A Multiple-Input OTA Circuit for Neural Networks

RUSSEL D. REED, MEMBER, IEEE, AND RANDALL L. GEIGER, SENIOR MEMBER, IEEE

Abstract—**An operational transconductance amplifier (OTA) circuit suitable for modeling neurons in VLSI implementations of artificial neural networks (NN) is described. It generates an output voltage which is a sigmoidal-like function of the linear sum of a number of weighted inputs. The weight of each input is individually controlled by a bias voltage which can be varied continuously and dynamically.**

RECENTLY, there has been an increase in interest in artificial neural networks for use in artificial intelligence applications. They seem to be especially useful in situations requiring pattern recognition or optimization with many simultaneous constraints. The self-programming properties of these networks (given appropriate learning rules) allow them to learn from example data sets even in the presence of noisy and conflicting data and their massive parallelism gives them a degree of fault tolerance [1].

Work in the area is still developing. Much is still unknown about how biological networks work and how artificial networks can be built with similar properties. Many different models with varying numbers of units, connection patterns and learning rules are still being explored [1], [2].

In order to simplify the analyses, most work uses the simplest possible model of a neuron: a node that sums a large number of weighted inputs and generates an output which is a function of that sum. The output function is most often a nonlinear monotonic increasing function, typically a binary step function or a sigmoid. Ensembles of these simplified neurons are connected together in various ways and trained to recognize certain patterns by adjusting the weight of each connection.

Working models of artificial neural networks have been demonstrated although, so far, they have been limited in size [3]–[7]. These circuits are much faster than software simulations running on conventional computers.

The circuit presented in the following paragraphs may be useful for VLSI implementations of neural networks. Acting as the analog of a single neuron, it generates an output voltage which is a sigmoidal-like function of the linear sum of a number of weighted input voltages. The weight of each input can be individually and continuously controlled by a bias voltage which can be varied dynamically. The inputs have a wide linear range and the number of inputs for each circuit can be large.

Fig. 1 shows the basic building block which represents one weighted input to the neuron. It sinks an output current, I, which is a linear function of the input voltage V_{gs1} and has a transconductance which is controlled by the bias voltage V_b. (The transconductance is defined as dI/dV_i and is the gain of the module.) When the outputs of a number of these blocks are connected to a common node, the currents sum according to Kirchoff's current law and an op amp can then be used to convert the current to an output voltage which is the weighted sum of the input voltages. The following paragraphs develop this in more detail.

In Fig. 1, when MOSFET $M1$ is biased in its active region, $V_{gs1} - V_{T1} > V_{ds1}$, the current I can be written as

$$I = \beta\left[(V_{gs1} - V_{T1}) - \frac{V_{ds1}}{2}\right]V_{ds1} \tag{1}$$

where $\beta = (\mu C_{ox})W/L$ is determined by the fabrication process and the size of the transistor. V_{T1} is the threshold voltage of $M1$. When $W_2/L_2 \gg W_1/L_1$ and M_2 is biased in its saturation region, $(V_{gs2} - V_{T2} < V_{ds2})$, then $V_{ds1} \approx V_b - V_{T2}$. If V_b is a constant voltage, and it is assumed that $V_T = V_{T1} = V_{T2}$, then I can be written

$$I = \beta(V_b - V_T)\left[V_i - \frac{V_b}{2} + \frac{V_T}{2}\right] \tag{2}$$

or

$$I = G(V_i - V_{offset}). \tag{3}$$

In other words, when $M1$ biased in its active (ohmic) region, I is a linear function of the input voltage $V_i = V_{gs1}$ and has a transconductance, G controlled by the bias voltage V_b.

Fig. 2 shows a graph of the response I versus V_i as a function of $V_b = V_B$ for a test chip which was fabricated in a 3-μm CMOS process [5]. When $V_{gs1} < V_b$, the response is nonlinear and I approaches 0. For $V_i > V_b$, the linearity of

Manuscript received July 8, 1988; revised November 2, 1988. This paper was recommended by Guest Editors R. W. Newcomb, and N. El-Leithy.

R. D. Reed was with the Texas A&M University, College Station, TX. He is now with World Instruments, Longview, TX.

R. L. Geiger is with the Department of Electrical Engineering, Texas A&M University, College Station, TX 77801.

IEEE Log Number 8826722.

