Current-Mode Subthreshold MOS Circuits for Analog VLSI Neural Systems

Andreas G. Andreou, *Member, IEEE*, Kwabena A. Boahen, *Member, IEEE*, Philippe O. Pouliquen, Aleksandra Pavasović, Robert E. Jenkins, *Member, IEEE*, and Kim Strohbehn, *Member, IEEE*

***Abstract*—This paper presents an overview of the *current-mode* approach for designing analog VLSI neural systems in subthreshold CMOS technology. Emphasis is given to design techniques at the device level using the current-controlled current conveyor and the translinear principle. Circuits for associative memory and silicon retina systems are used as examples. Where appropriate, we draw analogies to the physics and organization of information processing in the nervous system.**

I. Introduction

BIOLOGICAL information processing systems are compact, energy efficient, and excel at sensory perception and motor control—areas in which modern digital computers falter. Thus, it is not surprising that computer scientists and engineers in their quest to endow present-day computers with perceptual processing capabilities are studying the organization and physics of computation in the nervous system. Cowan and Sharpe give an excellent account of important developments in this field over the last 30 years [1].

Such studies yield models of computation that can be tested in several ways. However, Mead eloquently and convincingly argues in favor of a synthetic approach using silicon VLSI technology and analog circuits [2]. His group [3] and colleagues at Caltech [4] have demonstrated how an "opportunistic approach" using "garden variety" CMOS technology can result in effective electronic systems that solve difficult problems in computer vision.

We believe that just as the study of biological neural networks can be done at different levels [5], [6], the engineering (synthesis) of neurally inspired computing hardware requires a similar approach. In this paper we address the synthesis of analog VLSI neural systems, focusing on the *circuit level*. The physics and properties of silicon CMOS technology constrain our design methodology from the bottom end; the nature of the problem that we are solving imposes additional constraints from the top. We will show that an opportunistic approach at this level can yield circuits having complex functionality and small area with designs at the transistor level.

The use of complex building blocks such as transconductance amplifiers and of differential voltage signals is avoided. Thus, we emphasize the necessity for a design methodology that is free from previous notions of analog integrated circuit design and is tailored toward micropower analog collective computational systems. The design methodology presented here is based on current-mode (CM) subthreshold MOS circuits and is inspired by the organizational principles of biological neural circuits.

The paper is organized into five sections. Section II introduces the silicon CMOS technology. A subthreshold MOS model suitable for hand calculations is described along with three different bias arrangements for the device. Next we present data from fabricated chips that display the matching properties of MOS transistors in typical CMOS technologies. The current-mode approach in circuit design is described in Section III, where a current-controlled current conveyor and translinear circuits are introduced. In Section IV, we show how these can be employed in systems for associative processing and computer vision. A discussion of our design methodology and of the way in which it relates to actual biological microcircuits concludes the paper.

Manuscript received September 5, 1990; revised December 11, 1990. This work was supported in part by the Independent Research and Development program of Johns Hopkins University, Applied Physics Laboratory. Support was also received from the NSF in the form of a Research Initiation Award (MIP-9010364) to A. G. Andreou.

A. G. Andreou, K. A. Boahen, P. O. Pouliquen and A. Pavasović are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218.

R. E. Jenkins and K. Strohbehn are with the Applied Physics Laboratory, John Hopkins University, Laurel, MD 20707.

IEEE Log Number 9042381.

II. Subthreshold MOS Device Operation

Subthreshold CMOS technology has long been recognized as the technology of choice for implementing micropower digital and analog LSI circuits [7]. It offers the same advantages for the implementation of synthetic neural systems: high integration density, low power dissipation, and useful parasitic bipolar devices. In addition, it is easily accessible to engineers and scientists through silicon foundry services such as MOSIS [8].

A. Device Model

In this section, we present background material necessary for the understanding of the circuits described in later sections of the paper. For an n-type MOS transistor, the subthreshold current [3], [9] is given by

$$I_{ds} = I_0 e^{(1-\kappa)V_{bs}/V_T} e^{\kappa V_{gs}/V_T} (1 - e^{-V_{ds}/V_T} + V_{ds}/V_0) \quad (1)$$

where V_{gs} is the gate-to-source potential, V_{ds} is the drain-to-source potential, V_{bs} is the substrate (or well)-to-source potential (body effect), I_0 is the zero-bias current for the given device, $V_T = kT/q$ is 26 mV at room temperature, V_0 is the Early voltage (which is proportional to the channel length), and κ measures the effectiveness of the gate potential in controlling the channel current. All potentials are measured with respect to the source potential and are sign reversed for a p-type device.

Typical parameters for minimum-size devices (4 μm × 4 μm) fabricated in a standard digital 2 μm n-well process are $I_0 = 0.72 \times 10^{-18}$ A, $\kappa = 0.75$, and $V_0 = 15.0$ V. Thus the current

changes by a factor of 10 for an 80 mV change in V_{gs} or a 240 mV change in V_{bs} (up to about 100 nA, which is the limit of the subthreshold region). The model matches experimental data reasonably well [10] and is adequate for design simulations.

For devices in saturation (that is $V_{ds} \geq 4V_T$), neglecting the Early effect and the body effect,

$$I_{ds} = I_0 e^{\kappa V_{gs}/V_T}. \tag{2}$$

This simplified equation, containing only the dependence on V_{gs}, is sufficient for most circuit designs. On the other hand, (1) shows an explicit V_{bs} dependence that underlies the role of the substrate as another terminal which can control the drain-source current.

In the saturation region, the MOS transistor is a voltage-controlled current source with transconductance

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{\kappa I_{ds}}{V_T} \tag{3}$$

output conductance

$$g_{ds\text{sat}} = \frac{\partial I_{ds}}{\partial V_{ds}} = \frac{I_{ds}}{V_0}. \tag{4}$$

The MOS transistor itself can perform several useful circuit functions:

- In the common-source mode, it is an inverting amplifier with high voltage gain: $A = g_m/g_{ds\text{sat}}$ (see Fig. 1(a)). For our process $A = \kappa V_0/V_T \approx 430$.
- In the common-drain mode, it is a voltage follower with low output resistance: $1/g_m$ (see Fig. 1(b)).
- In the common-gate mode, it is a current buffer with low output conductance: $g_{ds\text{sat}}$ (see Fig. 1(c)).

