# Analog VLSI for Neural Networks

*Robert W. Newcomb and Jason D. Lohn*

## Introduction

One of the most promising strategies for implementing neural networks is through the use of electronic analog VLSI (Very Large Scale Integration) circuits. An analog circuit is one that processes a continuum of real-valued signals in continuous time, in contrast to a digital circuit which processes integer-valued (most often binary) signals in discretized time. VLSI refers to an integrated circuit design and manufacturing technology whereby hundreds of thousands to millions of active components (most often transistors) are placed on a chip on the order of 100 mm$^2$ in area and 0.5 mm thick. Because Artificial Neural Networks (ANNs) attempt to behave similarly to the brain with its millions of neurons, VLSI is the most appropriate presently available technology for their hardware implementations. Furthermore, both VLSI circuits and biological neurons are of the same class, that is, fundamentally analog.

Although during the 1980s digital ANNs held more interest, the first neural-like circuits were analog ones constructed by Dr. Otto Schmitt in the late 1930s using vacuum tube analog computer circuits (Schmitt, 1937). These were extended to rather cumbersome transistor circuits after the Second World War to obtain artificial neurons, in which much of the emphasis was placed on the initiation and propagation of action potentials. The circuits developed to accomplish these tasks relied on nonlinearities for implementing amplitude saturation, pulse repetition saturation, threshold effects, and dynamics for effecting time-domain changes on the action potentials (Reiss, 1964). But, because of the large size of the circuits used for just one neuron, very little was done to make full ANN systems until the advent of integrated circuits (ICs). In a number of research centers around the world during the mid-1960s, considerable interest began to develop in the design of analog IC neurons and systems built from them. Of importance to the signal processing capabilities in this development has been the recent emphasis on the synaptic combining of signals via weight-matrix summations, as opposed to the axon propagation of action potentials. Present analog ANNs consist of synaptic weights, implemented by amplifier gains; summation of the weighted signals, implemented by the use of Kirchhoff's laws (most conveniently the current law, denoted KCL); activation functions, realized by amplifier nonlinearities; and in many cases dynamics, via capacitors, for smoothly transitioning from an initial state to a desired equilibrium.

The workhorses of analog VLSI ANNS are the Differential Voltage Controlled Current Source (DVCCS) and capacitors. The DVCCs is used for making synaptic weights and activation functions, and capacitors are used for dynamics. A DVCCS takes a voltage difference as input, and gives an output current as a function of that difference. DVCCS gains can realize the weights when operating on small signals in a linear fashion and can also realize saturation nonlinearities when operating on large signals. In both cases, the DVCCS output currents can be conveniently summed by KCL. By using capacitors in conjunction with DVCCSs, any linear circuit can be realized (Bialko and Newcomb, 1971), so that any desired filtering of ANN signals is available. Along with the DVCCS and the capacitor, it is also convenient to have resistors for conversion of currents to voltage and voltage divisions, as well as devices for creating and scaling currents (called current sources and current mirrors, respectively). Except for passive resistors and capacitors (both of which are generally avoided in VLSI because of large area or nonideal characteristics), all of these devices can basically be constructed from VLSI transistors, which are discussed later in this article.

## Overview of an Analog ANN Implementation

First we present a complete ANN analog circuit to give an overview of the circuits discussed in later sections. Figure 1

