17

Design of a Bidirectional Associative Memory Chip

KWABENA A. BOAHEN and ANDREAS G. ANDREOU

17.1. INTRODUCTION

Bidirectional associative memories (BAMs) support heteroassociative storage and recall of binary patterns. A four-layer BAM network is shown in Fig. 17.1a. It has two input/output (I/O) layers, A and B, and two hidden layers, G and H. The pathways between layers are weighted projections that connect every unit in one layer to all units in the other. An input pattern at layer A recalls an associated pattern at layer B through the A ⇒ G ⇒ B pathway. Reciprocally, inputs at layer B elicit patterns at layer A via pathway B ⇒ H ⇒ A. This network supports truly bidirectional associations, i.e., two-way links between stored pattern pairs. Associative memory architectures and dynamics are reviewed in Chapter 1 and the BAM model is also discussed in Chapter 5 of this volume.

The recurrent pathways between layers A and B produce interesting dynamics in the BAM network. Stable reverberations occur when patterns in these layers reinforce each other, that is,

$$\mathbf{A}_j \rightarrow \mathbf{B}_j \rightarrow \mathbf{A}_j \ldots$$

for the jth association $(\mathbf{A}_j, \mathbf{B}_j)$. Consequently, the network remains in this state even after external inputs are removed. When a new input pattern is applied, the state of the network evolves to the stored pattern that best matches the input. Formally, using column vectors with ± 1 components to represent the patterns, the pattern $\mathbf{A}$, applied at layer A, recalls the association $(\mathbf{A}_p, \mathbf{B}_p)$ given by

$$\mathbf{A}_p^{\mathrm{T}}\mathbf{A} = \max_{j=1,\ldots,r} \mathbf{A}_j^{\mathrm{T}}\mathbf{A} \qquad (1)$$

where the max operation is over all stored associations.

Stored associations are encoded using a unary representation at the hidden layers (grandmother cells). A unique unit in each hidden layer is assigned to every association stored; this is essentially a sparse coding of the patterns present at the I/O layers. There is a one-to-one correspondence between weights and stored patterns; the weights of reciprocal connections between unit a_i (b_i) and units g_j and h_j are equal to the ith component of $\mathbf{A}_j$ $(\mathbf{B}_j)$. In other words, the weight vectors of hidden units g_j and h_j are simply $\mathbf{A}_j$ and $\mathbf{B}_j$, respectively. The folded architecture shown in Fig. 17.1b makes it possible to share one stored weight between two reciprocal connections.

The original BAM network proposed by Kosko (1988) had two layers and used a distributed representation. Our four-layer network is mathematically equivalent to Kosko's model (Boahen et al., 1989a, b). From a chip designer's point of view, however, a local representation offers three major advantages over a distributed one.

First, optimal storage efficiency of one information bit per storage is achieved. This is the appropriate measure of hardware efficiency for an implementation based on digital storage because the chip area is proportional to the number of storage cells. For reliable operation (Eq. 1) with random patterns, only up to $n/8$ associations can be stored in two-layer BAM with n neurons in each layer. This translates to an efficiency of $1/4(\log_2 n - 1)$ information bits per storage bit. (Each weight requires $\log_2(n/8) + 1$ bits.)

Second, no weight adjustment circuitry is required; pattern vectors are written directly to the weight storage cells. In a distributed memory, weights are updated by

$$m_{kl} \leftarrow m_{kl} + a_k b_l$$

for each new association. The overhead of this circuitry is especially severe if new associations are programmed only occasionally.

Third, introducing lateral inhibition within the hidden layers prevents recall performance from degrading as the number of stored patterns

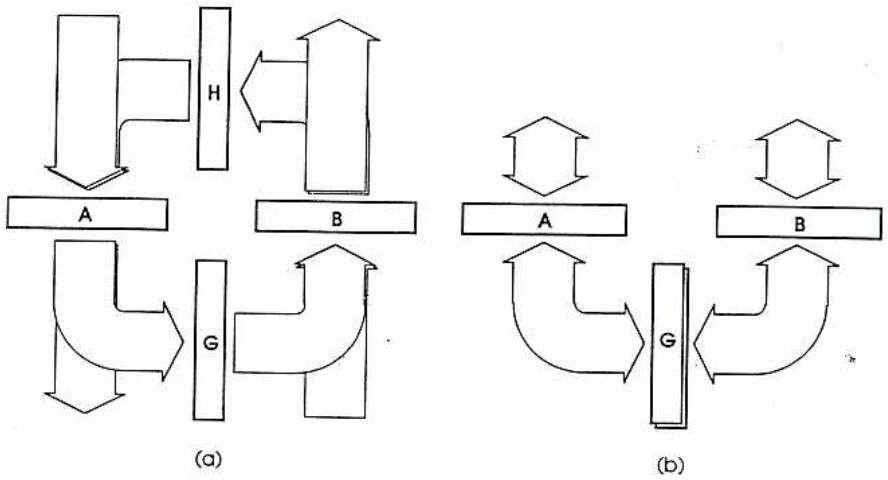


Fig. 17.1. (a) Four-layer bidirectional associative memory. The pathway (arrows) in the foreground allows inputs at layer A to produce outputs at layer B; the one in the background supports recall in the opposite direction. (b) Folded architecture: the hidden layers are merged and reciprocal connections are introduced.

increases. Inhibition enhances the contribution of the best match while reducing the contributions of other patterns. A four-layer BAM with this nonlinear expansion at the hidden layers is equivalent to a higher-order correlation two-layer BAM. However, a distributed higher-order network has very poor storage efficiency. For instance, a second-order network has one-third the storage efficiency of a first-order one, and a third-order network has only 1/15 (Peretto and Niez, 1986).

Poor fault tolerance is an often-cited disadvantage of local representations. Evidently there is a trade-off between storage efficiency and fault tolerance. Distributed representations achieve fault tolerance through redundant storage of information, hence their poor storage efficiency. Concomitantly, they are extremely robust against hard or soft faults in several memory locations (Sivilotti et al., 1985). Local representations can also achieve fault tolerance through redundancy; the user can trade storage efficiency for fault tolerance by storing more than one copy of each association.

Furthermore, robustness against certain failures in the hidden units and their communication lines is achieved by using a reentrant memory array. A hidden unit's operation is verified by performing a recall before committing an association to it. Each hidden unit is assigned an eligibility bit which captures its state when the programming (as opposed to recall) signal is asserted. This bit is used to select the weight storage location of the winning hidden unit for reentry; the desired new association overwrites the previously recalled one.

The four-layer BAM chip described here uses efficient circuit techniques to realize programmable reciprocal connections and lateral inhibition. A simple current-controlled current conveyor and a novel two-transistor reciprocal junction allow units to interact bidirectionally through a single line. This is accomplished without multiplexing by using *independent* voltage and current signals. A winner-take-all (WTA) circuit similar to the one proposed by Lazzaro et al. (1989) that requires just one communication line is used to implement lateral inhibition. Using these techniques, BAM chips that approach static RAM densities have been realized.

17.2. CIRCUIT TECHNIQUES

A unit's output is represented by a voltage signal; its inputs are represented by current signals. Since currents and voltages may be independently transmitted along the same line, these signal representations allow output and inputs to be communicated using just one line. Voltage output facilitates fan-out, while current input provides summation. Units interact bidirectionally through a reciprocal junction circuit using a simple current conveyor. These circuits were designed for subthreshold operation. The advantages of the subthreshold current-mode approach are outlined in Andreou and Boahen (1989).

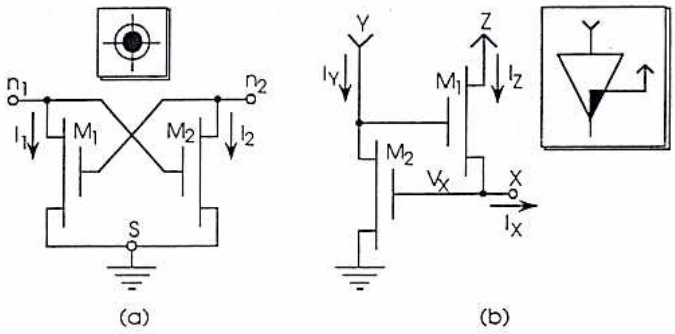


Fig. 17.2. (a) Reciprocal junction. Voltage signals applied at node n_1 or n_2 are converted to current signals. The lines that apply voltage signals bring back these currents. The circuit symbol is shown in the insert; the perpendicular lines correspond to nodes n_1 and n_2, (b) Current conveyor. The current, I_X, supplied to node X is conveyed to node Z. The voltage at X, V_X, is determined by the current I_Y. The conveyor's circuit symbol is in the insert; the black triangle depicts current buffering.

17.2.1. Subthreshold Device Model

A simple model for subthreshold conduction in the MOS transistor is given in Mead (1989):

$$I_{ds} = I_0\left(\frac{W}{L}\right)e^{\kappa V_{gb}/V_T}(e^{-V_{sb}/V_T} - e^{-V_{db}/V_T}) \quad (2)$$

where voltages are with reference to the local substrate or well and the channel length modulation is ignored. The drain current's dependence on the back-gate (well/substrate voltage) becomes explicit when voltages are referred to the source potential:

$$I_{ds} = I_0\left(\frac{W}{L}\right)e^{[(1-\kappa)V_{bs}]/V_T}\, e^{\kappa V_{gs}/V_T} \times \left(1 - e^{-V_{ds}/V_T} + \frac{V_{ds}}{V_0}\right) \quad (3)$$

I_0 is the zero-bias current for the given device, κ measures the effectiveness of the gate potential in controlling the channel current, and W and L are the channel width and length, respectively. V_0 is the Early voltage which is proportional to L. V_T $(=kT/q)$ is the thermal voltage which is equal to 26 mV at room temperature.[1] Typical parameters for minimum-size device (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. The current changes by a factor of 10 for an 80-mV change in V_{ds} or a 240-mV change in V_{bs}; this relation holds up to about 100 nA.

In the saturation region, $V_{ds} > 4V_T$ ($V_{dsat} \simeq$ 100 mV at room temperature), the device's output conductance and transconductance are given by

$$g_{dsat} = \frac{\partial I_{ds}}{\partial V_{ds}} = \frac{I_{ds}}{V_0} \quad (4)$$

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{\kappa I_{ds}}{V_T} \quad (5)$$

[1] This is the n-type device equation; the p-type version has voltage signs reversed.

The ratio $A = g_m/g_{dsat} = \kappa V_0/V_T$ measures the gain available from the device. It is independent of current and is 430 for the given parameter values.

17.2.2. Reciprocal Junction

The two-transistor circuit shown in Fig. 17.2a provides bidirectional interaction between units connected to nodes n_1 and n_2. It receives voltage inputs at these nodes and produces current outputs at the same nodes; each transistor acts like a synaptic junction. Interaction is turned on by grounding S, turned off by bringing S to V_{dd}, or modulated in a multiplicative fashion by applying an analog signal to the well.

When S is at V_{dd} one of the devices has a positive gate–source voltage and sources the current:

$$I_{off} = I_0\, e^{[\kappa(V_{n_1} - V_{n_2})]/V_T} = I_0 L$$

assuring $V_{n_1} > V_{n_2}$, without loss of generality, and ignoring the body effects. I_{off} is proportional to $L \equiv e^{\kappa V_{n_1}/V_T}/e^{\kappa V_{n_2}/V_T}$, the dynamic range of the current signals. For signals in the range 100 pA to 100 nA, $L = 10^3$, and $I_{off} \approx 1$ fA.

