical NPN bipolar transistors.††† The emitter of the vertical

† 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.

†† 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.

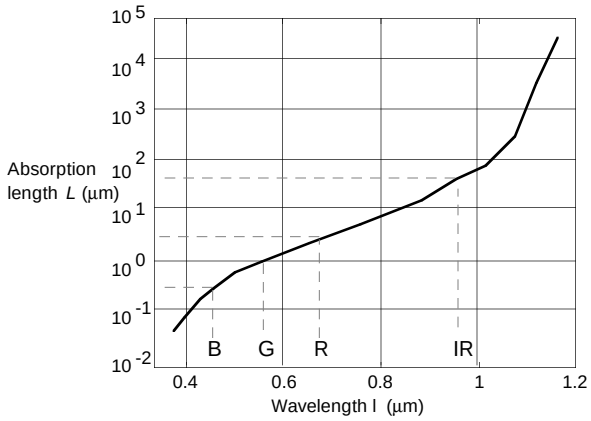


FIGURE 29 The photon absorption length as a function of photon wavelength at 300˚K. The absorption length is the distance over which 1/e of the incident photons are absorbed. Approximate wavelength of primary colors (B = blue, G = green, and R = red) from CIE color wheel, along with associated absorption 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 18). (Adapted from Dash and Newman.[3])

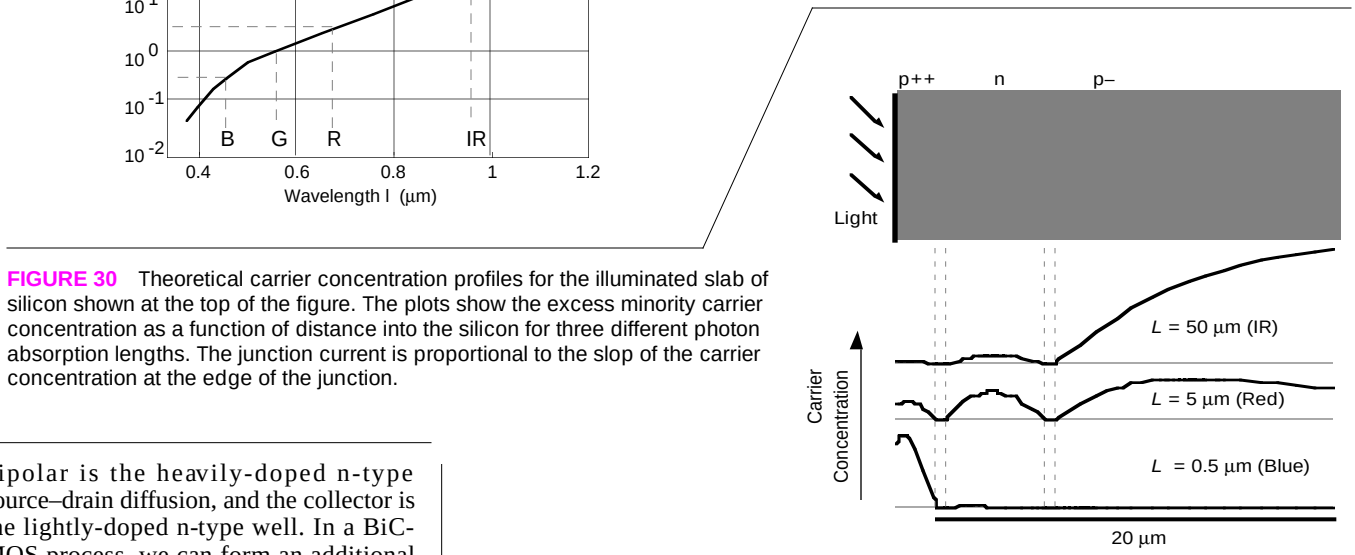


FIGURE 30 Theoretical carrier concentration profiles for the illuminated slab of silicon shown at the top of the figure. The plots show the excess minority carrier concentration as a function of distance into the silicon for three different photon absorption lengths. The junction current is proportional to the slop of the carrier concentration at the edge of the junction.

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 an additional four photodetectors: two photodiodes from the base to the emitter and collector, plus two vertical bipolars, one being the floating NPN, the other the parasitic PNP that uses the base diffusion as an emitter in a manner similar to the parasitic PNP discussed earlier.

Of these eight devices, we measured the six shown in Figure 31. We did not test the simplest photodiode from native diffusion to substrate, because we neglected to fabricate it on the same chip in the same configuration, making reliable results difficult to obtain. We expect it to behave similarly to the well–substrate photodiode. Also, we

††† 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 it lacks an 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.