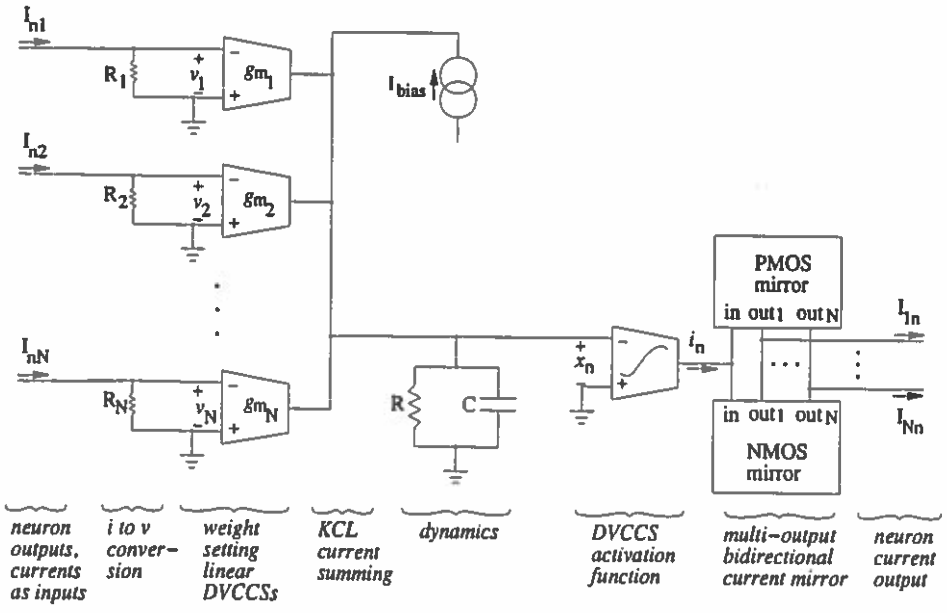


**Figure 1.** The $n$th neuron for a VLSI continuous-type ANN of $N$ neurons.

shows an analog circuit suitable for the VLSI realization of the continuous-time neuron equations which, for the $n$th neuron of a set of $N$, are (for simplicity of notation we omit subscripts $n$ on device parameters but not on the output currents and state)

$$C dx_n/dt + G x_n = \sum_{i=1,\ldots,N} g_{mi} \cdot R_i I_{ni} + I_{bias} \quad n = 1, \ldots, N \quad \text{(1a)}$$

$$i_n = g(x_n) \quad \text{(1b)}$$

In Equation 1, $g(\cdot)$ is any of the activation functions available (see Equations 2–4), with $i_n$ being its current output; $x_n$ is the $n$th neuron's state variable; $I_{ni}$ is the current output of the $i$th neuron, which is fed to the input of the $n$th neuron; the $g_{mi} \cdot R_i$ are the synaptic weights; and $I_{bias}$ is the bias input. With reference to Figure 1 and Equation 1, it will be shown below that VLSI circuits can be constructed to make this Hopfield class of analog neural networks, as well as any other analog ANN, such as ART 2 (see ADAPTIVE RESONANCE THEORY), pulsed Hebbian, and biological mimics. The nonlinear function $g(\cdot)$ in Equation 1b can be realized via a DVCCS (see Figure 5) exhibiting square-law, exponential, or sigmoidal processing. In this simple model of a neuron, these nonlinearities can be thought to correspond to the activation processes in the cell body. The weighted inputs from the synapses to the cell body can be thought to correspond to varying amounts of currents linearly summing, via KCL, at the input node to the left of the activation function DVCCS in Figure 1. On the right side of Equation 1a the weights $g_{mi} \cdot R_i$ are the current gains of DVCCSs operating as linear amplifiers with resistor inputs, the resistors (of resistance $R_i$) being used to convert the neuron output currents to voltages (constructed using direct layouts, see Figure 3, or DVCCS connections, depending upon their Ohmic value). The bias input $I_{bias}$, also on the right of Equation 1a, is constructed as a constant current source made of a transistor. Using a resistor-capacitor branch connected to this same input node of the activation function amplifier, we obtain the dynamics of the analog circuit, shown on the left side of Equation 1a incorporating the derivative. Because each neuron output current $i_n$ (which can be positive or negative) needs to be sent to each of the other neurons, it needs to be repeated $N$ times, this being accomplished by the bidirectional current mirror (see Figure 4B) in multioutput form on the right of Figure 1. This simply reproduces $N$ copies of the neuron output current $i_n$ irrespective of what load is presented to it. For adjustments, as may be needed for adaptive ANNs, the transconductances $g_{mi}$ can be made voltage variable by variation of the gain of the associated DVCCS (via the tail current introduced in Figure 5 below).