For proper operation in the on state, the devices must be in saturation (V_{n_1}, $V_{n_2} > V_{dsat}$). Then, for a small change in V_{n_1}, the changes in I_1 and I_2 are related by

$$\frac{\Delta I_1}{\Delta I_2} = \frac{g_{dsat_1}}{g_{m_2}} = \frac{1}{A}\frac{I_1}{I_2}$$

This gives $\Delta I_1/I_1 = \frac{1}{430}\Delta I_2/I_2$. Hence, changing V_{n_1} to increase I_2 by 100 percent causes a 0.23 percent change in I_1. So bidirectional communication is possible with less than −50 dB crosstalk.

17.2.3. Current Conveyor

A current-controlled current conveyor is shown in Fig. 17.2b. This circuit has a communication node X, a control node Y, and an output node

(a) Reciprocal junction. Voltage
plied at node n_1 or n_2 are converted
t signals. The lines that apply
gnals bring back these currents.
t symbol is shown in the insert; the
ular lines correspond to nodes n_1
) Current conveyor. The current,
ed to node X is conveyed to node
tage at X, V_X, is determined by the
. The conveyor's circuit symbol is
sert; the black triangle depicts
ffering.

$_{1}/g_{d\mathrm{sat}} = \kappa V_0/V_T$ measures the
n the device. It is independent
430 for the given parameter

l Junction

r circuit shown in Fig. 17.2a
onal interaction between units
s n_1 and n_2. It receives voltage
nodes and produces current
me nodes; each transistor acts
ction. Interaction is turned on
urned off by bringing S to V_{dd},
a multiplicative fashion by
g signal to the well.
V_{dd} one of the devices has a
rce voltage and sources the

$e^{[\kappa(V_{n_1}-V_{n_2})]/V_T} = I_0 L$

, without loss of generality, and
effects. I_{off} is proportional to
V_T, the dynamic range of the
or signals in the range 100 pA
$)^3$, and $I_{\mathrm{off}} \approx 1$ fA.
operation in the on state, the
in saturation ($V_{n_1}, V_{n_2} > V_{d\mathrm{sat}}$).
change in V_{n_1}, the changes in
ted by

$$\frac{1}{2} = \frac{g_{d\mathrm{sat}_1}}{g_{m_2}} = \frac{1}{A}\frac{I_1}{I_2}$$

$= \frac{1}{430}\Delta I_2/I_2$. Hence, changing
by 100 percent causes a 0.23
I_1. So bidirectional communi-
with less than -50 dB cross-

Conveyor

lled current conveyor is shown
is circuit has a communication
l node Y, and an output node

Z. The current I_X at the communication node is conveyed to the output node at a potential V_X determined by the control current I_Y. Traditional current conveyors (Smith and Sedra, 1968) use a voltage input to set V_X, i.e. $V_X = V_Y$. The current-controlled conveyor's operation is described by two simple relationships:

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

where the function $\mathcal{I}(V_X)$ converts V_X to a current using a transistor identical to M_2. Thus the communication node can receive one current I_X, and, at the same time, transmit a second current I_Y.

In this circuit, M_1 establishes V_X to make M_2's current equal I_Y. This is accomplished by negative feedback through M_2 which serves as an inverting amplifier. The conductance seen at node X is approximately Ag_{m_1}, i.e., the source conductance of M_1 times M_2's gain A. In addition, M_1 buffers I_X, transferring this current to its high-impedance drain terminal. This negative feedback arrangement is the core of Säckinger's regulated cascode circuit (Säckinger and Guggenbühl, 1990). In Boahen et al., 1989b, the authors proposed its use, in conjunction with the reciprocal junction, for two-way communication.

A current conveyor communicates bidirectionally with one or more conveyors through reciprocal junctions connected to its communication node. I_Y is the outgoing signal; it is replicated at each junction by V_X; and I_Z is the sum of all incoming signals. Changes in I_Z produce changes in V_X and therefore in the copies of I_Y. Small changes are related by

$$\frac{\Delta I_Y}{\Delta I_Z} = \frac{g'_{m_2}}{Ag_{m_1}} = \frac{1}{A}\frac{I_Y}{I_Z}$$

where the copy $I'_Y = I_Y + \Delta I_Y$ and g'_{m_2} ($=g_{m_2}$) is the transconductance of the device that mirrors I_Y. Hence, in percentages, the change in I'_Y is 430 times less than in I_Z, just as the previous case.

For large changes in I_Z, M_1's gate–source voltage changes by $(V_T/\kappa)\ln(I_{Z_2}/I_{Z_1})$, and therefore

$$\frac{\Delta I_Y}{I_Y} = \frac{1}{A}\ln\left(\frac{I_{Z_2}}{I_{Z_1}}\right)$$

This means I_Z can change by about a factor of 75 before I'_Y changes by one percent. Characteristics of minimum-sized versions of these circuits, obtained from chips fabricated through foundry services, are shown in Fig. 17.3.

17.2.4. Winner-Takes-All

In this circuit, shown in Fig. 17.4, m current conveyors compete for current supplied to a common line. This current, I_m, 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 (Lazzaro et al., 1989) to provide current outputs.

Input currents are supplied to control nodes, output currents are obtained from output nodes, and the communication nodes are connected together. Each conveyor sees a voltage source at its communication node—not a current source. If $I_Y < \mathcal{I}(V_X)$, M_2 enters the linear region, turning M_1 off (refer to Fig. 17.2b). 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 I_m.

When the inputs are very similar, the conversion from input to output is *exponential*, with the sum of the outputs normalized to I_m. In this case M_1 remains on and M_2 stays in saturation. The inputs develop voltage signals across M_2's drain conductance, $g_{d\mathrm{sat}_2}$; these voltages are converted exponentially to current by M_1. For example, a one-percent input difference produces a voltage difference of 0.15 V, so the corresponding outputs differ by a factor of 75.

If two conveyors, with inputs $I_Y \pm \frac{1}{2}\Delta I_Y$, are competing for I_m, their outputs are $I_Z \pm \frac{1}{2}\Delta I_Z$, where $I_Z = \frac{1}{2}I_m$ and the differential gain for small signals is

$$\frac{\Delta I_Z}{\Delta I_Y} = \frac{g_{m_1}}{g_{d\mathrm{sat}_2}} = A\frac{I_Z}{I_Y}$$

Hence, the small-signal differential gain is 430 for normalized inputs and outputs.

17.3. CIRCUIT DYNAMICS

Dynamic behavior of the current conveyor is determined by corner frequencies associated with its communication and control nodes. The communication node's capacitance arises from the interconnects and reciprocal junctions; capacitance is added at the control node to tailor the conveyor's response. A small-signal conveyor model is introduced and used to find optimal choices for the control node capacitance. It is also used to analyse a two-layer network of current conveyors and a WTA network. A simple small-signal transistor model consisting of a transconductance g_m, controlled by v_{gs}, in parallel with a conductance $g_{d\mathrm{sat}}$ is used here; it is

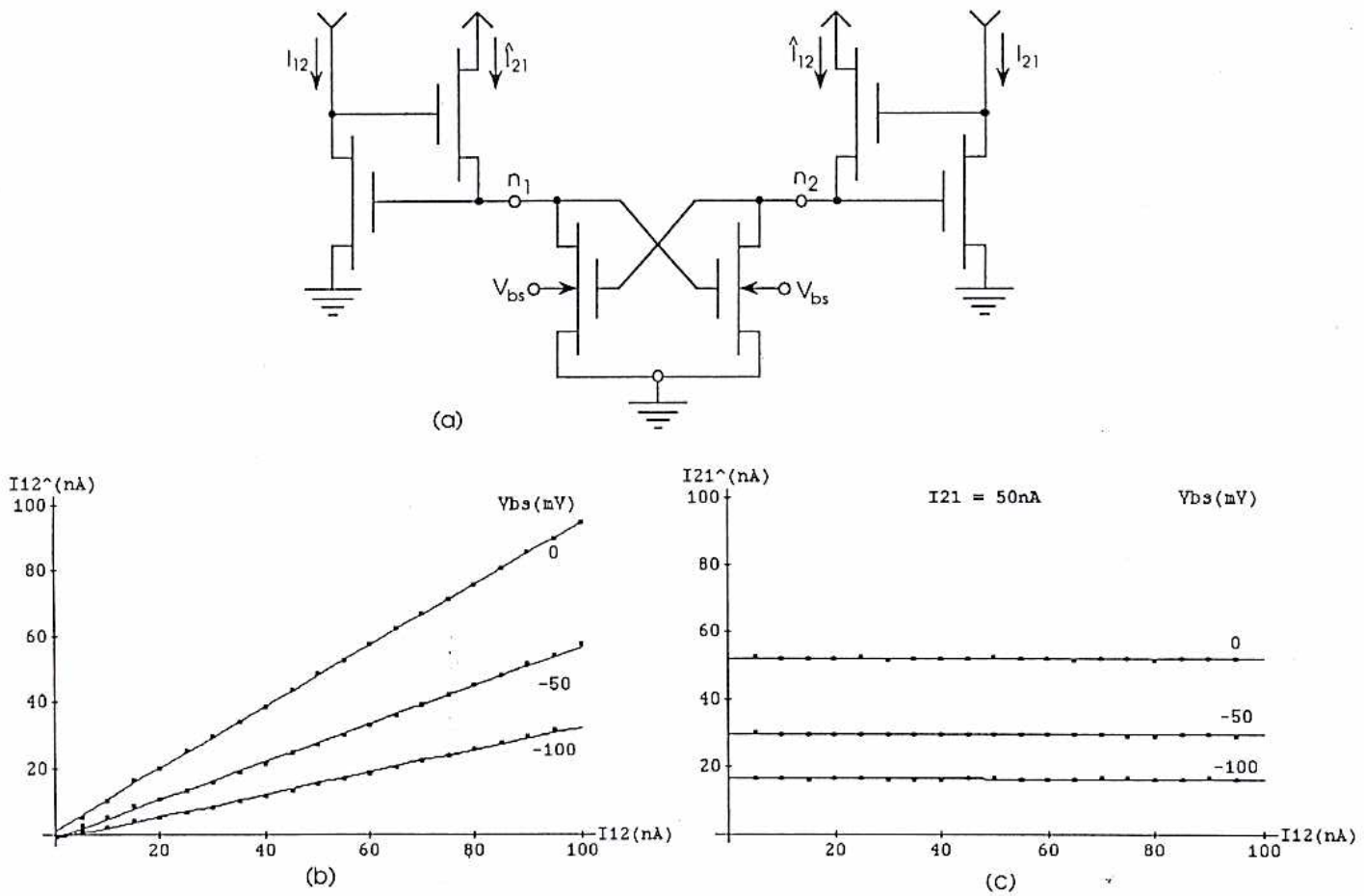
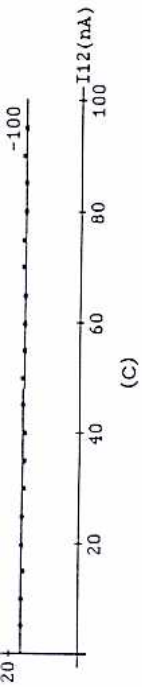


Fig. 17.3. (a) Two conveyors communicate through a reciprocal junction. They send signals I_{12} and I_{21} and receive signals $\hat{I}_{21}$ and $\hat{I}_{12}$. The weight is determined by the well voltage, V_{bs}. Experimental data were obtained by stepping I_{12} from 5 nA to 100 nA, with I_{21} held at 50 nA, for V_{bs} values of 0, -50 mV, and -100 mV. (b) $\hat{I}_{12}$ varied linearly with I_{12}; the slopes are 0.93, 0.57, and 0.33. (c) $\hat{I}_{21}$ 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 weighting is symmetric.



