Associative IC Memories with Relational Search and Nearest-Match Capabilities

Sateh M. S. Jalaleddine, *Member, IEEE,* and Louis G. Johnson, *Member, IEEE*

***Abstract*—An associative or content addressable memory (CAM) is crucial in data management and pattern recognition applications. In this paper we present the design and implementation of CAM's that execute relational ($<$, $>$, $\leq$, $\geq$) and nearest-match instructions. Implementation of a new relational search cell is presented. Direct performance comparison shows an order of magnitude improvement over existing designs with similar cell area. Then we present the design and implementation of a neural-inspired nearest-match CAM using a winner-take-all (WTA) network. An original approach to analyzing such neural-inspired CAM's is presented. A model which describes the behavior of the WTA network is derived to be utilized in the design and performance prediction of the network. Performance of the WTA network in differentiating between words with large bit mismatches is analyzed and an upper bound is set. Fully functional prototype chips have been fabricated through MOSIS using 2-μm double-metal CMOS technology. Theoretical, simulation, and physical chip measurements are in good agreement.**

I. Introduction

SUPERIORITY of associative or content addressable memory (CAM) in data management applications was proposed over three decades ago [1]. The inherent interconnect and size limitations constrain the use of CAM's in large database machines [2]. However, CAM's have been proposed for many important applications in which small-size CAM's (4 kb to a few hundred kilobits) can be utilized, for example, artificial intelligence [3]–[5], database accelerators [6]–[9], image pattern recognition [10], [11], packet switching in communication networks [12]–[14], data-flow computers [15]–[17], and multitask processing [18], [19]. In the last few years, the upward trend in the memory capacity of on-chip CAM's has been mainly due to technological advances (4-kb CAM using 3-μm double-metal CMOS process in 1985 [20] to 160-kb CAM using 0.8-μm triple-metal CMOS process in 1990 [9]). It should be noted that in addition to technological advances, architectural constraints have been imposed in [9] to achieve the reported CAM size. Such constraints require the use of common exact-match logic for every 16 b of SRAM as opposed to having matching logic for each bit.

Manuscript received August 27, 1991; revised February 3, 1992.

S. M. S. Jalaleddine is with the Electrical Engineering Department, Western Michigan University, Kalamazoo, MI 49008.

L. G. Johnson is with the Electrical and Computer Engineering Department, Oklahoma State University, Stillwater, OK 74078.

IEEE Log Number 9108187.

Traditionally, hardware CAM's involve implementation of the exact-match operation [21]–[26] whereby a memory search involves *exact* matching between an input word and stored ones. If no exact match is found, then a no-match flag is asserted. Recently, neural-inspired CAM's that exhibit outstanding features in executing the nearest-match operation have been reported [27]–[31]. In such CAM's, the constraint of searching for an *exact* match is *relaxed* to finding the nearest match (in Hamming distance sense). Thus, the memory search relaxes to a nearest match when there is no exact match. Digital techniques as opposed to neural-inspired approaches have also been employed to implement nearest-match CAM's [5], [6], [9], [32], [33]. These CAM's generally require several memory cycles to reach a decision while neural-inspired CAM's can evoke a decision in one cycle [34].

In this paper we present design/analysis and IC implementation of efficient CAM's that execute relational ($<$, $>$, $\leq$, $\geq$) and nearest-match instructions [35]. Section II describes the design and implementation of a new relational search CAM. Section III concerns the design, analysis, and implementation of a neural-inspired nearest-match CAM using a winner-take-all (WTA) network.

II. Relational Search CAM

The relation ($<$, $>$, $\leq$, $\geq$) between two words can be determined by implementing the carry function of a binary full subtractor (a binary full adder with one input complemented) and examining the carry-in and carry-out values. Carry generation at a given bit position (i) of a full subtractor, performing subtraction of an input word (D) from a stored word (S), can be formulated as

$$\begin{aligned} C_{i+1} &= S_i\overline{D}_i + (\overline{S_i \oplus D_i})C_i \\ &= (S_i \oplus D_i)S_i + (\overline{S_i \oplus D_i})C_i \\ &= (S_i \oplus D_i)\overline{D}_i + (\overline{S_i \oplus D_i})C_i. \end{aligned} \quad (1)$$

By virtue of the above carry function, the values of the carry-in (to the LSB) and carry-out (from the MSB) of the subtracter (S–D) determine the relation between unsigned words S and D as follows:

Carry-in (LSB)	Carry-out (MSB)	Relation between stored (S) and input data (D)
0	0	$S \leq D$
1	0	$S < D$
0	1	$S > D$
1	1	$S \geq D$

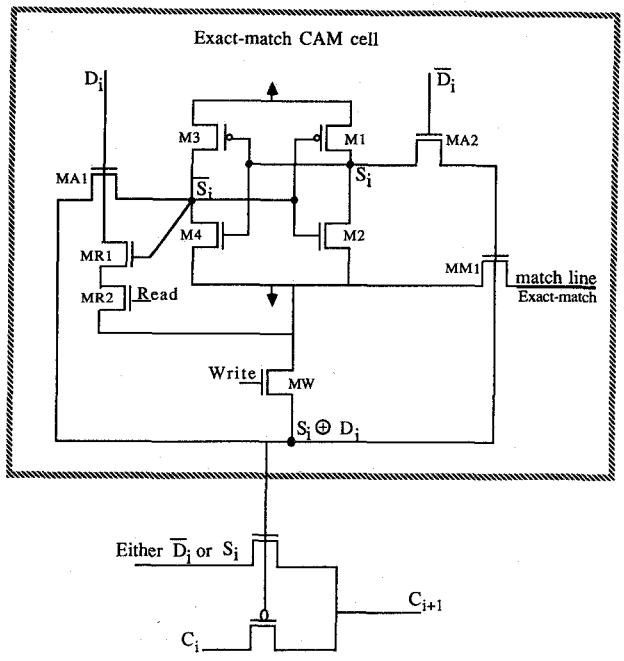


Fig. 1. Relational search CAM cell.

The realization of a practical relational search CAM calls for efficient circuit implementation of the bit cell, which is presented next.

