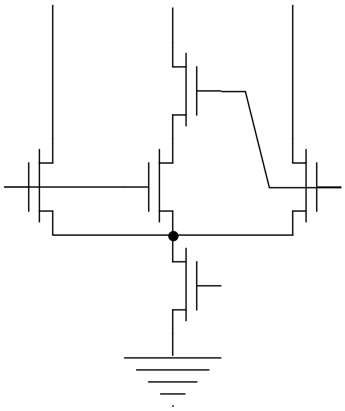
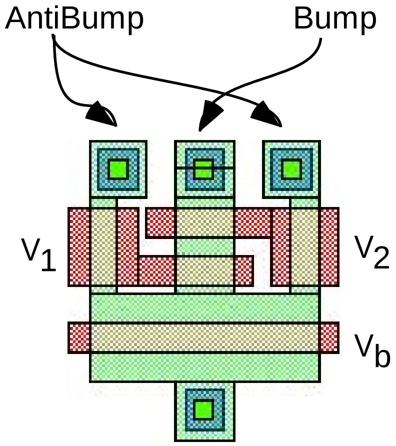
CALIFORNIA INSTITUTE OF TECHNOLOGY
COMPUTATION AND NEURAL SYSTEMS PROGRAM

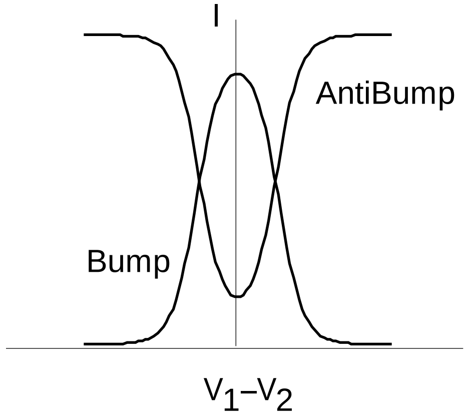
MAY 24, 1993 CNS MEMO 26

BUMP CIRCUITS
for computing similarity and dissimilarity of analog voltages

T. Delbrück







PASADENA, CALIFORNIA 91125

ABSTRACT

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

CONTENTS

BUMP CIRCUITS

for computing similarity and dissimilarity of analog voltages

Tobias Delbrück
139-74 Caltech
Pasadena CA 91125
tobi@pcmp.caltech.edu

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

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

This discussion suggests that it would be useful to have a circuit primitive that computes a measure of the similarity, or distance, between inputs. In this report, we discuss a class of such circuits—which we call **bump** circuits.

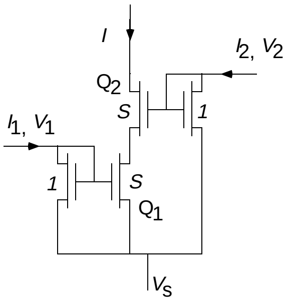


FIGURE 1 The simple current-correlator. *S* is the strength ratio between the transistors in the middle leg to the transistors in the outer legs. *I* is only large when both I_1 and I_2 are large.

SIMPLE CURRENT-CORRELATOR

Carver Mead recognized that in subthreshold operation, the circuit in Figure 1 computes an interesting generalized measure of the correlation of the two input currents I_1 and I_2. We will refer to this circuit as the **simple current-correlator**. Intuitively, the series-connected transistors act as a sort of analog logical AND combination. If either of the gate voltages on these series-connected transistors is low, then the output current is shut off; conversely, if both of the input voltages are high, then the output current is large. In the intermediate regions, the circuit computes a kind of product of the input currents.

This work has been published in part as T. Delbrück, ““Bump” circuits for computing similarity and dissimilarity of analog voltages,” in *Proc. of International Joint Conference on Neural Networks,* vol. 1, pp. I-475--479, 1991. These circuits have been patented as U.S. Patent 5,099,156, plus continuations in part. A previous CNS memo (number 10) was an incomplete version of the current report.

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

We will analyze the simple current-correlator in the subthreshold operating region, where the transistor current is exponential in the terminal voltages. To compute the mathematical form of the response of the simple current correlator, we use the transistor law for subthreshold operation [13]:

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

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

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

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

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

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

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

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

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

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

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

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

SIMPLE BUMP CIRCUIT

The first bump circuit, which we will refer to as the **simple bump circuit**, is shown in Figure 2a. The input is the differential voltage $\Delta V = V_1 - V_2$; the output, plotted versus ΔV in Figure 2b, is the current I_{out}. We can see that I_{out} becomes large only when the inputs are close together.