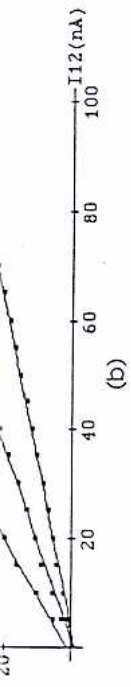


Fig. 17.3. (a) Two conveyors communicate through a reciprocal junction. They send signals I_{12} and I_{21} and receive signals $\hat{I}_{21}$ and $\hat{I}_{12}$. The weight is determined by the well voltage, V_{bs}. Experimental data were obtained by stepping I_{12} from 5 nA to 100 nA, with I_{21} held at 50 nA, for V_{bs} values of 0, -50 mV, and -100 mV. (b) I_{12} varied linearly with I_{12}; the slopes are 0.93, 0.57, and 0.33. (c) $\hat{I}_{21}$ 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 weighting is symmetric.

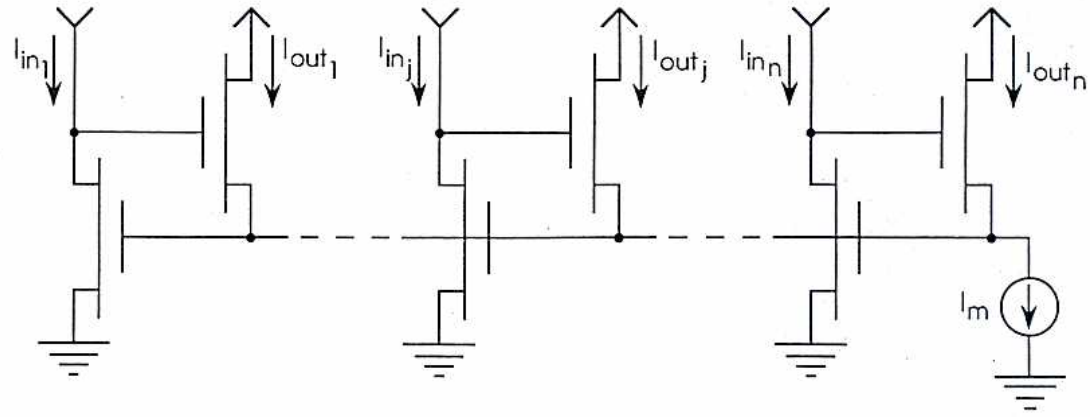


Fig. 17.4. Winner-takes-all (WTA) circuit. m current signals are applied to m current conveyors. The conveyor that has the largest input conveys the current I_m supplied to a common line; the other conveyors have zero output.

assumed that the communication and control node capacitances are much larger than the intrinsic device capacitances.

17.3.1. Small-Signal Conveyor Model

The signals i_X, i_Y, and v_Z are viewed as the conveyor's inputs; they produce output signals v_X, v_Y, and i_Z (see Fig. 17.2b). Its small-signal behavior is described by[2]

$$\begin{bmatrix} v_X \\ v_Y \\ i_Z \end{bmatrix} = \begin{bmatrix} Z_X & Z_{XY} & A_{XZ} \\ Z_{YX} & Z_Y & A_{YZ} \\ A_{ZX} & A_{ZY} & Y_Z \end{bmatrix} \begin{bmatrix} i_X \\ i_Y \\ v_Z \end{bmatrix}$$

The outputs of interest are v_X, which communicates i_Y, and i_Z which should equal i_X. Ignoring the dependence on v_Z, we have

$$\begin{bmatrix} v_X \\ i_Z \end{bmatrix} = \begin{bmatrix} Z_X & Z_{XY} \\ A_{ZX} & A_{ZY} \end{bmatrix} \begin{bmatrix} i_X \\ i_Y \end{bmatrix}$$

The transimpedance Z_{XY} converts the outgoing signal i_Y to a voltage, while the incoming signal i_X appears at the supply node with a gain of A_{ZX}. Ideally, $Z_{XY} = 1/g_{m_2}$ and $A_{ZX} = 1$. The impedance Z_X seen at the communication node and the gain A_{ZY} from the control node to the supply node produce interference between incoming and outgoing signals; they should equal zero. In practice, all the parameters are finite and frequency dependent. They are characterized by corner frequencies $\omega_X = g_{m_1}/C_X$ and $\omega_Y = g_{m_2}/C_Y$, associated with nodes X and Y. The gain provided by the voltage follower M_1 has dropped to one-half at ω_X; the inverting amplifier M_2 has unity gain at ω_Y.

In the s-domain, the conveyor's small-signal parameters are given by

$$\left.\begin{aligned} Z_X(s) &\approx \left(\frac{1}{Ag_{m_1}}\right)\left(\frac{1+A\rho s}{1+\rho s+\rho s^2}\right) \\ Z_{XY}(s) &\approx \left(\frac{1}{g_{m_2}}\right)\left(\frac{1}{1+\rho s+\rho s^2}\right) \end{aligned}\right\} \quad (6)$$

$$\left.\begin{aligned} A_{ZX}(s) &\approx -\frac{1+\rho s}{1+\rho s+\rho s^2} \\ A_{ZY}(s) &\approx \left(\frac{g_{m_1}}{g_{m_2}}\right)\left(\frac{s}{1+\rho s+\rho s^2}\right) \end{aligned}\right\} \quad (7)$$

where the complex frequency variable s is in units of ω_X and $\omega_Y = (1/\rho)\omega_X$; the approximations $A \gg 1$ and $\rho A \gg 1$ were used. Poles occur at $s = \frac{1}{2}[-1 \pm \sqrt{(1-4/\rho}]$; if $\rho < 4$ they are complex. Optimal conveyor behavior occurs when $\rho = 2$; Bode plots of the transfer functions are shown in Fig. 17.5.

17.3.2. Conveyor Responses

The impedance seen at the communication node, $Z_X(j\omega)$, equals $1/Ag_{m_1}$, at DC as shown in the previous section. Above $\omega_X/\rho A$, or ω_Y/A, its magnitude increases because M_2's gain starts decreasing. When ω_X is exceeded, C_X causes the impedance to decrease. It reaches a maximum magnitude of $1/g_{m_1}$ at $\hat{\omega}_X = \omega_X/\sqrt{\rho}$. At the

[2] In our notation, Z_Q or $1/Y_Q$ is the impedance seen at node Q, Z_{QR} is a transimpedance at node Q controlled by node R, and A_{QR} is a voltage or current gain from R to Q. By convention, positive current flows into the circuit.

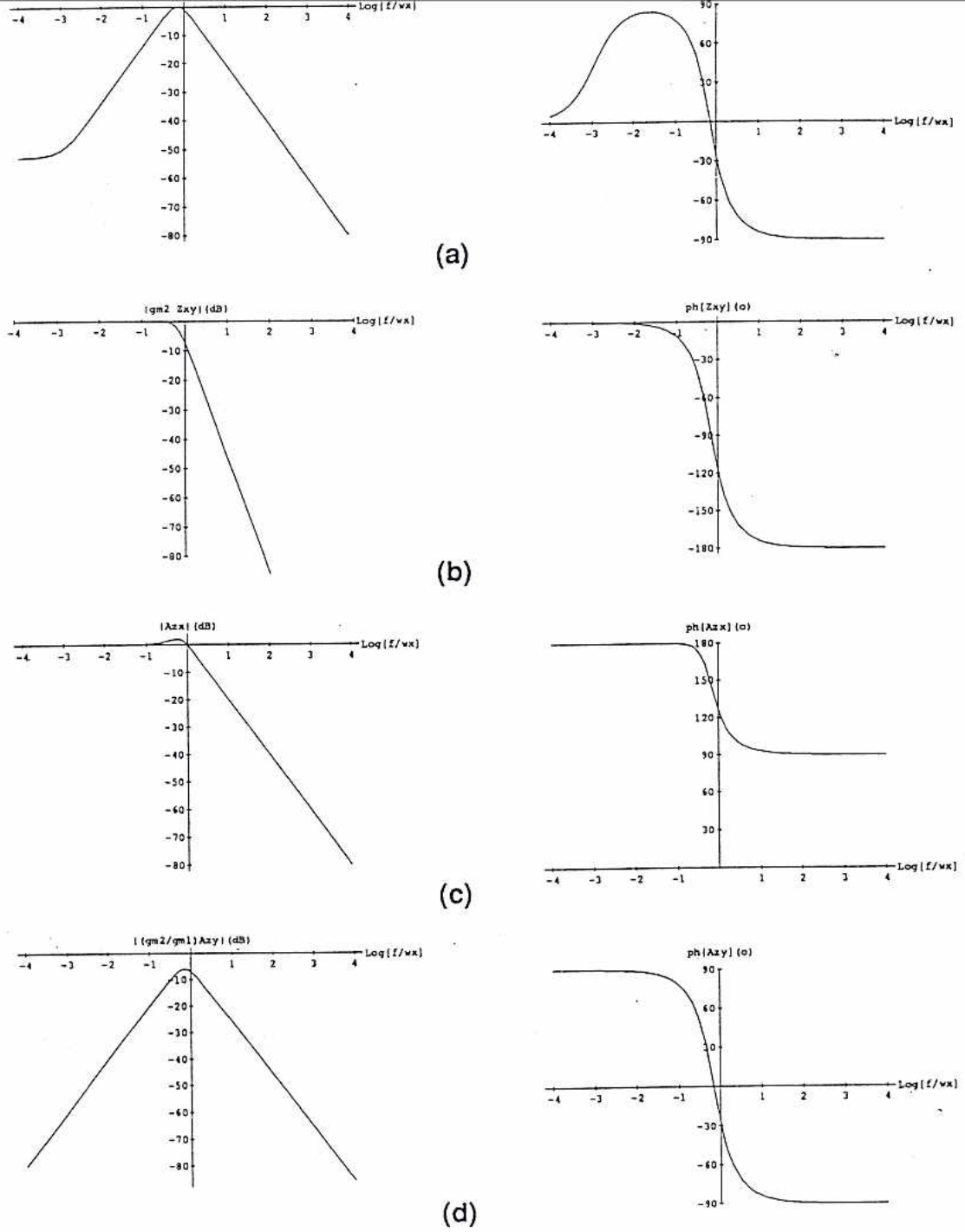
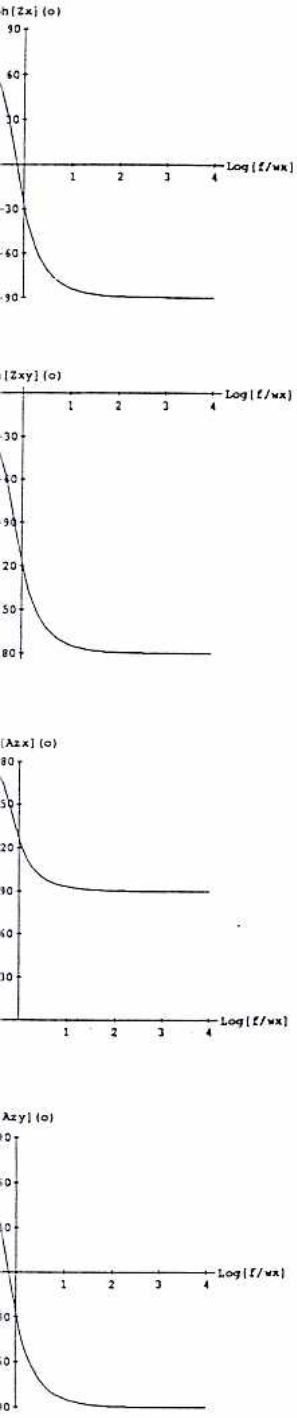


Fig. 17.5. Conveyor transfer functions for $\rho = 2$ and $A = 430$; both magnitude (left) and phase (right) are shown. The frequency scale is in units of ω_X. (a) Z_X, the communication node impedance. (b) Z_{XY}, the transimpedance from control node to communication node. (c) A_{ZX}, the current gain from communication node to supply node. (d) A_{ZY}, the current gain between control and communication nodes.



and phase (right) are shown.
(b) Z_{XY}, the transimpedance
ication node to supply node.