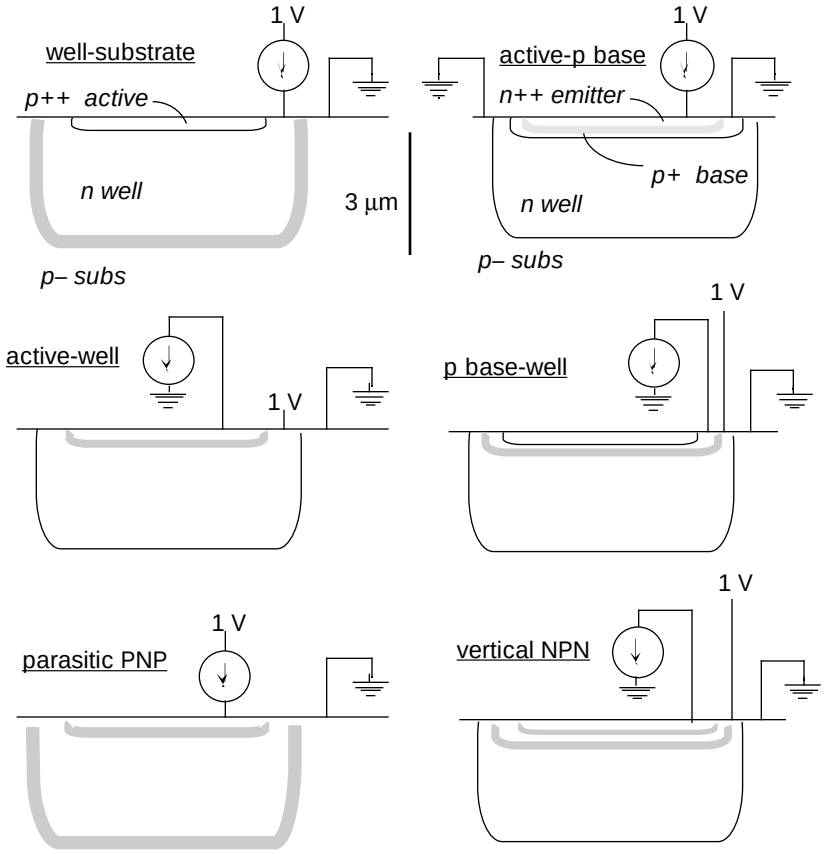


FIGURE 31 The structures and biasing setups used to measure the quantum efficiencies of the devices. The upper four devices are photodiodes and the lower pair are phototransistors. In the text, we refer to the elements by the underlined names in the figure. The active junctions are shaded for each device. A scale bar is shows approximate dimensions (horizontal dimension not to scale).

