Analog Components for the VLSI of Neural Networks

Simon Y. Foo, Lisa R. Anderson and Yoshiyasu Takefuji

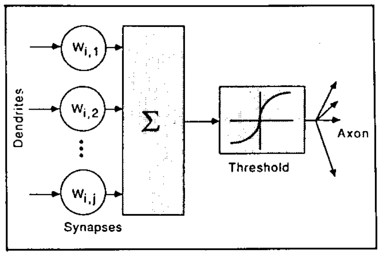
Artificial neural networks (ANNs) are attempts to mimic, at least partially, the structure and functions of brains and nervous systems. The human brain contains billions of biological neurons whose manner of interconnection allows us to reason, memorize, and compute. Advances in VLSI technology and a demand for "intelligent" machines have created a strong resurgence of interest in emulating neural systems for real-time applications.

The same factors have spurred research on artificial intelligence (AI) over the last few years. Current AI technology based on knowledge-based expert systems has relied heavily on symbolic manipulations. The approach's major limitation is that the knowledge base is a static set of rules cast by human experts. At the inevitable error--prone interface between the human experts and the AI programmers, the programmers must cope with fuzzy information.

Artificial neural networks, on the other hand, are trained by successive examples in a real-world environment. As the ANNs adapt to the changes in their environment, they develop their own internal rules. One advantage of ANNs is their ability to handle fuzzy or incomplete data. Current neural network models include:

- Hopfield Networks
- Hamming Networks
- Widrow's Adaline
- Rosenblatt's Single-layer Perceptrons
- Werbos's Backward Error Propagation for Multi-layer Perceptrons
- Carpenter and Grossberg's Adaptive Resonance Theory (ART)
- Hinton and Sejnowski's Boltzmann Machines
- Kohonen's Self-Organization Feature Map
- Fukushima's Neocognitrons
- Kosko's Bidirectional Associative Memory (BAM)

Artificial neural networks can be implemented with simple analog devices



1. Functional model of an artificial neuron.

In particular, ANNs employ an enormous number of communication links among the processing elements (PEs) to perform distributed parallel processing (PDP). Because of the robust (or fault-tolerant) nature of ANNs, a few degraded or nonfunctional PE's will not greatly affect the overall operation of the neural network. The speed and robustness of ANNs make

8755-3996/90/0700-0018$1.00 © 1990 IEEE

them very attractive for a variety of applications, such as pattern recognition, robotic control, and combinatorial optimization.

An artificial neuron can be modeled as a multi-input nonlinear thresholding device with weighted interconnections, or synapses (Fig. 1). The cell body of an electronic neuron is represented by a nonlinear amplifier (e.g., a high-gain amplifier), while the synapses are represented by variable resistors (Fig. 2). The dynamics of each neuron is governed by an ordinary first-order differential equation (or difference equation for discrete-time systems) which describes the motion of the neural network. By applying Kirchoff's current law at the input node of the amplifier, the differential equation describing the time evolution of the analog circuit is:

$$C_i\frac{dU_i}{dt}=\sum_{j=1}^{N} G_{ij}V_j-\frac{U_i+\Delta U_{io}}{\rho_i}$$

$$V_i=f(U_i)$$

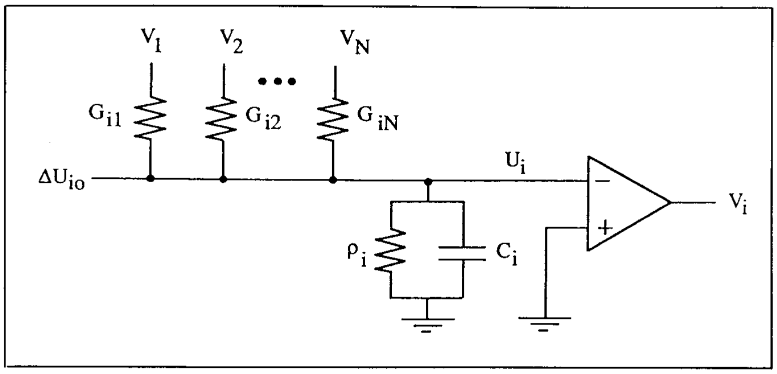
where U_i is the input voltage to the ith amplifier; ΔU_{io} is the external noise to the ith amplifier; V_i is the output voltage of the ith amplifier; C_i is the input capacitance of the ith amplifier; r_i is the input resistance of the ith amplifier; G_{ij} is the conductance of the resistive interconnect between the ith and jth amplifiers; f is the transfer function of the nonlinear amplifier; and N is the number of neurons in the network.

As shown earlier, the basic electronic neuron has two main components: resistive interconnects (synapses) and a processing element (cell body). Signals received from other neurons in the form of potentials across the resistive interconnects are collected by summing currents. Each synaptic weight or resistive interconnect is modeled as a passive resistor with conductance G. Based on its input neural voltages U, the PE produces an output signal V according to its nonlinear transfer function f, and the output signal V is then propagated to other neurons.