B. Device Matching

Traditional analog integrated circuits depend on good matching between components. Therefore, large devices operating above threshold are used to reduce mismatch to very low levels. In contrast, to achieve VLSI densities we must employ transistors that have small geometries, typically ($4\ \mu m \times 4\ \mu m$). This together with their operation in the subthreshold region, makes the drain current strongly dependent on variations of fabrication process parameters, in particular I_0. Characterization of the fabrication process and the matching properties of the basic devices is thus of paramount importance because it provides the necessary information for designing working systems.

In the following section we summarize our work in this area, and present experimental data applicable to n-well and p-well standard industrial processes. We have characterized large, dense transistor arrays in the subthreshold region of operation [11], [12]. More than 150 000 transistors have been tested using an automated data acquisition system. Measurements exposed three factors affecting matching: edge effects, striation effects, and random variations (Fig. 2).

The edge effect manifests itself as a dependence of the transistor current on its position with respect to the surrounding structures. N-type transistors surrounded by other n-type transistors have a larger drain current than identically designed and biased transistors on the edges of the arrays. The opposite is true for p-type transistors. Variations in transistor currents caused by the edge effect typically range from 5% to 15% for n-type transistors and from 20% to 50% for p-type transistors.

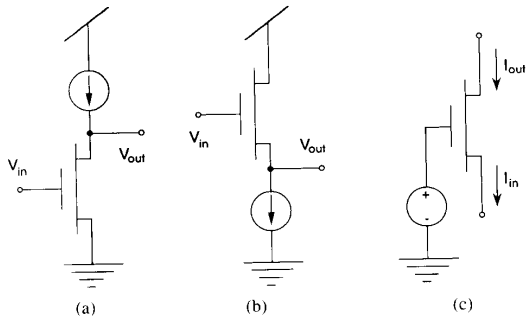


Fig. 1. The three modes of operation of a transistor: (a) high gain inverting amplifier (common source), (b) voltage follower (common drain), and (c) current buffer (common gate).

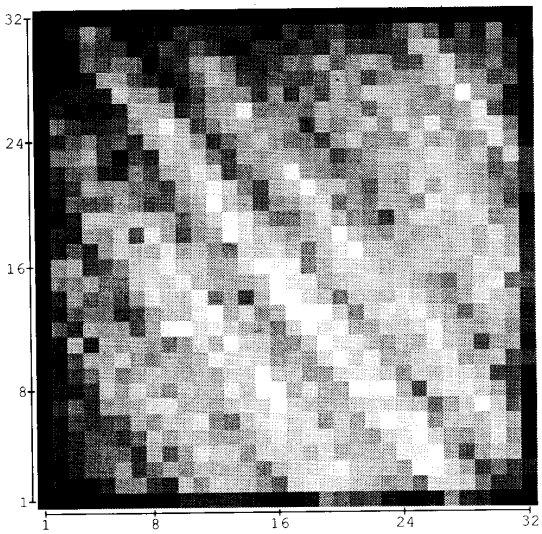


Fig. 2. Density plot of currents in a 32 × 32 array of 16 μm × 16 μm n-channel transistors fabricated in a p-well process. Each transistor is represented by a square; the current is coded by the shade of gray where the minimum and maximum values are represented by black and white, respectively.

The striation effect exhibits itself as a sinusoidal spatial variation in transistor current. The amplitude is about 30% of the average current and the spatial period varies slowly from 100 μm to 200 μm for p-type devices in the p-well process.

The random variation follows a Gaussian distribution (Fig. 3). Fig. 4 shows the dependence of its normalized standard deviation on transistor size. Each data point represents measurements from approximately 1000 transistors. The standard deviation of the current is proportional to the current and inversely proportional to the length of square devices (i.e., devices whose length and width are the same):

$$\sigma_I = \sigma_0 \frac{I_{ds}}{L} \tag{5}$$

where σ_0 is the proportionality constant for the given device type, I_{ds} is the nominal device current, and L is the length of the device (square geometry).

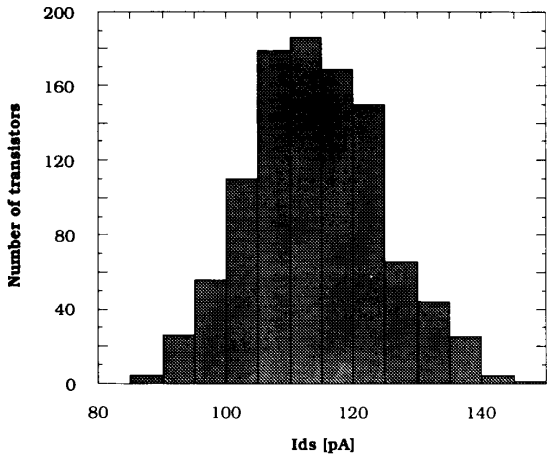


Fig. 3. Histogram for the subthreshold current of 1024 n-channel MOSFET's fabricated in an n-well process.

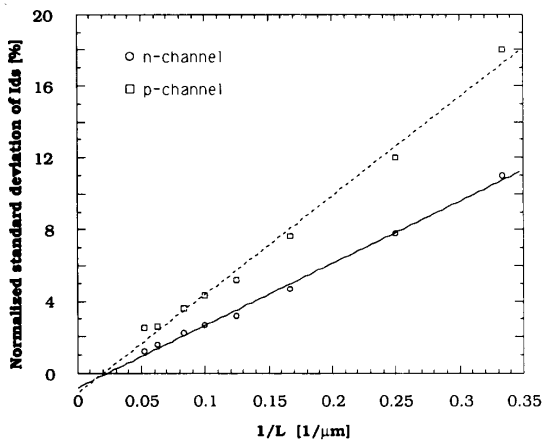


Fig. 4. Dependence of normalized standard deviation of subthreshold drain current on transistor size. They all have square geometries with channel length L and were fabricated in the same process as those of Fig. 2.

Edge effects and striation effects present significant problems in large-scale analog computational systems, since they cannot be reduced by increasing transistor area. The striation effect is especially damaging, since the orientation of the striations is not known *a priori*. The edge effect can be cancelled by symmetrical placement of transistors.

At the system level, random variations are addressed by distributing computations over a large number of elements. We can reduce the normalized standard deviation of the currents from matched current sources by a factor of $\sqrt{N}$ by using N small devices for each current source. Furthermore, we can position our N small devices in a way that minimizes the edge and striation effects. Finally, we can take advantage of this distribution of the signal over many devices to implement parallel computation. Thus, our systems are carefully designed to simultaneously perform parallel distributed processing and reduce matching problems.

