RITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT unlimited
DECLASSIFICATION/DOWNGRADING SCHEDULE	
PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ~~F49620-88-C-0112~~ AFOSR-TR- 90 1068
NAME OF PERFORMING ORGANIZATION California Institute of Technology 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AFOSR: ~~M~~adelene Weinberger
6c. ADDRESS (City, State, and ZIP Code) Pasadena, California 91125	7b. ADDRESS (City, State, and ZIP Code) Air Force Office of Scientific Research Bolling Air Force Base, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION ~~DARPA~~/AFOSR 8b. OFFICE SYMBOL (If applicable) NE	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-88-C-0112
8c. ADDRESS (City, State, and ZIP Code) Bldg 410 Bolling AFB DC 20332	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F; PROJECT NO. 6485 Darpa; TASK NO.; WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)

An Optoelectronic Realization of Neural Network Models

12 PERSONAL AUTHOR(S)
A. Agranat, C. Neugebauer, V. Leyva, and A. Yariv

TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final Technical Report	FROM 8-1-88 TO 2-28-90	90-10-12	20

16 SUPPLEMENTARY NOTATION The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
neural networks
artificial intelligence

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This research project is aimed at developing silicon based implementations of neural network models. The main advantages of our approach are its use of standard, present day technology and its highly parallel memory access due to the use of optics.

Two different embodiments of the electronic part of the neural processor have been realized. A phototransistor based network using standard ~~CMOS~~ technology has been built and tested.

The CCD version of the optoelectronic architecture has been fabricated and tested, proving the viability of this architecture. All electronic loading has been explored and offers possibilities of rugged, compact systems.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
NAME OF RESPONSIBLE INDIVIDUAL Craig	22b. TELEPHONE (Include Area Code) (202) 767-4931
	22c. OFFICE SYMBOL NE

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

Table Of Contents

Final Technical Report

Optoelectronic Realizations of Neural Network Models

Reporting Period: August 1, 1988 to February 28, 1990

Principal Investigator: Amnon Yariv

Contract Number : F49620-88-C-~~0122~~ 0112 RPA Order No.: 6485

Program Code 8D10

I. Introduction:

Massively parallel systems, such as neural networks, are most efficiently implemented on highly parallel hardware.[1] Parallel neural network hardware requires parallel memory access. Optics provides a natural means of achieving this parallel memory access without sacrificing flexibility, since in optics, information can be easily manipulated and transmitted in two dimensions.

Efficient electronic implementations of neural networks rely on local weight storage to perform the basic synaptic accumulation function (a matrix-vector multiplication) and avoid the parallel memory access problem.[2]

Electronic networks that learn require weight modification -- done either internally by some built in learning rule (inflexible) or externally, in which case the weight changes must be time multiplexed (slow). A hybrid optoelectronic approach toward neural networks is proposed which combines the parallel access of optics with the speed and real-world compatibility of electronics.

In the basic synaptic weighting operation (a matrix-vector multiplication), the weight matrix, $\underline{W}$, contains the connection strengths between input and output neurons, which are represented as input and output vectors ($\underline{V}$ and $\underline{I}$, respectively). Since the input and output vectors have only O(N) elements, electronic means of communication can suffice. The weight matrix, however, is $O(N^2)$, requiring either costly wiring, inflexible local memory, or parallel optics for efficient operation. The ideal implementation would equalize the access times associated with vector or matrix modifications.

The optoelectronic approach presented here[3] consists of a spatial light modulator (SLM), which produces a two dimensional pattern of light intensities corresponding to the weight matrix, $\underline{W}$, and a electronic neural processor (NP), which performs the matrix-vector multiplication. The pattern on the SLM is imaged onto the NP integrated circuit, where an array silicon photodetectors convert the light intensities into currents and perform the matrix-vector multiplication. Note that the NP IC deals only with vectors (O(N) communication task done electronically) while matrix communication ($O(N^2)$

task) is handled through optics. Thus vector and matrix access times are equalized by confining matrices to optics and vectors to electronics. From a compatibility perspective, this approach uses optics only where completely necessary, making the system more attractive for real-world applications than all optical schemes.