A. Relational Search CAM Cell

We take advantage of the fact that the EXCLUSIVE-OR function ($S_i \oplus D_i$) is the basic function of an exact-match CAM cell. A compact circuit implementation of (1) using an exact-match CAM cell, which produces a relational CAM cell, is shown in Fig. 1 [35]. Thus, the addition of only two pass transistors allows a CAM cell to execute relational search operations. It should be noted that any exact-match CAM cell (see, for example, [36]) can be employed to provide the $S_i \oplus D_i$ function. The exact-match CAM cell shown in Fig. 1 is a capacitive-access cell, which is discussed elsewhere [37]. The stored datum in the CAM cell is S_i while D_i is the corresponding input datum to be compared with S_i.

B. Relational CAM Chip Implementation

The construction of a relational search CAM requires cells to be cascaded to form data words. Cascading relational search CAM cells forms a ripple carry chain for each word. A large number of bits in a word causes long carry chains and consequently longer signal propagation delay. Reducing carry-chain delay can be accomplished by inserting buffers every K cells. It is shown in [35] that inserting a buffer every three cells ($K = 3$) results in the shortest delay attainable through the carry chain without resorting to lookahead carry techniques. It is also shown that choosing $K = 4$ is a better compromise among area, power consumption, and carry-chain delay. A fully functional 4-word $\times$ 8-b chip has been fabricated through MOSIS using 2-μm double-metal CMOS technology. The chip implementation included a buffer inserted every four cells. The measured delay time for a relational search is 45 ns including the I/O pad delays. However, the delay due to the carry chain (8-b word) itself (excluding I/O pad delays) is around 15 ns. This carry-chain delay is based upon excellent agreement among typical data (I/O pad delays) provided by MOSIS along with theoretical and simulation results [35].

C. Comparison with Existing Designs

Published relational search CAM cell designs either require extensive hardware area (logic gates) in each cell [38], [39] or are implemented by an iterative algorithm using an off-chip processor [5] which is time consuming. For direct comparison between our implementation and existing ones, each design is evaluated on the basis of 1) the number of transistors needed in addition to a RAM cell using CMOS technology, 2) the implementation technology, 3) the number of bus (control) lines that are needed for each word (as transistor feature size is shrinking, the number of interconnect lines should be minimized to avoid significant performance limitations), and 4) the time performance of a relational search based on a 32-b word length. For achieving the same basis of comparison, the relational search time of our implementation is extrapolated from the measured 8-b time to a 32-b time. The number of transistors needed in addition to a RAM cell for our relational CAM cell is five transistors, i.e., two pass transistors for the carry-function implementation, one transistor for the buffer (cascaded two CMOS inverters) inserted every four CAM cells, and two transistors needed to implement the EXCLUSIVE-OR function as is the case in a standard CAM cell [36] (the pull-down transistor for exact-match line discharge is not necessary for relational

TABLE I
COMPARISON BETWEEN THIS RELATIONAL SEARCH CAM IMPLEMENTATION AND EXISTING DESIGNS

Design	# of Transistors Needed in Addition to a RAM Cell	# of Bus Lines That Extend a Word	Implementation Technology	Time Performance Per 32-b Word
Ramamoorthy *et al.*, 1978 [38]	36	2	Not implemented	Not estimated
Hurson, 1986 [39]	32	4	Not implemented	Estimated 125 ns
Ogura *et al.*, 1989 [5]	3	1	1.2-μm CMOS 2-metal	Based on IC measurements 5440 ns
The relational search CAM presented in this paper	5	1	2-μm CMOS 2-metal	Based on IC measurements 90 ns

search). All designs require logic gates for each word to provide relational information processing at the word level. Such word overhead is not included in our comparison since it is required for all designs and is relatively negligible for even small-size word CAM's. The comparison results summarized in Table I indicate that the time performance of our chip is at least an order of magnitude better than existing designs with a comparable cell area.

It should be emphasized that size of the transistors (pass transistors) in our implementation can be of a minimum size as opposed to the designs in [38] and [39] where logic gate implementations may require larger transistors to maintain reasonable propagation speed. Moreover, our design performance results are based on physical silicon implementation in contrast to the estimated claims in some designs [38], [39].

III. NEAREST-MATCH CAM

The theme of this section is a CAM that finds the nearest Hamming distance match among stored data to an input word. Unlike the relational search CAM, the bits in each word are interpreted as having equal weight. Moreover, the power of a relational search CAM lies in its ability to find *all the words* that satisfy a relational criterion to an input word in one cycle. On the other hand, the nearest-match CAM possesses the capability of finding *a word* that best matches an input word in one cycle.

We present an implementation of a winner-take-all (WTA) network [40] which is similar in operation to the neural-inspired Hamming net [41]. Both networks dynamically find the nearest match between an input and stored words, but the Hamming net requires each word to be connected to all the other words while the WTA network requires only one connection among all the words. This makes the area of the Hamming net $O(n^2)$ while the WTA area is only $O(n)$, where n is the total number of words. We have presented the silicon implementation and analysis of a nearest-match CAM similar to the Hamming net elsewhere [34], [42]. The subject of this section concerns a nearest-match CAM with $O(n)$ WTA network.

A subthreshold circuit implementation of an $O(n)$ WTA network has been implemented by Lazzaro *et al.* [43] for use in an image processing system. The basic circuit operation of their implementation is that the winner word corresponds to the circuit with the largest weight currents. Hence, the weight current for a given word must be proportional to the number of matching bits. CAM performance is usually determined by examining how well a CAM can differentiate between a winner and a close loser word where the number of mismatched bits is small compared to the total number of bits in the word. Thus, Lazzaro *et al.*'s WTA circuit is not suitable for CAM implementation since small amounts of mismatch correspond to small percentage changes in the match current.

In the following subsections our WTA circuit implementation is presented. Then a simple transistor model is introduced that allows a closed-form solution of the WTA circuit behavior. The transistor model is then used to derive a model for the WTA winner and loser words. The winner word expressions are used to determine the WTA circuit design parameters. The loser word expression is utilized for determining the resolution of the WTA network in differentiating between words with large bit mismatches with the input word. Then SPICE simulation results are presented that show good agreement with the theoretical analysis. Finally, a fully functional prototype chip is presented. Chip test results show good agreement with theoretical and SPICE simulation results.