At the circuit level, we bias the transistors at a constant I_{ds} rather than a constant V_{gs} [7]. This is so because the variability of I_0 is much larger than the variability of κ [11]. Biasing at a constant current and thinking in terms of *current-domain* [10] signals is the essence of the *current-mode* approach in circuit design.

III. Circuit Techniques: The Current-Mode Approach

There is no widely accepted definition for the current-mode approach in circuit design [13]. By current-mode, we shall refer to circuits that use both currents and voltages (like every electronic circuit) but where signals are represented as currents and voltages play only an incidental role.

Translinear circuits [14] form a large subclass of current-mode circuits, [13, ch. 2]. Designs based on the translinear principle enable complex computations to be performed in the analog domain without the explicit use of differential voltage signals.

A. The Translinear Principle

The translinear principle can be stated as follows:

> In a closed loop containing an equal number of oppositely connected translinear elements, the product of the current densities in the elements connected in one direction is equal to the corresponding product for elements connected in the opposite direction.

Using this principle, computationally powerful current-mode circuits can be synthesized. A translinear element is simply a physical device with a linear relationship between transconductance and current. Traditionally, translinear circuits have been built using bipolar transistors. However, the MOS transistor is also a translinear element when operated in the subthreshold region (see (3)); the absence of a base current makes it an ideal one! The resulting translinear circuits can perform both linear and nonlinear operations on current inputs, including products, quotients, and power terms with fixed exponents [14]. The current mirror is a trivial example of a translinear circuit, and our silicon optical motion detector [15] uses the translinear multiplier/divider in Fig. 5 to compute the correlation I_Z between two signals, I_X and I_Y. By applying the translinear principle around the loop indicated by the arrows, we find

$$I_W I_Z = I_X I_Y \quad \text{or} \quad I_Z = \frac{I_X I_Y}{I_W} \tag{6}$$

where I_W normalizes the result.

The above relation can also be derived by summing the voltages around the loop GND-A-B-C-GND (conservation of energy):

$$V_1 + V_2 + V_3 + V_4 = 0.$$

Replacing the gate-source voltages for M_1, M_2, M_3, and M_4 with their respective drain-source currents through (2) and assuming the same κ and I_0 for all devices, we obtain

$$\frac{V_T}{\kappa} \ln\left(\frac{I_X}{I_0}\right) + \frac{V_T}{\kappa} \ln\left(\frac{I_Y}{I_0}\right) - \frac{V_T}{\kappa} \ln\left(\frac{I_W}{I_0}\right) - \frac{V_T}{\kappa} \ln\left(\frac{I_Z}{I_0}\right) = 0$$

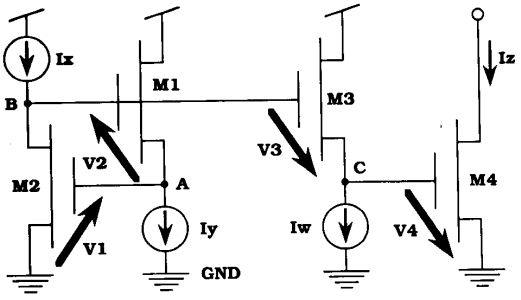


Fig. 5. A simple translinear circuit that performs a normal product computation.

or

$$\ln\left(\frac{I_X I_Y}{I_0}\right) = \ln\left(\frac{I_Z I_W}{I_0}\right)$$

from which (6) readily follows.

Yet another way of looking at the function of the circuit in Fig. 5 is as a log–antilog block. Transistor M_1 does the log-ing and M_4 the antilog-ing. The other two transistors serve as voltage level shifters; this is equivalent to *normalization* in the current domain.

The translinear property of subthreshold MOS transistors is also useful for analyzing the dynamics (temporal response) of large-scale collective computational systems [16].

B. The Current Conveyor

Smith and Sedra introduced the concept of a current conveyor as an ingenious hybrid voltage/current three-port device. It is a versatile building block for analog signal processing applications designed to replace the operational amplifier [17].

Although the original implementation used five bipolar transistors, the current conveyor can perhaps be most easily explained by considering a single device. A transistor can transfer a current from a high-conductance to a low-conductance node (see Fig. 1(c)) or a voltage from a high-impedance to a low-impedance node (see Fig. 1(b)). These two characteristics can be exploited simultaneously (see Fig. 6(a)) such that the device acts as a voltage follower (node X will follow voltage changes at node Y) and conveys the current at X to the low-conductance node Z. This dual role, obtained in this case using only a single transistor, captures the essence of the current conveyor.

Current conveyors can easily interact with one another as shown in Fig. 6(b). Here, node X follows the greatest input voltage V_{Yi}, turning off all other current conveyors. The tail current I_X is then entirely conveyed to the output node Z_i identifying the ith input voltage as being the greatest.

A two-transistor current-controlled current conveyor is shown in Fig. 7(a). The authors first proposed its use in a scheme for two-way communication over a single line [10], [18], [20] (see Section IV) and as a nonthresholding neuron in an earlier associative memory design [19]. As with the standard current conveyor [17], the current-controlled current conveyor has a communication node X, a control node Y, and a supply node Z. However, the potential at node X is determined by the control current I_Y. Thus, the current-controlled current conveyor's operation can be described by two simple relationships:

$$I_Z = I_X \qquad \mathcal{I}(V_X) = I_Y \tag{7}$$

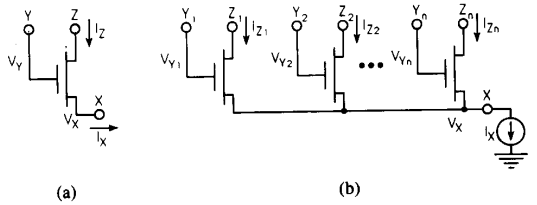


Fig. 6. (a) A single transistor current conveyor circuit. In this configuration, Y is a voltage input node, Z is a current output node, and X is a hybrid voltage/current node. (b) N single-transistor current conveyors interact through a single line to perform a voltage input winner-takes-all function. The two-transistor version of this circuit is the well-known differential pair.