II. Technical Discussion:

A. The CCD Neural Processor

Architecture:

The operation of the CCD-NP occurs in two distinct parts, namely the loading operation and the computing operation. In the current designs, loading can be accomplished by either shining a two dimensional pattern of intensities on the light sensitive CCD matrix elements (optical loading) or electrically demultiplexing the entire matrix through a few pins (electrical loading). Both schemes serve the same purpose -- that being to load the CCD matrix with charge in proportion to the desired matrix elements of the vector matrix multiply, W_{ij}. The optical arrangement necessary is shown in Figure 1. Note that the CCD NP acts as a standard CCD imager in this mode. The electrical loading is similar to the readout method of imagers -- only reversed. Instead of reading a matrix out line by line to form video, video rate information is loaded *into* the chip, using the exact same CCD structure clocked backwards.

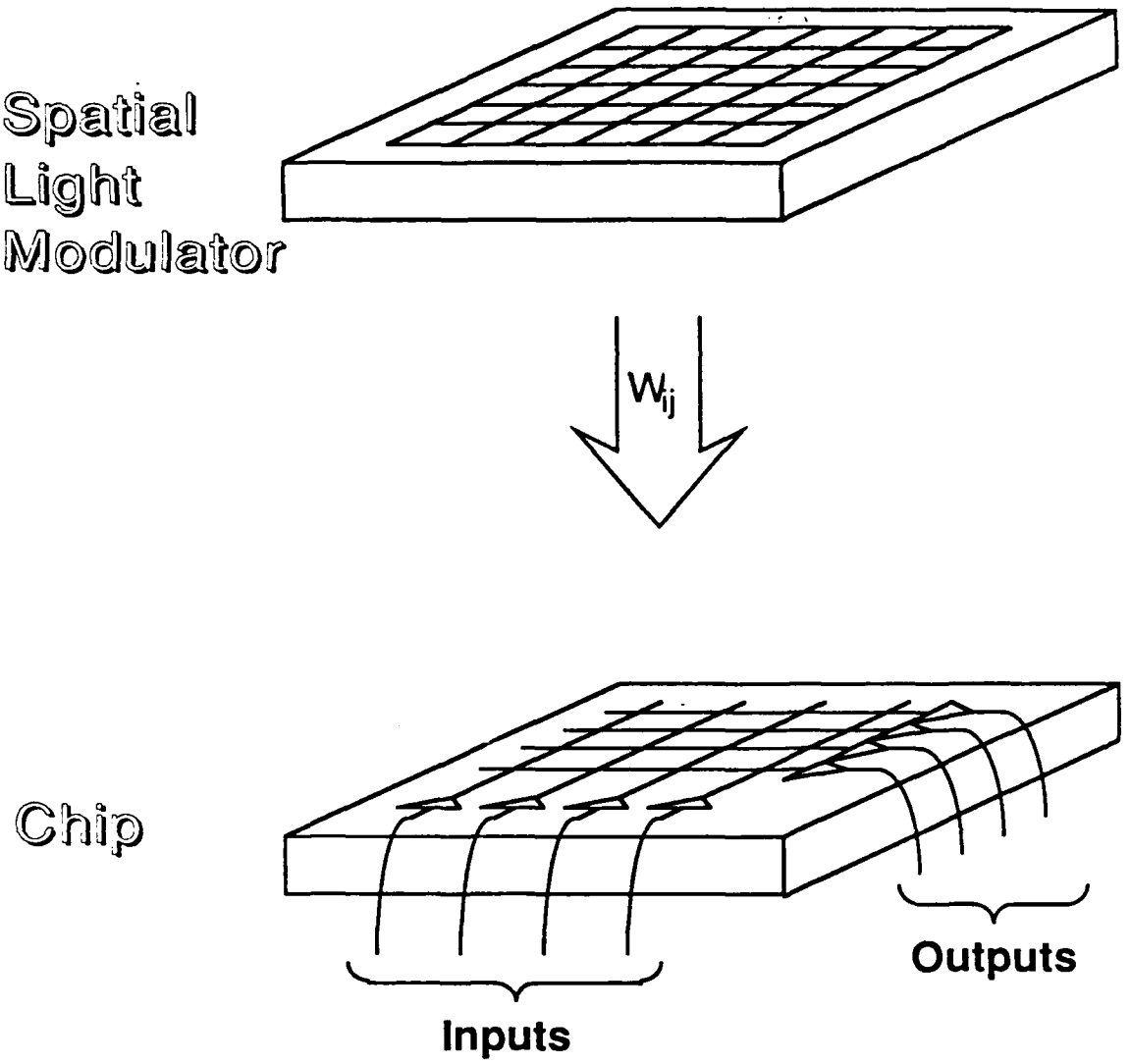

Figure 1. Optoelectronic Architecture

Once the matrix of charge (corresponding to the W_{ij}'s) is in place, the device computes the product of the input vector, V_j, and accumulates the output vector, I_i. The CCD NP computes these function in a semiparallel fashion, i.e. it takes N clock

cycles to compute a vector matrix multiplication ($O(N^2)$). A single row of the CCD NP is shown in Figure 2. The single row contains N charge packets that continually

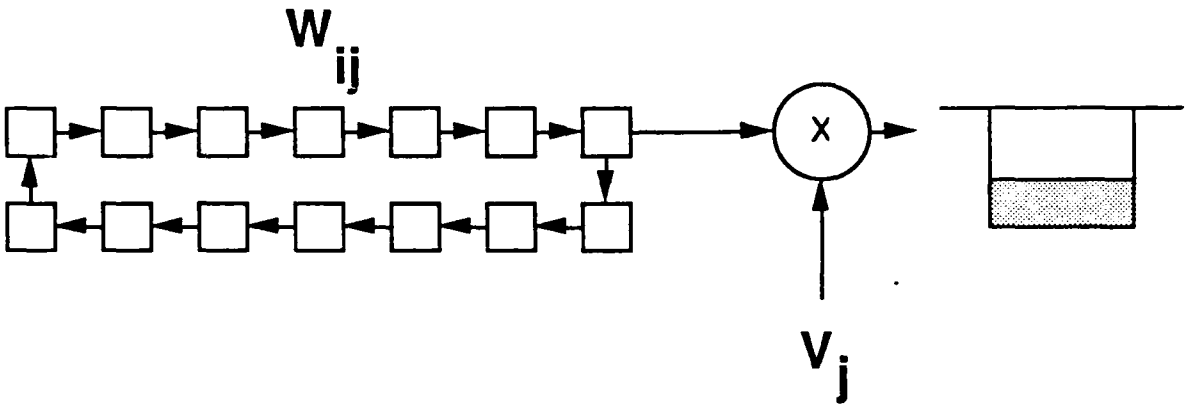

Figure 2. Single Row of CCD NP

revolve around the ring. During the first clock cycle, the multiplier to the right of the ring has two coincident values, the first matrix element of that row and the first input vector element. During this clock cycle, these values are multiplied and added to an accumulated result in the well to the right of the multiplier. Subsequent clock cycles rotate the charge around the ring one position and multiply the matrix element by the proper input vector element. Note that the vector element needed at each clock cycle is the same for all rows of the CCD NP.