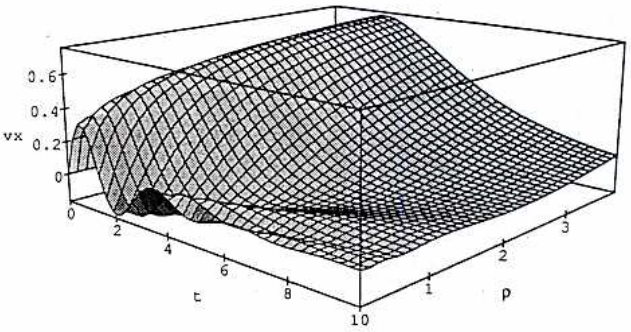


Fig. 17.6. Response of voltage at node X to a current step applied there at $t = 0$. $v_X(t)$ is plotted for values of ρ from 0.2 to 3.9; amplitude is in units of i_u/g_{m_1}, where i_u is the height of the step input, and time is in units of $1/\omega_X = C_X/g_{m_1}$. The settling time is large for both small and large values of ρ; it is minimized round 2.

maximum, $Z_X(j\omega)$'s phase is zero and its magnitude is A times its DC value, producing a 53-dB peak (see Fig. 17.5a).

Ringing occurs in the voltage at node X when $Z_X(s)$ has complex poles. A unit current step generates the voltage signal

$$v_X(t) = \frac{1}{Ag_{m_1}} + \frac{e^{-\frac{1}{2}\omega_X t}}{g_{m_1}\sqrt{(4/\rho - 1)}} \times \{2 \sin \omega_0 t - [\sqrt{(4/\rho - 1)} \cos \omega_0 t + \sin \omega_0 t]/A\}$$

$$\approx \frac{1}{Ag_{m_1}} + \frac{2e^{-\frac{1}{2}\omega_X t} \sin \omega_0 t}{g_{m_1}\sqrt{(4/\rho - 1)}} \qquad (8)$$

where $\omega_0 = (\omega_X/2)\sqrt{(4/\rho - 1)}$. The responses for various values of ρ are shown in Fig. 17.6. The settling time is minimized around $\rho \approx 2$.

The transimpedance from the control node to the communication node, $Z_{XY}(j\omega)$, equals $1/g_{m_2}$ at low frequencies. At high frequencies, C_Y shunts i_y and the follower's gain decreases, so there is a 40-dB per decade roll-off. The break-frequency equals $\hat{\omega}_X$. A peak appears in the response if $\rho < 2$; when $\rho = 2$ the response is maximally flat and settling time is minimized (see Fig. 17.5b).

The current gain between the communication and output nodes, $A_{ZX}(j\omega)$, is unity for frequencies up to ω_X; but the response peaks just before the break frequency. When $\rho = 2$, the peak is 2 dB and occurs at $0.55\omega_X$. Above ω_X, C_X shunts i_X, and so the gain decreases at 20 dB per decade (see Fig. 17.5c). Note that $(1 + A_{ZX})i_X$ gives the current in C_X; equating the voltage across C_X to $Z_X i_X$, yields the relationship

$$A_{ZX} = Z_X g_{m_1} s - 1$$

Since s is the ratio between ω and ω_X, $g_{m_1}s$ gives the admittance of C_X at ω.

The current gain from the control node to the output node, $A_{ZY}(j\omega)$, is zero at DC. For AC signals, the transimpedance Z_{XY} develops voltage signals at the communication node which produce currents in C_X; these currents appear at the supply node. In fact

$$A_{ZY} = Z_{XY} g_{m_1} s$$

C_X introduces a zero at DC, so the gain initially increases at 20 dB per decade, reaches $g_{m_1}/\rho g_{m_2}$ $(= C_X/C_Y)$ at $\hat{\omega}_X$, and then rolls off 20 dB per decade (see Fig. 17.5d).

To summarize, the current conveyor provides a bidirectional communication channel as illustrated in Fig. 17.7. Its bandwidth is $\omega_X/\sqrt{\rho}$ for outgoing signals (I_Y to I'_Y), and ω_X for incoming ones (I_X to I'_X). When $\rho = 2$, the outgoing channel is maximally flat and settling time is minimized for both channels. The interference between these channels, i.e., the fraction of the other signal that is added, can be as high as g'_{m_2}/g_{m_1} $(= I'_Y/I_X)$ for the outgoing one and C_X/C_Y for the incoming one; this happens at the frequency $\omega_X/\sqrt{\rho}$. These reflections introduce undesired feedback paths

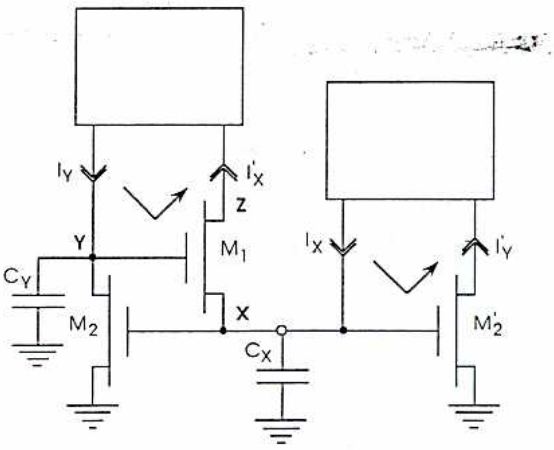


Fig. 17.7. Bidirectional communication channel between two current-mode circuits (boxes). The bent arrows depict channel interference, or reflections, caused by capacitances at the conveyor's communication and control nodes.

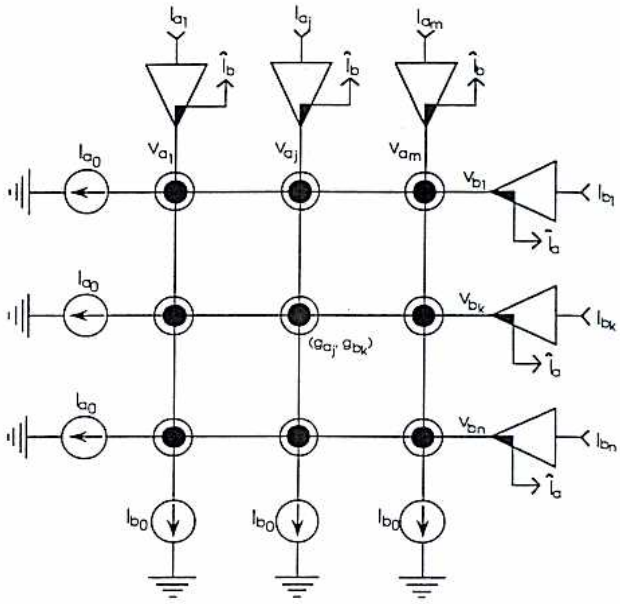


Fig. 17.8. Two-layer network. There are m current conveyors in the first layer and n in the second; they are fully connected by an $n \times m$ array of reciprocal junctions. Interference between incoming and outgoing signals at the conveyors' communication nodes produces signal loops. It is shown that the fixed bias currents I_{a_0} and I_{b_0} are necessary to obtain a loop gain less than unity.

around circuits that interface bidirectionally with the conveyor.

17.3.3. Network's Loop Gain

A fully connected two-layer network of current conveyors and reciprocal junctions is shown in Fig. 17.8. Conveyors in layer A send out currents $I_{a_1}, \ldots, I_{a_m}$; those in layer B transmit $I_{b_1}, \ldots, I_{b_n}$. Thus, all conveyors in layer A receive the same current

$$\hat{I}_b = I_{b_0} + \sum_{k=1}^{m} I_{b_j}$$

Similarly, all conveyors in layer B receive $\hat{I}_a$, the sum of I_{a_j} plus I_{a_0}; I_{a_0} and I_{b_0} are fixed bias currents. The reciprocal junction connecting conveyors a_j and b_k has two small-signal transconductances controlled by a_j and b_k, respectively: $g_{a_j} = \kappa I_{a_j}/V_T$ and $g_{b_k} = \kappa I_{b_k}/V_T$.

Conveyor a_j presents a small-signal impedance Z_{a_j} to the network. At resonance, $Z_{a_j} \to 1/G_a$, where $G_a = \kappa \hat{I}_b/V_T$ is the transconductance of its buffering device. In this case, all layer A conveyors present the same impedance to the network; they also see the same network admittance. This symmetry dictates that they develop the same signals v_a.

The signals developed at layer B will be

$$v_b = -\sum_{j=1}^{m} \frac{g_{a_j}}{G_b} v_a$$

assuming resonance occurs there too. These signals feed back to layer A, producing

$$v'_a = -\sum_{k=1}^{n} \frac{g_{b_k}}{G_a} v_b$$

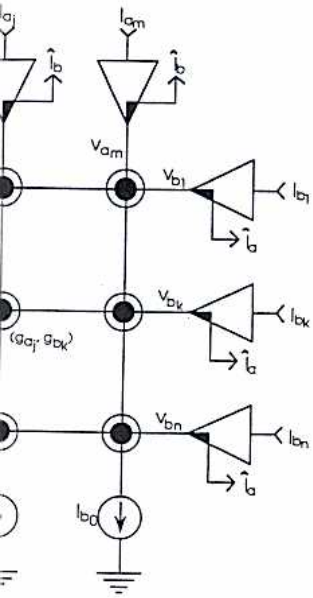
Therefore, the loop gain $A_L \equiv v'_a/v_a$ is

$$A_L = \frac{1}{G_a G_b} \sum_{j=1}^{m} g_{a_j} \sum_{k=1}^{n} g_{b_k} = \frac{\sum_{j=1}^{m} I_{a_j}}{I_{a_0} + \sum_{j=1}^{m} I_{a_j}} \frac{\sum_{k=1}^{n} I_{b_k}}{I_{b_0} + \sum_{k=1}^{n} I_{b_k}} \tag{9}$$

This is the largest value for A_L, for the worst-case condition where all conveyors resonate at the same frequency and they are fully connected. The fixed bias currents, I_{a_0} and I_{b_0}, ensure that $A_L < 1$.