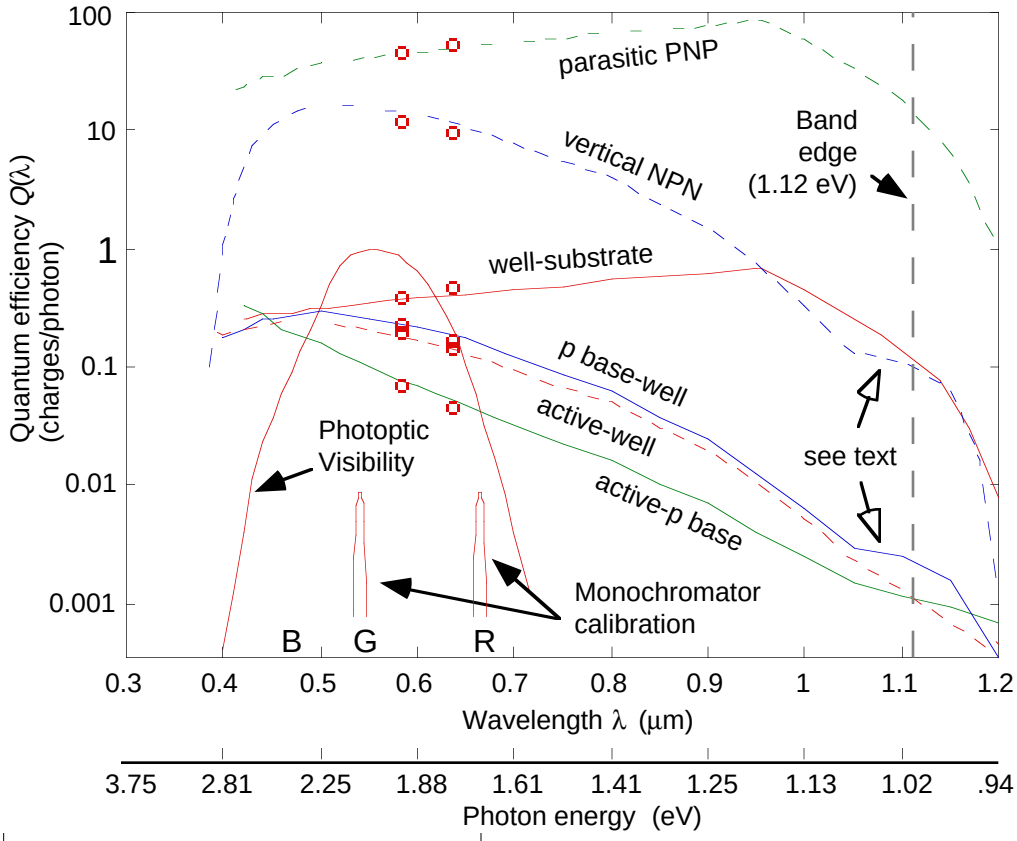


FIGURE 32 Measured spectral quantum efficiencies $Q(\lambda)$ versus photon wavelength and energy. The quantum efficiency is the number of collected charges per incident photon; it is larger than one for the phototransistors because they have built-in current gain. Each curve is labeled with the name of the device as shown in Figure 31. The small circles are the absolute calibration points measured with discrete LEDs. The photopic visibility curve shows the relative visibility of photons under photopic conditions; this curve is normalized to 1 at its maximum. The primary colors, according to the CIE color chart, are labeled on the wavelength axis. A shelf, marked with a hollow arrow, appears around the band edge for both the vertical NPN phototransistor and for the p base-well photodiode that forms the base to collector junction (page 19). The monochromator calibration curves show the spectral line width of the monochromator and the calibration wavelengths.

did not test the parasitic PNP that uses p–base as emitter because we forgot about it, although almost certainly its behavior is similar to the usual parasitic PNP that uses active as emitter.

A scale bar in Figure 31 shows the approximate vertical dimensions of the junctions of this 2-μm feature size BiCMOS process, 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} cm^{-3}. The p+ base is boron-doped, with a junction depth of 0.45–0.5 μm, and a surface concentration of $1–2 \cdot 10^{17}$ cm^{-3}. The n-well is phosphorous-doped with a junction depth of approximately 3 μm, and a surface concentration of $3\text{-}4 \cdot 10^{14}$ cm^{-3}. The p-type substrate has a doping of $3–4 \cdot 10^{14}$ cm^{-3}.†

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

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:

$$Q(\lambda) = \frac{\text{\# collected charges}}{\text{\# incident photons at wavelength}\lambda} \quad (33)$$

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 [6].

THE SPECTRAL RESPONSES

Figure 32 shows the absolute quantum efficiency versus photon wavelength. Figure 33 shows the same data on a linear scale. Figure 32 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 and the photodetectors that collect light-generated carriers from a diffusion-limited volume defined by the diffusion length of minority carriers. The diffusion-limited detectors are more sensitive to longer wavelengths.
- The response spectra are broad band. All the detectors cover 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 formed inside the well are most sensitive to a wavelength around 500 nm, while the deep well-substrate junction 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 at near-infrared photon energy, to a low of about 0.3 for the shallow, volume-limited junctions.
- All the quantum efficiencies are approximately 0.3 at the blue end of the spectrum.
- 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 parasitic PNP response is very similar in shape to the response of the well-substrate junction, but is completely unlike the response of the shallow active-well junction. A similar observation can be made for the vertical NPN. This observation simply means that most of the base current in the phototransistors comes across 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 high and low intensities. Figure 34 shows the current gain for the bipolar

† These parameters come from the specifications for the MOSIS 2–μm n-well BiCMOS process.