Implementation:

A number of complete designs have been submitted to Ford Aerospace for fabrication. These include

1. Optical Loading, Floating Gate Sensing
2. Optical Loading, Diffusion Sensing
3. Electrical Loading, Floating Gate Sensing
4. Electrical Loading, Diffusion Sensing

All chips submitted contain 256 by 256 matrices. The details of the fabrication process made it optimal to divide the optically sensitive parts from the electrically loaded parts. The electrically loaded parts did not need an extra metalization layer (the light shield on the optical parts) and thus the yield is expected to be higher for these parts.

The choice of sensing structures boiled down to two choices, floating gate or diffusion sensing. The sensing structure is used to nondestructively sense the matrix of charge as it revolves past the multiplier in Figure 2 so that the multiplication process does not adversely affect the matrix charge. This nondestructive sensing also allows the matrix to be used many times for vector matrix multiplication without reloading.

The chosen floating gate structure is shown in Figure 3. The circuit consists of a precharge transistor attached to a floating gate which is positioned over the ring

of charge in Figure 2. When a charge packet in the ring (proportional to the matrix element W_{ij}) passes beneath the sensing gate (which has been precharged to a fixed level) the floating gate will experience a proportional change in voltage, equal to

$$\Delta V = Q/C = W_{ij}/C$$

where Q is the sensed charge and C is the capacitance of the floating gate.