Intuitively, the simple bump circuit operates as follows. The currents I_1 and I_2 through the two legs of the differential pair are comparable only when

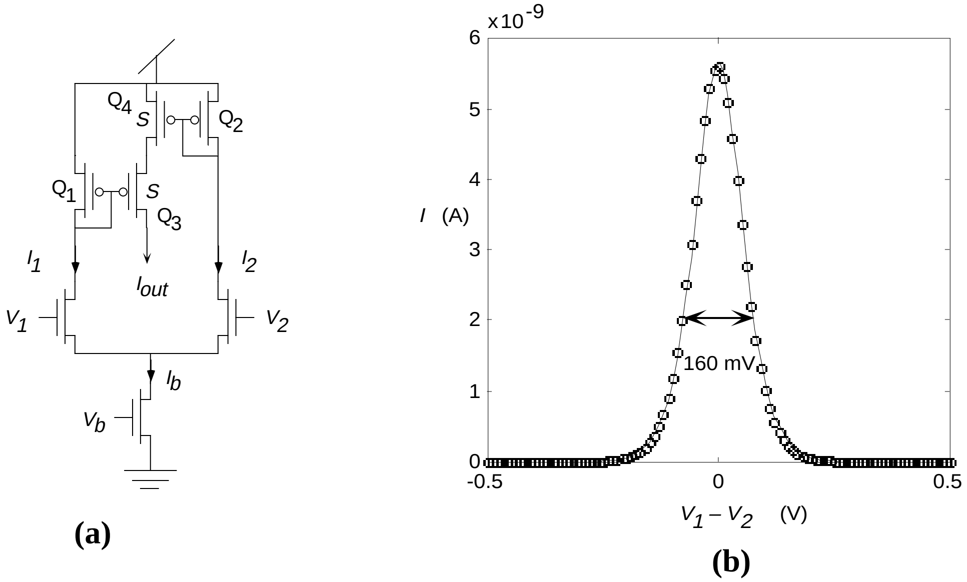


FIGURE 2 The simple bump circuit and response. **(a)** Circuit. The similarity output is I_{out}. The transistors Q_{1-4} are the simple current-correlator. **(b)** The output current from the circuit in (a) as a function of the differential input voltage. The solid curve is a fit of the form in Equation 6. The double arrow shows the width of the bump, measured at *1/e* below the maximum.

the differential input ΔV is near zero. When ΔV is larger than a few units of kT/q, the current in one of the two legs shuts off. The transistors Q_{1-4} form the simple current correlator shown in Figure 1. Thus, if ΔV is large, I_{out} is zero. If $\Delta V = 0$, I_{out} takes on its maximum value.

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

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

$$I_1 = \frac{I_b}{1 + e^{-\kappa\Delta V}} \quad . \qquad (5)$$

Using this result in Equation 3 and simplifying, we obtain

$$\begin{aligned} I_{out} &= I_b \frac{S}{4} \operatorname{sech}^2\left(\frac{\kappa\Delta V}{2}\right) \\ &= \frac{I_b}{\frac{4}{S}\cosh^2\frac{\kappa\Delta V}{2}} \quad , \qquad (6) \end{aligned}$$

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

By including four additional transistors in the simple bump circuit in Figure 2a, we can add a wide-range transconductance output [13] that computes $I_1 - I_2$. The new circuit takes a differential voltage and produce both a transconductance and a bump output. Because the addition is obvious, we do not show it here.

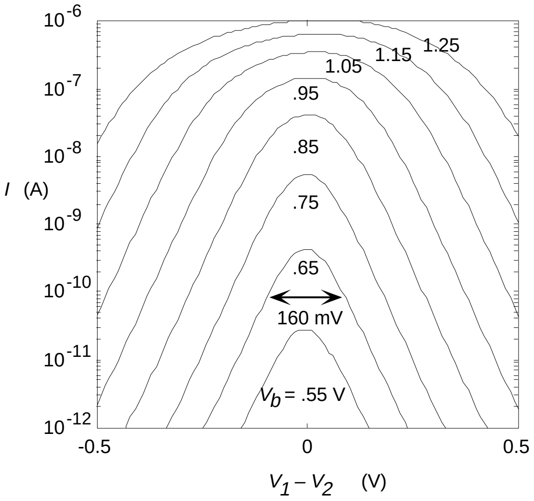


FIGURE 3 Simple bump circuit output as function of differential input, for various bias voltages. In these measurements, the input to V_1 was held constant at approximately 2V, and V_2 was swept around V_1. The double arrow shows the width of the bump, measured at *1/e* below the maximum.

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

Simple bump bias behavior

Results for different bias currents are shown in Figure 3. The form of the response is invariant under subthreshold biasing condition. Above threshold, the width of the bump grows and becomes somewhat unsymmetrical. The direction of the above-threshold asymmetry is such that the maximum output current occurs when I_1 is slightly larger than I_2. By duplicating the correlating transistors with swapped connections, we could symmetrize the response above threshold.

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

$$\Delta V_{1/e} = \frac{4.34}{\kappa} \qquad (7)$$

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

BUMP-ANTIBUMP CIRCUIT

Figure 4a shows the **bump-antibump** circuit. This circuit is more flexible than the simple bump circuit because there are three outputs: I_1, I_2, and I_{mid}, shown in Figure 4b. Output I_{mid} is the bump output. Outputs I_1 and I_2 behave like rectifier outputs, becoming large only when the corresponding input is sufficiently larger than the other input. If I_1 and I_2 are combined, they form the antibump output—the complement of the bump output.