**Transistors and VLSI Layouts**

The key circuit component in analog VLSI is the transistor, a three- or four-terminal device that can behave as a switch in digital circuits and as an amplifier in analog ones. Transistors may be fabricated using a variety of technologies: BJT (bipolar junction transistor), MOS (metal oxide semiconductor), CMOS (complementary MOS), and others. For neural network implementations, MOS and BJT devices have been used the most; when both occur together, the process is called BiCMOS and is the most prevalent present-day analog VLSI technology.

Figure 2 shows the circuit symbols for those transistors of most interest to ANN VLSI, along with a top view of an IC layout of each. The fabrication details can be found in Geiger, Allen, and Strader (1990); however, for our purposes, it is enough to know only a few aspects of their operation.

In the MOS transistor the drain current, $I_D$, which flows from the outside into the drain, D, and then through the device to the source, S, is controlled by the voltage at the gate, G, with respect to the source, $V_{GS}$, when the latter is "above" threshold voltage, $V_{th}$. As the threshold voltage can be used as a fine control on the ANN weights, we note that it is dependent on the bulk-to-source voltage, $V_{BS}$, where the bulk material (B) is that of the substrate into which the transistor is embedded. The two types of MOS transistors, NMOS and PMOS, are distinguished by their conduction mechanisms, with the currents and voltages of the latter being ideally the negative of the former in the complementary case desired for CMOS fabrications. Since the channel can be formed by enhancing or depleting charge, we have enhancement- and depletion-mode transistors of each of the NMOS and PMOS types; the distinction is that the threshold voltages of depletion-mode devices are generally of opposite sign to those of enhancement-mode devices. Depletion-mode transistors are not as common in analog VLSI because of the extra fabrication steps needed, but they can be used to obtain more flexible designs. MOS transistors can be operated such that between the drain and source a resistor is seen whose value depends on the gate-to-source voltage, giving a voltage-variable resistor useful for adaptation. More commonly the MOS transistor is operated in its saturation mode where, instead of a resistor between drain and source, a current source is seen. This current source depends on the gate-to-source voltage in a square-law fashion, conveniently allowing for quadratic weights. By operating an MOS transistor at very low (subthreshold) gate-to-source voltages, exponential behavior is obtained; subthreshold operation is convenient for low power designs but is not too robust. In