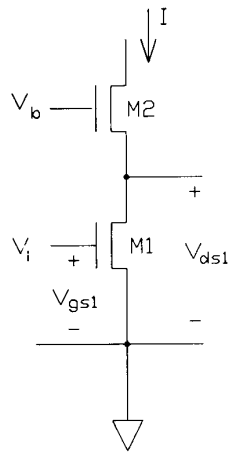


Fig. 1. MOSFET structure representing one weighted input. The input voltage V_{gs1} is modulated by the bias voltage V_b.

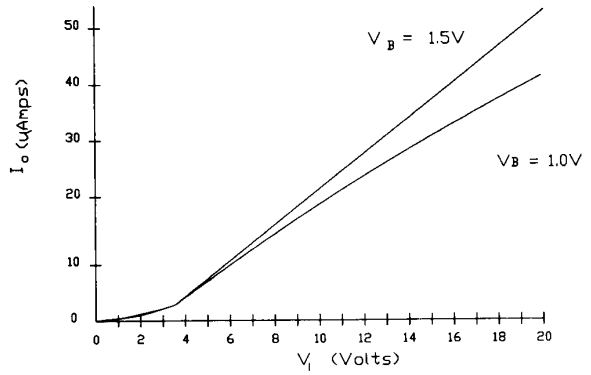


Fig. 2. Output current versus input voltage for one weighted link.

the response is most strongly dependent on the ratio $(W2/L2)/(W1/L1)$ with larger ratios giving a more linear response but generally requiring more circuit area. For the circuit tested, $W1/L1 = 3/30$ μm and $W2/L2 = 300/10$ μm giving a ratio $(W2/L2)/(W1/L1) = 300$. These are much larger than the minimum size devices because the original circuit was designed for linearity and frequency response rather than for area efficiency. In a neural network application, the optimizations would be made in favor of area efficiency because of the large number of synapses required. Simulations with $W1/L1 = 5/25$ μm and $W2/L2 = 100/5$ μm and a ratio of 100 show good linearity and might be a reasonable design starting point.

When a number of these blocks are connected to a common node, the currents sum according to Kirchoff's current law. When two of these modules are combined with a unity gain current mirror (as shown in Fig. 3 for two inputs on each module), the output current will be

$$I_{\text{out}} = \sum_i \pm G_i(V_i - V_{\text{offset}_i}) \tag{4}$$

with the sign of each term depending on if the block is connected to the input or output side of the current mirror. The circuit of Fig. 3 is thus recognized as a four-input

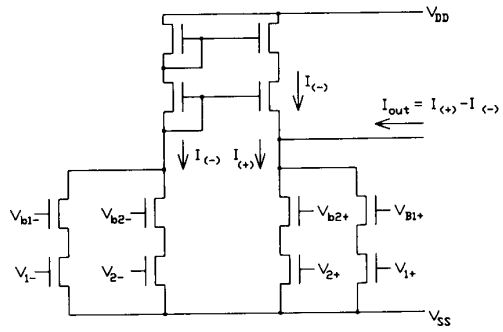


Fig. 3. Four-input multiple-input OTA (MIOTA).

OTA. Note that the voltages V_i and V_{bi} have been referred to the lower supply voltage. With the typical analog CMOS supply voltages of ± 5 V and $V_{\text{bi}} = -3.5$ V, for example, the input voltage has a linear range between -3.5 and $+5$ V, a span of 8.5 V.

The circuit in Fig. 3 was originally designed for use in conventional adaptive signal processing applications with special emphasis placed on obtaining a wide linear range and multiple inputs. The transconductance is controlled by the bias voltage in a continuous fashion and is used to tune the circuit. In order to remove the offset voltage which is undesirable in most applications, the groups of modules on either side of the current mirror are made symetrical in size and bias voltage.

Certain optimizations of this circuit can be made for use in neural networks. First, the number of inputs can be greatly increased. Second, symetrical excitory and inhibitory inputs are not required. This will give the circuit an overall offset but, typically, one or more inputs would be dedicated to setting the node threshold and these would be adjusted to account for the offset. Finally, since neural networks do not require a very linear response, the relative sizes of the input transistors can be reduced to save circuit area. If the linearity requirement is removed altogether and $M1$ is operated in the subthreshold region ($V_{gs1} < V_T$) then V_b will still control the transconductance and the structure becomes similar to the links used in [4]. This will minimize current consumption but places a restriction on the allowable input voltages.