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

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

$$I_b = I_1 + I_2 + I_{\text{mid}}, \qquad (8)$$

we can compute the current I_{mid}:

$$I_{\text{mid}} = \frac{I_b}{1 + \frac{4}{S}\cosh^2\frac{\kappa\Delta V}{2}} \qquad (9)$$

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

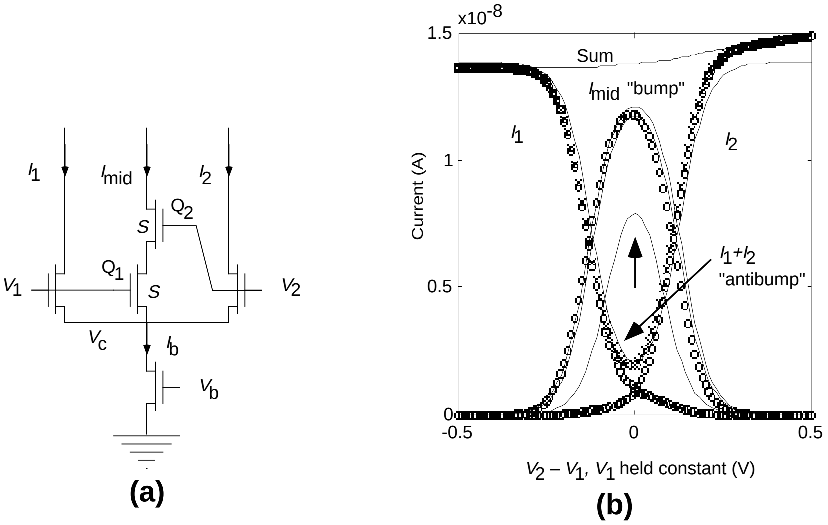


FIGURE 4 **(a)** The bump-antibump circuit. **(b)** Output characteristics of bump-antibump circuit. The plots show data points along with theoretical fits of the form given in the text. The curve pointed to by the arrow shows the fit that would result from using the *S* ratio 5.33 derived from the drawn layout geometry, before any process correction. The two theoretical curves shown for I_{mid} are the result of computing the fit using the best numerical fit to the entire curve (*S*=22.4), or using the ratio of maximum to minimum current in I_1+I_2 (*S*=28). The two numerically fit curves are virtually indistinguishable and clearly different that the theoretical curve derived from the layout geometry. The MOSIS fabrication service supplied width and length reduction parameters for this run of the order of 0.5µm. Using these parameters, we compute an *S* ratio of 6.6, still far short of the observed behavior. The curve labeled Sum is I_1+I_2+I_{mid}. The slope on the Sum curve is due to the drain conductance of the bias transistor.

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

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

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

Figure 4(b) shows measured representative operating curves for this circuit. We can see that the theoretical form for I_{mid} fits the data quite well, with one important exception. This exception is a discrepancy between the measured and expected value for *S*, the ratio of transistor strengths between the middle and the outer legs of the circuit. This effect is also seen in bump circuits fabricated on a different chip, as shown in Figure 5. The bump-antibump circuits acts as though the *S* ratio is much larger than it has any right to be, given the drawn layout of the circuit. This effect is fortuitous because it means that the designer who wants to use these bump circuits need not use inconvenient and bulky layout in order to achieve a large width and size for the bump response, which has generally been the desired profile for most designs to date. Starting on page 9, we examine this discrepancy in detail. From a practical perspective, designers can examine the measured data in Figure 5, which shows operating curves from eight bump-antibump circuits with different *S* ratios and layout sizes, to determine approximately the correct layout to use for a particular application.

Bump-antibump circuit bias behavior

Figure 6 shows the response of the bump output of the bump-antibump circuit for different bias currents. Below threshold, the width of the bump is a constant, independent of the bias current. Above

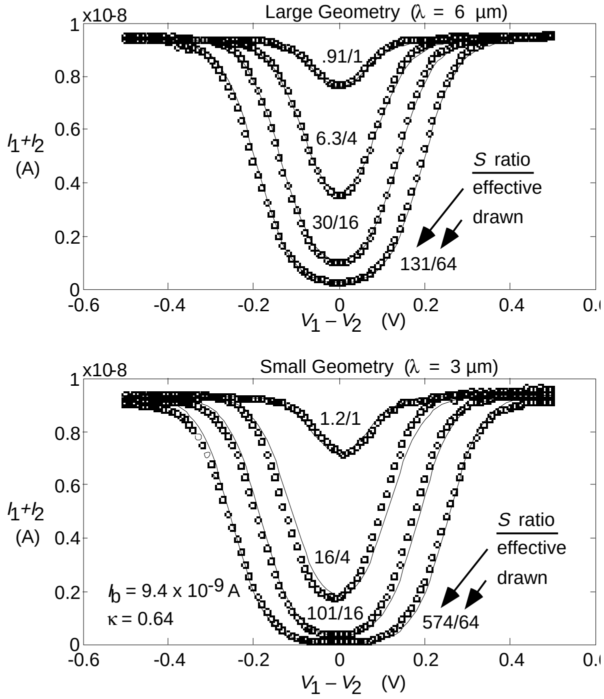


FIGURE 5 Antibump outputs from 8 bump-antibump circuits with different geometries. The solid curves show the theoretical fits derived from Equation 9. The numbers beside each curve are the actual and expected *S* ratios for the circuit. The different bump circuits in each set of graphs all had transistors of the same width in the outer legs, and transistors of the same length in the middle leg. For top set of curves, minimum transistor dimension was 6 μm, correlating transistors had widths 6, 12, 24, and 48 μm, and outer transistors had lengths 6, 12, 24, and 48 μm. For bottom set of curves, all transistor dimensions were halved. The discrepancy between measured and expected *S* values are larger for the circuits with smaller dimensions. We fit these curves by minimizing the total squared error, using a single common I_b and κ,.