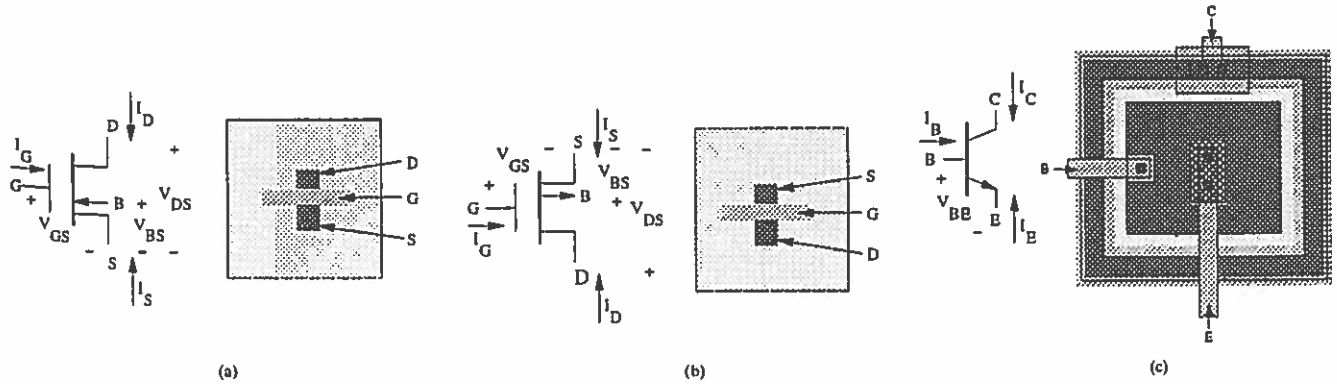


**Figure 2.** VLSI transistors and layouts. *A*, NMOS. *B*, PMOS. *C*, NPN BJT.

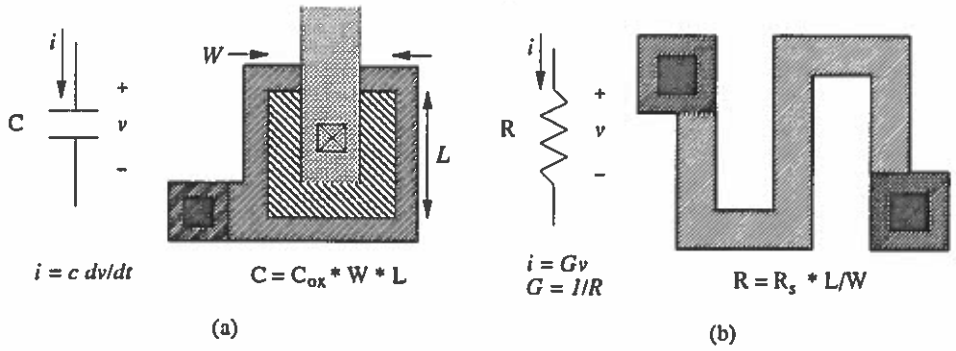


**Figure 3.** Passive components and layouts. *A*, MOS capacitor. *B*, Snake resistor.

all of these cases, the drain current is proportional to the width-to-length ratio, $W/L$, of the channel; this acts as a design parameter that is very easily set in a VLSI layout. For bipolar transistors, exponential nonlinearities are obtained, and the VLSI design parameter is the emitter area to which the collector current is proportional. For more details on these devices, see Geiger et al. (1990).

Besides the active transistors, passive capacitors are used to obtain dynamics needed for realization of those ANNs that are described by differential equations, and for the derivatives needed for backpropagation. Several types of capacitors are available in VLSI. The primary one is realized by an oxide between two conductive layers (presently polysilicon, but possibly metal on top and doped silicon on the bottom), as shown in Figure 3A. These capacitors are linear time-invariant capacitors and satisfy $i = Cdv/dt$ with capacitance $C = C_{ox}WL$ for which $C_{ox}$ is the capacitance per unit area of the (gate) oxide used. The area, $WL$, of the polysilicon plate serves as a design constant. Unfortunately, to obtain capacitance of useful values requires considerable area, and, hence, capacitors often take up a good portion of analog VLSI neural networks. At times, one also needs linear resistors, for transformation of currents to voltages or for biasing, in which case the most common means of VLSI implementation is via strips of polysilicon, often in snake form to optimize layout (Figure 3B). The conductance $G$ is given through the sheet resistance $R_s$ (in ohms/square micron, a material constant) by $G = W/(LR_s)$, in which $L$ is the length (distance between contact pads) and $W$ is the width of the polysilicon. These resistors also take up considerable area and thus are avoided, but for small values of resistance they are sometimes invaluable (for values of 10 to 100 ohms). Larger-valued resistors are constructed using transistors.

### Primary Circuits

Two of the key components in an ANN are the weights and the nonlinear activation functions. A weight can be realized by a DVCCS operating in its linear region, while a sigmoidal nonlinear activation function can be realized by operating a DVCCS over its full nonlinear range. We consider, as background, current sources, current mirrors, and resistors constructed as diode-connected transistors. These are all used for biasing the transistors, that is, setting the modes of operation of transistors, while the current mirrors and sources are used for various adjustments, as in adapting weights.