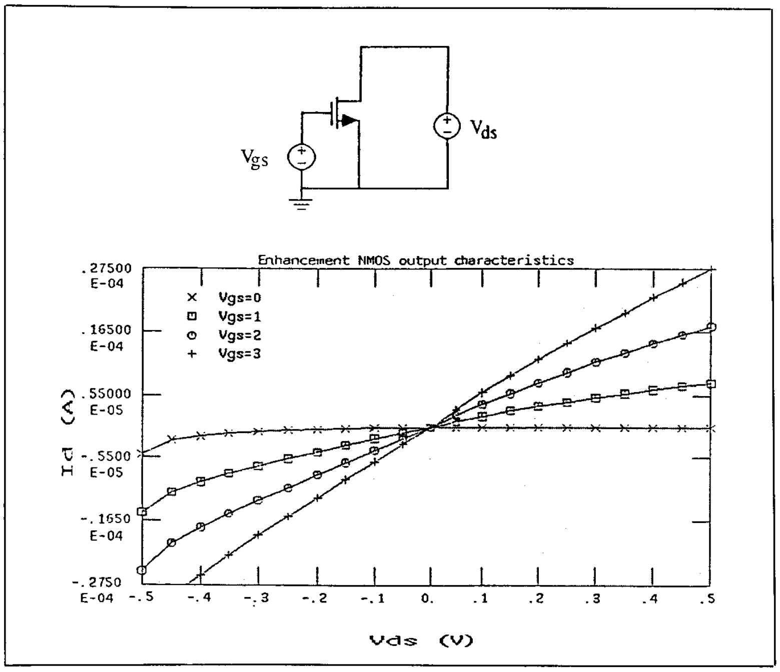
Building Blocks

One of the most important aspects of neural networks is their learning capability, whereby synaptic strengths between neurons are adaptively changed according to an algorithm. Such learning could be supervised (e.g., Hopfield's) or unsupervised (e.g., Kohonen's). Learning algorithms requiring high resolution in interconnection strengths need high precision circuitry for weight adjustments. Such complex circuits, in turn, requires more silicon area. In general, analog circuits are used where only moderate precision is required (even though high precision analog circuits can be built at the expense of more silicon area). Conversely, digital circuits are used for high resolution weight representation. For example, the back-propagation learning algorithm requires at least an 8-bit weight representation for a large problem of practical interest. Therefore, analog circuits are most appropriate for learning algorithms with high fault-tolerance and requiring moderate or low precision, while digital circuits are used for high-resolution learning algorithms.

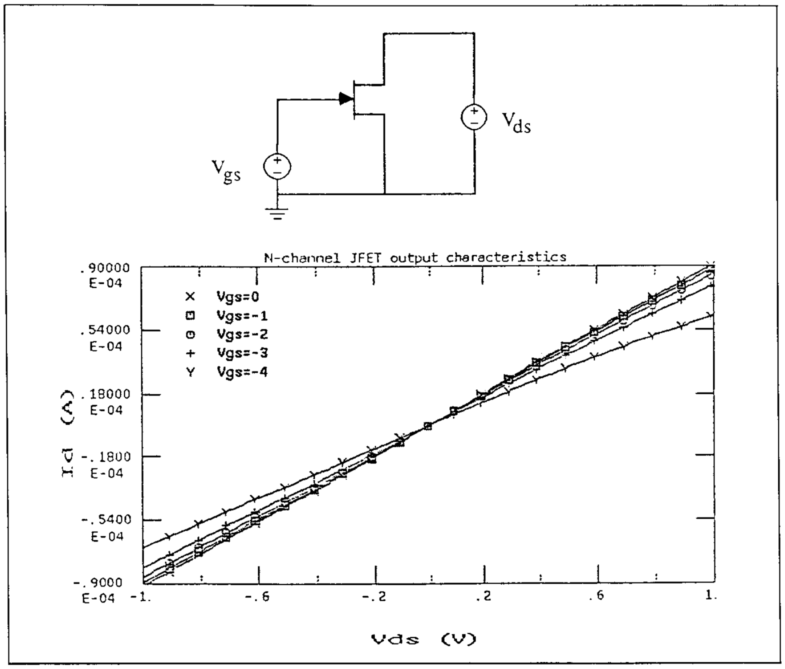
The neural components can be either deterministic or stochastic, leading to deterministic and stochastic neural networks. There are basically four methods for building stochastic neural networks, i.e., through: (1) stochastic I/O function in the PE, (2) stochastic synaptic strength, (3) external input noise, and (4) a combination



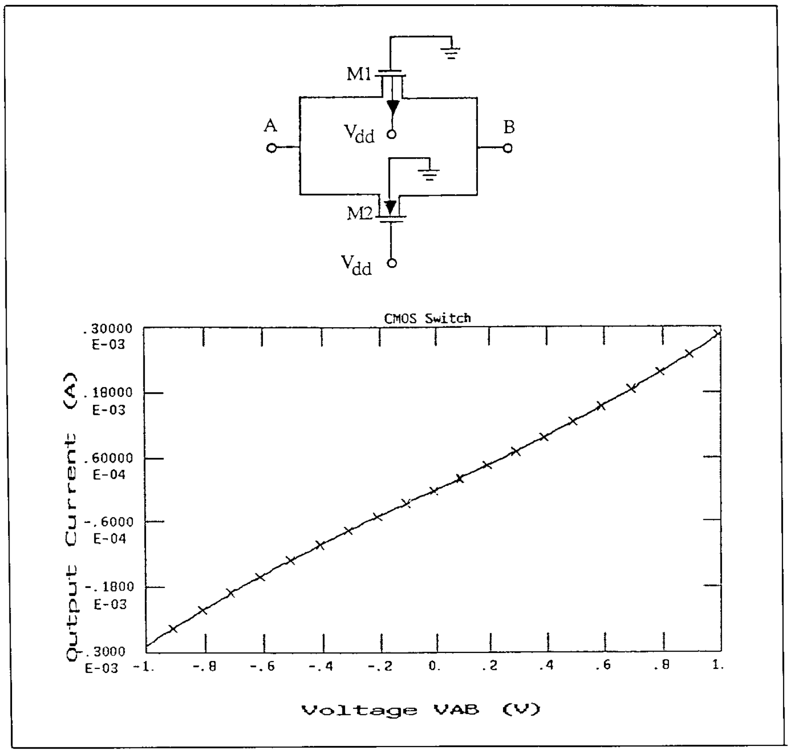
2. Circuit schematic of an artificial neuron.



3. An enhancement-mode n-channel MOSFET and its characteristics for small V_{ds}.



4. An n-channel JFET and its linear characteristics for small V_{ds}.



5. A CMOS switch and its characteristics in the active region.

of the above. However, within the scope of this paper, we shall focus on deterministic conductance circuits and deterministic PEs. These devices are analyzed using the Simulation Program with Integrated Circuit Emphasis (SPICE) program.