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

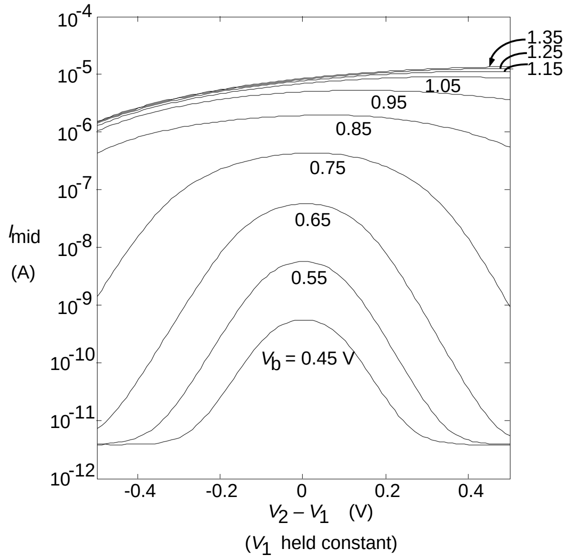


FIGURE 6 Effect of bias level on bump-antibump bump output. The curves plot the output current I_{mid} as a function of the differential input voltage $V_2 - V_1$. In this measurement, V_1 was held constant and V_2 was swept around V_1. The different curves are for different bias voltages. Up to a bias voltage of 0.55 V, the curves are simply shifted on a semi-logarithmic axis, indicating that the circuit is operating in the subthreshold region. (The bias transistor on this bump circuit is very strong, so the behavior goes above-threshold for a relatively small bias voltage.) When the circuit goes above threshold, the bump widens out and eventually becomes asymmetrical.

BUMP TRANSCONDUCTANCE AMP

By adding a current mirror to the bump-antibump circuit we can produce an output consisting of the difference current $I_{out} = I_1 - I_2$ (Figure 7). We will refer here to the resulting circuit as the **bump amplifier.** The output from a number of these bump amplifiers is shown in Figure 8, where they are compared with the theoretical output from a simple transconductance amplifier [13]. We can see that there is a flattened region in the center of the response curve that is due to the current flowing through the center leg of the circuit. The circuit in Figure 7 differs from a transconductance amplifier only in the addition of the two correlating transistors.

The rationale for the bump amplifier came from our observations that voltage offsets and current mismatches often dominate the behavior of

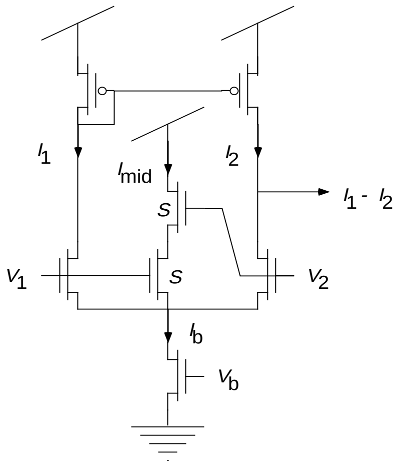


FIGURE 7 Bump amplifier.

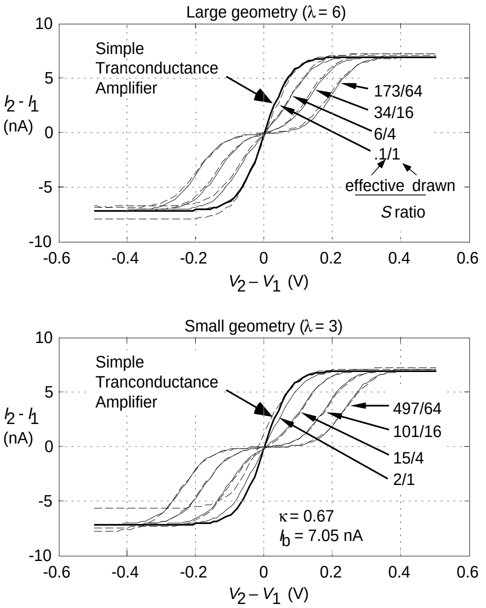


FIGURE 8 Responses of bump amplifiers of various *S* ratios. Dotted curves are data, solid curves show fits of the form Equation 11, and thick curves show theoretical simple transconductance amplifier response. Numbers next to curves show fitted and intended *S* values for curves. For top set of curves, minimum transistor dimension was 6 µm, correlating transistors had widths 6, 12, 24, and 48 µm, and outer transistors had lengths 6, 12, 24, and 48 µm. For bottom set of curves, all transistor dimensions were halved. Effects of offsets are visible on some of the curves, and these offsets occur primarily on the side of the response where the output current is mirrored.

produce a transconductance element that would ignore small voltage offsets of the input voltage, while retaining a monotonic saturating output characteristic for larger inputs.