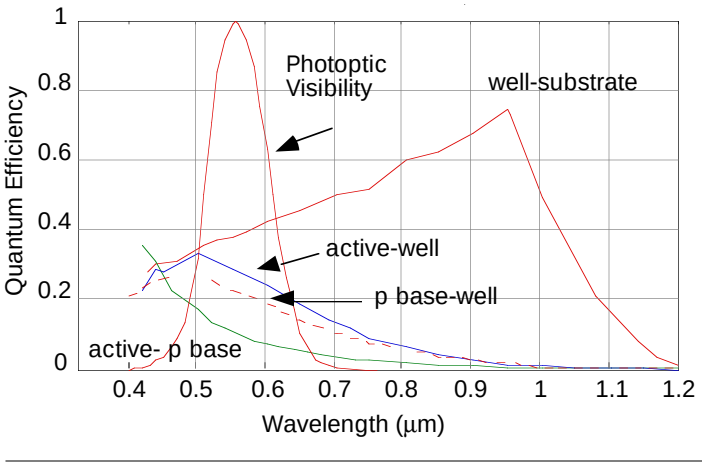


FIGURE 33 Absolute quantum efficiencies plotted on a linear scale. Data are a subset of Figure 32.

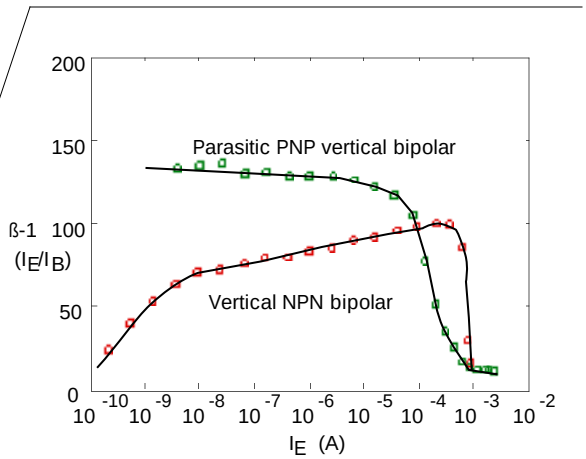


FIGURE 34 Bipolar transistor current gain as a function of emitter current. Ordinate is ratio of emitter current to base current. Data is from n-well, double-poly, 2 μm feature size, BiCMOS MOSIS technology. Collector to emitter voltage was held at 1 V. Transistor dimensions in μm are given in the table below.

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

transistors as a function of emitter current, measured by base-current injection. These current gains 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 one 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.
- There is an interesting shelf in the spectral response of the well to p–base junction right around the band edge that can also be seen in the NPN vertical bipolar response. 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

People frequently ask how much current to expect from a given size of photodiode or phototransistor. Since light intensity varies 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. Under these conditions, each 10 μm by 10 μm photodiode area, with quantum efficiency 0.5, generates a current

$$1\frac{J/s}{m^2}\times\frac{eV}{1.6\times10^{-19}J}\times\frac{\text{quantum}}{2.5eV}\times(10\mu m)^2\times0.5 = 10^8\frac{\text{quanta}}{s} = 25pA \quad (34)$$

Typical moonlight is about 3 decades less light, and hence a photocurrent of only 25 fA (i.e. 25x10^{-15} A), or approximately 2.5x10^5 charges/second, or 4000 in 1/60 s. In sunlight, the irradiance can be as much as 1 kW/m^2, corresponding to a photocurrent of 25 nA. Optics and scene reflectance properties reduce these numbers by about a factor of 10 under typical conditions.

PREVIOUS WORK

The precursor for the work described here is an optical mouse system built by Dick Lyon at Caltech. He used precharged photodiodes to produce a digital signal at a time inversely proportional to intensity. The system incorporated gain control by precharging of the photonode after it had discharged to a given level, and not at after a fixed integration time.

Mead's original logarithmic photoreceptor [19] used a single parasitic vertical phototransistor that feeds into a series of Darlington-connected lateral bipolar transistors. A feedback arrangement converts the final current into a voltage that is logarithmic in the intensity. This receptor was used in the Tanner and Mead optical flow chip. There was no good reason for using the lateral bipolar transistors.

The large area required by the lateral bipolar transistors quickly led to more compact designs. The early Mahowald and Mead silicon retinas, and subsequently many other Mead lab projects, used a simple logarithmic photoreceptor that consists of a parasitic vertical bipolar phototransistor, with a series pair of diode-connected MOS transistors as a load [13,20]. This simple design has the virtue that the layout is trivial and the receptor requires no bias controls. The signal produced by the receptor has a gain of 200–300 mV/decade of intensity. The two main problems with this receptor which led to the development of the adaptive devices are

1. A poor matching between different receptors. Signals output from neighboring receptors differ as much from offsets as from true signals caused by the scene.
2. A slow time-response, rendering the receptor useful only under bright lighting conditions and making all time-responses strongly intensity-dependent.