Variable Linear Conductance Devices

As discussed earlier, synaptic weights can be simulated by variable linear resistors. Nonprogrammable neural networks with fixed-value resistors are relatively small and easy to fabricate, but they have very limited applications. Resistors as small as 0.25 μm x 25 μm have been built [5]. With a density of 4 resistors per square micrometer, a total of 4 x 10^8 resistors can be packed into a 1 cm^2 chip. Research into the design and implementation of programmable interconnection weights include a binary interconnection circuit by Graf and Jackel [6]. Their VLSI chip has 4,416 digital circuits and it performs an evaluation of 44 billion connections per second (cps). However, the digital interconnection circuits occupy excessive silicon area, and they also require digital-to-analog converters (DAC) at the outputs. This extra overhead makes the digital approach unattractive for VLSI of ANNs.

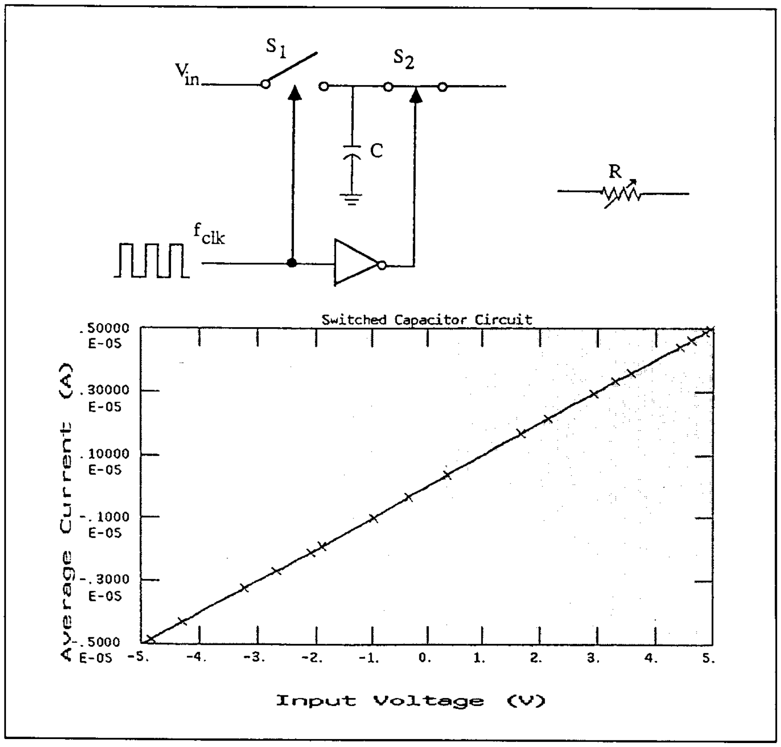
We know that a human brain has about 100 billion neurons, and that each neuron (or nerve cell) is typically connected to approximately 10,000 other neurons. Since biological neurons respond only at the millisecond time scale (much slower than transistors), it is apparent that the computation abilities of our brains arise from the large number of neurons and the huge number of interconnection links. Due to the physical limitation of a two-dimensional silicon wafer, electronic neural networks have to rely heavily on "simple" models of neurons and synapses.

Perhaps the simplest active devices for simulating the variable resistor are the junction field-effect transistors (JFETs) and the metal-oxide-semiconductor field-effect transistors (MOSFETs). Both the JFET and the MOSFET act as voltage-controlled linear resistors for small values of V_{ds} (drain-to-source voltage), operating in the active region. Figure 3 shows an n-channel enhancement-mode MOSFET and its characteristics. The output resistance of the device is controlled by the input gate-to-source voltage, V_{gs}, and in this case, the resistance ranges from 16.5 kΩ to 370 kΩ. As V_{gs} increases, the resistance decreases.

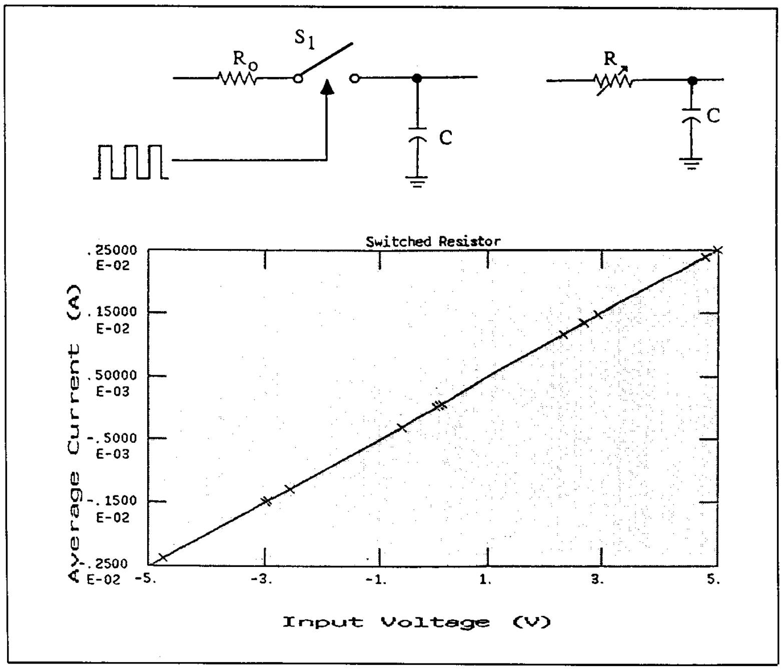
The MOSFET behaves as a linear resistor if the output voltage lies between approximately -0.5 volt to +0.5 volt. This is a feasible range since we intend to keep the operation of the neural network in the millivolt range so as to reduce power dissipation.