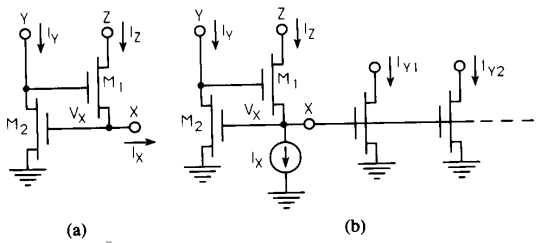


Fig. 7. (a) The two-transistor current-controlled current conveyor. (b) Application of the current-controlled current conveyor to fan-out copies of the current I_Y.

where the $\mathcal{I}(\,)$ is the function that relates the output V_X to the control current I_Y. For subthreshold operation this function is the *logarithm* but it continuously transforms into a *square root* as the current increases and the device operates above threshold.

When node Z is connected to the power supply, the two-transistor current-controlled current conveyor is the simplest form of a *trans-resistance* amplifier—a single transistor is the simplest form of a *trans-conductance* amplifier.

At the system level it is important to note that, by using transistors identical to M_2, we can mirror the control current I_Y by collecting their gates to node X (see Fig. 7(b)). This facilitates *fan-out*, distribution of an output signal (current) in the form of a voltage, as well as *fan-in*, the aggregation of input signals which are currents; both can be done simultaneously over a single physical line.

IV. System Applications

Let us now consider the concepts and circuits described earlier in the context of larger systems—associative memories and silicon vision chips with a focus at local circuits.

The bidirectional associative memory (BAM) model was chosen as a test-bed for silicon associative processing ideas. It is a good example of an analog VLSI system whose organization can be "loosely" related to some neural computation. On the other hand, the model for outer-plexiform retinal processing is truly biologically inspired. This distinction is necessary because the definitions for synthetic "neurons" and "synapses" in the two systems refer to very different structures.

The discussion of the BAM circuits is more quantitative, emphasizing the circuit properties of the current-controlled current conveyors. The discussion of the silicon vision system, provides on the other hand, a more "gentle" transition from circuits to systems.

A. Associative Memory

Two different BAM architectures have been implemented. Both use unary signal representations ("grandmother cells") and have been discussed in detail elsewhere [18], [19]. The circuit techniques that follow have been employed in the second generation, four-layer BAM architecture [18] to realize programmable reciprocal connections and lateral inhibition at nearly static RAM densities. They also serve as examples of three types of *interactions* between assemblies of current-controlled current conveyors.

1) Bidirectional Junction: In a neuronal circuit, the interaction between neurons is mediated by specialized structures called synapses [21]. A neuron receives its inputs from other neurons through synaptic junctions which may have different efficacies. In a typical VLSI implementation the synapses are implemented as a two-dimensional array with neurons on the periphery [18], [19]. This is because $O(N^2)$ synapses are required in a network with N neurons.

The two transistor circuit[1] shown in Fig. 8(a) allows two-way interaction between synthetic neurons (current-controlled current conveyors) connected to nodes n_1 and n_2 (as in Fig. 8(b)). Each node receives a voltage input and produces a current at the other node; each transistor behaves like a synaptic junction. Interaction is turned on by grounding S, turned off by pulling S up to V_{dd}.

If the interaction is turned off (S is at V_{dd}) and V_{n1} is greater than V_{n2}, M_1 is shut off, and M_2 sources the current:

$$I_{\text{off}} = I_0 e^{\kappa(V_{n_1} - V_{n_2})/V_T} \equiv I_0 \epsilon.$$

Thus, the current leakage I_{off} is proportional to ϵ, the dynamic range of the current signals. For signals in the range of 100 pA to 100 nA, for instance, $\epsilon = 10^3$, and $I_{\text{off}} = 1$ fA in the worst case.

In our bidirectional communication scheme, synthetic neuron B (on the right in Fig. 8(b)) receives the current signal I_{12} from neuron A as $\hat{I}_{12}$. Simultaneously, synthetic neuron A receives the current signal I_{21} from neuron B as $\hat{I}_{21}$. Small changes in the current signal I_{12} causes slight changes in the received current $\hat{I}_{21}$. These variations are related by

$$\frac{\Delta \hat{I}_{21}}{\Delta I_{12}} = \frac{g_{d\text{sat}}}{g_m} = \frac{1}{A}\frac{\hat{I}_{21}}{I_{12}}$$

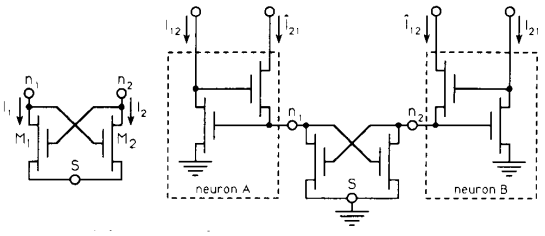
as before. Hence, bidirectional communication is possible with less than -50 dB cross-talk.

If the two transistors forming the bidirectional junction are fabricated in an isolated well, the well voltage can be used to modulate the interaction between neurons. The circuit of Fig. 8(b) was fabricated in a standard, 2 μm CMOS process using 4 μm $\times$ 4 μm transistors. Experimental data, including the modulation of the well voltage (V_{bs}), are shown in Fig. 9.

Well modulation has also been successfully employed in a current-mode implementation of the Herault–Jutten neural network model [22]. This system separates, in real time, unknown mixed signals such as speech, through a Hebbian-like unsupervised learning algorithm.

The bidirectional junction communication scheme is by itself also a two-input, two-output translinear circuit!

[1]In one of our previous papers [10], this was referred to as a reciprocal junction. For reasons that will become clear in the discussion of the retina circuits, this naming convention is not used anymore.



(a) (b)

Fig. 8. (a) Bidirectional junction. A voltage signal applied to node n_1 is converted to a current signal at node n_2, and vice versa. (b) Two synthetic neurons communicate through a bidirectional junction. They send current signals I_{12} and I_{21} and receive signals $\hat{I}_{21}$ and $\hat{I}_{12}$.

2) Winner-Takes-All: In the current-mode winner-takes-all (WTA) circuit, shown in Fig. 10, N current conveyors compete for current supplied to a common line. The current, I_x, is steered to the output of the conveyor with the largest input current; all other outputs are zero. This is an adaptation of Lazzaro and Mead's original circuit [23] to provide current outputs. Input currents are supplied to the control nodes Y, output currents are obtained from the supply nodes Z, and the communication nodes X are connected together.