A current source can be constructed as the drain to source, of current $I$, of an MOS transistor operating in its saturation region (see Geiger et al., 1990, p. 49) with a voltage source of voltage $V$ attached gate-to-source. We note that (1) the current I can be adjusted by varying the above-threshold voltage $V$; (2) current sources of one polarity are changed into ones of opposite polarity by reversing the attachment points or by interchanging NMOS and PMOS; and (3) one needs to maintain the saturation mode of operation (by application of sufficient voltage across the current source nodes). If the transistor is operated in its ohmic region (see Geiger et al., 1990:49), with small $V_{DS}$, then the same circuit gives a voltage-variable resistor of conductance $G(V)$, which is useful for making small-area resistors (10–1000 ohms) as well as adaptive adjustments.

Figure 4A shows current mirrors which allow the current in one section of an ANN to determine that in another, perhaps for adjusting weights. These mirrors use a diode connection of one transistor to set the gate-source voltages for the input and output transistors to be equal. The current mirrors of Figure 4A allow current to flow in only one direction. However, by placing a P-mirror on top of an N-mirror, as in Figure 4B, we can get a bidirectional current mirror. By replacing all of the MOS devices by BJTs, similar BJT devices can be constructed. Furthermore, by placing several output transistors on one input transistor, multiple-output current mirrors are easily constructed, and these are of considerable use for distributing current in current-mode VLSI ANNs (as in Figure 1).

Figure 5 shows the basic configuration of a DVCCS. The tail current, $I_T$, is steered between $I_1$ and $I_2$ by the differential pair consisting of identical transistors $T_1$ and $T_2$ (of NMOS or NPN types), with the steering controlled by the voltage difference of the input voltages, $V_d = V_1 - V_2$. The difference of the transistor currents, $I_d = I_1 - I_2$, is designed to be a function of $V_d$ and $I_T$, independent of any loads or the current mirror. To obtain the current output as this difference, the current mirror is used along with KCL at the output node so that $I_{out} = -I_d$. The function of $I_{out}$ versus $V_d$ realized depends upon the NMOS or NPN transistors and their modes of operation used to form the current difference. In all cases, the gates/bases are the leads to the left (in $T_1$) and right (in $T_2$), the drains/collectors are at the top, and the sources/emitters are at the bottom. In practice, there is some loading by whatever is attached, in which case other current mirrors are attached for isolation.

For possible nonlinearities of $I_d$ versus $V_d$, there are several design alternatives. For NMOS we can obtain the sigmoidal function

$$I_d = (KW/L)[(2I_T/(KW/L)) - V_d^2]^{1/2} V_d \qquad V_d^2 < [I_T/(KW/L)] \tag{2}$$

which is linearized to

$$I_d = g_m V_d \qquad g_m = [2I_T(KW/L)]^{1/2} \tag{3}$$

Typical orders of magnitude are $O(I_T) = 10^{-3}$, $O(K) = 10^{-4}$, $10^{-2} < O(W/L) < 10^{2}$, giving $10^{-5} < O(g_m) < 10^{-2}$ over a

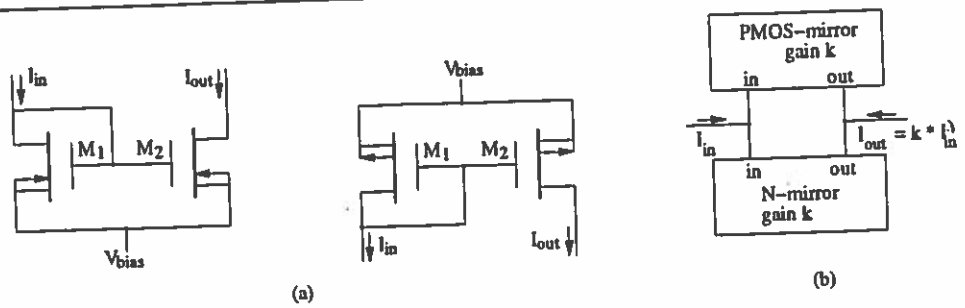


**Figure 4.** MOS current sources resistors and current mirrors. *A*, NMOS and PMOS unidirectional current mirrors, $I_{out} = ((L_2/W_2)/(L_1/W_1))I_{in}$. *B*, Bidirectional; current source.