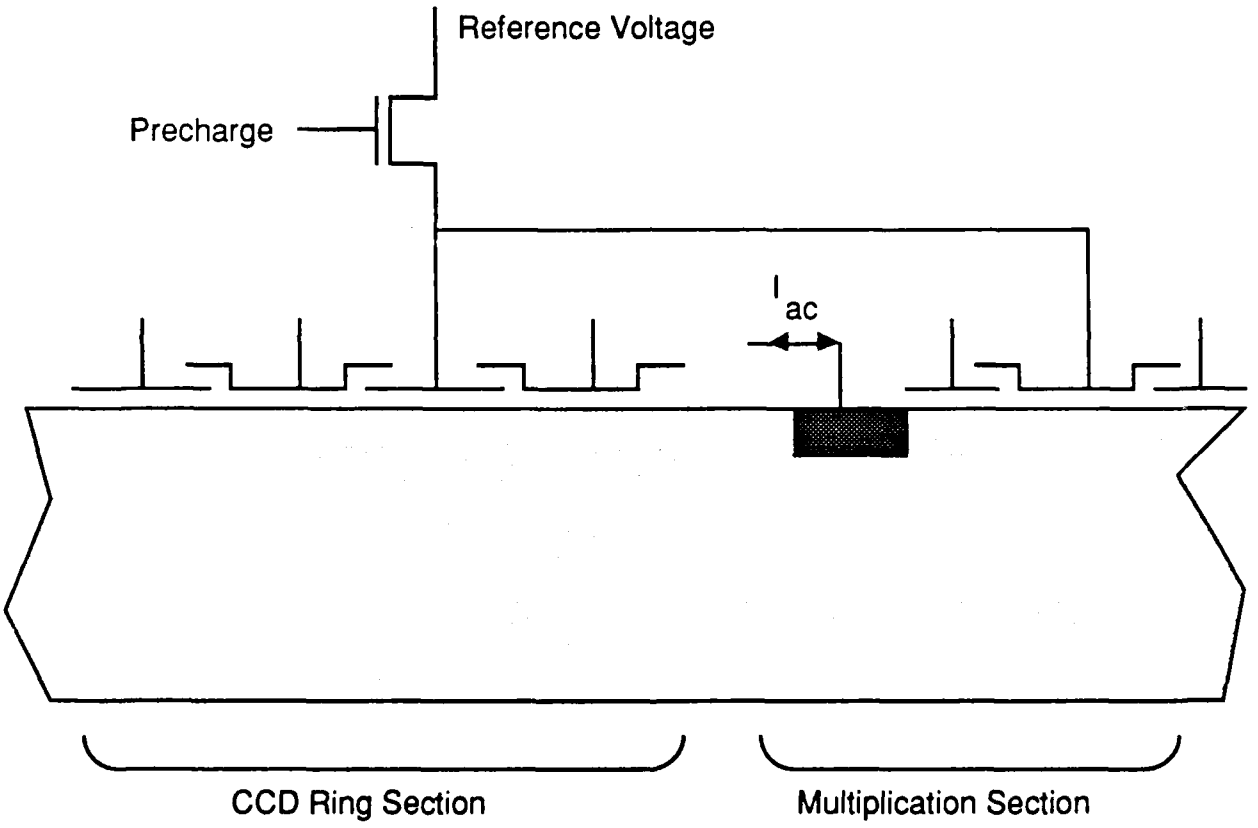

Figure 3. Floating Gate Charge Sensing.

The diffusion sense device works in a similar fashion (it has the same equation) but does not have a precharge transistor, as shown in Figure 4. The diffusion sits in the path of the rotating charge ring and senses the charge by capacitive sensing. While the diffusion sensing method should have lower noise, the floating gate offers a number of practical advantages, primarily that the floating potential (i.e. the precharge voltage) can be adjusted.