Each conveyor sees a voltage V_X at its communication node. Consequently, for each conveyor, if $I_{Yi} < \mathcal{I}(V_X)$ (see (7)), M_2 enters the linear region ($V_{ds} < 4V_T$), turning M_1 off. Otherwise, M_1 adjusts V_X to set $\mathcal{I}(V_X) = I_Y$. Thus the conveyor with the largest input sets the voltage on the common line and conveys the current I_X to its supply node.

When two input currents are very similar (for instance $I_{Y1} \approx I_{Y2}$), the conversion from input to output is exponential. In this case, M_2 stays in saturation and M_1 remains on. I_{Y2} develops a voltage signal across M_2's drain conductance $g_{d\text{sat}}$, which is converted exponentially to current by M_1, and similarly for I_{Y1}. The sum of the output currents ($I_{Z1} + I_{Z2}$) is equal to I_X. The differential gain for small signals is

$$\frac{\Delta I_Z}{\Delta I_Y} = \frac{g_m}{g_{d\text{sat}}} = A\frac{I_Z}{I_Y}$$

where $I_{Y1} = I_Y + \Delta I_Y$, $I_{Y2} = I_Y - \Delta I_Y$, $I_{Z1} = I_Z + \Delta I_Z$, and $I_{Z2} = I_Z - \Delta I_Z$. Hence, the small-signal differential gain is 430 for normalized inputs and outputs—a 1% input difference produces a voltage difference of 0.15 V, so the corresponding outputs differ by a factor of 75.

3) Pyramidal Neurons: This local circuit arrangement from the middle neuron layer of the BAM is chosen as an example to address two issues. First, we show how complementary current-controlled current conveyors (employing n-type and p-type transistors) can be connected; second, we discuss a metastability problem associated with such connection and our solution to it.

Cortical pyramidal neurons are arranged in layers with projections from one layer to the next. Their apical dendrites receive incoming signals while their axons carry outgoing signals. Pyramidal cell axons have collaterals that branch back and contact basal dendrites of neighboring cells; these lateral interactions are mainly inhibitory (see Fig. 11(a)).

Fig. 11(b) shows our pyramidal cell circuit. The circuit consists of an n-type synthetic neuron (M_1, M_2), a p-type WTA cell (M_3, M_4), and an extra transistor (M_5) to keep the circuit

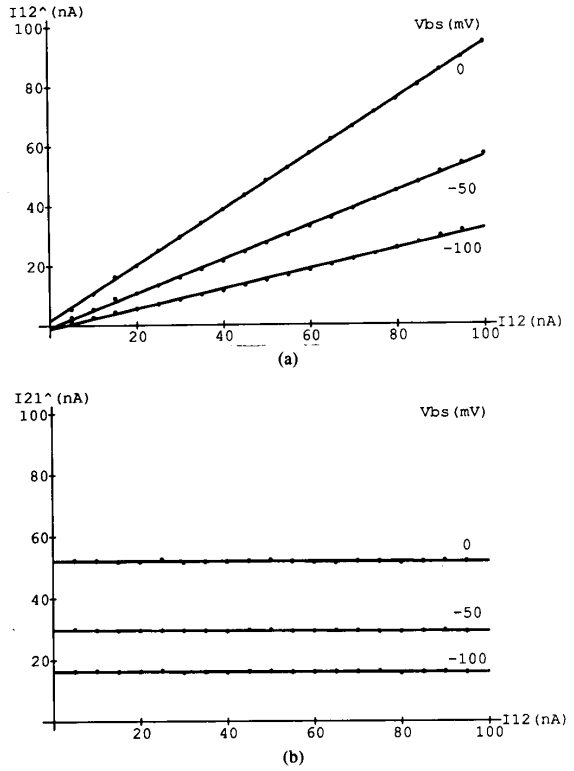


Fig. 9. (a) Experimental data for the bidirectional communication scheme obtained by stepping I_{12} from 5 nA to 100 nA, with I_{21} held at 50 nA, for V_{bs} values: 0, -50 mV, and -100 mV. $\hat{I}_{12}$ is directly proportional to I_{12}; the slopes are 0.93, 0.57, and 0.33. (b) Simultaneously, $\hat{I}_{21}$ remained constant; the slopes are less than -0.004. $\hat{I}_{21}/I_{21}$ equals 1.04, 0.59, and 0.33, showing that the modulation is symmetric.

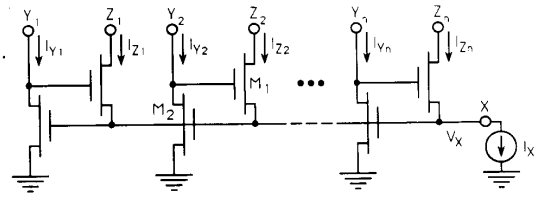


Fig. 10. Winner-takes-all (WTA) circuit. N current signals are applied to N current-controlled current conveyors. The conveyor that has the largest input conveys the current I_X supplied to the common line; the other conveyors have zero output.

in the proper mode of operation when it is first turned on. This is necessary, because when an n-type and a p-type current-controlled current conveyor feed each other, their current buffering devices (M_1 and M_3) can act as common-source amplifiers. This creates an undesirable positive feedback loop causing these devices to enter the linear region, thus driving X_V and X_L to the power supply rails. M_5 prevents this by keeping the conveyor's buffer (M_1) in saturation. Although M_5 *shunts* part of the incoming current, a fixed fraction of the input is always passed to the WTA cell [18].

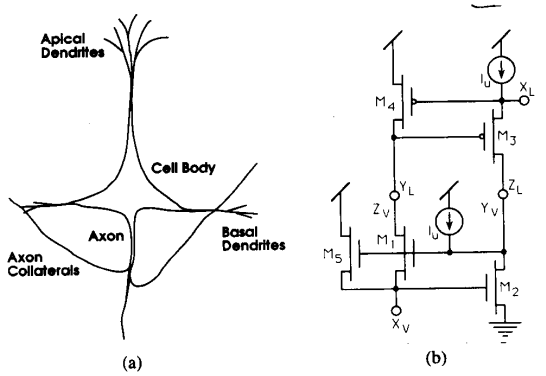


Fig. 11. Pyramidal neuron circuit of the bidirectional associative memory. Demonstrates how current-controlled current conveyors of different transistor type can interact vertically through their control and output nodes.