A. Winner-Take-All Network

Unlike Lazzaro *et al.*'s WTA circuit, our circuit operates by selecting the minimum weight current word as the winner and the MOS transistors are biased in strong inversion. We set the weight current to the amount of mismatch between the input and stored word. Hence, the weight current has a large percentage change between words with a small number of mismatches, which allows better performance in most CAM applications. Similar to Lazzaro *et al.*'s [43] circuit, our circuit (shown in Fig. 2) uses source followers for interconnecting the words in the network. The weight current of each word is determined by the number of MC_{ij} transistors that are on. Whenever an input datum mismatches a stored one, then transistor MC of that CAM cell will be on ($V_{S \oplus D}$ is high).

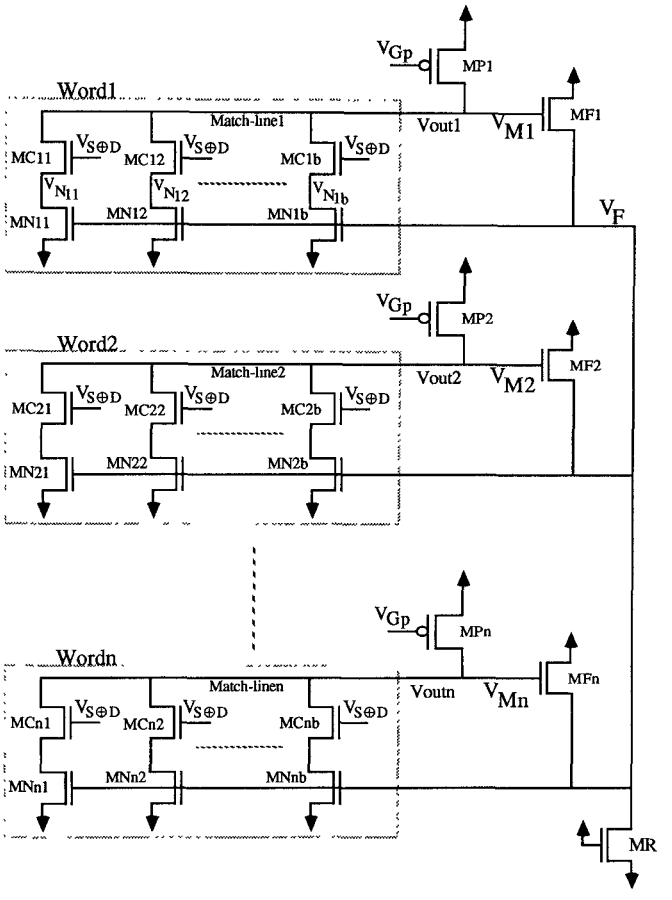


Fig. 2. Winner-take-all network (b-bits by n-words) circuit implementation.

The principal operation of our WTA circuit is that the winner word (lowest pull-down conductance) sets the feedback voltage V_F common to all the words. The conductance of the series transistors *MC* and *MN* relative to *MP* determines the output voltage of each word, V_{out_i}. Increasing the conductance of *MC* and *MN* while keeping the conductance of *MP* unchanged results in lower output voltage. Hence, a word with more bit mismatches has more *MC* transistors turned on, which results in a higher pull-down conductance. The output voltage of the winner is higher than the rest of the words since its pull-down current (conductance) is the smallest. Consequently, the source follower of the winner word drives the feedback voltage V_F to a higher voltage than would have been possible from any of the other loser words. Since the loser words are now driven by a higher feedback voltage than the one required to regulate their output voltage level, their output voltages fall even further below the winner output voltage. The loser words with higher pull-down currents will have their outputs closer to zero.

B. Transistor Model

A simple quasi-empirical MOS model is qualitatively sufficient to allow reliable prediction of the WTA circuit behavior, as will be shown by comparison with SPICE simulations presented in a later section. We employ a simple short-channel MOS model that allows a closed-form solution of circuit operation to be achieved. The simple MOS model for an nMOS is

$$I_D = G(V_{GS} - V_T)\frac{V_{DS}}{V_c + V_{DS}}, \qquad V_{DS} < V_c \quad (2a)$$

$$I_D = G'(V_{GS} - V_T)(1 + \lambda V_{DS}), \qquad V_{DS} \geq V_c \quad (2b)$$

$$I_D = 0, \qquad V_{GS} < V_T \quad (2c)$$

where

- V_c carrier velocity saturation voltage equal to $L_{\text{eff}}E_c$,
- G conductance of a transistor operating in the ohmic region,
- G' saturation region transistor conductance $= G/2$,
- λ drain barrier lowering/channel-length modulation factor,
- V_{GS} gate-to-source voltage,
- V_{DS} drain-to-source voltage,
- V_T the extrapolated strong inversion threshold voltage,
- L_{eff} effective channel length,
- W_{eff} effective channel width.

The transistor parameters G, V_T, V_c, and λ are determined from empirical fits to transistor currents for a particular process. Modeling a pMOS device is achieved by having voltage polarities reversed.

C. Derivation of Winner Word Expressions

An analytical solution for the WTA network in Fig. 2 is possible since the common feedback voltage V_F depends only on the winner word. In this section, we derive expressions for V_F, V_{Mw}, and V_{Nw} of the winner word. V_{Nw} is defined such that $V_{Nwj} = V_{Nw}$, where V_{Nwj} is the drain-to-source voltage of MN_{wj} and is assumed to be the same in each bit position j, assuming all MN_{ij} and MC_{ij} transistors are matched.

For the winner word circuit to function most efficiently, all transistors except *MR* should operate in saturation. *MF* of the winner word (MF_w) operates in saturation (*MF* transistors in all loser words are cut off) and sets the feedback voltage V_F which, in turn, regulates the output voltage (V_{Mw}) through the amplifier consisting of the MN_{wj}, MC_{wj}, and MP_w transistors. Transistors MP_j should operate in saturation so that an equal pull-up current to all the words is supplied by setting a common bias (V_{Gp}) to the gates of transistors MP_i. Biasing one word with a bigger pull-up current than the rest of the words can be viewed as increasing the weight of that word, which is out of the scope of our present application of the WTA. In the present WTA CAM application, the word weight is determined *only* by the number of pull-down transistors that are turned on (mismatched bits). Finally, MN_{wj} and MC_{wj} transistors operate in the saturation region since the winner output voltage (V_{Mw}) should be at a high voltage level.