The depletion-mode device has the same characteristics as the enhancement-mode device except that V_{gs} is negative. Similarly, an n-channel JFET also acts as a voltage-controlled linear resistor for small values of V_{DS} (Figure 4). When V_{gs} varies from 0 to -4 volts, the resistance varies from 11 kΩ to 15 kΩ. The linear region lies approximately from -1.0 volt to +1.0 volt. Thus the JFET has a wider dynamic range than the MOSFET. It is also possible to utilize MOS switches and floating active resistors for simulating variable linear resistors. Both devices can be implemented using a single MOSFET. However, a more practical approach is the CMOS switch (transmission gate) based on a pair of complementary MOSFETs, as shown in Figure 5. The resistance of the CMOS switch is controlled by the voltage at the bulk of transistor Ml and the gate voltage of transistor M2. The linear resistive region lies approximately from -1.0 volt to +1.0 volt. In our example shown in Figure 5, the output resistance is approximately equal to 3.7 kΩ. The CMOS switch has a major advantage over a single MOSFET switch or a floating active resistor. In particular, the problem of clock feedthrough is eliminated through the parallel connection of the n- and p-channel devices, which require opposing clock signals. Consequently, the dynamic range is greatly increased as a result of the complementary devices. In our previous approach, we used the analog variable resistor techniques for simulating the synaptic strength [3]. Basically, there are four types of a variable resistor: a switched-capacitor circuit, a switched-resistor, a switched-ladder resistor, and a voltage-controlled resistor. The purpose of these devices is to linearly vary the flow of current controlled by a clock pulse or an input voltage.

In a switched-capacitor circuit, the value of resistance R depends on an input clock frequency f_{clk} and a capacitor C, where $R = 1/(f_{clk}C)$. As the clock frequency and capacitance increase, the resistance decreases. Figure 6 shows the switched-capacitor circuit and its characteristics. In this example where f_{clk} = 1 kHz and C = 1 nF,



6. A switched-capacitor circuit and its SPICE output.



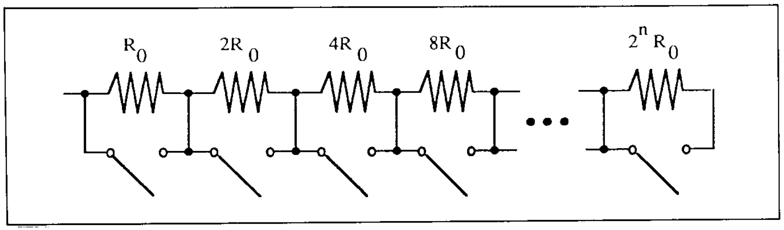
7. A switched-resistor circuit, its equivalent representation, and its SPICE output.

the resistance is approximately equal to 1 MΩ with an input voltage range of -5.0 to +5.0 volts.

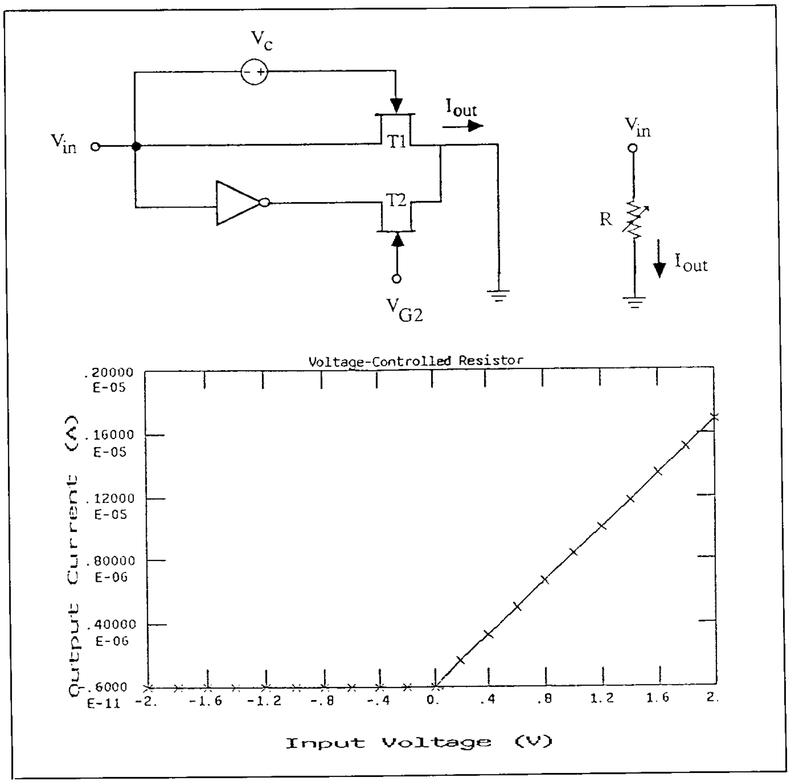
A switched-resistor circuit is composed of a fixed R_o resistor, an analog switch, and a capacitor. The value of the resistance R is determined by the ratio $R = R_o/d$ where d is the duty cycle of the switch. As the duty cycle decreases, the value of R increases. Figure 7 shows the switched-resistor circuit, its equivalent representation, and its output characteristics. In our example where R_o = 1 kΩ and d = 0.5, the resistance is approximately equal to 2 kΩ with an input voltage range of -5.0 to +5.0 volts.

A switched-ladder resistor is composed of n analog switches in parallel with nR_o to $2^{n-1}R_o$ resistors in series, as shown in Figure 8. The total resistance is controlled by the analog switches. As a result, there are $(2^{n+1} - 1)$ possible values ranging from 0 to $(2^{n+1} - 1)R_o$ ohms. For example, if all of the switches are turned "off", the total resistance from one end to the other becomes $(2^{n+1} - 1)R_o$ ohms.

Analog devices with high accuracy can be built at the expense of larger silicon area and higher power dissipation. For example, we could utilize an elaborate and precision voltage-controlled linear resistor introduced by Czarnul [2]. This circuit is based on a matched pair of FETs (Fig. 9). The linear resistance is determined by the gate voltage V_{G2} of transistor T_2 and the floating voltage source V_c of transistor T_1. The output current I_o is determined by $I_o = \beta V_{in}(V_c - V_{G2})$, where β is the transconductance parameter dependent on the fabrication and the geometry of the transistors. The output of this device is linear in the range of 0 to +2 volts, where the resistance is approximately equal to 1 MΩ.



8. A switched-ladder resistor network.



9. Czarnul's precision voltage-controlled linear resistor and its SPICE output.

Similar approaches include a CMOS voltage-controlled linear resistor with a wide dynamic range by Youssef, Newcomb, and Zaghloul [4]. A pair of complementary enhancement-mode MOS transistors is used to offset the nonlinearity present in the circuit. Figure 10 shows the voltage-controlled linear resistor circuit and its characteristics. Although this circuit provides high resolution resistance with a wide dynamic range, the circuit requires several external voltage sources which makes it expensive to implement. This overhead is hardly justified since neural networks are fault-tolerant requiring only moderate accuracy.