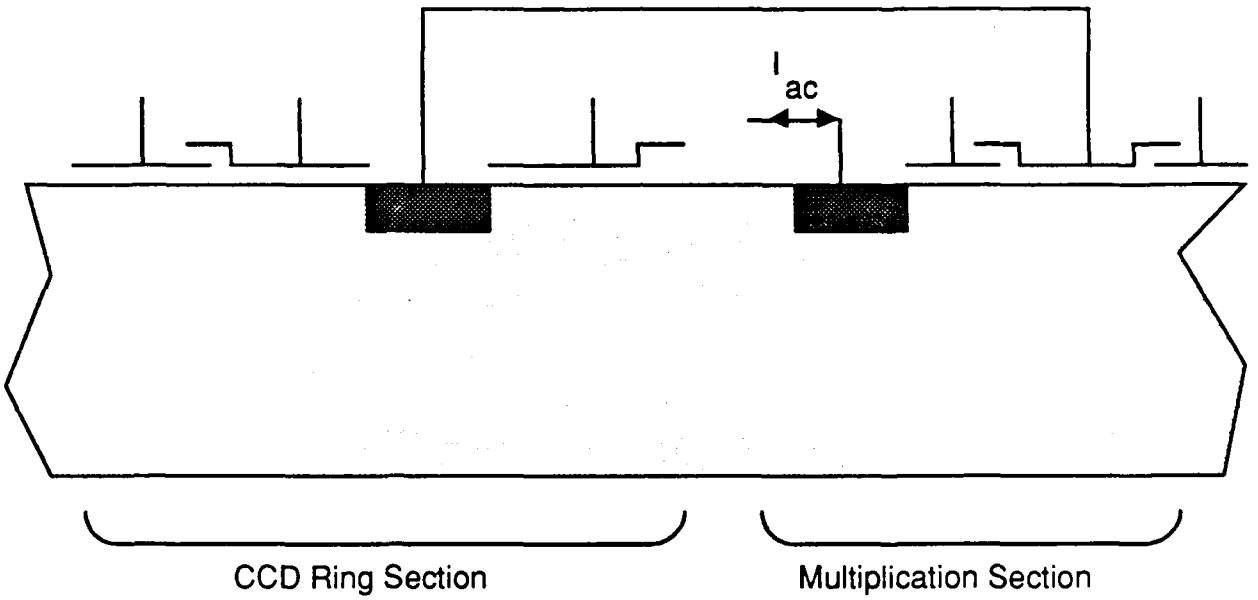

Figure 4. Diffusion Sensing of Charge Packet.

All the chips fabricated have a nondestructive charge domain multiplier circuit that is used to sense the matrix elements. This multiplication unit can be used in either a binary or analog fashion, only the clocking differentiating the modes.

Test Results:

The preliminary tests performed on the CCD-NP measured its accuracy and speed as a matrix-vector multiplier and assessed the ability of the matrix storage rings to act as a short term memory. The linearity of the matrix-vector multiplication is shown in Figures 5 and 6. In Figure 5, the results of multiplying a constant matrix by a binary vector are shown as a function of the number of 'on' elements in the vector. Similarly, in Figure 6 the results of multiplying a binary matrix by a constant vector are shown as a function of the number of 'on' columns in the matrix. A third test (diagramed in Figure 7) consists of multiplying a 'half on' matrix by a 'half on' vector given as a function of the phase between the 'on' parts of the matrix and vector. As expected, a triangle shaped output is observed (Figure 8). The maximum frequency used was 1.5 MHz (which gives approximately 0.5×10^9 interconnection updates/second). The tests were repeated for many iterations yielding a decay rate of the matrix contents of ~6% after 300 milliseconds of continuous operation.