The analysis above ignores a number of second order error sources such as device size mismatches and differences in V_T due to process variation and source–substrate bias. These effects are not thought to be critical in these applications; none of them cause the response to become nonmonotonic, for instance, and, because neural network training is an error tolerant process, it can compensate for certain circuit variations and nonlinearities.

Fig. 4 shows how the circuit can be used to implement an artificial neuron. The operational amplifier and resistor convert the output current to a voltage which, for equally sized devices, is given by

$$I_{\text{out}} = \sum_i \pm \beta(V_{\text{bi}} - V_T)(V_i - V_{\text{offset}_i}) \tag{5}$$

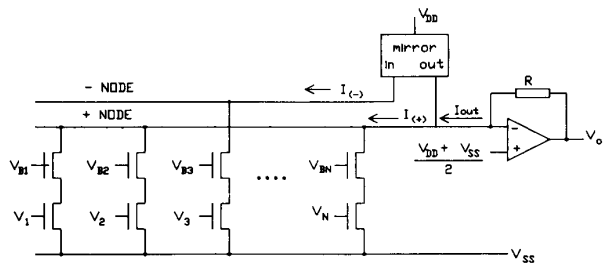


Fig. 4. MIOTA circuit modeling one neuron.

and

$$V_o = RF(I_{\text{out}}) \tag{6}$$

where $F(I_{\text{out}})$ is a nonlinear function describing the saturation behavior. (6) holds when each input $V_i > V_{\text{bi}}$. When $V_i < V_{\text{bi}}$, the response becomes nonlinear and dI_x/dV_i approaches 0. When the magnitude of I_{out} is greater than some value, the output voltage V_o will saturate at the supply voltage. These natural limiting effects give the circuit response a sigmoidal shape. ("Sigmoidal" is used here in the sense of any smooth "S" shaped curve; it does not refer to a specific function.)

When R is large, $F(I_{\text{out}})$ will saturate at smaller currents and the response will approach a binary step function. External constant inputs can bias the circuit to effectively shift the threshold as needed. When R is small, $F(I_{\text{out}})$ will be a more smoothly varying sigmoidal function.

The sign in (6) is determined by whether the input is excitory or inhibitory. Excitory signals such as V_1 and V_2 draw current from the positive node; inhibitory signals such as V_3 draw current from the negative node. The weight of each input, (i), is controlled by its bias voltage V_{bi}.

This circuit appears to be useful for VLSI implementations of neural nets because each weighted link can be realized with just two MOSFET transistors and all inputs are high impedances which respond to voltages rather than currents. It has advantages over typical op amp voltage summing circuits because the weight of each input is continuously controllable with a bias voltage rather than being determined by a fixed resistor or being switched in discrete steps. Also, the limited input impedance of each resistor in the typical op amp summing circuit means that a node driving a large number of these inputs would be required to source a relatively large current.

Regular arrays of these circuits can be laid out in a crossbar arrangement to create large VLSI networks. The crossbar arrangement is a common one and has been used by Hopfield and Tank [9], among others, to connect every input to every summing node. This paper does not address the problem of how the $N \times M$ bias voltages needed to control the weights of a circuit with N inputs and M outputs would be generated and stored. One method would be to store the voltage on a capacitor which is periodically refreshed by another system addressing the capacitors in a row/column fashion. The external system would be responsible for setting and adjusting the weights of the network. Another method would use EPROM or EEPROM cells optimized to store analog voltages.

This synaptic circuit might also be used if, as in biological networks, pulses with a firing rate proportional to the signal are used rather than dc voltages. The circuit produces a current proportional to the weighted sum of the input signals which could be integrated on a capacitance. The output pulses could then be generated by a simple op amp circuit or by neuristor circuits of the type described in [10]. $M2$ of Fig. 1 would still act to control the relative strength of each input pulse.