Table I summarizes the performance of the variable linear resistors. The approximate conductance is calculated using the linear least squares data fitting method over the input voltage range. The inverse of conductance yields the output resistance of the device. The error margin is calculated as the average error in the approximation of the data.

Deterministic Processing Elements

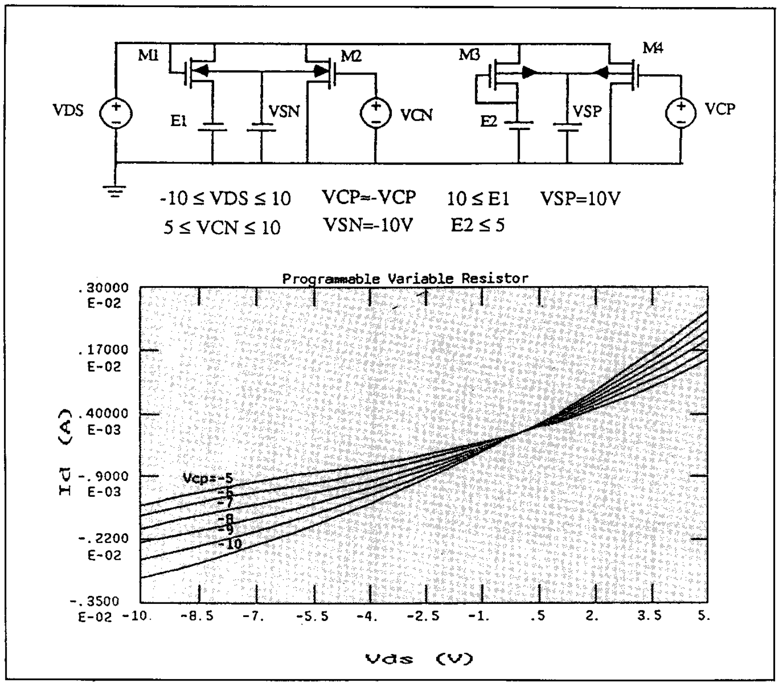
The processing elements are the key components of a neural network. The PEs collect and process all the incoming signals propagated from other neurons through the synapses. Based on a nonlinear activation function, the neuron may "fire" if the sum of input signals exceeds a certain threshold, or it may be turned "off" if not. In general, there are three basic nonlinear transfer functions for artificial neurons, i.e., high-gain limit, linear threshold, and sigmoid (Fig. 11). The high-gain limit (or step) function used for configuring associative memory in the Hopfield neural networks, Hamming networks, and Boltzmann machines can be easily implemented by an analog comparator. Figure 12 shows an analog comparator and its transfer function, where the threshold of the step function is controlled by the reference voltage V_{ref}. Such a neuron based on an analog comparator is often called a two-state neuron.

The linear threshold activation function is commonly used in the single-layer feedforward networks such as ADALINE and single-layer perceptrons. The linear function with a fixed threshold can be easily implemented by a noninverting operational amplifier circuit (Fig. 13). The voltage gain of the amplifier is determined by $A_v = (R_1 + R_2)/R_1$. The bounds of the voltage gain are defined by means of the lower saturation point (LSP) and the upper saturation point (USP), respectively, where LSP = $-V_{cc}R_1/(R_1 + R_2)$, and USP = $V_{cc}R_1/(R_1 + R_2)$.

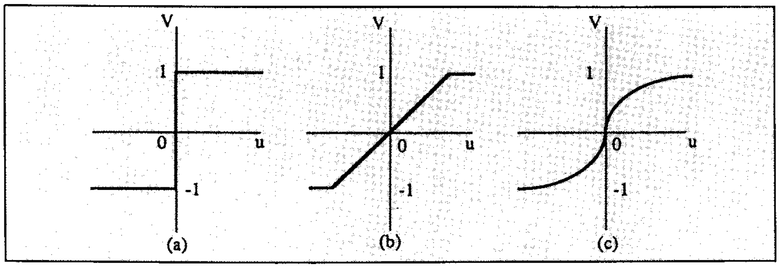
A twin cascaded inverting amplifier can be used to implement a linear function with adjustable threshold (Fig. 14). The output voltage of the circuit is determined by $V_o = V_i(R_2/R_1) - V_{ref}(1 + R_2/R_1)$ and the threshold θ is determined by $q = V_{ref}(R_1 + R_2)/R_2$. The LSP and USP of the adjustable threshold circuit are given by LSP = $V_{ref}(R_1 + R_2)/R_2 - V_{cc}(R_1/R_2)$ and USP = $V_{ref}(R_1 + R_2)/R_2 + V_{cc}(R_1/R_2)$. The linear circuit with adjustable threshold allows considerable flexibility, since not only is the threshold adjustable, the range of linearity is also adjustable.

Another nonlinear function widely used in the sigma-pi and Hopfield neural networks is the continuous sigmoid function. In multi-layer feedforward associative networks (or backpropagation networks), the learning process depends on the delta rule [7] which requires a differentiable and nondecreasing function such as the sigmoid transfer function $V = 0.5 \tanh(\lambda U)$, where λ is the voltage gain, and U and V are input and output voltages, respectively. An approximate sigmoid function with a fixed gain can be realized by a high-gain inverting amplifier in cascade with a unity gain inverting amplifier (Fig. 15a).