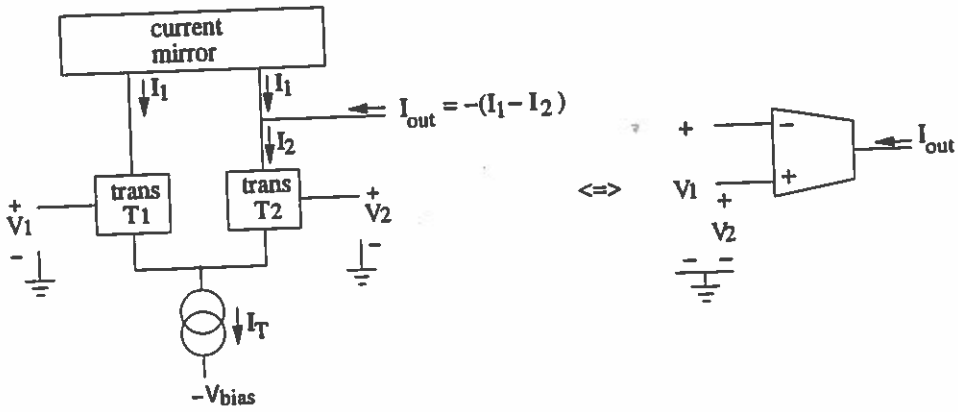


**Figure 5.** DVCCS. Basic configuration and circuit symbol, $T = M$ (when MOS) or $Q$ (when BJT).

limited range of input $V_d$. For NPN transistors in Figure 5, or NMOS in the subthreshold range,

$$I_d = I_T \tanh[V_d/(2V_T)] \tag{4}$$

where $V_T \approx 0.025$ is the thermal voltage at room temperature. This characteristic is quite nicely sigmoidal, leading to the BJT DVCCS having considerable importance for VLSI construction of ANN activation functions, especially for backpropagation circuits.

In some instances, it is necessary to convert the output of a DVCCS into a voltage (producing a Differential Voltage Controlled Voltage Source, DVCVS), as when voltage output for an activation function is desired. This can be accomplished by directing DVCCS output current into a resistor. However, one of the best ways to do this is to attach the gates of a CMOS pair to the output of the DVCCS. Since the CMOS pair allows no current at its input, the DVCCS can of course no longer act as a current source, but its output voltage is determined by other factors (specifically the channel-length modulation effect through the Early voltage). Other voltage amplifiers are available in the literature (Geiger et al., 1990), but high-gain operational amplifiers are not generally reasonable in VLSI for ANNs.

Since the DVCCS and the capacitor are sufficient to generate all linear circuits, we can construct many of the components of ANNs using them; see Kardontchik (1992) for filter examples. However, ANNs also require nonlinearities and, as we have seen, several nonlinearities are available, such as square-law and sigmoidal tanh ones. To build other nonlinearities, it is convenient to have multipliers, which can also be constructed from the DVCCS. An excellent multiplier is based upon the four-quadrant Gilbert multiplier (Geiger et al., 1990, p. 737), which uses two DVCCSs with the transistors of their tail current sources forming another differential pair. Assuming linear operation in the saturation region of the transistors and inputs $V_x$ and $V_y$ bounded by $V_x^2 \ll 2I_1(KW/L)$, $\ll 2I_2(KW/L)$, $V_y^2 \ll 2I_T(KW/L)$, with $K, I_1$ and $I_2$ as in Equation 2, the Gilbert multiplier gives $I_{out} \approx V_x V_y$. This Gilbert multiplier has successfully been used to multiply voltage-determined weights with neuron output voltages (Linares et al., 1993, p. 446). Dividers are possible but even less recommended than multipliers in VLSI.