The circuit has two communication nodes, X_V and X_L. X_V is analogous to both the apical dendrites and the axon of a cortical pyramidal neuron in that it carries both the incoming and outgoing excitatory signals. X_L mediates lateral inhibitory interactions, mimicking the cortical cell's basal dendrites and recurrent axon collaterals.

A pyramidal neuron contributes a unit current I_U to the WTA competition and also has a quiescent current of I_U. Therefore, its outgoing signal $\mathcal{G}(V_{X_V})$ is I_U in the quiescent state but increases to $(m + 1)I_U$ (where m is the number of competing cells) when it is winning (that is, when the input current at X_V exceeds all other input currents).

B. Silicon Vision

Our device-level design methodology and current-domain signal representations can also be used for implementing analog VLSI silicon retinas [3, ch.15]. However, unlike the circuits for associative memory, here there is a direct analogy between the actual biological circuits and the silicon counterparts.

Fig. 12(a) illustrates interactions [24] between cells in the outer-plexiform layer of a retina. No attempt is being made here to reverse-engineer the retina of any particular species; the turtle retina, though, has similar connections in the outer-plexiform. This simple structure, consisting of just two types of neurons, gives the well-known center/surround receptive field. In engineering terms, this system can be thought as a linear *spatial band-pass filter*, whose impulse response is the difference of two Gaussians (DOG) or the Laplacian [6]. However, both the biological and the silicon systems are nonlinear and do much more than that!

The photoreceptors are activated by light; they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells, in turn, suppress the activity of the receptors through inhibitory chemical synapses. The receptors and horizontal cells are electrically coupled to their neighbors by gap junctions. These electrical synapses allow ionic currents to flow from one cell to another, and are characterized by a certain conductance per unit area.

Mapping of outer-plexiform retina processing onto silicon can be done in several ways [25]. In a design that does not employ differential voltage signals [20], chemical synapses may be im-

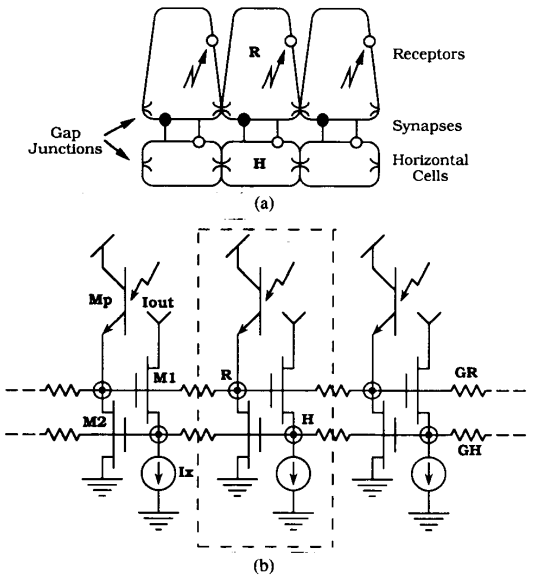


Fig. 12. (a) Illustration of cells and synapses in the outer-plexiform layer of a retina. Light excites the photoreceptors. (b) Current-mode electronic implementation of outer-plexiform retinal processing.

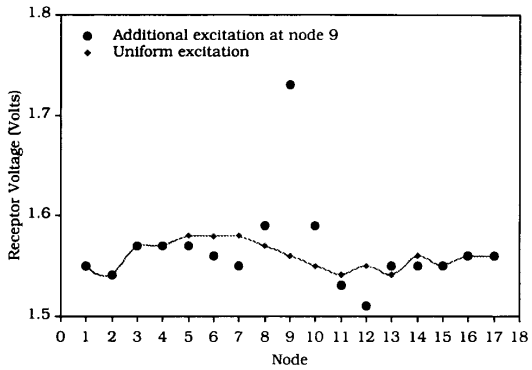


Fig. 13. Experimental results from one-dimensional 17 node outer-plexiform silicon retina. This circuit has been prototyped out of large MOS transistor arrays ($W:L = 800:4\ \mu m$). Thus the coupling conductances and the currents are appropriately scaled, but the circuit still operates in subthreshold ($GH = 2$ mS and $GR = 40\ \mu S$). The bias currents I_X are set to a nominal value of 30 μA and under uniform illumination the receptors are supplied with 10 μA currents. The two sets of data correspond to "uniform" illumination (diamonds) and "delta function" excitation (filled dots). In the latter case, receptor node 9 is excited with an additional current of 20 μA.

plemented using nonlinear transconductances while gap junctions are realized using resistances (Fig. 12). Nodes (equipotential regions) in the top layer correspond to receptors R while those in the lower layer represent the horizontal cells H. The two layers are coupled vertically using our two-transistor current-controlled current conveyor circuit and laterally using two resistive networks. A parasitic bipolar transistor is used to transduce light into current. It sources current to the receptor nodes while M_2 sinks current from those nodes; these opposing effects correspond to excitation and inhibition. M_1 sources current (excites) the horizontal cell nodes. The bias current I_X at the source of device M_1 sets its transconductance. For subthreshold operation, the voltages encode photocurrents logarithmically, allowing a large dynamic range.

Fig. 13 shows experimental data from a fabricated circuit that verifies its function and robustness in individual transistor mismatch. The appearance of the on-center/off-surround organization is evident. The activation neighborhood can be controlled by adjusting the space constants to the two resistive networks (conductance of gap junctions). When $GR = 0$ (absence of gap junctions between receptors), the system looses its ability to develop a truly "Mexican-hat" response and has a local winner-takes-all behavior, i.e., activation of a single node, with a small neighborhood of inhibition.

This system is used as a front end in a number of different applications. Depending on the application, the output of the circuit can be taken from different points. For example, a current-output signal is provided at the drain of device M_1. This encodes the difference between the activation in the two resistive networks and is independent of the light intensity. The output of the circuit can also be taken as the voltage at the receptor terminals; this, however, provides a signal that is dependent on the absolute light intensity.