After determining each transistor operation region of the winner word, we derive expressions for V_F, V_{Mw}, and V_{Nw}. The sum of currents at nodes *Mw*, *Nwj*, and *F* for

the winner word (see Fig. 2) can be expressed by the following differential equations:

$$C_{Mw}\frac{dV_{Mw}}{dt} = I(MP_w) - \sum_{j}^{m_w} I(MC_{wj}) \tag{3a}$$

$$C_{Nwj}\frac{dV_{Nwj}}{dt} = I(MC_{wj}) - I(MN_{wj}) \tag{3b}$$

$$C_F\frac{dV_F}{dt} = I(MF_w) - I(MR) \tag{3c}$$

where C_{Mw}, C_{Nwj}, and C_F are the parasitic capacitances at nodes Mw, Nwj, and F, respectively. Using the transistor model (2) where the transistor operation regions are assumed as discussed in the previous paragraph, (3) can be written as

$$C_{Mw}\frac{dV_{Mw}}{dt} = G'_P(V_{SGp} - |V_{Tp}|)[1 + \lambda_p(V_{dd} - V_{Mw})] - \sum_{j}^{m_w} G'_{Cwj}\{[V_{G\,\mathrm{on}} - V_{Nwj} - V_{Tn}(V_{Nwj})] \cdot [1 + \lambda n(V_{Mw} - V_{Nwj})]\} \tag{4a}$$

$$C_{Nwj}\frac{dV_{Nwj}}{dt} = G'_{Cwj}\{[V_{G\,\mathrm{on}} - V_{Nwj} - V_{Tn}(V_{Nwj})] \cdot [1 + \lambda n(V_{Mw} - V_{Nwj})]\} - G'_{Nwj}(V_F - V_{Tn})(1 + \lambda n V_{Nwj}) \tag{4b}$$

$$C_F\frac{dV_F}{dt} = G'_F[V_{Mw} - V_F - V_{Tn}(V_F)] \cdot [1 + \lambda_n(V_{dd} - V_F)] - G_R\frac{(V_{dd} - V_{Tn})(V_F)}{V_{cR} + V_F} \tag{4c}$$

where

- m_w number of mismatched bits in the winner word,
- $V_{G\,\mathrm{on}}$ the ON (high) voltage at node $S \oplus D$ in the CAM cell which indicates a mismatched bit,
- $V_{Tn}(V_{SB})$ threshold voltage with body effect where the source-to-body voltage is V_{SB},
- V_{cR} carrier velocity saturation voltage for transistor MR.

Since transistors MN_{wj} and MC_{wj} for each CAM cell are assumed identical (match-line pull-down transistors in each CAM cell), the summation term can be replaced by m_w, $G'_{Cwj} = G'_C$, $G'_{Nwj} = G'_N$, and $V_{Nwj} = V_{Nw}$. Moreover, assuming λ effects are negligible, (4) yields in the steady state

$$V_F = \frac{G'_P}{m_w G'_N}(V_{SGp} - |V_{Tp}|) + V_{Tn} \tag{5}$$

$$V_{Mw} = V_{Tn}(V_F) + \left[1 + \frac{G_R}{G'_F}\frac{(V_{dd} - V_{Tn})}{(V_{cR} + V_F)}\right]V_F \tag{6}$$

$$V_{Nw} = [V_{G\,\mathrm{on}} - V_{Tn}(V_{Nw})] - \frac{G'_P}{m_w G'_C}(V_{SGp} - |V_{Tp}|) \tag{7}$$

where $V_{Tn}(V_{Nw})$ is the threshold voltage of transistor MC with $V_{SB} = V_{Nw}$. As can be seen from (5) and (6), V_F and V_{Mw} decrease as m_w increases. This effect is anticipated since larger m_w means larger pull-down conductance which causes V_{Mw} to decrease.

Having derived expressions for V_{Mw}, V_F, and V_{Nw}, the next step is to examine the range of parameters where these equations are valid. Since the WTA circuit works best when the transistors operate in the regions assumed in deriving the above equations, the range of parameters where the equations are valid will be our design criterion for determining circuit parameters. For (5), (6), and (7) to be valid, transistors MP, MF, MN, and MC operate in the saturation region and transistor MR operates in the ohmic region. Thus

$$V_{dd} - V_{Mw} \geq V_{cP} \quad (MP \text{ saturated})$$

$$V_{dd} - V_F \geq V_{cF} \quad (MF \text{ saturated})$$

$$V_{Nw} \geq V_{cN} \quad (MN \text{ saturated})$$

$$V_{Mw} - V_{Nw} \geq V_{cC} \quad (MC \text{ saturated})$$

$$V_F < V_{cR} \quad (MR \text{ ohmic}).$$

Determining transistor size ratios that satisfy the above inequalities is achieved by: 1) substituting in the above inequalities the expressions for V_{Mw}, V_{Nw}, and V_F as given in (5)–(7); 2) setting $V_{Gp} = 3$ V, which is a conservative value for keeping MP in strong inversion; 3) setting $m_w = 1$ so that the network is able to pick out a winner word with only one bit mismatch; and 4) using typical transistor parameters based on MOSIS 2-μm CMOS process ($V_{Tn} = 0.81$, $|V_{Tp}| = 0.76$ V, $V_{Tn}(V_F) = V_{Tn}(V_{Nw}) = 1.2$ V, $\mu_n/\mu_p = 2.5$, $V_c = 1$ V, and $V_{G\,\mathrm{on}} = 3.6$ V), which yields $G'_P/G'_n \leq 1.1$ and $G_R/G'_F \leq 4/9$ where G'_P and G'_F are the saturation region conductances of transistors MP_i and MF_i. $G'_n = G'_N = G'_C$ is the saturation region conductance of transistors MN_{ij} and MC_{ij}. G_R is the ohmic region conductance for transistor MR. At this point, constraints on circuit parameters have been determined which limit the relative sizes of transistors MP, MF, MN, MC, and MR. Our analytical expressions and limits on circuit parameters derived in this section are in excellent agreement with results obtained using SPICE simulation shown in Fig. 3.