## Applications

Neural modeling implementations using analog VLSI have produced numerous systems. In this section, we briefly highlight some of these applications and invite the interested reader to learn more by way of the following references: Mead (1989), Zornetzer (1990), and Sánchez (1992, 1993). Other applications include constrained optimization (Tank and Hopfield, 1986), Fourier transform computations (Culhane, in El-Leithy and Newcomb, 1989), oscillators (Linares, in El-Leithy and Newcomb, 1989), Hebbian learning (Meador, Watola, and Nintunze, 1991), A/D conversion (Yuh, in Sánchez and Newcomb, 1993), data compression (Fang, in Sánchez and Newcomb, 1992), pattern recognition (Salam and Wang, 1991), and fuzzy controllers (Yamakawa, in Sánchez and Newcomb, 1993).

Analog VLSI circuits have been applied to the task of modeling neurophysiological phenomena (see SILICON NEURONS). One approach in which nerve cell characteristics were modeled in hardware is the silicon neuron of Mahowald and Douglas (1991). In the circuits comprising the silicon neuron, the neuron's ability to self-generate electrochemical impulses is emulated. Other approaches have been taken by Moon (in Sánchez and Newcomb, 1992). Circuit realizations for a set of low-level electrochemical processes occurring within synapses have also been constructed. Using dynamics derived from actual neurophysiological data, second-messenger chemical "pools" (Hartline, in El-Leithy and Newcomb, 1989) were simulated (Tsay, 1993) using VLSI analog multipliers and DVCCSs.

One of the best matches to date between analog VLSI circuitry and a biologically based application is the silicon retina of Mead (1989). This chip implements the first stages of invertebrate retinal processing and produces signals similar to those found in real retinas. Another silicon retina implementation which includes tuned pixels is discussed by Delbrück (in Sánchez and Newcomb, 1993).

ANN associative memories (such as the Hopfield net) store patterns in weights such that when a "noisy" pattern is presented, the complete pattern is produced from the memory. Boahen and his co-workers describe the implementation of a three-layer, 46-neuron heteroassociative memory (in El-Leithy and Newcomb, 1989). Using current-mode circuits operating in subthreshold conduction, the chip contains a regular array of cells, each cell containing two synapses and a 1-bit weight memory cell. Inverters are used for thresholding neurons, current sources are used in the bias circuit, and a multiplier circuit is used in the synapse. A class of adaptable associative memories can also be realized using DVCCSs, incorporating Gilbert multipliers for transconductance weights (Linares, in Sánchez and Newcomb, 1993).

## Discussion

Compared to digital technology, analog VLSI offers the ANN world the distinct advantages of speed and real-time processing, though it suffers from relatively large size requirements and lack of standard cells. It also offers the ability to make continuous and speedy adjustments for adaptive neural networks and those needing efficient calculations of derivatives, as in backpropagation ANNs. Although the absolute error for analog components is typically larger than 5%, the relative precision can usually be controlled to be under 0.1% when implemented in VLSI. Roughly, this is the equivalent of 8-bit digital resolution at hundreds to thousands of mHz. When working with the primary circuits discussed here, such as the DVCCS and the current mirrors, voltage and current differences matter most, so that it is the relative tolerance that is critical. In any event, because ANNs are by conception fault tolerant, precision is not usually of concern.

VLSI neurons can have their dimensions comparable to those of biological neurons with considerably faster signal processing. However, real neurons take full advantage of their 3D nature, whereas most present-day VLSI structures are essentially planar. Although connection wires can be routed under other wires in multiple metal VLSI constructs, and there do exist some prototype 3D processes, the technology is still quite limited. Pulse-coded ANNs are amenable to a mixture of analog and digital realizations; the action potentials can be standardized and then realized by digital pulses, while the synaptic effects can be most conveniently realized by analog devices, since real-valued weights are involved.

A large number of other devices of interest to specialized areas of ANNs are presently available for VLSI. Among such devices are charge-coupled devices (CCDs), possibly for axon-like propagation or enzyme-effect mimicking, floating-gate devices for long-term storage of weights, and JFETs for less delicate fabrications. It should be noted that the MOS devices take minimal area but they are subject to damage by static charge that can puncture the very thin gate oxide. For the future, there are the very small resonant tunneling devices, which use a different substrate than the silicon of present VLSI, and molecular devices, which probably show the greatest long-range potential because of their minimal size and general signal handling.