The SeeHear chip [22], designed by Neilson, Mahowald, and Mead, was the first analog VLSI vision circuit that used the idea of adaptation. In that chip, the simple nonadapting logarithmic receptors from the early Mahowald-Mead silicon retinas were used as input to Mead's hysteretic differentiator circuit [18]. The problem with this arrangement is that the uncoupling of the receptor and adaptation is inefficient in terms of transistor count, and forgoes the speedup advantages of the active feedback arrangement.

Delbrück and Mead [8] built a preliminary version of an integrated adaptive receptor that uses two stages of high-gain amplifica-

tion, one at the phototransducer, and the other in the feedback loop. This design had a large transistor count, lacked any speedup advantage of the active feedback, and used an inferior adaptive element.

Delbrück and Mead [7] built an adaptive silicon retina for sensing time derivatives of the image contrast, using the same feedback arrangement as described here, but lacking the speedup advantage of using a photodiode and a cascode, and using an inferior adaptive element. This silicon-retina chip uses the scanning mechanism to reset the pixel output every time the pixel output is sampled.

Mead[17] built an adaptive silicon retina that uses ultraviolet light to move charge onto and off floating gates for storage of analog mismatch compensation. A phototransistor with a drain-type load produces high voltage gain at the input, and this high gain point is used in the UV adaptation. Unfortunately, making high gain at the input is incompatible with high speed operation, and inherently prevents the type of voltage-clamping speedup we have discussed. Subsequent work has tended away from UV adaptation, because of technical problems with shining high-energy photons onto the chip and making the circuit function correctly at the same time.

Mahowald[12] incorporated the feedback arrangement described here into a silicon retina with interesting network feedback. Her receptor does not have the speedup advantage of the active feedback, and uses an inferior compressive adaptive element, leading to very unsymmetrical adaptation rates for bright- and dark-going illumination changes. Mahowald's chip with the receptor improvements we have discussed here is in fabrication.

Mann[15] developed several adaptive photoreceptor circuits that are more flexible than the ones described here, in that they have an adjustable temporal passband. These receptors also use U.V. mismatch compensation. They use a larger number of components than the receptor described here. They also lack the advantage of active speedup and use inferior adaptive elements.

None of the previous photoreceptors were satisfactorily characterized, in the sense that no one bothered to measure simple engineering metrics like usable dynamic range and sensitivity.

Up until lately, continuous-time analog photoreceptors have not received a great deal of commercial attention. Nearly all effort is concentrated on sampled imagers like CCD video cameras. Photoreceptors are used in very specialized applications like optical repeaters, where high integration density and wide dynamic range are not so important compared with response speed, and where special fabrication technologies are available that allow use of tricks like PIN photodiodes and avalanche multiplication.

RELATION TO BIOLOGICAL PHOTOTRANSDUCTION

In the introduction we discussed functional characteristics of biological photoreceptors and how they inspired the silicon version. Three characteristics stood out: The adaptive properties, the gain control resulting in illumination independent contrast response, and the invariance of the response time to illumination.

In the silicon receptor, the adaptation state is stored as a charge on a capacitor. In rods and cones, the adaptation state is (at least partially) stored as the internal calcium concentration. The full story of photoreceptor adaptation mechanisms is too complex for discussion here, but the idea that there is an independent state variable for adaptation state is similar in the silicon and biological receptors.

More germane is the invariance of response time to illumination, because here we see some functional differences. How does the response-time invariance come about? The enormous gain generated in rods comes from 3–4 amplification stages, each with modest gain. 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.

This multistage gain mechanism contrasts with the silicon receptor, which has only a two-stage amplifier, one of which has fixed gain. As shown earlier, this gain mechanism results in a fast response time, but the response time is also inversely proportional to the intensity. We have not elected to try to build a silicon receptor with multistage gain, in analogy with the biological mechanism, because the response time is adequately fast already, and the log generating mechanism in subthreshold operation is so convenient. All the gain control in the silicon receptor is done right at the input stage, by the exponential I-V relationship at the source of the feedback transistor. The logarithm gotten for free, courtesy of Boltzman, is apparently not used for this same purpose in biological receptors. The obvious first guess, a membrane channel with diode rectifying characteristics, is not what the cones or rods use.

SUMMARY

We have discussed a photoreceptor circuit that is useful for continuous-time phototransduction. It consists of a logarithmic input stage coupled to an adaptive feedback circuit. The feedback circuit produces an output with high gain for transient—and presumably interesting—signals, and has low gain for static signals, including circuit offsets. The feedback also speeds up the response by clamping the input node voltage.