D. Derivation of Loser Word Expressions

The main goal in this section is to determine the resolution of the WTA circuit in differentiating between words with large bit mismatches to the input word. An analytical expression is obtained for the loser voltage (V_{ML}) using the MOS model (2). The requirement that transistor MF

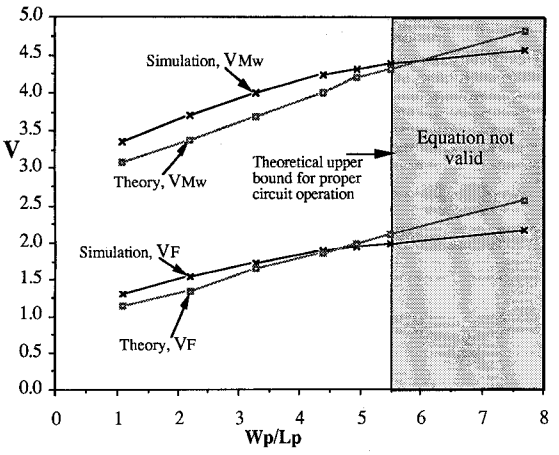


Fig. 3. Effects of pull-up transistor W_p/L_p on the feedback voltage (V_F) and winner voltage (V_{Mw}) where $m_w = 1$: theory (equations (5) and (6)) versus SPICE simulation.

of the loser word should be cut off is used as the criterion for proper loser word circuit operation. This is consistent with our assumption that only the winner word determines the feedback voltage V_F. Using this criterion along with the V_{ML} expression and V_F as given in (5), we arrive at an upper limit for bit mismatches between the winner and a loser word such that the WTA network still functions properly.

Denote m_L as the number of bit mismatches of a loser word with the input word where $m_w < m_L$. As discussed earlier, the output of a loser word (V_{ML}) is expected to be at a low voltage level. Thus, transistors *MN* and *MC* are assumed to operate in the ohmic region. As was the case in the winner word, *MP* should stay in saturation. Writing KCL equations at nodes *M* and *N* for a loser word using the MOS model (2) yields

$$G'_P(V_{SGp} - |V_{Tp}|)[1 + \lambda_p(V_{dd} - V_{ML})] - \sum_j^{m_L} G_{CLj}\left\{[V_{G\,\mathrm{on}} - V_{NLj} - V_{Tn}(V_{NLj})] \cdot \frac{V_{ML} - V_{NLj}}{V_{cc} + V_{ML} - V_{NLj}}\right\} = 0 \tag{8}$$

$$G_{CLj}\left\{[V_{G\,\mathrm{on}} - V_{NLj} - V_{Tn}(V_{NLj})]\frac{V_{ML} - V_{NLj}}{V_{cc} + V_{ML} - V_{NLj}}\right\} - G_{NLj}(V_F - V_{Tn})\left(\frac{V_{NLj}}{V_{cN} + V_{NLj}}\right) = 0 \tag{9}$$

where

m_L number of mismatched bits in a loser word,
$V_{G\,\mathrm{on}}$ the ON voltage at node $S \oplus D$ in the CAM cell,
$V_{Tn}(V_{SB})$ threshold voltage with body effect where the source-to-body voltage is V_{SB}.

As is the case in the winner word voltage derivation, we assume $G_{CLj} = G_C$, $G_{NLj} = G_N$, $V_{NLj} = V_{NL}$, and $\Sigma_j^{m_L} G_{CLj} = m_L G_C$. Manipulating the above equations gives

$$\frac{V_{NL}}{V_{cN} + V_{NL}} = \frac{G'_P(V_{SGp} - |V_{Tp}|)[1 + \lambda_p(V_{dd} - V_{ML})]}{m_L G_N(V_F - V_{Tn})}. \tag{10}$$

In order to arrive at a simple expression for V_{NL} using (10), 1) substitute V_F (set by the winner word) without neglecting the λ effects since the feedback voltage accuracy is important in determining the loser voltage, and 2) use typical MOSIS transistor parameters values and apply first-order approximations in calculating the λ effects by assuming V_{Mw}, V_{Nw}, and V_{ML} at their typical values of 4, 3, and 0 V, respectively. Equation (10) yields

$$V_{NL} = V_{cN}\frac{0.71m_w}{(m_L - 0.71m_w)}. \tag{11}$$

Substituting (11) in (8) and (9) and manipulating, with λ_p effect approximated to first order as $(1 + \lambda_p V_{dd})$, gives

$$V_{ML} = V_{cN}\frac{0.71m_w}{(m_L - 0.71m_w)} + V_{cc}\left\{\frac{1}{\dfrac{m_L G_C\{V_{G\,\mathrm{on}} - V_{Tn}(V_{NL}) - V_{cN}[0.71m_w/(m_L - 0.71m_w)]\}}{G'_P(V_{SGp} - |V_{Tp}|)(1 + \lambda_p V_{dd})} - 1}\right\}. \tag{12}$$

The above equation shows the loser voltage increases as m_w increases. Since the winner voltage decreases as m_w increases, there is an implied limitation on the ability of the WTA circuit to detect large bit mismatches. The major constraint on the loser word circuit to operate properly (its output voltage stays low enough) is that transistor *MF* stays cut off (since V_F should be set by the winner word). Hence, for *MF* to be cut off, the inequality

$$V_{ML} - V_F < V_{Tn}(V_F) \tag{13}$$

should be satisfied.

The resolution of the WTA circuit in differentiating between words with large bit mismatches is such that (13) is obeyed. Substituting V_{ML} and V_F expressions in (13), using typical MOSIS transistor parameter values, and assuming large m_L and m_w, (13) can be approximated as

$$m_L \geq 1.1m_w. \tag{14}$$

Thus, reliable resolution of a winner and loser word requires that a loser be 10% more different from the input word than the winner is.