**Road Map:** Implementation of Neural Networks
**Background:** I.1. Introducing the Neuron; I.3. Dynamics and Adaptation in Neural Networks
**Related Reading:** Silicon Neurons; Digital VLSI for Neural Networks

## References

Bialko, M., and Newcomb, R. W., 1971, Generation of all finite linear circuits using the integrated DVCCS, *IEEE Trans. Circuit Theory*, CT-18(6):733–736.

El-Leithy, N., and Newcomb, R. W., Eds., 1989, Special issue on neural networks, *IEEE Trans. Circuits Sys.*, 36.

Geiger, R. L., Allen, P. E., and Strader, N. R., 1990, *VLSI Design Techniques for Analog and Digital Circuits*, New York: McGraw-Hill. ◆

Graf, H. P., and Jackel, L. D., 1989, Analog electronic neural network circuits, *IEEE Circuits Devices Mag.*, 5:44–55.

Kardontchik, J. E., 1992, *Introduction to the Design of Transconductor-Capacitor Filters*, Boston, MA: Kluwer.

Lee, B. W., and Sheu, B. J., 1991, *Hardware Annealing in Analog VLSI Neurocomputing*, Norwell, MA: Kluwer.

Linares, B., Sanchez, E., Rodriguez, A., and Huertas, J., 1993, A CMOS analog adaptive BAM with on-chip learning and weight refreshing, *IEEE Trans. Neural Netw.*, 4:445–455.

Mahowald, M., and Douglas, R., 1991, A silicon neuron, *Nature*, 354: 515–518.

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

Meador, J., Watola, D., and Nintunze, N., 1991, VLSI implementation of a pulse Hebbian learning law, *Proc. 1991 IEEE Int. Sympos. Circuits Sys.*, Singapore, pp. 1287–1290.

Mueller, P., van der Spiegel, J., Blackman, D., et al., 1989, Design and fabrication of VLSI components for a general purpose analog neural computer, in *Analog VLSI Implementation of Neural Systems* (C. A. Mead and M. Ismail, Eds.), Boston: Kluwer.

Reiss, R. F. Ed., 1964, *Neural Theory and Modeling*, Stanford, CA: Stanford University Press.

Salam, F., and Wang, Y., 1991, A real-time experiment using a 50-neuron CMOS analog silicon chip with on-chip digital learning, *IEEE Trans. Neural Netw.*, 2:461–464.

Sánchez, E., and Newcomb, R. W., Eds., 1992, 1993, Special issues on neural network hardware, *IEEE Trans. Neural Netw.*, 3, 4.

Schmitt, O. H., 1937, An electrical theory of nerve impulse propagation and mechanical solution of the equations of nerve impulse propagation, *Am. J. Physiol.*, 119:399–400.

Tank, D. W., and Hopfield, J. J., 1986, Simple "neural" optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit, *IEEE Trans. Circuits Sys.*, CAS-33:533–541.

Taylor, G. W., 1978, Subthreshold conduction in MOSFET's, *IEEE Trans. Electron Devices*, ED-3:337–350.

Tsay, S.-W., 1993, Implementation of Hartline pools and neural-type cells by VLSI circuits, in *Advances in Control Networks and Large-Scale Parallel Distributed Processing Models*, vol. 2, (M. Fraser, Ed.), Norwood, NJ: Ablex.

Zornetzer, S. F., Davis, J. L., and Lau, C., Eds., 1990, *An Introduction to Neural and Electronic Networks*, San Diego: Academic Press. ◆

# The Handbook of Brain Theory and Neural Networks

EDITED BY
Michael A. Arbib



A Bradford Book
THE MIT PRESS
Cambridge, Massachusetts
London, England
1995