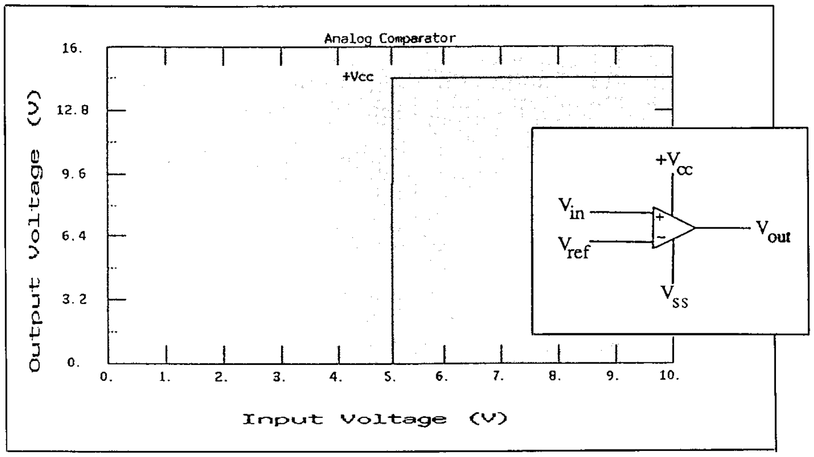
A more useful sigmoid circuit with variable gain control uses an unbuffered voltage comparator with a positive feedback loop and double cascaded inverters with negative feedback loops (Fig. 15b). The unbuffered comparator provides an approximate sigmoid transfer function, while the negative feedback loops of the inverters act as gain controls. Therefore, the gain of the sigmoid function can be increased by increasing the ratio $r = R_8/R_7 = R_{10}/R_9$, i.e., the gains of the inverters. Figure 16 (top) shows a detailed schematic of the variable-gain sigmoid circuit. SPICE simulations of the variable-gain sigmoid circuit based



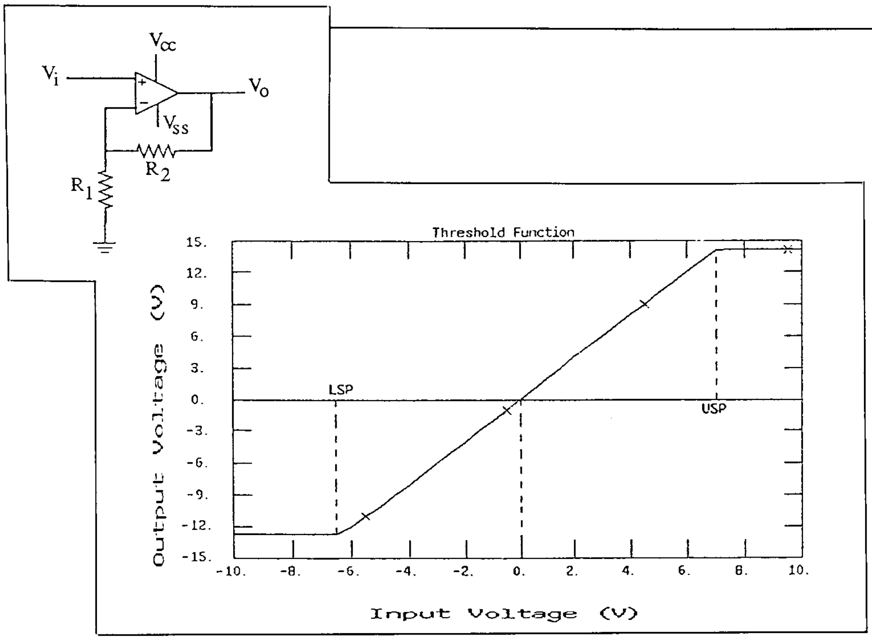
10. A CMOS voltage-controlled linear resistor with a wide dynamic range, and its SPICE output.



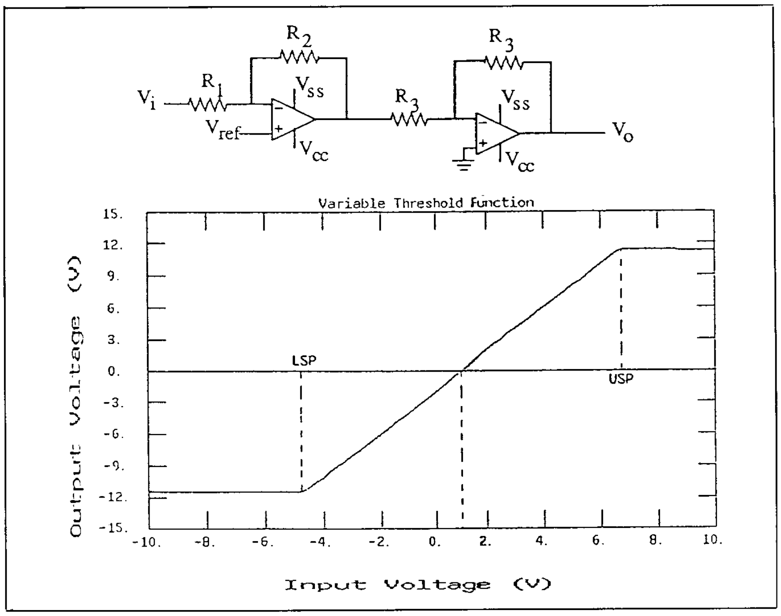
11. Nonlinear transfer functions for artificial neurons based on (a) high gain limit or step, (b) linear threshold, and (c) sigmoid functions.