E. Comparison Between Analytical Analysis and SPICE Simulation

The SPICE level 2 MOS model provided by MOSIS, for a typical 2-μm CMOS process, is used in all our simulation runs. To verify our analysis as compared with SPICE simulations, the following circuit parameters were employed: $V_{dd} = 5$ V, $V_{Tn} = 0.81$ V, $|V_{Tp}| = 0.76$ V, $\mu_n/\mu_p = 2.5$, $\gamma_n = 0.55\ V^{0.5}$, $\lambda_n = 0.04\ V^{-1}$, $\lambda_p = 0.05\ V^{-1}$, $2|\phi_F| = 0.6$ V, $V_{Gp} = 3$ V, $V_{Tn}(V_F) = V_{Tn}(V_{Nw}) = V_{Tn}(1.5\text{ V}) = 1.2$ V, $W_N/L_N = W_C/L_C = 4/2$, $V_c = 1$ V, $W_R/L_R = 4/4$, $W_p/L_p = 8/2$, and $W_F/L_F = 12/2$. The effect of m_w and m_L on V_{Mw} and V_{ML}, and effects of transistor W/L ratios on circuit performance as predicted by the theoretical analysis have been exhaustively verified [35]. Here we present some of these results.

Fig. 3 shows theoretical and SPICE results for V_{Mw} and V_F as W_p/L_p varies from 2/2 to 14/2. Notice that the *MP* transistor starts to operate in the ohmic region as W_p/L_p is increased beyond 11/2 (V_{Mw} greater than 4.25 V), which is why the theory and SPICE simulation do not agree in this region. Fig. 4 shows theoretical and simulation results for the winner voltage V_{Mw} and the feedback voltage V_F as a function of m_w. Good agreement is obtained between our theory and SPICE for a wide range of m_w. It is very important to predict V_{Mw} for large m_w so that circuit design and performance estimation can be adequately achieved as in (14). Fig. 5 shows the loser voltage as a function of m_L when $m_L = m_w + 1$ as determined analytically by (12) and by SPICE simulation. Both theoretical and simulation results are in good agreement, especially for large m_L which is the region most crucial in determining the circuit resolution performance.

We now consider the WTA performance in detecting large bit mismatches using SPICE as compared to the theoretical upper bound of m_L set by (14). Fig. 6 shows SPICE simulation of a 16-word × 30-b WTA showing only the winner word with m_w bit mismatches and four loser words with m_L of $m_w + 1$, $m_w + 2$, $m_w + 3$, and $m_w + 4$. The difference voltage $V_{ML} - V_F$ is less than $V_{Tn}(V_F)$ as long as the upper limit constraint on m_L (14) is obeyed. Moreover, notice that the difference between the winner voltage and the runner-up loser voltage, as constrained by (14), never decreases below 1 V for the total range of m_L shown in Fig. 6. Such winner and runner-up voltage difference is sufficient for differentiating winner from loser words.

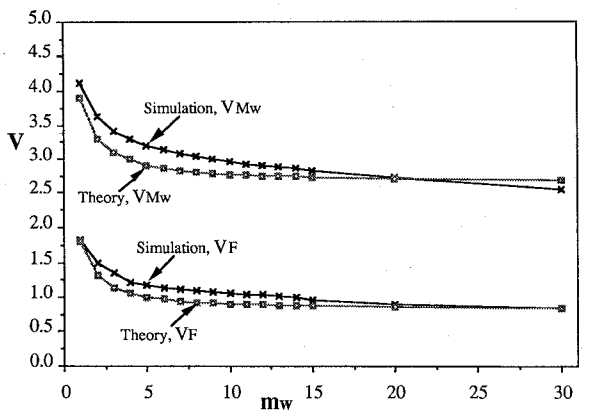


Fig. 4. Feedback voltage (V_F) and winner voltage (V_{Mw}) as a function of m_w: theory (equations (5) and (6)) versus SPICE simulation.

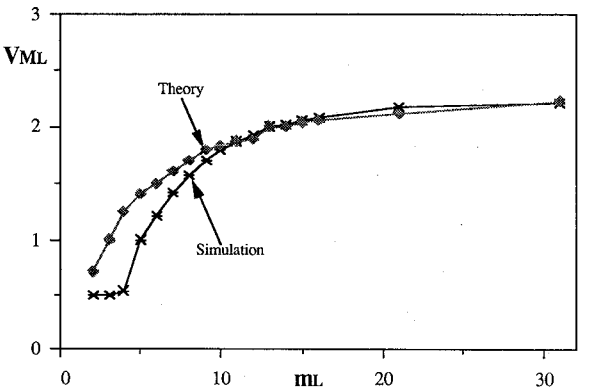


Fig. 5. Loser voltage (V_{ML}) as a function of m_L where $m_L = m_w + 1$: theory (equation (12)) versus SPICE simulation.

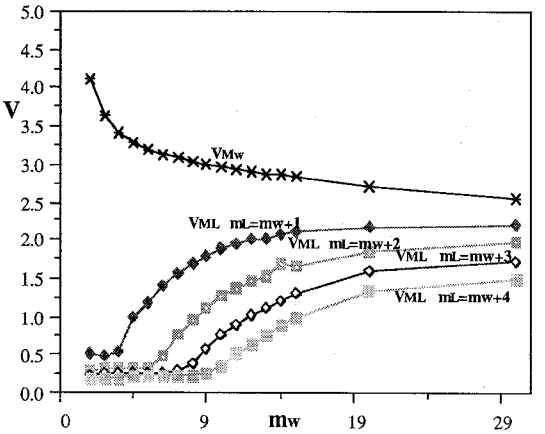


Fig. 6. SPICE simulation of a 16-word × 30-b WTA CAM showing the winner word with m_w bit mismatches and four loser words with m_L of $m_w + 1$, $m_w + 2$, $m_w + 3$, and $m_w + 4$.