17.3.4. Winner-Takes-All Response

As a conveyor's input approaches the winning cell's input, the current I_m switches from the winning cell to this cell's output (see Fig. 17.4). When their inputs are the same, say equal to I_Y, their outputs equal $\frac{1}{2}I_m$; therefore, their small-signal parameters are identical. For the inputs $I_Y \pm \frac{1}{2}\Delta I_Y$, the outputs are $\frac{1}{2}I_m \pm \frac{1}{2}\Delta I_Z$, where $\frac{1}{2}\Delta I_Z$ is the current that flows from one conveyor to the other. This current flows across Z_X to cancel the effect of the inputs which act through Z_{XY}.



e occurs there too. These layer A, producing

$$-\sum_{k=1}^{n} \frac{g_{bk}}{G_a} v_b$$

gain $A_L \equiv v_a'/v_a$ is

$$g_{aj} \sum_{k=1}^{n} g_{bk}$$

$$\frac{I_{aj}}{I_{aj}} \frac{\sum_{k=1}^{n} I_{bk}}{I_{b0} + \sum_{k=1}^{n} I_{bk}} \quad (9)$$

ue for A_L, for the worst-case conveyors resonate at the hey are fully connected. The and I_{b0}, ensure that $A_L < 1$.

s-All Response

ut approaches the winning rent I_m switches from the cell's output (see Fig. 17.4). e the same, say equal to I_Y, $\frac{1}{2}I_m$; therefore, their small- e identical. For the inputs s are $\frac{1}{2}I_m \pm \frac{1}{2}\Delta I_Z$, where $\frac{1}{2}\Delta I_Z$ ows from one conveyor to it flows across Z_X to cancel its which act through Z_{XY}.

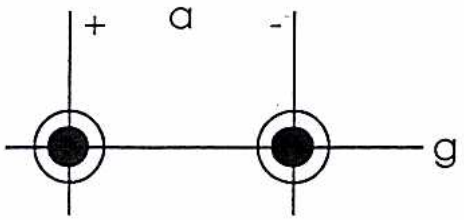


Fig. 17.9. Excitatory or inhibitory reciprocal weight using a pair of junctions which weight signals by 0 (off state) or 1 (on state).

Therefore,

$$\frac{\Delta I_Z}{2} = \frac{Z_{XY}}{Z_X} \frac{\Delta I_Y}{2}$$

Hence, the differential gain is

$$A(s) = \frac{Z_{XY}}{Z_X} = \frac{g_{m_1}}{g_{m_2}} \frac{A}{1 + A_S} \quad (10)$$

Here, $g_{m_1} = \kappa I_m/2V_T$, $g_{m_2} = \kappa I_Y/V_T$, and s is in units of ω_Y; the response is first-order and independent of ω_X. The gain equals A at DC, for normalized inputs and outputs as shown in the previous section, rolls off 20 dB per decade above ω_Y/A, and becomes unity at ω_Y.

17.4. BAM CIRCUITS

I/O units and hidden units perform nonlinear operations on their inputs. An I/O unit has two states, corresponding to the pattern values ± 1; its state is determined by comparing its excitatory and inhibitory inputs. A hidden unit also has two states; its state is determined by comparing its input with all other units in its layer. These comparisons are performed by WTA cells; a current conveyor supplies the input and communicates the outputs. This current conveyor/WTA combination is named after the cortical pyramidal cell because their functions are similar.

Programmable reciprocal weights are realized with a pair of reciprocal junctions and a dual rail scheme (Sivilotti et al., 1985) as shown in Fig. 17.9. For an excitatory (inhibitory) weight the junction on the excitatory (inhibitory) line is turned on; the other junction is turned off. In the forward direction (a to g), analog inputs are applied as voltages on either the excitatory or inhibitory lines, depending on their sign. If the relevant junction is on, i.e., the input is excitatory and the weight is excitatory or they are both inhibitory, the input appears as a current on the common line. In the reciprocal direction (g to a), positive analog inputs applied as voltages on the common line appear as current on the excitatory line if the weight is excitatory, otherwise they appear on the inhibitory line. This circuit, together with a flip-flop that stores the junctions' complementary states, is called a BAM cell.

17.4.1. Pyramidal Cell

Cortical pyramidal cells are arranged in layers with projections from one layer to the next. Their apical dendrites receive incoming afferent signals while their axons carry outgoing efferent signals. Pyramidal cell axons have collaterals that branch back and contact basal dendrites of neighboring cells; these lateral interactions are mainly inhibitory. Figure 17.10; shows a cortical pyramidal cell and the pyramidal cell circuit. The circuit has two communication nodes, X_Y and X_L: X_Y carries afferent and efferent signals, as do apical dendrites and axons of cortical pyramidal cells. X_L mediates lateral inhibitory interactions, mimicking the cortical cell's basal dendrites and recurrent axon collaterals. Inputs extrinsic to the network are applied at node Y_Y; the voltage at Y_Y is used to monitor the inputs and the WTA output of this cell.

The circuit consists of an n-type current conveyor (M_1, M_2) and a p-type WTA cell (M_3, M_4). The extra transistor (M_5) is needed to stabilize the circuit. When a WTA cell and a current conveyor feed each other, their current-buffering devices, M_1 and M_3, can act as common-source amplifiers. An unstable positive feedback loop results; the devices enter the linear region and X_Y and X_L go the the rails. The extra device prevents this by driving node X_Y to keep the conveyor's buffer (M_1) in saturation. Although M_5 shunts part of the incoming current, a fixed fraction of the input is passed to the WTA cell.

A pyramidal cell contributes a current I_U to the WTA competition and has a quiescent current of I_U. Therefore, its efferent signal $\mathcal{I}(V_{X_v})$ is 1 (in units of I_U) in the quiescent state but increases to $(m + 1)$ when it is winning, where m is the number of competing cells. Note that the winner's output exceeds the combined output of all the other cells.

Given the capacitances at the communication nodes X_Y and X_L, the capacitances at Y_Y and Y_L are chosen to optimize the current conveyor's

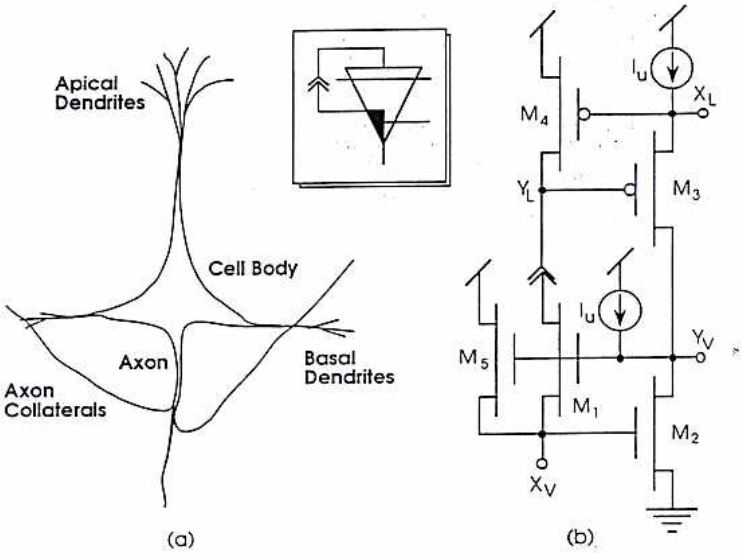


Fig. 17.10. (a) Cortical pyramidal cell. Inputs are received at the apical and basal dendrites; the cell's output is communicated by its axon. (b) Pyramidal cell circuit. An n-type current conveyor communicates inputs and outputs through node X_V and a p-type WTA cell realizes lateral inhibition through node X_L. The circuit symbol includes the current conveyor's control and supply nodes and the WTA's input, as well as the communication nodes.

response and to ensure that the feedback path around the WTA cell has less than unity gain. For the conveyor,

$$\rho_V = \frac{\omega_{X_V}}{\omega_{Y_V}} = \frac{I_{X_V}}{I_{Y_V}} \frac{C_{Y_V}}{C_{X_V}} \tag{11}$$

This relation is used to determine C_{Y_V} given the ratio I_{X_V}/I_{Y_V} and the desired value for ρ_V.

The loop transmission is given by

$$A_L(j\omega) = A_{ZY_L}(j\omega/\omega_{Y_L}) A_{ZY_V}(j\omega/\omega_{X_V})$$

where A_{ZY_L}, the WTA's differential gain, is given by Eq. (11) and A_{ZY_V}, the conveyor's current gain from Y to Z, is given by Eq. (9). These functions have break-points at $\omega_1 = \omega_{Y_L}/A$ and $\omega_2 = \omega_{X_V}/\sqrt{\rho_V}$, respectively. If $\omega_2 \gg \omega_1$, the loop response increases at 20 dB per decade below ω_1, is flat between ω_1 and ω_2, and rolls off 40 dB per decade above ω_2. Between ω_1 and ω_2, the asymptotic approximation for the gain is

$$A_L = \frac{g_{m_1} g_{m_3}}{g_{m_2} g_{m_4}} \frac{\omega_{Y_L}}{\omega_{X_V}} = c \frac{I_{Z_L}}{I_{Y_V}} \frac{C_{X_V}}{C_{Y_L}}$$

where c is the fraction of I_{X_V} that is passed to the WTA cell and $I_{Y_V} = I_{Z_L} + I_u$. Since $I_{Z_L} = \frac{1}{2}(mI_u)$, we have

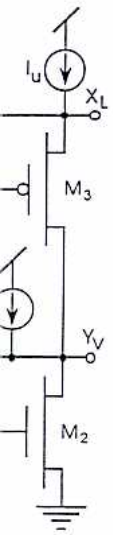
$$A_L = \frac{cm}{m+2} \frac{C_{X_V}}{C_{Y_L}} \tag{12}$$

Given the desired value for A_L and the number of competing cells, m, this relation is used to find C_{Y_L}.

17.4.2. BAM Cell

The BAM cell has two modes of operation: *recall* and *programming*. In the recall mode, the reciprocal junctions mediate interactions between I/O units and hidden units through the $A\pm$ lines and the H line (see Fig. 17.11a). In the programming mode, one synapse is turned on, and the other is turned off, by setting the flip-flop appropriately; the $A\pm$ lines serve as bit lines and the H line is used as a word line. The BAM cell circuit is identical to a conventional SRAM cell except for two extra devices, M_1 and M_3, in the reciprocal junctions. These transistors attempt to write zeroes to unselected cells. Nevertheless, SRAM-like operation may be achieved by sizing the pull-ups.