Summary

A multiple-input OTA circuit has been presented which may be useful in VLSI implementations of neural networks. It generates an output voltage which is a sigmoidal function of the linear sum of a large number of input voltages, each input having a weight which is set by an externally controllable bias voltage. Large numbers of these cells can be fashioned in regular arrays. It appears to be efficient because each weighted connection is implemented with only two MOSFET transistors.

References

[1] *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. James L. McClelland and David E. Rumelhart, Eds., Cambridge: MIT Press, 1986.

[2] *Proc. IEEE First Int. Conf. on Neural Networks*, San Diego, Ca, June 21–24, 1987.

[3] Hans P. Graf, Lawrence D. Jackel, and Wayne E. Hubbard, "VLSI implementation of a neural network model," *Computer*, March 1988.

[4] James Hutchinson, Christof Koch, Jin Luo, and Carver Mead, "Computing motion using analog and binary resistive networks," *Computer*, Mar. 1988.

[5] M. A. Sivilotti, M. R. Emerling, and C. A. Mead, "VLSI architectures for implementation of neural networks," in *Proc. Amer. Inst. of Physics Conf. on Neural Networks for Computing*, pp. 408–413, Snowbird, UT, 1987.

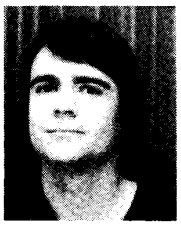
[6] A. Moopen, H. Langenbacher, A. P. Takoor, and S. K. Khanna, "Programmable synaptic chip for electronic neural networks," in *Proc. Amer. Inst. of Physics Conf. on Neural Information Processing Systems*, pp. 564–572, Denver, CO, 1987.

[7] H. P. Graf, L. D. Jackel, R. E. Howard, B. Straughn, J. S. Denker, W. Hubbard, D. M. Tennant, and D. Schwartz, "VLSI implementation of a neural network memory with several hundreds of neurons," *Proc. Amer. Inst. of Physics Conf. on Neural Information Processing Systems*, pp. 182–187, Denver, CO, 1987.

[8] Russell D. Reed, "A multiple-input operational transconductance amplifier with a wide linear range," Masters Thesis, Texas A&M Univ., Aug. 1986.

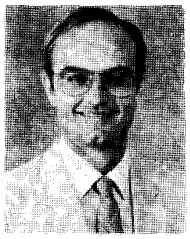
[9] David W. Tank and John J. Hopfield, "Simple 'neural' optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit," *IEEE Trans. Circuits Syst.*, vol. CAS-33, May 1986.

[10] R. W. Newcomb, "MOS Neuristor Lines," *Constructive Approaches to Mathematical Models*. Coffman/Fix, Eds. New York: Academic, pp. 87–111, 1970.



Russell D. Reed (S'79–M'84) received the B.S. and M.S. degrees in electrical engineering from Texas A&M University in 1981 and 1986, respectively.

Mr. Reed is a member of INNS. He is currently with World Instruments in Longview, Texas.



Randall L. Geiger (S'75–M'77–SM'82) received the B.S. degree in electrical engineering and the M.S. degree in mathematics from the University of Nebraska, Lincoln, in 1972 and 1973, respectively, and the Ph.D. degree in electrical engineering from Colorado State University, Fort Collins, in 1977.

He joined the Department of Electrical Engineering at Texas A&M University in 1977 and currently holds the rank of Professor. His technical research interests include monolithic linear IC design, telecommunications, filter design, biotechnology and VLSI circuits. He received the Myril B. Reed Best Paper Award at the 1981 Midwest Symposium on Circuits and Systems for presentation of a paper titled "Switched Resistor Filters". He is co-founder and past president (1982–1985) of World Instruments Inc., a firm specializing in microprocessor based instrumentation. He served as Conference Chairman at the 1983 UGIM Conference, has been a member of the Midwest Symposium Steering Committee since 1980 and is a Registered Professional Engineer in the State of Texas. He has served as session chairman and organizer at the IEEE International Symposium on Circuits and Systems and the Midwest Symposium, was an Associate Editor of the IEEE Transactions on Circuits and Systems from 1983 to 1985, and is currently a Member of the Administrative Committee of the IEEE Circuits and Systems Society and the Circuits and Systems Society Editor of the *IEEE Circuits and Devices Magazine*.