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

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

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

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

Figure 8 shows the measured responses of bump amplifiers with different values of *S*, along with fits of the form given in Equation 11. As for the bump-antibump circuit, the fitted value for *S* is much larger than the drawn values.

Dick Lyon has observed that the bump amplifier may be used as a transconductance amplifier with an increased linear input-range, compared with the simple transconductance amplifier (personal communication). This device may be useful in filter circuits in which one amplifier must saturate at a larger voltage than another. It may also be useful in circuits that must decrease power consumption in the middle of the operating range. By choosing a certain value of *S*, we can make the response of the bump amplifier maximally linear. To find the desired intermediate value of *S*, we can minimize the “acceleration” of *I* at $\Delta V = 0$, by setting the third derivative of *I* with respect to ΔV to zero, and solving for *S*. It turns out that the desired value of *S* is 2, independent of κ. By examining Figure 8, we can see that small variations in *S* do not greatly affect the linearity of the response. For example, the curve with an *S* ratio of 6 is difficult to distinguish from a straight line passing through the origin.

The input range of this modified amplifier is larger than that of a simple transconductance amplifier. We can think of the expanded range as arising from an adaptive bias current. For small input, the effective differential-pair bias current is small, because part of the total bias current I_b is

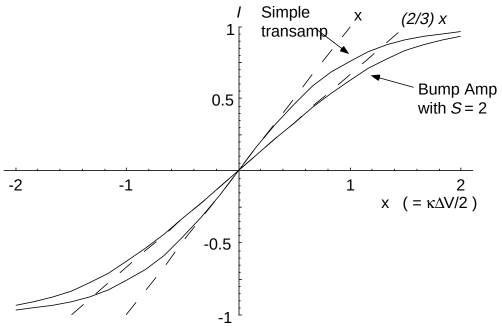


FIGURE 9 Theoretical comparison of simple transconductance response with maximally linear bump amplifier response. Solid curves show the simple transconductance amplifier and bump amplifier responses. Dashed curves show line through origin with same slopes as curves. The slope of the bump amplifier response is 2/3 that of the simple transconductance amplifier, and the deviation from linear behavior occurs at a larger voltage.

supplied by I_{mid}, and hence is stolen from the differential pair. For larger input, I_{mid} becomes smaller, and the effective differential-pair bias current becomes larger, increasing the linear input range. We can compute the increase in the linear range by doing a Taylor expansion of Equation 11, centered around $\Delta V = 0$, using the value $S = 2$ computed earlier for the optimally linear case:

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

A Taylor expansion of the simple transconductance amplifier response given by Equation 12 is

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

From these expansions, we can see that the bump amplifier has a transconductance that is 2/3 that of the simple transconductance amplifier for the same saturation current. Also, the term of order ΔV^3 has vanished, leaving only a residual of order ΔV^5. Figure 9 shows a theoretical plot comparing the bump amplifier and the simple transconductance amplifier responses.

Name	Author(s)	Description
Motion chip.	Delbrück [3]	Antibump circuit used to measure energy in delay line activity, by computing distance of delay line voltage from reference level.
Stereopsis	Mahowald [12]	Bump circuit used in conversion of value encoding (voltage) to place encoding (localized current).
Bump fuse	Liu & Harris [10]	Antibump circuit used to measure voltage difference to break resistive fuse at a controllable threshold.
Spike classifier	Kerns & Watts [5]	Bump circuit used to classify spike height.
Focus	Tobi Delbrück [3]	Antibump circuit used to measure energy in spatial pattern, by comparing image to smoothed reference image.
Herroult-Jutten network	Cohen and Andreou [2]	Bump amplifier configuration used in Gilbert multiplier to approximate a cubic term.
Cochlea	Lyon et al. [11]	Bump amplifier used to stabilize filter circuit.
Image Compression	Tawel [15]	Simple bump circuits used to classify pixel blocks for vector quantization.
Gradient Descent	Kirk, et al. [7][8]	4-input bump circuit used as target function for on-chip gradient descent implementation.

TABLE 1 Applications of bump circuits.

APPLICATIONS OF BUMP CIRCUITS

Table 1 licsts examples of applications of bump and antibump circuits in systems. In several cases, we can think of part of the system behavior as a specialized RBF network. The systems in Table 1 fall into three categories:

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

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

ELECTROSTATICS OF SUBTHRESHOLD CHANNEL

Analysis of the bump-antibump response curves for different *S* ratios tells us that the circuits show a larger *S* ratio than we expect from the drawn layout (Figure 5). Either short wide transistors are stronger, or long narrow transistors are weaker, than we expect from the drawn layout. This effect is large, and therefore of practical importance.