The receptor can be fabricated in an area of about 70 by 70 μm^2 in a 2–μm single poly CMOS process. The useful dynamic range, which we arbitrarily define as the range over which the bandwidth is at least 60 Hz, extends down to moonlight illumination levels, or at least 1.5 decades more than a plain logarithmic detector without active feedback.

The receptor noise in a plain logarithmic detector is almost exactly what is predicted by a simple noise theory that assumes that the noise is purely shot noise, with a bandwidth set by the conductance of the feedback and the capacitance of the input node. The noise in the adaptive receptor is within a factor of two of this value.

TECHNICAL INNOVATIONS

The three technical improvements in the receptor are as follows:

1. Using a photodiode, instead of the phototransistor used in previous designs, increases the dynamic range by at least a decade, without degrading signal quality.
2. Using a cascode transistor in the feedback amplifier yields another 0.5–1 decade dynamic range.
3. Using a new adaptive element increases the adaptation time constant and the symmetry of the receptor response to light- and dark-going transients, compared to previous designs, which used elements that were very susceptible to the effects of parasitic minority carriers.

We have also discussed two theoretical topics in which we have developed a much clearer understanding:

1. We quantitatively understand the physical origin and dominant sources of electronic noise in continuous-time logarithmic receptors.
2. We quantitatively understand the limits on response speed that effectively limit the lower end of the receptor dynamic range.

CONCLUSION

Operations that require spatiotemporal computation on noisy input data in real time are a natural fit to analog VLSI. We have shown how to build a practical photoreceptor circuit that is suitable as a front-end transducer for analog VLSI vision systems. The receptor is usable under photopic vision conditions. The continuous transduction process leads to intensity-invariant total noise that is within a factor of two of photon counting noise. Photoreceptors of this general type will undoubtedly be used in commercial vision systems, where designers will not be able to resist the capability to link low-power, continuous-time computation together with transduction.

ACKNOWLEDGEMENTS

We acknowledge support from the ONR and by MOSIS, the ARPA silicon foundry.[4] David Van Essen inspired this detailed study (especially of the noise properties). Rahul Sarpeshkar made a crucial observation about the relationship between shot and thermal noise. David Standley and Dick Lyon provided helpful comments about the manuscript. Micah Siegel and Sanjoy Mahajan pointed out that κ variations affect the gain vs. intensity characteristics and their measurement discrepancies inspired the discovery of the importance of the Miller effect in Q_{fb}.

REFERENCES

1. D.A. Baylor, G. Matthews, K.-W. Yau, "Two components of electrical dark noise in toad rod outer segments," *J. Physiol.* vol. 309, pp. 591–621, 1980.
2. 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.
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. 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.
5. T. Delbrück, "Silicon retina with Correlation-Based, Velocity-Tuned Pixels," *IEEE Transactions on Neural Networks,* vol. 4, no. 3, pp. 529–541, May 1993.
6. T. Delbrück, *Investigations of Analog VLSI Phototransduction and Visual Motion Processing,* Ph.D. Thesis, Dept. of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 91125, 1993.
7. 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.
8. 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.
9. 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.
10. M.A. Mahowald, *An Analog VLSI System for Stereoscopic Vision,* Kluwer Academic Publishers: Norwell, Massachusetts, ISBN 0-7923-9444-5, 1994.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. C.A. Mead, "Neuromorphic Electronic Systems," *Proceedings of the IEEE,* vol. 78, pp. 1629–1636, 1990.
17. 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.
18. C.A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison--Wesley, 1989.
19. 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.
20. C.A. Mead and M.A. Mahowald, "A Silicon Model of Early Visual Processing," *Neural Networks,* vol. 1, pp. 91–97, 1988.
21. J.R. Meyer-Arendt, "Radiometry and Photometry: Units and conversion factors," *Applied Optics,* vol. 7, pp. 2081–2084, 1968.
22. 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.)
23. R.A. Normann and I. Perlman, "The effects of background illumination on the photoresponses of red and green cones," *J. Physiol.,* vol. 286, pp. 509–524, 1979.
24. A. Rose, *Concepts in Photoconductivity and Allied Problems,* InterScience Publishers (div. of John Wiley), p. 104, 1963.
25. A. Rose, *Vision: Human and Electronic,* Plenum Press: New York, 1973.
26. R. Sarpeshkar, T. Delbrück, and C.A. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits and Devices*, vol. 9, no. 6, pp. 23–29, 1993.
27. 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.
28. S.M. Sze, *Physics of Semiconductor Devices, Second Edition,* New York: John Wiley & Sons, chapter 13 and appendix I, 1981.
29. 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.

INDEX