A selected cell is shown in Fig. 17.11b. M_3, M_6, and M_7 are omitted; these devices are off. M_3 remains off, M_6 and M_7 reinforce the new state when they turn on. Assume the n-type devices are the same size and the saturation current of the pull-ups is R_K times that of the n-types



l dendrites; the cell's output is
nmunicates inputs and outputs
$_L$. The circuit symbol includes
he communication nodes.

$$\frac{cm}{m+2}\frac{C_{X_V}}{C_{Y_L}} \qquad (12)$$

lue for A_L and the number
this relation is used to find

o modes of operation: *recall*
In the recall mode, the
s mediate interactions
d hidden units through the
ine (see Fig. 17.11a). In the
one synapse is turned on,
d off, by setting the flip-flop
± lines serve as bit lines and
a word line. The BAM cell
a conventional SRAM cell
devices, M_1 and M_3, in the
These transistors attempt to
elected cells. Nevertheless,
n may be achieved by sizing

own in Fig. 17.11b. M_3, M_6,
; these devices are off. M_3
M_7 reinforce the new state
Assume the n-type devices
d the saturation current of
times that of the n-types

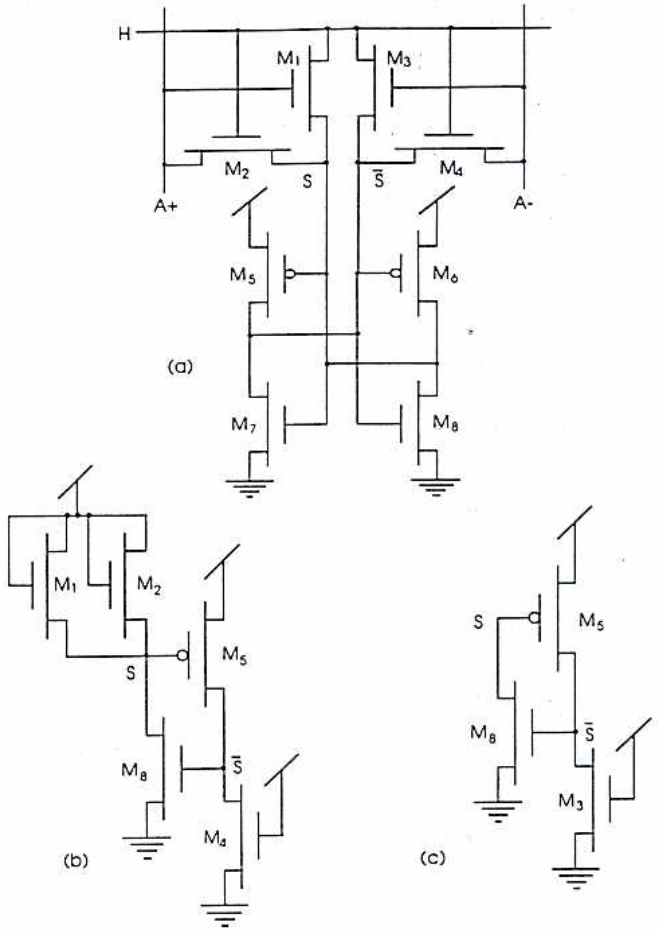


Fig. 17.11. (a) BAM cell circuit. It consists of a pair of reciprocal junctions and a pair of cross-coupled inverters. (b) A selected cell (H high) with $A+$ high and S low. Devices that are off are omitted. The access devices M_1 and M_2 together pull S up to reduce the current in M_5 so that M_4 can bring $\bar{S}$ low. (c) An unselected cell (H low) with both $A+$ and S low. Node S stays at GND, so M_3 cannot overcome M_5.

for equal gate-to-source voltages. Observing that M_1, M_2, and M_5 have the same V_{gs}, it follows that M_5 has $R_K/2$ times the current in M_8. M_4's current, which equals M_8's, will exceed M_5's if $R_K < 2$. An unselected cell is shown in Fig. 17.11c. The extra access device, M_3, attempts to upset the cell. To prevent this, M_5's saturation current must exceed that of M_3; this requires that $R_K > 1$. This analysis also applies if the states of S and $A+$ are reversed—simply interchange the roles of $A+$ and $A-$.

SPICE simulations confirmed that correct operation is possible with ratios that vary by a factor of 2; for 3/3 (width/length) n-type devices, the cell worked correctly for p-type sizes ranging from 7/2 to 13/2. The BAM cell layout is shown in Fig. 17.12. Scalable design rules were used; the separation between p- and n-types is 10λ. The resulting cell area is $38\lambda \times 38\lambda$.

17.5. BAM NETWORK

A BAM network is built using pyramidal cells and BAM cells as shown in Fig. 17.13. The I/O units communicate with external circuitry through their vertical control nodes; unidirectional currents are used as input and the differential voltages serve as outputs. This network finds the hidden unit whose weight vector best matches the input pattern and recalls the association assigned to that unit. The recurrent pathway sustains the new state after external inputs are removed; it also produces hysteresis. Network design involves specifying values for fixed bias currents and pyramidal cell capacitances. SPICE simulations confirmed the network's operation and verified the design procedure; the delays obtained agree with theoretical predictions.

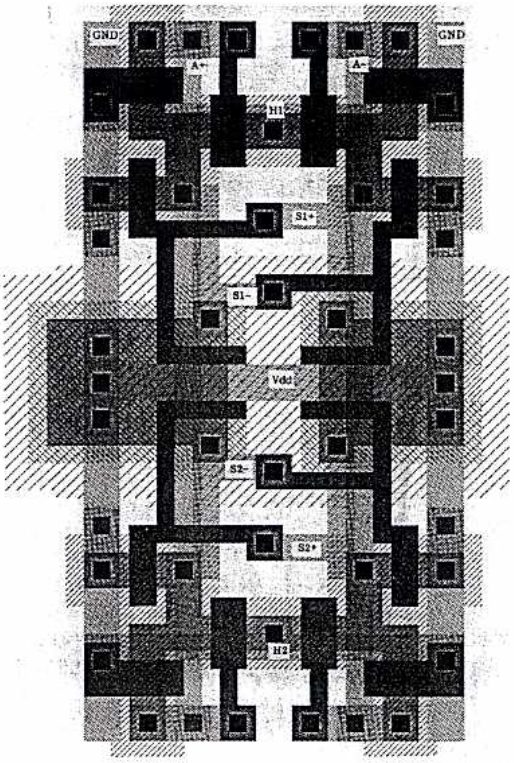


Fig. 17.12. BAM cell layout (two cells are shown). GND, $A+$ and $A-$ run vertically in metal-2; V_{dd} and H run horizontally in metal-1. Contacts to these lines are shared by adjacent cells.

17.5.1. Operation

I/O units supply input to a hidden unit through connecting BAM cells. The winning cell in an I/O unit has an output of $3I_u$; the complementary cell's output is I_u. Hence, if an I/O unit's state agrees with the BAM cell's, it contributes 3 (in units of I_u), else its contribution is only 1. Accordingly, the hidden unit whose weight vector *best* matches the I/O layer's state gets the largest input. This unit is picked by winner-takes-all competition.

Hidden units feed their output to either the excitatory or the inhibitory input of an I/O unit, depending on the connecting BAM cell. If the winning unit's BAM cell is in the $+1(-1)$ state, it contributes $(r+1)I_u$ to the excitatory (inhibitory) input, where r is the number of hidden units. In contrast, no more than $(r-1)I_u$, the total output of the remaining units, is supplied to the complementary input. Accordingly, the I/O layer's state *exactly* matches the winning hidden unit's weight vector.

Analog-valued input patterns are presented by supplying currents to the I/O units. For positive and negative values, unidirectional currents are applied to excitatory and inhibitory pyramidal cells, respectively. Their outputs equal the sum of these currents and the intrinsic ones (from WTA cells). Therefore, the input pattern and the currently recalled one are both projected to the hidden layer. The network changes states only if the former is larger in magnitude.

Specifically, let $\mathbf{A}_0$ be the currently recalled vector, $t\mathbf{A}_1$ be the input vector, pumped up t times, and $\mathbf{A}_2$ be the weight vector that best matches $\mathbf{A}_1$. These vectors have ± 1 components; their units are given by the outputs of the I/O units' WTA cells, i.e., $2I_u \cdot \mathbf{A}_1$ forces a transition from $\mathbf{A}_0$ to $\mathbf{A}_2$ if $\mathbf{A}_2^T(\mathbf{A}_0 + t\mathbf{A}_1) > \mathbf{A}_2^T(\mathbf{A}_0 + t\mathbf{A}_1)$ or

$$t(d_{01} - d_{12}) > d_{02}$$

where d_{jk} $(=d_{kj})$ is the Hamming distance

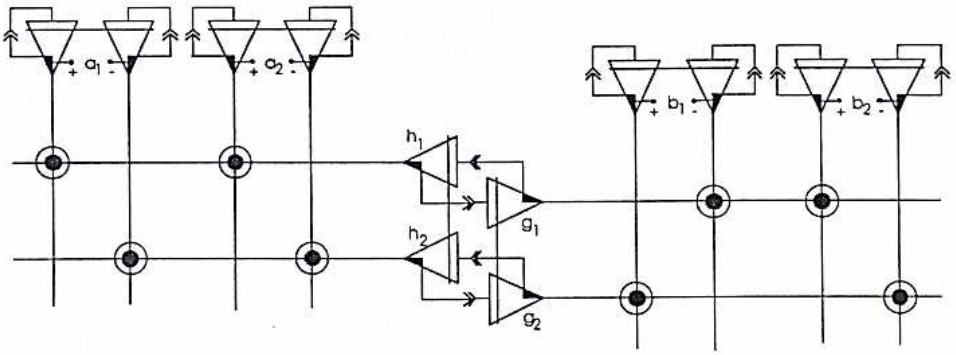


Fig. 17.13. A BAM network with two units in each layer. I/O units have two competing pyramidal cells that compare their excitatory and inhibitory inputs; hidden units have just one and compete with each other for output. A hidden unit's input is buffered by a unit in the other hidden layer. BAM cells connect these units together; only junctions that are "on" are shown. The associations stored are $([+1\ +1]^T, [-1\ +1]^T)$ (first row) and $([-1\ -1]^T, [+1\ -1]^T)$ (second row).

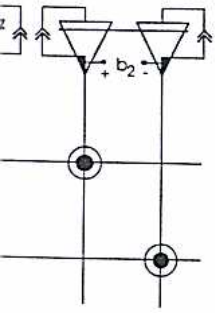
l cell's, it contributes 3 (in
s contribution is only 1.
en unit whose weight vector
layer's state gets the largest
picked by winner-takes-all

their output to either the
bitory input of an I/O unit,
nnecting BAM cell. If the
cell is in the $+1(-1)$ state,
I_u to the excitatory (inhibi-
the number of hidden units.
e than $(r-1)I_u$, the total
ng units, is supplied to the
t. Accordingly, the I/O
natches the winning hidden

t patterns are presented by
the I/O units. For positive
unidirectional currents are
and inhibitory pyramidal
eir outputs equal the sum
the intrinsic ones (from
, the input pattern and the
are both projected to the
vork changes states only if
magnitude.
be the currently recalled
put vector, pumped up t
e weight vector that best
tors have ± 1 components;
by the outputs of the I/O
$2I_u \cdot A_1$ forces a transition
$A_0 + tA_1) > A_2^T(A_0 + tA_1)$

$d_{12}) > d_{02}$

the Hamming distance



npeting pyramidal cells that
e with each other for output.
ect these units together; only
) (first row) and $([-1\ -1]^T$,

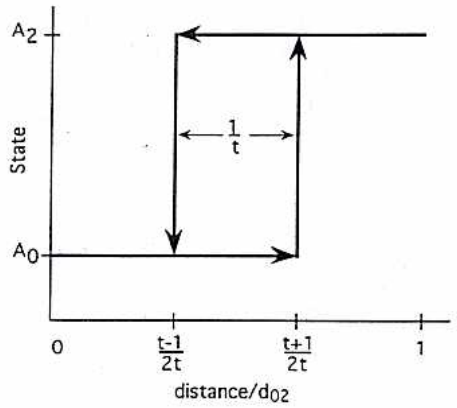


Fig. 17.14. Network recall hysteresis. Two possible states $\mathbf{A}_0$ and $\mathbf{A}_2$ are shown on the vertical axis; d_{02} is the Hamming distance between them. The Hamming distance of the input vector from $\mathbf{A}_0$ is shown on the horizontal axis; its components are t times greater than inputs fed back by the recurrent pathway.

between $\mathbf{A}_j$ and $\mathbf{A}_k$. This condition cannot hold if $d_{12} > d_{01}$ or $t \leq 1$ ($|d_{01} - d_{12}| \leq d_{02} \leq d_{01} + d_{12}$). If $\mathbf{A}_1$ is initially equal to $\mathbf{A}_0$ and, one by one, the bits that differ between $\mathbf{A}_1$ and $\mathbf{A}_2$ are flipped, a state transition occurs when

$$\frac{d_{12}}{d_{01}} = \frac{t-1}{t+1}$$

For instance, when $t = 3$, the input's distance from the target must be half that to the current vector. In the absence of feedback, however, these distances need only be equal. Effectively, the recurrent pathway introduces hysteresis over a region $(1/t)d_{02}$, as shown in Fig. 17.14. It also biases recall in favor of a vector similar to the current one. This behavior is useful in applications where the network interfaces directly with analog environmental sensors.