12. An analog comparator and its output characteristics.



13. A noninverting amplifier with fixed threshold, and its characteristics.



14. A linear circuit with adjustable threshold, and its characteristics.

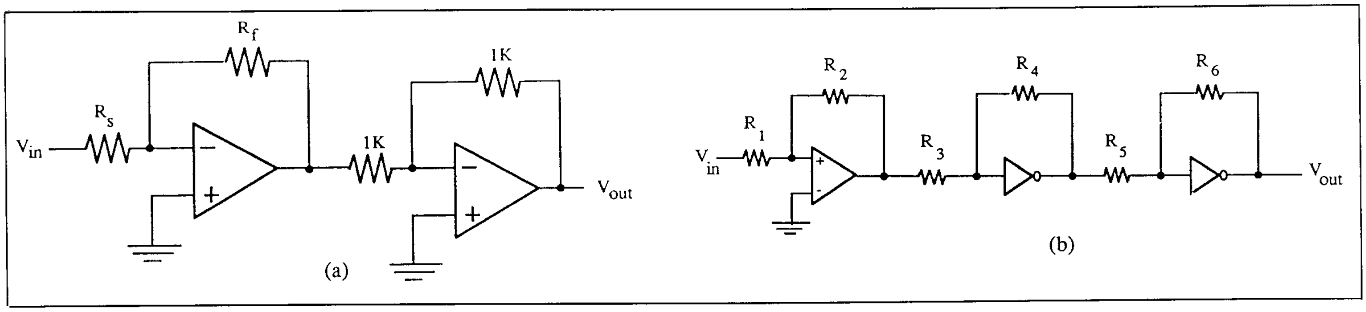
CMOS operational amplifiers were performed for r = 100, 250, 500, and 1000. The results are plotted in Figure 16 (bottom). It is observed that as r increases, the sigmoid curve approaches a high gain limit function; lowering r results in a gentler slope sigmoid curve. When r = 0 (i.e., zero gain), the output voltage is zero regardless of input.

Why Simple Analog Circuits?

First, analog circuits are faster than digital implementations in terms of speed-to-amount of hardware ratio. Second, analog circuits provide us a better understanding of the true analog nature of biological neural networks. Third, since these analog components require less circuitry, it is possible to pack more components onto a single VLSI chip. The simplicity of these devices also make them very attractive for rapid VLSI prototyping. Fourth, since the analog conductance devices and processing elements operate in small voltage swings, they dissipate less power, and thus reduces the problem of heat transfer.

A survey of current approaches seems to indicate a tradeoff between the complexity of the circuits and their size in silicon. If more functionality (e.g., higher resolution, learning capabilities) is desired, fewer circuits can fit onto a single chip. The solution usually depends on the application of such a neural network. Current analog neural networks have computational speeds of 10^9 to 10^{11} interconnections per second (ips), a much higher rate than digital circuits can achieve, according to recent DARPA studies. Although board-level emulators are more flexible and easier to program than the special-purpose neural hardware, they only operate at a speed of 10^6 to 10^7 ips.

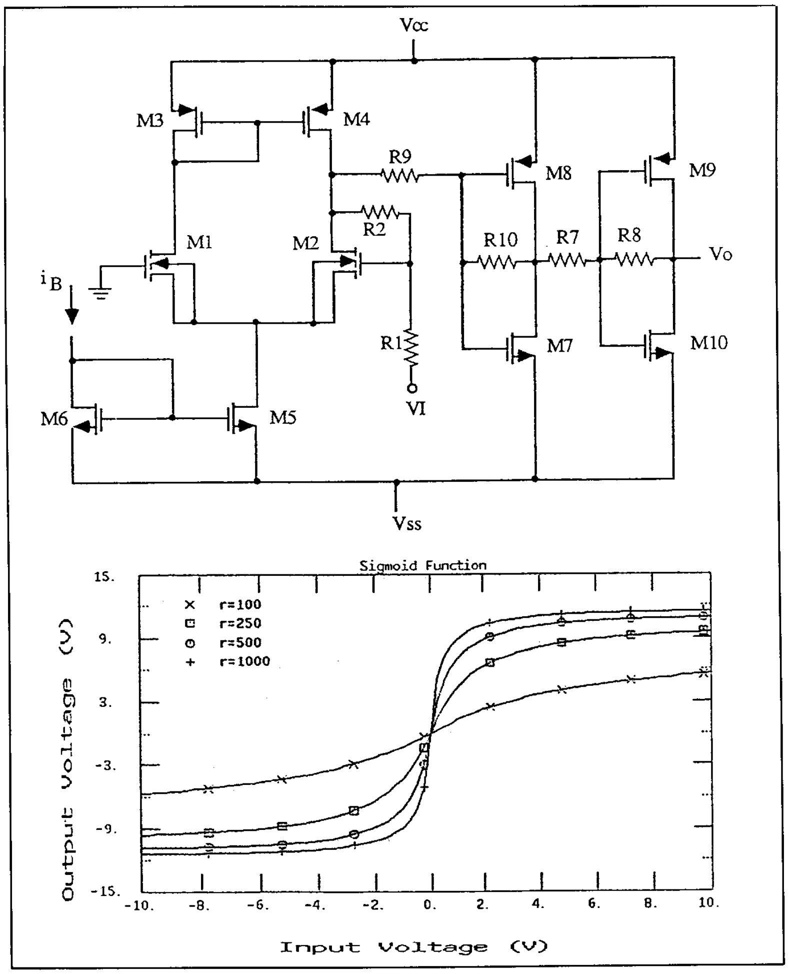
Many researchers are also looking at



15. Sigmoid function with (a) fixed gain control and (b) variable gain control.

optical computing to eliminate the problems faced by metallic interconnects. Optical neurocomputers are more accurately called electro-optic neural networks because they utilize considerable electronics, especially in the I/O. For example, Caulfield et. al. [8] introduced a hybrid optically programmed electronic (HOPE) neural network. The network is programmable through a Page Oriented Holographic Memory (POHM) which could store up to 10^6 bits (or 1000-by-1000 pixel image) with random access time as low as 1 nanosecond. Advantages of optical computing will include higher switching speeds, reduced mutual interference, no fan-out constraints, and no capacitive loading effects as in the case of metallic interconnects. Currently, the switching speed of optical computers is on the order of 10^{-12} to 10^{-15} seconds, compared to 10^{-9} to 10^{-11} seconds in electronic computers. Similarly, the communication bandwidth of optical computers is on the order of 1 to 10^2 gigabits per second, while that of electronic computers is on the order of 10 to 10^3 megabits per second. The revolution in optical computing—and its potential benefits—was succinctly stated by Professor Kai Hwang of University of Southern California, a renowned expert in computer architecture. He predicted that the next generation artificial neural systems (ANS) will be based on 3D wafer scale integration (WSI) technology with optical interconnections.