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

Figure 10 shows a conceptual model of the three-dimensional electrostatics around a MOS transistor channel. By conceptual, we mean that the potentials are what we expect, based on the geometry of the transistor and experience with similar structures. We have not explicitly computed these potentials starting from electrostatic theory. Part (a) of the figure shows a mesh plot of the surface potential in and around a transistor channel. The source-drain regions are at the lowest potential. The bulk covered with thick oxide is at the highest surface potential. The channel is at an intermediate potential. Part (b) of the figure conceptually shows what happens to the potential across the channel of the transistor, perpendicular to the flow of current, for different transistor widths.

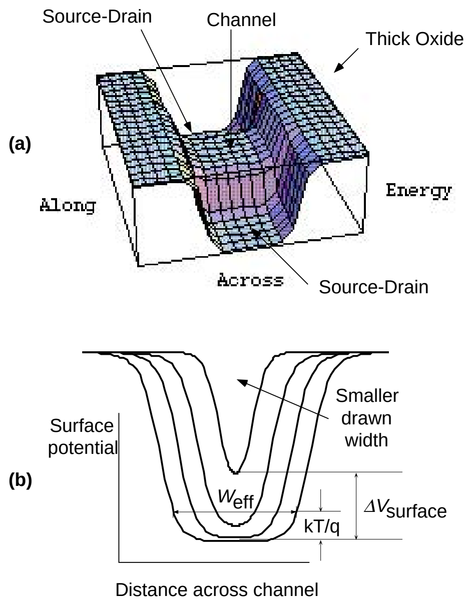


FIGURE 10 Conceptual plots of surface potential in and around transistor channel.

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

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

The fringing field causes a bowl-shaped potential across the channel. In subthreshold, the concentration of carriers is exponential in the potential, and hence the effective width of the channel is smaller than the drawn width. This effect is larger for channels with smaller width, both because the fractional effect of a fixed width is larger, and because the bottom of the surface potential starts to lift off its wide-channel minimum. Dave Kewley (personal communication, [6]) has designed devices that rely on this effect to modulate the flow of current, through the explicit use of side gates on the transistor.

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

We distinguish the across-channel effect just discussed from the usual dimensional effects. It is well-known that the physical *length* of a transistor channel is smaller than the drawn layout, because lateral diffusion of the source-drain implants causes the implant to extend under the gate region. The typical distance is about a quarter micron in

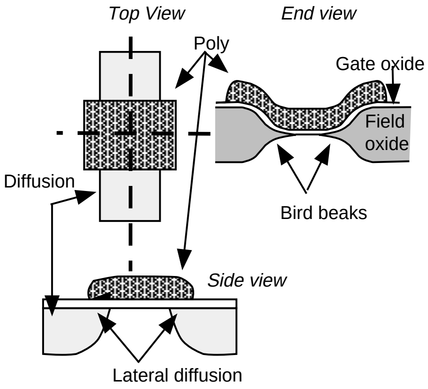


FIGURE 11 Lateral diffusion and oxide encroachment effects in transistor fabrication. We show three views of a transistor: from the top, a cross-section along the channel, and a cross-section across the channel.

the 2 µm process used in these chips. The length is further reduced by the finite width of the depletion regions surrounding the source and drain. Modulation of the drain depletion region by drain voltage results in the Early effect. It is also well-known that the physical *width* of a transistor is smaller than the drawn layout, because the process of field oxide growth eats under the masking nitride, causing **bird beaks**. These bird beaks make the oxide thicker, in regions where the drawn geometry indicates thin oxide. The thickened oxide effectively makes the channel narrower. The bird beaks are typically on the order of an eighth to a quarter micron wide. These two effects are schematically illustrated in Figure 11.

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

1. Data from the bump-antibump circuits tells us the scale of the effect is much larger than can be accounted for by the usual dimensional reductions. This data, however, does not disambiguate along-channel and across-channel effects.
2. Data about threshold voltages for various sizes of transistors tells us that the effect is mainly seen in narrow transistors, rather than in short transistors.
3. Data that we measured from various sizes of transistors confirms the dominance of the effect in narrow transistors, and shows that the effect depends on the transistor operating regime. In other words, the transistor size, relative to other transistors, is a function of whether the transistor operates in weak, moderate, or strong inversion.

To support our conclusions, we shall discuss these three items in order.

Bump Circuit Data

Figure 12 shows the relation between the intended S ratio and the actual value measured by fitting the antibump response curves in Figure 5. To understand these data, we will model the effect of the width and length reductions on the strength of the transistor. Hence, we will *pretend* that the effects are only due to a dimensional reduction in the size of the transistor channel. This analysis will be sufficient for phenomenological understanding, and later we will discuss the physics underlying the behavior, and the validity of the assumption that a simple dimensional reduction can account for the observed behavior.