F. IC Chip Test Results

A 16-word × 12-b nearest-match CAM IC chip was fabricated through MOSIS using 2-μm double-metal CMOS technology. Transistor ratios employed in the chip are: $W_p/L_p = 9/6$, $W_F/L_F = 24/2$, $W_R/L_R = 4/4$, and $W_N/L_N = W_C/L_C = 4/2$. For testing purposes, each input and output of the CAM circuit is connected to an external pin. This allowed direct measurement of the analog voltages on the match lines for each word. Thus, the size of the prototype CAM is limited by the number of pins available on the chip. Our prototype CAM is designed for implementation using a 40-pin MOSIS TinyChip. A microphotograph of the chip is shown in Fig. 7. The chip was tested and its correct functionality has been fully verified [35]. Good agreement among theoretical, SPICE, and physical chip measurements has been achieved for the winner and loser voltages as shown in Fig. 8. The mea-

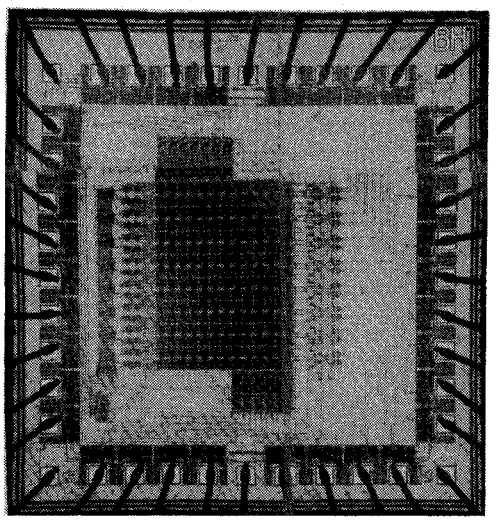


Fig. 7. Microphotograph of the 16-word × 12-b prototype WTA CAM chip.

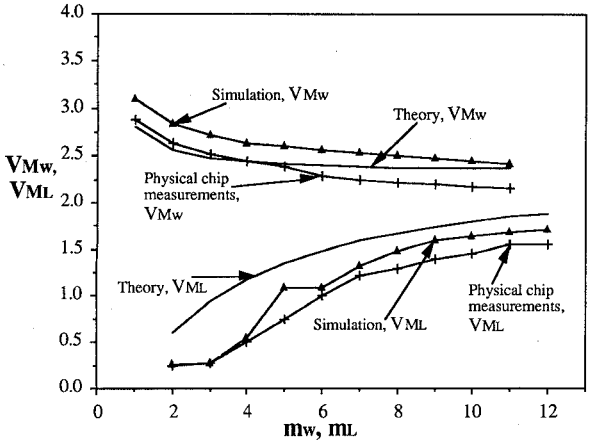


Fig. 8. Theoretical, SPICE, and physical chip measurements of the WTA winner and loser voltages, where $m_L = m_w + 1$.

sured convergence of the circuit was less than 90 ns in the worst case of $m_w = m_w + 1$.

IV. Conclusion

We have presented the design and implementation of CAM's that execute relational and nearest match instructions. The first part described the design and implementation of a new relational search CAM. A fully functional prototype chip was fabricated and tested. Direct performance comparison between our relational search CAM and existing ones was presented. The chip performance showed an order of magnitude improvement over existing designs with similar cell area. The second part concerned the design, analysis, and implementation of a neural-inspired nearest-match CAM using a winner-take-all (WTA) network. An original approach to analyzing such neural-inspired CAM's was presented. A model that describes the behavior of the WTA network was derived to be utilized in the design and performance prediction of the network. Performance of the WTA network in differentiating between words with large bit mismatches was analyzed and an upper bound was set. A fully functional prototype chip has been fabricated and tested. Theoretical, SPICE simulation, and physical chip measurements are in good agreement. All chips have been fully functional from first fabrication runs.

References

[1] A. G. Hanlon, "Content-addressable and associative memory systems: A survey," *IEEE Trans. Electron. Comput.*, vol. EC-15, pp. 509–512, Aug. 1966.

[2] G. Z. Qadah, "Database machines: A survey," in *Proc. 1985 AFIPS Nat. Comput. Conf.*, vol. 54, pp. 211–223.

[3] J. G. Bonar and S. P. Levitan, "Real-time LISP using a content addressable memory," in *Proc. 10th Int. Conf. Parallel Processing*, Aug. 1981, pp. 112–117.

[4] J. Naganuma *et al.*, "High-speed CAM-based architecture for a Prolog machine (ASCA)," *IEEE Trans. Comput.*, vol. 37, pp. 1375–1383, Nov. 1988.

[5] T. Ogura *et al.*, "A 20-kbit associative memory LSI for artificial intelligence machines," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1014–1020, Aug. 1989.

[6] H. Yamada *et al.*, "A high-speed string-search engine," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 829–834, Oct. 1987.

[7] M. Hirata *et al.*, "A versatile data string-search VLSI," *IEEE J. Solid-State Circuits*, vol. 23, pp. 329–335, Apr. 1988.

[8] J. P. Wade and C. G. Sodini, "A ternary content addressable search engine," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1003–1013, Aug. 1989.

[9] M. Motomura *et al.*, "A 1.2-million transistor, 33-MHz, 20-b dictionary search processor (DISP) ULSI with a 160-kb CAM," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1158–1165, Oct. 1990.

[10] S.-I. Chae *et al.*, "Content-addressable memory for VLSI pattern inspection," *IEEE J. Solid-State Circuits*, vol. 23, pp. 74–78, Feb. 1988.

[11] S. R. Jones *et al.*, "A 9-kbit associative memory for high-speed parallel processing applications," *IEEE J. Solid-State Circuits*, vol. 23, pp. 543–548, Apr. 1988.

[12] D. Bursky, "Content-addressable memory does fast matching," *Electron. Design*, vol. 36, no. 27, pp. 119–121, Dec. 8, 1988.

[13] H. Yamada *et al.*, "Real-time string search engine LSI for 800-Mbit/sec LANs," in *Proc. Custom Integrated Circuits Conf.*, May 1988, pp. 21.6.1–21.6.4.

[14] A. J. McAuley and C. J. Cotton, "A self-testing reconfigurable CAM," *IEEE J. Solid-State Circuits*, vol. 26, pp. 257–261, Mar. 1991.

[15] H. Kadota *et al.*, "An 8-kbit content-addressable and reentrant memory," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 951–957, Oct. 1985.

[16] G. A. Uvieghara *et al.*, "An on-chip smart memory for a data-flow CPU," *IEEE J. Solid-State Circuits*, vol. 25, pp. 84–94, Feb. 1990.

[17] H. Takata *et al.*, "A 100-mega-access per second matching memory for a data-driven microprocessor," *IEEE J. Solid-State Circuits*, vol. 25, pp. 95–99, Feb. 1990.

[18] H. Bergh, J. Eneland, and L.-E. Lundstrom, "A fault-tolerant associative memory with high-speed operation," *IEEE J. Solid-State Circuits*, vol. 25, pp. 912–919, Aug. 1990.