In general, a biological neuron can be modeled as a very high fan-in/fan-out processing element, whereby a neuron is typically connected to several thousands of neurons. Consequently, the wiring of the large number of resistive interconnects on a two-dimensional surface of a silicon wafer represents the major bottleneck for implementing electronic neural networks. Current efforts to achieve such high interconnectivity or to remedy this problem are based on several approaches discussed above. In this paper, we briefly described several possible analog circuits for building electronic neural networks based on discrete and integrated devices. We believe that the moderate resolution of these "simple" analog components are sufficient for implementing neural networks such as Hopfield's model and Fukushima's Neocognitron paradigm. This research also represents a step toward fulfilling the need for real-time neural computing through the VLSI of simple analog components.



16. Detailed schematic of the sigmoid circuit with variable gain, and its SPICE output for various r's.

Table I

Switch Type	Voltage Range (V)	Resistance (Ω)	Average Error
Switched Capacitor	-5.0 to +5.0	1 M	1.4 E-12
Switched Resistor	-5.0 to +5.0	2 K	6.6E-10
CMOS Switch	-1.0 to +1.0	3.7 K	7.2 E-6
Enhancement-NMOS			
Vgs=0	-0.5 to +0.5	.37 M	8.4 E-7
Vgs=1	-0.5 to +0.5	48.2 K	1.1E-6
Vgs=2	-0.5 to +0.5	24.5 K	1.1E-6
Vgs=3	-0.5 to +0.5	16.5 K	1.1E-6
N-Channel JFET			
Vgs=0	-1.0 to +1.0	.37 M	8.4 E-7
Vgs=-1	-1.0 to +1.0	48.2 K	1.1E-6
Vgs=-2	-1.0 to +1.0	24.5 K	1.1E-6
Vgs=-3	-1.0 to +1.0	16.5 K	1.1E-6
Voltage Controlled Resistor	0.0 to 2.0	1.2 M	1.1E-9

Acknowledgment

The authors would like to thank Dr. H. John Caulfield of Center for Applied Optics, University of Alabama in Huntsville, Alabama, and Dr. Harold Szu of Naval Research Laboratory, Washington, DC, for their encouragement and invaluable discussions. ***CD***

References

1. J. J. Hopfield and D. W. Tank, "Neural Computation of Decisions in Optimization Problems," Biological Cybernetics, No. 52, pp. 141-152,1985.

2. Z. Czarnul, "Design of Voltage-Controlled Linear Transconductance Elements with a Matched Pair of FETs," IEEE Trans. on Circuits and Systems, CAS-33, No. 10,1986.

3. S. Foo, Y. Takefuji, and T. J. Harrison, "Analog Components for Electronic Neural Networks," 1st Florida Annual Microelectronics Conference, Boca Raton, FL, May 10-11, 1989.

4. H. Youssef, R. Newcomb, M. Zaghloul, "A CMOS Voltage-Controlled Linear Resistor with a Wide Dynamic Range," Proc. of 21st Southeastern Symposium on System Theory, Tallahassee, FL, March 26-28, 1989.

5. L. D. Jackel, R. E. Howard, H. P. Graf, B. Straughn, and J. S. Denker, "Artificial Neural Networks for Computing," J. Vac. Sci. Technology, Vol. B61, 1986.

6. H. P. Graf and L. D. Jackel, "VLSI Implementations of Neural Network Models," in Concurrent Computing, New York: Plenum, 1988.

7. J. L. McClelland, D. E. Rummelhart, and the PDP Research Group, "Parallel Distributed Processing, Volumes I and II," MIT Press, 1986.

8. H. J. Caulfield, J. Kinser, and S. K. Rogers, "Optical Neural Networks," Proceedings of the IEEE, 1989.

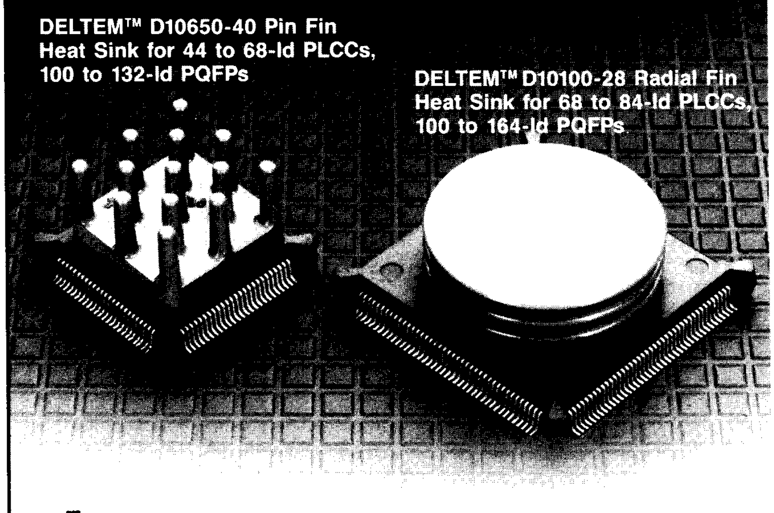
Suggested Further Reading: H. P. Graf and L. D. Jackel, "Analog Electronic Neural Network Circuits," IEEE Circuits and Devices Magazine, July 1989; C. Mead, "Analog VLSI and Neural Systems," Addison-Wesley, 1989; N. H. Farhat, "Optoelectronic Neural Networks and Learning Machines," IEEE Circuits and Devices Magazine, September 1989; L. E. Atlas and Y. Suzuki, "Digital Systems for Artificial Neural Networks," IEEE Circuits and Devices Magazine, November 1989.

Biography

Simon Y. Foo [M] is Assistant Professor at the Department of Electrical Engineering in the FAMU/FSU College of Engineering, Florida State University, Tallahassee, Florida.

Lisa R. Anderson was a student at the Department of Electrical Engineering in FAMU/FSU's College Engineering.

Yoshiyasu Takefuji [M] is Assistant Professor at the Department of Electrical Engineering and Applied Physics at Case Western Reserve University in Cleveland, Ohio.



DELTEM™ D10650-40 Pin Fin Heat Sink for 44 to 68-ld PLCCs, 100 to 132-ld PQFPs
DELTEM™ D10100-28 Radial Fin Heat Sink for 68 to 84-ld PLCCs, 100 to 164-ld PQFPs