We can make a simple model for the effect of width and length reductions on S. The effect of *length* reductions are significant only on the *short* transistors in the circuit, namely, the transistors in the middle leg. Similarly, the effect of *width* reductions are significant only in the *narrow* transistors in the circuit, namely, the transistors in the outer legs of the circuit. Hence, our simple model includes only those geometrical reductions shown in Figure 13. In the limit of large S, the effective S is related to the drawn S by

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

Equation 15 does *not* include the effect of the width and length reductions on the long dimensions of the transistors, because these effects are only significant for small S, and including these effects makes the expression much uglier. If we *include* these effects, we obtain the fitted curves in Figure 12, which are accurate fits even for small S. The slopes of these fitted curves, along with the

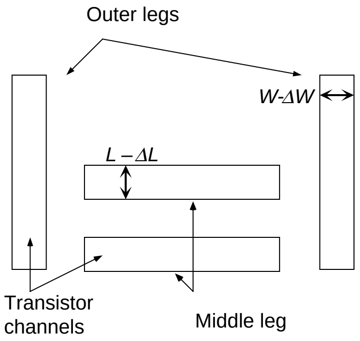


FIGURE 13 Geometrical model of length and width reductions in bump circuits. Rectangles show transistor channels.

drawn dimensions of the transistors, lets us conclude that

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

The data indicate an effect that is much larger than we would expect from the half micron scale for the size of the bird beaks and lateral diffusion. In four of the bump circuits, the small transistor dimensions are only 3 µm, so Equation 16 is a substantial effect. The circuits that gave us the data in Figure 12 confound possible length and width variations, because we varied both length and width equally in the layout. As a result, we cannot say what are the relative influences of width and length effects.

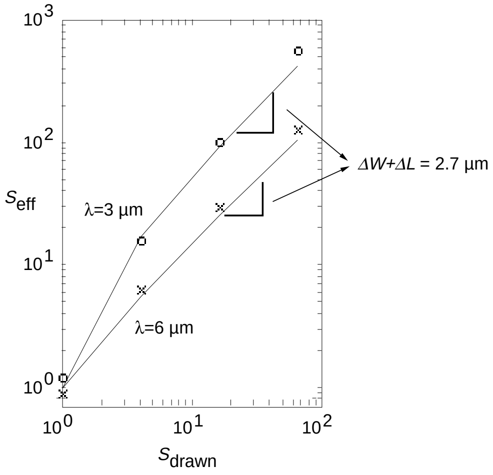


FIGURE 12 The relation between *S* intended by the layout geometry and the measured *S*. These data were derived from the curves in Figure 5. The top data are from the bump circuits with minimum dimension 3 µm, while the bottom data are from the circuits with minimum dimension 6 µm. The solid lines represent a fit to these data assuming that the discrepancy is due to a width reduction *ΔW* and a length reduction *ΔL* in all the transistors. The result of the fit show that *ΔW+ΔL* is approximately 2.7µm.

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

TABLE 2 Threshold voltage behavior vs. transistor size, from MOSIS parameters.

MOSIS test parameters

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

The threshold voltage measurements are shown in Table 2 for the run that produced the chip that resulted in the data in Figure 5 and Figure 12. We can see that the threshold voltage for the narrowest transistor (3 µm wide by 2 µm long) is 94 mV higher than the threshold voltage for a wide transistor of the same length (18 µm wide by 2 µm long). A larger threshold voltage translates to a weaker transistor, for subthreshold operation. A threshold shift of 94 mV means that the current, at a given subthreshold gate bias voltage, is about

$\exp(\kappa(94\text{mV})/V_T) \approx 15$ times smaller. Hence, the narrow transistor is about 15 times weaker than the layout would predict given identical threshold voltages. If we wanted to account for the threshold voltage shift by assuming a geometrical reduction in the width of the transistor, we would need to assume a width reduction of 2.8 µm, because then the effective width would be 0.2 µm, 15 times smaller than the drawn width of 3 µm.

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

Our own transistor measurements

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

The results of these measurements are shown in Figure 14. The transistor strengths are roughly proportional to the *W/L* ratio. However, there is a systematic bias-level effect that is not explained by a constant geometrical factor. The effect is strongest in the subthreshold operating region and becomes smaller, approaching a constant level, for above-threshold operation. Comparing the effect for the 4 wide by 20 long transistor with the 20 wide by 20 long transistor, we see that the 4 wide transistor is about 0.5 times a strong as drawn geometry would predict, for subthreshold operation. For above-threshold operation, the effect reduces to a factor of about 0.8. The width reduction needed to explain this effect makes a corresponding variation from 2 µm in subthreshold to about 0.8 µm above-threshold. In contrast, by comparing the 30 wide by 6 long transistor with the 20 wide by 20 long transistor, we see again that the length effect is smaller and much more constant.