17.5.2. Design

A network with n units per I/O layer and r units per hidden layer is connected by two $n \times r$ BAM cell arrays. Each BAM cell supplies 0, I_u, or $(r+1)I_u$ to an I/O unit and I_u or $3I_u$ to a hidden unit; it also adds capacitances C_A and C_H to their communication nodes. Therefore, an I/O unit receives 0 to $2rI_u$ and has rC_A capacitance at its vertical communication node; hidden units get between nI_u and $3nI_u$ and have nC_H. Knowing the input ranges and the communication node capacitances, fixed bias currents and control node capacitances may be specified.

Table 17.1. Control node capacitances for pyramidal cells. n and r are the number of units in the I/O and hidden layers; c is the fraction of the input passed to the WTA cell; C_A and C_H are the capacitances per BAM cell on the A and H lines.

Unit	I/O	Hidden
C_{Y_V}	$3C_A/2$	rC_H
C_{Y_L}	crC_A	$2cnC_H$

Choosing a fixed bias of rI_u for the hidden layers and zero for the I/O layers gives a network loop gain of two-thirds (from Eq. 9). The first choice limits the variation of the I/O units' inputs to a factor of 3—from rI_u to $3rI_u$. The hidden units' inputs vary over a similar range: from nI_u to $3nI_u$. The second choice was made to avoid adding current sources at the hidden layer; these sources would have to be well matched to avoid producing errors in picking the maximum.

The chosen control node capacitances are given in Table 17.1. For the I/O units, the ratio I_{X_V}/I_{Y_V} varies from $\frac{2}{3}r$ to $2r$ and $C_{X_V} = rC_A$. So, with $C_{Y_V} = \frac{3}{2}C_A$, ρ_V varies between 1 and 3; see Eq. (11). Choosing $C_{Y_L} = crC_A$ gives a loop gain of one-half since $m = 2$ for the I/O units; see Eq. (12). For the hidden units, I_{X_V}/I_{Y_V} ranges from $2n/(r+1)$ to $2n$, assuming a winning unit receives at least $2nI_u$ and a losing one gets at most $2nI_u$. Given that $C_{X_V} = nC_H$, choosing $C_{Y_V} = rC_H$ means ρ_V varies from 2 to $2r$, assuming $r \gg 1$; see Eq. (11). Choosing $C_{Y_L} = 2cnC_H$ gives a loop gain of less than one-half; see Eq. (12).

Unfortunately, except for C_{Y_V} in the I/O units, the control node capacitances must be scaled up as the network's size increases. MOS capacitors are used on the chip; the control node potential is large enough to strongly invert the channel. In 2-μm CMOS technology with 400-Å gate oxide, the MOS capacitor has about 0.9 fF/μm², $C_A \approx$ 20 fF, and $C_H \approx 30$ fF. Therefore, each BAM cell adds 55 μm² to the area of the capacitors; this represents a 4 percent overhead.

17.5.3. Speed

Delays arise from the conveyors' incoming and outgoing channels, and the WTA cells. The former may be modeled as first-order low-pass with corner frequencies at ω_X and $\omega_X/\sqrt{\rho_V}$, respectively. Their rise times are given by $t_r = 2.2/\omega_c$, where ω_c is the corner frequency. So

the delay due to the conveyor is

$$t_{cc} = 2.2(1 + \sqrt{\rho_V})\frac{C_{X_V}}{g_{m_1}} = 2.2(1 + \sqrt{\rho_V})\frac{V_T}{\kappa v}$$

where $v = I_{X_V}/C_{X_V}$ is the slew rate; it is between I_u/C_A and $3I_u/C_A$ for the I/O units (replace C_A with C_H for the hidden units) and is independent of the network's size. To obtain the longest delay, let $v \equiv I_u/C_A$. For $I_u = 40$ nA and $C_A = 20$ fF, $v = 2\text{V}/\mu\text{S}$, and with $\rho_V = 2$, the delay is only 92 nsec! Such speed is possible because of the small voltage signals, typically about 50 mV.

For the WTA, the current available to slew the input voltage is the difference between the input and the present maximum. This signal, ΔI, must produce a two-diode-drop voltage swing, ΔV. For the I/O units the capacitance is $C_{Y_L} = crC_A$ and, typically, $\Delta I = crI_u$, so the slew rate is v. Since ΔV is about 2 V, the delay, $t_{WTA} = \Delta V/v$, is 1 μsec. The same relationship holds for the hidden units with $v = I_u/2C_H$. WTA circuits that require only 200 mV swings are being developed to reduce the delay to 100 nsec.

Summing delays due to the conveyor and the WTA gives 1.1 μsec and 2.1 μsec for I/O and hidden units, respectively. Therefore, signals take 6.4 μsec to propagate around the network. The external inputs are removed after this period and the outputs may be read. The recall time is dominated by the WTA's delay; faster circuits could reduce it to 1 μsec, with 40 nA unit current.

17.5.4. Simulations

The network in Fig. 17.13 was simulated using SPICE;[3] the results are shown in Fig. 17.15. The SPICE deck was automatically extracted from the layout and included all the parasitic capacitances; all devices were 4 μm × 4 μm. Pyramidal cell capacitances were determined according to the previous section, with $C_A = 37.5$ fF and $C_H = 51.5$ fF; these values include capacitances due to the I/O and hidden units themselves. The unit current was 40 nA and the I/O units received a fixed bias of 80 nA; the supply was 5.0 V.

Initially, g_1 and h_1 are winning and the states of a_1 and b_1 are $+1$ and -1, respectively. At $t = 5$ μsec, 300 nA is applied to a_1's inhibitory pyramidal cell. To see what happens, start at the top left-hand corner of Fig. 17.15 and follow the arrows; refer to Fig. 17.13 for the connectivity.

Moving down the left side, the top graph shows a_1's output voltages; its inhibitory output increases (dotted line). The reciprocal junction connecting a_1 to a_2 replicates the signal and g_2's input increases at $t = 5.06$ μsec (dotted line in next graph). Now, g_2's input exceeds g_1's and their output voltages change to reflect this (next graph). The reciprocal junctions send g_1's and g_2's outputs to b_1's inhibitory and excitatory inputs, respectively; its inputs change at $t = 7.26$ μsec (next graph).

Moving up the right side, the bottom graph shows b_1's outputs. They are sent to the H units by the same reciprocal junctions; inhibitory and excitatory outputs go to h_1 and h_2, respectively. At $t = 9.25$ μsec, h_2's input increases and h_1's decreases (next graph). Now h_2 starts winning and the H units' outputs change accordingly (next graph). a_1's excitatory and inhibitory inputs come from h_1 and h_2, respectively; they change at $t = 11.55$ μsec (next graph).

Back at the beginning, a_1's outputs respond and are sent to layer G at $t = 13.56$ μsec; it took 8.50 μsec for signals to propagate around the network. g_2's input increases further while g_1's decreases. Therefore, g_2 continues to win after the external input is removed at $t = 15$ μsec. The I/O units have a 2.0-μsec delay, and the hidden units have 2.3 μsec. Theoretically, with v equal to 1.06 V/μsec and 0.78 V/μsec, for I/O and hidden units, respectively, their current conveyor delays are 140 nsec and 195 nsec. For the WTAs, $\Delta V = 1.9$ V, $\Delta I = 40$ nA, with $c = \frac{1}{2}$, and $C_{Y_L} = 41$ fF, which gives a delay of 1.95 μsec. The hidden units have $\Delta I = 80$ nA and $C_{Y_L} = 88$ fF; this gives 2.07 μsec. Therefore, the expected delays are 2.09 μsec and 2.26 μsec; these figures are in agreement with the SPICE results.

Observe that pairs of signals in the same row of Fig. 17.15 are actually communicated through the same node. Therefore, a sudden change in either signal produces a glitch in the other. When the voltage changes, capacitive currents are produced; this explains the positive spike in a_1's inhibitory input at $t = 5$ μsec. When the current changes, the voltage deviates while the conveyor accommodates the new input; this explains the negative spike in h_2's output at $t = 5$ μsec. Notice that this spike is replicated in a_1's input, producing the opposing spike.

The differential outputs at the I/O unit's vertical control nodes are shown in Fig. 17.16; ±75 mV differential signals appear at the outputs depending on the unit's state. The transitions are caused by changes in the WTA's outputs and the network's inputs, except for those at $t = 5$ μsec and $t = 15$ μsec, which are due to the extrinsic input. Initially, the A layer's state is $[+1\ +1]^T$ and the B layer's is $[-1\ +1]^T$; recalling the association in the first row. After the external

[3] Berkeley SPICE, version 3C1, with the BSIM model (Level 4).

plicates the signal and g_2's
= 5.06 μsec (dotted line in
's input exceeds g_1's and
change to reflect this (next
l junctions send g_1's and
inhibitory and excita-
ely; its inputs change at
ph).
t side, the bottom graph
ey are sent to the H units
junctions; inhibitory and
to h_1 and h_2, respectively.
input increases and h_1's
). Now h_2 starts winning
s change accordingly (next
nd inhibitory inputs come
ectively; they change at
aph).
ing, a_1's outputs respond
at $t = 13.56$ μsec; it took
o propagate around the
creases further while g_1's
g_2 continues to win after
moved at $t = 15$ μsec. The
sec delay, and the hidden
eoretically, with v equal to
/μsec, for I/O and hidden
r current conveyor delays
5 nsec. For the WTAs,
nA, with $c = \frac{1}{2}$, and
s a delay of 1.95 μsec. The
= 80 nA and $C_{Y_L} = 88$ fF;
Therefore, the expected
d 2.26 μsec; these figures
the SPICE results.
f signals in the same row
y communicated through
fore, a sudden change in
glitch in the other. When
capacitive currents are
the positive spike in a_1's
5 μsec. When the current
viates while the conveyor
v input; this explains the
tput at $t = 5$ μsec. Notice
eplicated in a_1's input,
g spike.
tputs at the I/O unit's
are shown in Fig. 17.16;
nals appear at the outputs
state. The transitions are
e WTA's outputs and the
t for those at $t = 5$ μsec
are due to the extrinsic
ayer's state is $[+1+1]^T$
$[-1+1]^T$; recalling the
row. After the external