[19] L. R. Tamura *et al.*, "A 4-ns BiCMOS translation-lookaside buffer," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1093–1101, Oct. 1990.

[20] T. Ogura, S.-I. Yamada, and T. Nikaido, "A 4-kbit associative memory LSI," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1277–1281, Dec. 1985.

[21] R. Igarashi, T. Kurosawa, and T. Yaita, "A 150-nanosecond associative memory using integrated MOS transistors," in *ISSCC Dig. Tech. Papers*, Feb. 1966, pp. 104–105.

[22] J. T. Koo, "Integrated-circuit content-addressable memories," *IEEE J. Solid-State Circuits*, vol. SC-5, pp. 208–215, Oct. 1970.

[23] J. L. Mundy *et al.*, "Low-cost associative memory," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 364–369, Oct. 1972.

[24] R. M. Lea, "Low-cost high-speed associative memory," *IEEE J. Solid-State Circuits* (Corresp.), vol. SC-10, pp. 179–181, June 1975.

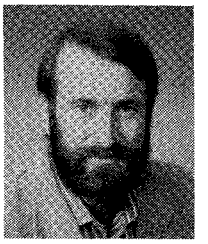
[25] H. Shin and M. Malek, "A Boolean content addressable memory and its applications," *Proc. IEEE* (Corresp.), vol. 73, pp. 1142–1144, June 1985.

[26] L. Chisvin and R. J. Duckworth, "Content-addressable and associative memory: Alternatives to the ubiquitous RAM," *IEEE Computer*, vol. 22, pp. 51–64, July 1989.
[27] M. Sivilotti, M. R. Emerling, and C. A. Mead, "A novel associative memory implemented using collective computation," in *Proc. Chapel Hill Conf. VLSI*, 1985, pp. 329–342.
[28] A. Moopenn, J. Lambe, and A. P. Thakoor, "Electronic implementation of associative memory based on neural network models," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, pp. 325–331, Apr. 1987.
[29] R. E. Howard *et al.*, "An associative memory based on an electronic neural network architecture," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 1553–1556, July 1987.
[30] K. A. Boahen *et al.*, "A hetroassociative memory using current-mode MOS analog VLSI circuits," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 747–755, May 1989.
[31] H. P. Graf and L. D. Jackel, "Analog electronic neural network circuits," *IEEE Circuits Devices Mag.*, vol. 5, pp. 44–49, July 1989.
[32] H. Liu and K. Fu, "VLSI arrays for minimum-distance classification," in *VLSI for Pattern Recognition and Image Processing*. New York: Springer-Verlag, 1984, pp. 45–63.
[33] L. T. Clark and R. O. Grondin, "A pipelined associative memory implemented in VLSI," *IEEE J. Solid-State Circuits*, vol. 24, pp. 28–34, Feb. 1989.
[34] S. M. S. Jalaleddine and L. G. Johnson, "Integrated circuit associative memory based on neural mutual inhibition," *Proc. Inst. Elec. Eng.*, Pt. G, *Circuits, Devices, Syst.*, to be published in 1992.
[35] S. M. S. Jalaleddine, "Integrated circuit implementation of associative memories with neural inspired best match and relational search capabilities," Ph.D. dissertation, Oklahoma State Univ., Stillwater, Dec. 1990.
[36] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A System Perspective*. Reading, MA: Addison-Wesley, 1985, p. 351.
[37] S. M. S. Jalaleddine and L. G. Johnson, "Capacitive-access CAM cell and its read/write circuit implementation," *Electron. Lett.*, vol. 27, no. 10, pp. 876–878, May 9, 1991.
[38] C. V. Ramamoorthy, J. L. Turner, and B. W. Wah, "A design of a fast cellular associative memory for ordered retrieval," *IEEE Trans. Comput.*, vol. C-27, no. 9, pp. 800–815, 1978.
[39] A. R. Hurson, "VLSI/space complexity of an associative parallel join module," in *Proc. 1986 Int. Conf. Parallel Processing*, pp. 379–386.
[40] L. G. Johnson and S. M. S. Jalaleddine, "MOS implementation of winner-take-all network with application to content-addressable memory," *Electron. Lett.*, vol. 27, no. 11, pp. 957–958, May 23, 1991.
[41] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4–22, 1987.
[42] L. G. Johnson and S. M. S. Jalaleddine, "Parameter variations in MOS CAM with a mutual inhibition network," *IEEE Trans. Circuits Syst.*, vol. 38, no. 9, pp. 1021–1029, Sept. 1991.
[43] J. Lazzaro *et al.*, "Winner-take-all networks of O(N) complexity," in *Advances in Neural Information Processing Systems*, I. D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989.
[44] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*. New York: McGraw-Hill, 1987.

Sateh M. S. Jalaleddine (S'84–M'90) was born in Saida, Lebanon, on August 19, 1962. He received the B.S. and M.S. degrees in electrical engineering from the University of Tulsa, Tulsa, OK, in 1985 and May 1988, respectively, and the Ph.D. degree in electrical and computer engineering from Oklahoma State University, Stillwater, in 1990.

From January 1984 to December 1987 he held Teaching/Research Assistantships and Scholarships in the Electrical Engineering Department at the University of Tulsa. From August 1987 to December 1990 he was a Graduate Research/Teaching Associate in the School of Electrical and Computer Engineering at Oklahoma State University. In 1989 he was a Lecturer teaching VLSI design. From January to August 1991 he was a Visiting Assistant Professor in the School of Electrical Engineering and Computer Science at the University of Oklahoma, Norman. Since August 1991 he has been an Assistant Professor in the Electrical Engineering Department at Western Michigan University, Kalamazoo. His research interests include associative processing, hybrid electronic design, high-performance IC memories, device modeling, neural networks, and parallel computer architectures.

Dr. Jalaleddine is a member of Tau Beta Pi, Eta Kappa Nu, and several societies of the IEEE.



Louis G. Johnson (M'79) received the B.S., M.S., and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1967, 1969, and 1973, respectively.

He worked at General Dynamics, Pomona Division, from 1974 to 1978. Since 1978 he has been with the School of Electrical and Computer Engineering at Oklahoma State University, Stillwater, where he is currently Associate Professor. His current research interests include VLSI design, machine vision, digital signal processing, and neural networks.