If a current output signal is required, it may be advantageous

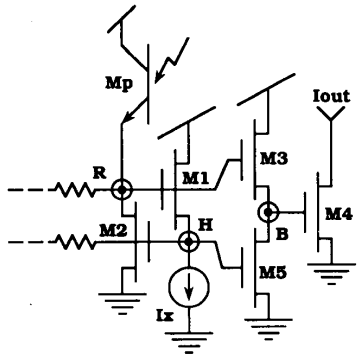


Fig. 14. Alternative output circuit that employs the translinear principle to convert a floating differential voltage signal to one referenced to GND. Transistor M_4 encodes the same current signal as M_1. This circuit is replicated in a two-dimensional array.

to use the alternative scheme (Fig. 14) that *couples* the outer-plexiform to other circuitry, such as inner-plexiform neurons. This provides for a better current signal and prevents undesirable interactions between two subsystems. It also demonstrates how the translinear principle is employed to convert a differential voltage signal V_{RH}, not referenced to the ground, to one that is referenced to ground, V_B. The operation of this circuit follows the discussion of the translinear multiplier/divider in Section III.

V. Discussion

Our design methodology is based on a few important principles. These are summarized below, where we also attempt to justify our earlier statement that these principles were drawn as analogies from known physics and organization of information processing in the nervous system.

- *Computations are carried out in the analog domain.* This results in simpler, more effective, and more compact functional blocks that interact comfortably with the "real world." Outer-plexiform retinal processing is a good example of how simple analog primitives can perform a massively parallel vision computation.
- *Systems are designed with power dissipation and area efficiency as prime engineering constraints.* The high level of parallelism attainable with VLSI analog computation imposes serious limits on the amount of power that each subcircuit can be allowed to dissipate. This is why we operate devices with current in the nA and even in the pA range. Subthreshold MOS currents are comparable to ionic currents in cell membranes, which range from a few picoamperes to a few microamperes. We also employ transistors of minimum size to achieve high integration density.
- *Random variations in transistor characteristics do not affect system performance.* Subthreshold MOS operation and small geometry devices result in poor matching in the characteristics of individual transistors. However, we are not concerned with accuracy or matching in the basic elements because our design methodology at both the circuit and the system level compensates for that. This is true for biological systems; they perform well despite the much poorer precision of their neurons and synaptic connections. The better matching in the basic silicon elements also implies that one needs to be *eclectic* at the local circuit level on how far the *neural paradigm* can be used.
- *Economize on interconnects*: Both at the local circuit level and at the architectural level, we are concerned with interconnects. Global communications are expensive not only in *space* but also in *energy*. Where possible, computation must be done with the smallest amount of global interactions. The biologically inspired outer-plexiform system is a good example of how complex parallel processing can be achieved through only nearest neighbor interactions.
- *Exploit the physics of basic devices:* In our silicon circuits, we have exploited the translinear property of subthreshold MOS transistor for synthesis and analysis. Unlike traditional analog design, optimizing circuits by careful sizing of the devices is not necessary. This is because the transconductance is determined by the current signals and does not depend on the geometry. Longer channels are used to avoid the Early effect and large devices are employed in a few critical parts of the system to improve accuracy in the computation (current matching), for example, devices M_2 in the WTA circuit. Translinear circuits are a classical example of how a rich form for circuit design emerges from the properties of the basic units. There is also a direct analogy between the device physics of voltage-gated ionic channels in excitable membranes and that of the MOS transistor in the subthreshold region [3], [27]. Their operation is based on Boltzmann's law; thus both exhibit the same exponential dependence of current on voltage, although this dependence is sometimes steeper in ionic channels because of correlated charge control [26].
- *Local negative feedback*: The current-controlled current conveyor is a circuit that depends strongly on *local negative feedback* for its function. It is interesting to note that in the biological outer-plexiform circuit, the local coupling of the two ionic channels, the excitatory on the horizontal cells and the inhibitory on the receptors, also form a local negative feedback loop that determines the temporal properties in the outer-plexiform. This coupling between two cells through an excitatory and an inhibitory synapse is referred to as a reciprocal junction or reciprocal arrangement [21]. However, one should be careful in making such analogies; it is highly unlikely that there exist a *single* biological structure that performs the function of transistor M_1 in the current-controlled current conveyor.
- *Current-domain signal representation*: The circuits described in this paper employ unidirectional current signals. Our previous experience with bidirectional current mode circuits for associative memories [9] was not very encouraging. There were systematic mismatches in currents because of extra mirroring operations in devices of different type [18]. This mirroring operation not only consumes more space; it also makes the architecture less fault tolerant (it includes unnecessary circuitry). Of course, differential current signal representations require an extra communication line; the bidirectional communication scheme described earlier was employed to solve the problem on silicon.

The dynamics in our circuits, an important topic, have not been discussed in this paper. A systematic methodology for compensating the circuits and ensuring for stability, much like the techniques suggested by Wyatt and Standley [28], would help the designer of these systems.

We have used the two-layer retina as an example of how to map biological microcircuits directly into silicon. Chemical synapses do not present a problem; they are nicely implemented with transconductances. The "Mexican-hat" response is obtained using only nearest neighbor interactions, which minimizes wiring in an analog VLSI implementation. Its response is robust over a wide range of operating conditions. The importance of this receptive field is evident from the fact that it is used in the very first brain tissue, the outer-plexiform of the retina. Similar receptive fields have been found in other parts of the brain. Through evolution, nature has discovered the simplest, most reliable, and most efficient architectures to accomplish the task at hand—properties that anyone wishing to build large-scale analog information processing systems treasures.

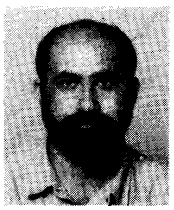
ACKNOWLEDGMENT

M. Cohen and S. Ayyar provided excellent comments during the preparation of the document. The authors would like to thank Prof. C. Mead of Caltech and T. Sejnowski of UCSD for encouraging this research. They also wish to thank Prof. E. Sànchez-Sinencio, guest editor, for providing an opportunity to present this work in this special issue.