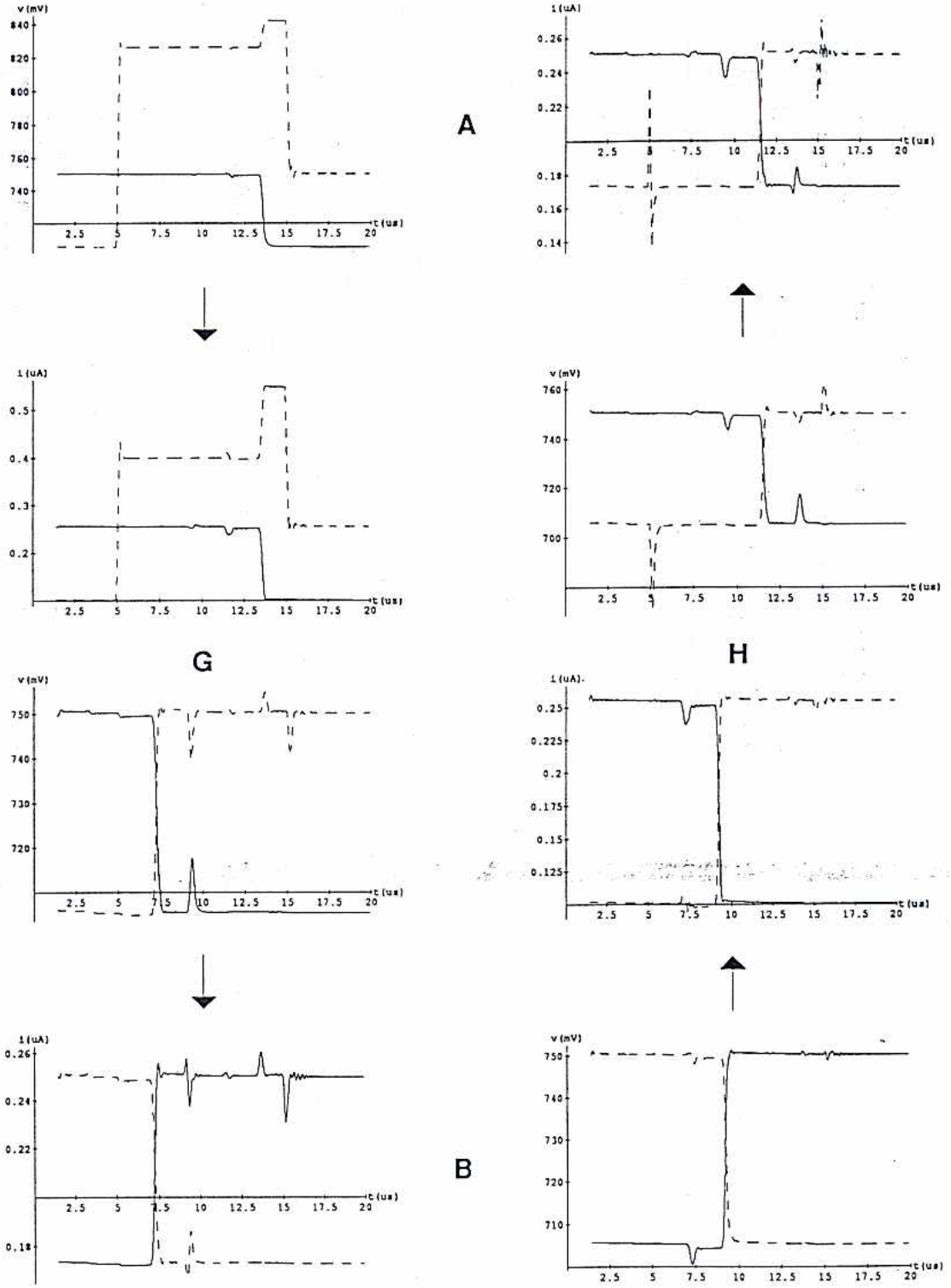


Fig. 17.15. SPICE simulation. Input currents and output voltages of I/O units a_1 and b_1 are at the top and the bottom, respectively; their inhibitory signals are in dotted lines. The hidden units are in the middle; g_2's and h_2's signals are in dotted lines.

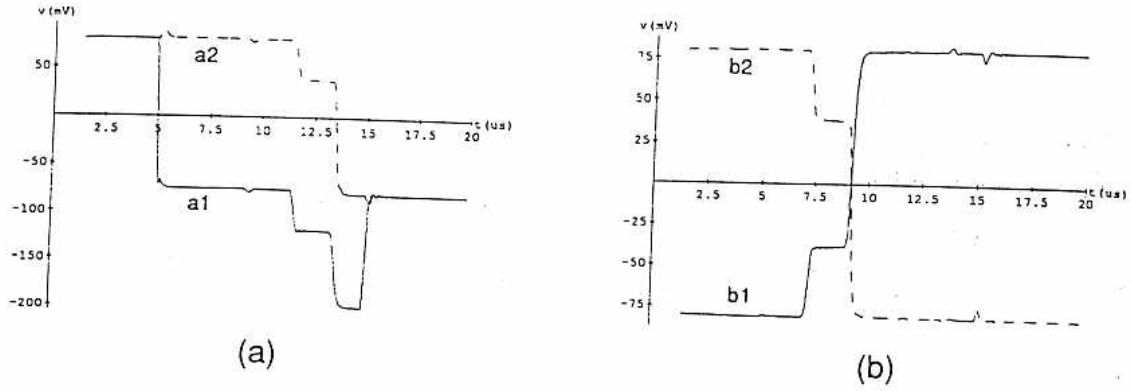


Fig. 17.16. SPICE simulation (continued). Differential voltages measured at the vertical control nodes of the I/O units. (a) A units, (b) B units.

input is removed they are $[-1\ -1]^T$ and $[+1\ -1]^T$, respectively; recalling the association stored in the second row.

17.6. SUMMARY

A design for a bidirectional associative memory chip has been developed; this second-generation design evolved from our first attempt (Boahen et al., 1989a) and from the experience of others (Sivilotti et al., 1985; Graf and de Vegvan, 1987; Verleysen et al., 1989). A complete test chip including latches for reentrant programming and custom digital I/O pads that interface directly with the I/O units is now in fabrication. This chip has 11 224 transistors and 40 pads; the die area is 2.2 mm × 2.2 mm in 2-μm technology.

Architectural issues were addressed at two levels. First a network architecture with good storage and recall performance was proposed. Second, an efficient, scalable, and fault-tolerant chip architecture was developed. Large-scale integration of analog circuits was deal with at three levels: signal representations, devices, and circuits. Representations that use interconnects efficiently and minimize the effect of their capacitance were chosen. A region of device operation with minimal power consumption and high transconductance was chosen. Minimalistic circuits were sought to perform communication and computation in a large network.

The effect of device mismatch in large associative memories and the related issue of optimum device geometry and placement need to be addressed. Random fluctuations in the device parameters are bound to average out across the synaptic arrays; however, the units themselves may have to be scaled up to obtain more precision. The effect of systematic variations across the chip also needs to be investigated. Characterization of these process variations (Pavasović, 1990) is invaluable in this respect.

ACKNOWLEDGMENTS

This research was funded by the Independent Research and Development program of the Applied Physics Laboratory; we thank Robert Jenkins for his personal interest in this work. Aleksandra Pavasović kindly provided the device data and Philippe Pouliquen made several useful suggestions.

REFERENCES

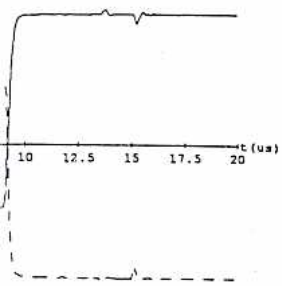
Andreou, A. G. and Boahen, A. K. (1989). "Synthetic Neural Circuits using Current-Domain Signal Representations," *Neural Computation*, Vol. 1. MIT Press, Cambridge, MA, pp. 489–501.

Boahen, K. A., Pouliquen, P. O., Andreou, A. G., and Jenkins, R. E. (1989a). "A Heteroassociative Memory using Current-Mode Circuits and Systems," *IEEE Trans. Circuits*, **36**(5), 747–755.

Boachen, K. A., Andreou, A. G., and Pouliquen, P. O. (1989b). "Architectures for Associative Memories using Current-Mode Analog MOS Circuits," *Advanced Research in VLSI: Proc. Dec. Caltech Conference on VLSI*, C. L. Seitz, ed. MIT Press, Cambridge, MA.

Graf, H. P. and de Vegvar, P. (1987). "A CMOS Implementation of a Neural Network Model," *Advanced Research in VLSI: Proc. of the 1987 Stanford Conference*, P. Losleben, ed. MIT Press, Cambridge, MA.

Kosko, B. (1988). "Bidirectional Associative Memories," *IEEE Trans. Systems, Man, and Cybernetics*, **18**, 49–60.



(b)

al control nodes of the I/O

f systematic variations
eeds to be investigated.
ese process variations
aluable in this respect.

ed by the Independent
ment program of the
atory; we thank Robert
l interest in this work.
ndly provided the device
quen made several useful

n, A. K. (1989). "Synthetic
rent-Domain Signal Repre-
utation, Vol. 1. MIT Press,
–501.
P. O., Andreou, A. G., and
Heteroassociative Memory
cuits and Systems," *IEEE*
7–755.
A. G., and Pouliquen, P. O.
for Associative Memories
log MOS Circuits," *Advan-*
oc. Dec. Caltech Conference
MIT Press, Cambridge, MA.
ar, P. (1987). "A CMOS
Neural Network Model,"
VLSI: Proc. of the 1987
Losleben, ed. MIT Press,

tional Associative Memo-
ms, Man, and Cybernetics,

Lazzaro, J., Ryckebusch, S., Mahowald, M. A., and Mead, C. A. (1989). "Winner-Take-All Networks of O(n) Complexity," in *Advances in Neural Information Processing Systems*, Vol. 1, D. S. Touretzky, ed. Morgan Kaufmann, San Mateo, CA, pp. 703–711.

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

Pavasović, A. (1990). "Subthreshold Operation of MOSFET Devices in Analog VLSI Circuits," Ph.D. Dissertation, Johns Hopkins University.

Peretto, P. and Niez, J. J. (1986). "Long Term Memory Storage Capacity of Multiconnected Neural Networks," *Biological Cybernetics*, **54**, 53–63.

Säckinger, E. and Guggenbühl, H. (1990). "A High-Swing, High Impedance MOS Cascode Circuit," *IEEE Solid-State Circuits*, **25**(1), 289–298.

Sivilotti, M., Emerling, M., and Mead, C. A. (1985). "A Novel Associative Memory Implemented using Collective Computation," in *Advanced Research in VLSI: Proc. 1985 Chapel Hill Conf.*, H. Fuchs, ed. Computer Science Press, Rockville, Maryland.

Smith, K. C. and Sedra, S. A. (1968). "The Current Conveyor—A New Circuit Building Block," *IEEE Proc.*, **56**, 1368–1369.

Verleysen, M., Sirletti, B., Vandemeulebroecke, A. M., and Jespers, P. G. (1989). "Neural Networks for High-Storage Content-Addressable Memory: VLSI Circuit and Learning Algorithm," *IEEE J. Solid-State Circuits*, **24**(3), 562–569.

Associative Neural Memories

Theory and Implementation

Edited by

MOHAMAD H. HASSOUN

New York Oxford
OXFORD UNIVERSITY PRESS
1993