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

We can summarize the transistor strength effects in terms of a model where the subthreshold channel potential has a bowl-like shape across the channel. The radius of curvature of the edges of the bowl, in subthreshold, is on the order of a micron by a thermal voltage. In other words, about a micron from the channel edge the channel poten-

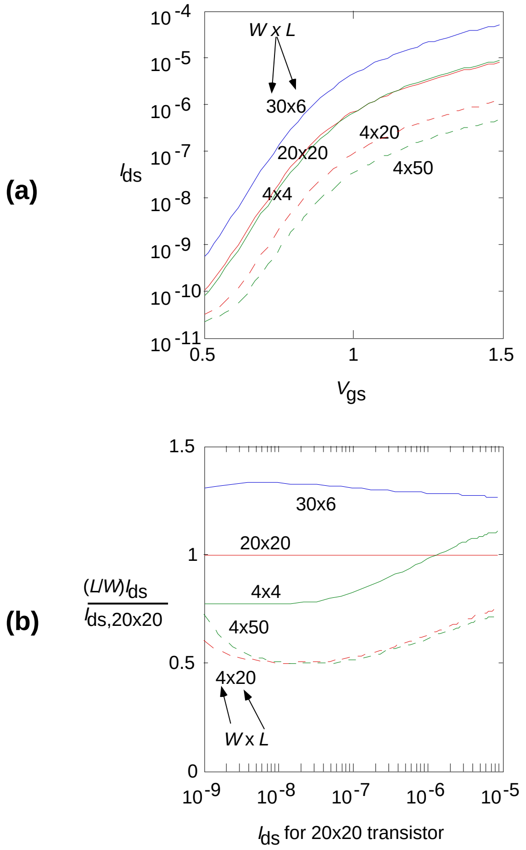


FIGURE 14 Effect of transistor sizing on transistor strength. **(a)** shows the drain current as a function of the gate-source voltage, for various drawn transistor sizes. **(b)** shows the ratio of the transistor current, normalized by the drawn *W/L* ratio for the transistor, to the current in the 20 x 20 transistor. These curves are plotted vs. the current in the 20 x 20 transistor. If the transistor current scaled exactly like the drawn transistor *W/L* ratio, then all the curves in (b) would be identical, i.e., unity. For all these data, the source of the transistor was held at the bulk potential (grounded).

tial is about a thermal voltage higher than the middle of the channel. For wide channels, the effect appears as an effective width reduction. For narrow channels, the effect appears as an increase in the threshold voltage, or a decrease in I_0.

SUMMARY

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

ACKNOWLEDGEMENTS

John Harris had the idea of calling these circuits “bump” circuits. Dave Kewley, Dick Lyon and Carver Mead contributed significantly to this report, in terms of a critical analysis of the transistor strength discrepancy. Dick Lyon analyzed the bump amplifier configurations and realized they would be useful in filter circuits. Doug Kerns provided contextual information about commercial split-gate mixer circuits. Dave Kewley showed me his own measurements of transistor width and length. Dick Lyon and Dave Kewley provided critical comments on earlier versions of this report. Bhusan Gupta supplied the test chip used to compare different sized transistors.

REFERENCES

1. L.A. Akers and J.J. Sanchez, “Threshold voltage models of short, narrow and small geometry MOSFETS: a review,” *Solid State Electronics*, vol. 25, pp. 621–641, 1982.
2. M.H. Cohen and A.G. Andreou, “MOS circuit for nonlinear Hebbian learning,” *Electronics Letters,* vol 28, pp. 809–810, 1992.
3. T. Delbrück, “Analog VLSI predictive visual motion processing,” *IEEE Trans. on Neural Networks*, in press, 1993.
4. J. Hertz, A. Krogh, R.G. Palmer, *Introduction to the theory of Neural Computation,* Boston, MA: Addison Wesley, 1991.
5. D.A. Kerns and L. Watts, personal communication, 1992.
6. D. Kewley, personal communication, 1992.
7. D. Kirk, *Accurate and precise computation using analog VLSI, with applications to computer graphics and neural networks,* Ph.D. Thesis, California Institute of Technology, Caltech-CS-TR-93-??, 1993.
8. D. B. Kirk, D. Kerns, K. Fleischer, A.H. Barr, “Analog VLSI implementation of multi-dimensional gradient descent,” *Neural Information Processing Systems,* vol. 5, 1993 (in press).
9. E.H. Li, K.M. Hong, Y.C. Chen, K.Y. Chan, “The narrow-channel effect in MOSFETS with semi-recessed oxide structures,” *IEEE Trans. on Electron Devices,* vol. 37, pp. 692–701, 1990.
10. S.C. Liu and J.G. Harris, “Edge-Enhancing Resistive Fuse,” SPIE Technical Symposia on Optical Engineering and Photonics in Aerospace Sensing, Orlando, FL, vol. 1473, pp. 185-193., 1991.
11. D. Lyon, personal communication, 1992.
12. M.A. Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function,* Ph.D. Thesis, California Inst. of Tech., Dept. of Computation and Neural Systems, Pasadena CA, 91125, 1992.
13. C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison--Wesley, 1989.
14. M. Sivilotti, *Wiring Considerations in Analog VLSI Systems, with Application to Field-Programmable Networks*, Ph.D. Thesis, California Institute of Technology, Dept. of Computer Science, Pasadena, CA, June 1991.
15. R. Tawel, “An analog processor for image compression,” *IEEE J. Neural Networks, Special Hardware Issue* (in press).
16. P.P. Wang, “Device characteristics of short-channel and narrow-width MOSFETS,” *IEEE Trans. on Electron Devices,* vol. ED-25, pp. 779–786, 1978.