REFERENCES

[1] J. D. Cowan and D. H. Sharp, "Neural nets," *Quart. Rev. Biophys.*, vol. 21, 3, pp. 365–427, 1988.
[2] C. A. Mead, "Silicon models of neural computation," in *Proc. IEEE First Int. Conf. Neural Networks*, vol. I (San Diego, CA), 1987.
[3] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.
[4] C. Koch, "Analog VLSI circuits for computer vision," *Neural Comput.*, vol. 1, no. 2, pp. 184–200, 1989.
[5] W. Reichardt and T. Poggio, "Visual control of orientation and behavior in the fly: Part I," *Quart. Rev. Biophys.*, vol. 9, pp. 311–375, 1976.
[6] D. Marr, *Vision*. New York, NY: W. H. Freeman, 1982.
[7] E. Vittoz, "Micropower techniques," in *VLSI Circuits for Telecommunications*, Y. P. Tsividis and P. Antognetti, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1985.
[8] *MOSIS User Manual*, USC/Information Science Institute, The MOSIS Service, 4676 Admiralty Way, Marina del Ray, CA 90292-6695.
[9] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based

on weak inversion operation," *IEEE J. Solid-State Circuits*, vol. SC-12, pp. 224–231, June 1977.
[10] A. G. Andreou and K. A. Boahen, "Synthetic neural circuits using current domain-signal representations," *Neural Comput.*, vol. 1, pp. 489–501, 1989.
[11] A. Pavasović, "Subthreshold region MOSFET mismatch analysis and modeling for analog VLSI systems," Ph.D. dissertation, Johns Hopkins University, Baltimore, MD, 1990.
[12] A. Pavasović, A. G. Andreou, and C. R. Westgate, "Characterization of CMOS process variations by measuring subthreshold current," in *Non-destructive Characterization of Materials*, vol. IV, C. O. Ruud and R. E. Green, Eds. New York: Plenum Press, 1991.
[13] C. Toumazou, F. J. Lidgey, and D. G. Haigh, Eds., *Analogue IC Design: The Current-Mode Approach* (IEEE Circuits and Systems Series), vol. 2. London: Peter Peregrinus, 1990.
[14] B. Gilbert, "Translinear circuits: A proposed classification," *Electron. Lett.*, vol. 11, no. 1, pp. 14–16, 1975.
[15] A. G. Andreou and K. Strohbehn, "Analog VLSI implementations of the Hassenstein–Reichardt–Poggio models for vision computation," in *Proc. 1990 IEEE Symp. Systems, Man and Cybernetics* (Los Angeles, CA), Nov. 1990.
[16] K. A. Boahen and A. G. Andreou, "Design of a bidirectional associative memory chip," Electrical and Computer Engineering, Tech. Rep. JHU/ECE90-06, Johns Hopkins University, 1990.
[17] K. C. Smith and A. S. Sedra, "The current conveyor—A new circuit building block," *Proc. IEEE*, vol. 56, pp. 1368–1369, Aug. 1968.
[18] K. A. Boahen, A. G. Andreou, P. O. Pouliquen, and A. Pavasović, "Architectures for associative memories using current-mode analog MOS circuits," in *Proc. Decennial Caltech Conf. VLSI*, 1989.
[19] K. A. Boahen, P. O. Pouliquen, A. G. Andreou, and R. E. Jenkins, "A heteroassociative memory using current-mode MOS analog VLSI circuits," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 747–755, May 1989.
[20] A. G. Andreou, "Synthetic neural systems using current-mode circuits," in *Proc. IEEE 1990 Int. Symp. Circuits and Syst.* (New Orleans), May 1990.
[21] G. M. Shepherd, *The Synaptic Organization of the Brain*. New York, NY: Oxford University Press, 1979.
[22] M. Cohen, Master's thesis, Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD, 1991.
[23] J. Lazzaro, R. Ryckebusch, M. A. Mahowald, and C. A. Mead, "Winner-take-all networks of $O(n)$ complexity," in *Advances in Neural Information Processing Systems*, vol. 1, D. S. Touretzky, Ed. Los Altos, CA: Morgan Kaufmann, 1989.
[24] J. E. Dowling, *The Retina, an Approachable Part of the Brain*. Cambridge, MA: Harvard University Press, 1987, ch. 3.
[25] A. G. Andreou and K. A. Boahen, "Can you get the Mexican-hat function using two resistive networks?" JHU Electrical and Computer Engineering Tech. Rep., JHU/ECE 88-08, Johns Hopkins University, 1988.
[26] B. Hille, *Ionic Channels of Excitable Membranes*. Sunderland, MA, Sinauer Associates, 1984, pp. 54–57.
[27] C. A. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, pp. 1629–1636, Oct. 1990.
[28] J. L. Wyatt, Jr., and D. L. Standley, "Criteria for robust stability in a class of lateral inhibition networks coupled through resistive grids," *Neural Comput.*, vol. 1, no. 1, pp. 58–67, Spring 1989.



Andreas G. Andreou (M'90) received the M.Sc. and Ph.D. degrees in electrical engineering and computer science from Johns Hopkins University, Baltimore, MD, in 1983 and 1986 respectively.

At present, he is assistant professor of electrical and computer engineering at Johns Hopkins. His research interests are in the areas of device physics, analog VLSI, and neural computation.

Dr. Andreou is a member of Tau Beta Pi.

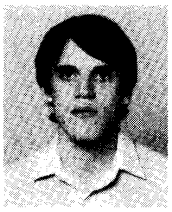
*

Kwabena A. Boahen (S'86–M'89) is a Ph.D. student at Caltech after having completed a B.S./M.S.E. degree in electrical and computer engineering at John Hopkins University, Baltimore, MD.

His current research interests are analog VLSI design and testing, with applications in synthetic neural and sensory systems.

Mr. Boahen is a member of Tau Beta Pi.

*



Philippe O. Pouliquen is a Ph.D. graduate student of electrical and computer engineering at the Johns Hopkins University, Baltimore, MD. His current research interests are instrumentation, and analog VLSI.

*

Aleksandra Pavasović received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Johns Hopkins University, Baltimore, MD, in 1988 and 1990 respectively. Her Ph.D. dissertation research was focused on the matching properties of MOS transistors in subthreshold.

*

Robert E. Jenkins (M'86) received the B.S. in engineering and the M.S. in physics from the University of Maryland.

He is a lecturer in the School of Engineering and a principal staff engineer at the Applied Physics Laboratory (APL) at Johns Hopkins. He is also a member of the program committee for the part-time master's degree program in electrical engineering. At APL, he has over twenty years of experience in computing systems and is currently leading the Space Dept. IR&D program.

*

Kim Strohbehn (S'78–M'79) received the M.Sc. and Ph.D. degrees in electrical engineering from Iowa State University, Ames, in 1977 and 1979 respectively.

He is a senior staff engineer at the Applied Physics Laboratory (APL) of the Johns Hopkins University. His research interests are in VLSI, astronomical signal processing, and control.