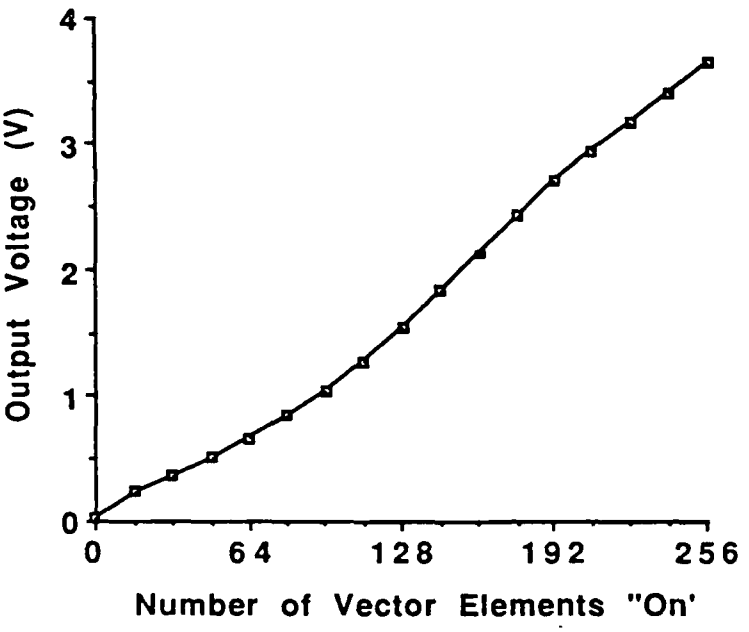

Figure 5. Vector Linearity.

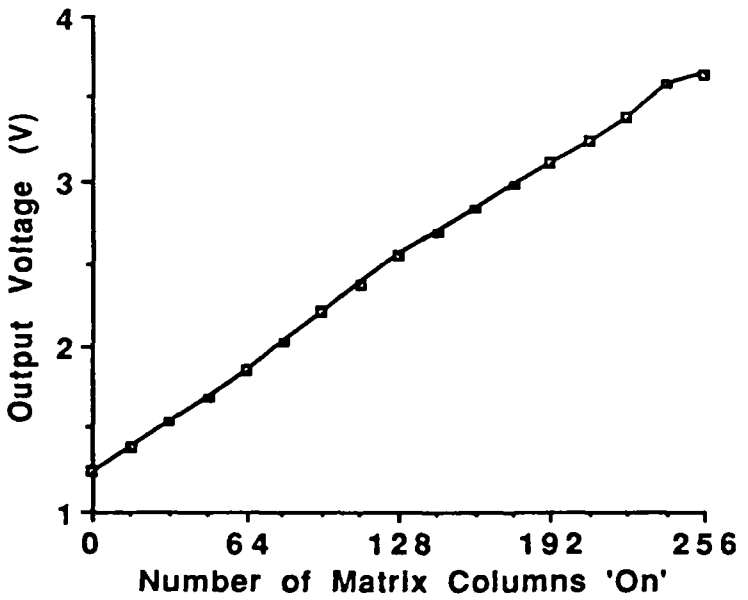

Figure 6. Matrix Linearity.

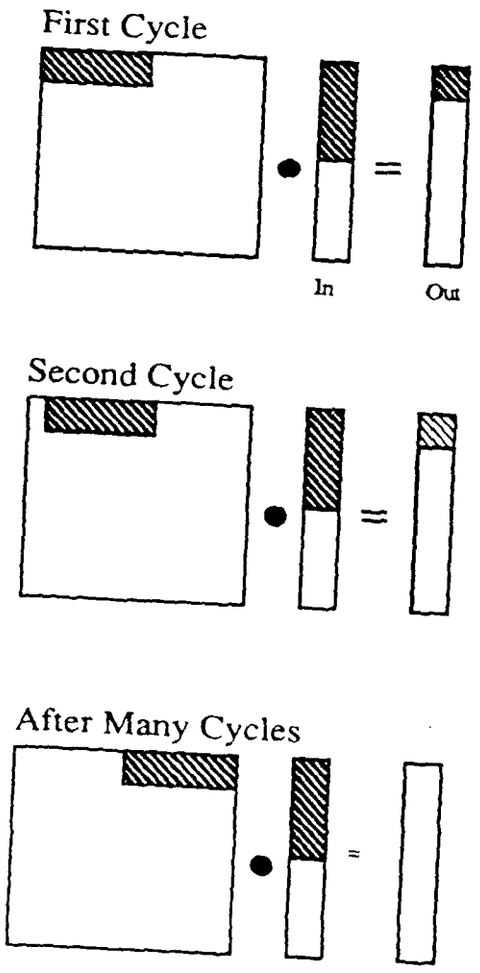

Figure 7. Experimental input for triangle wave test.

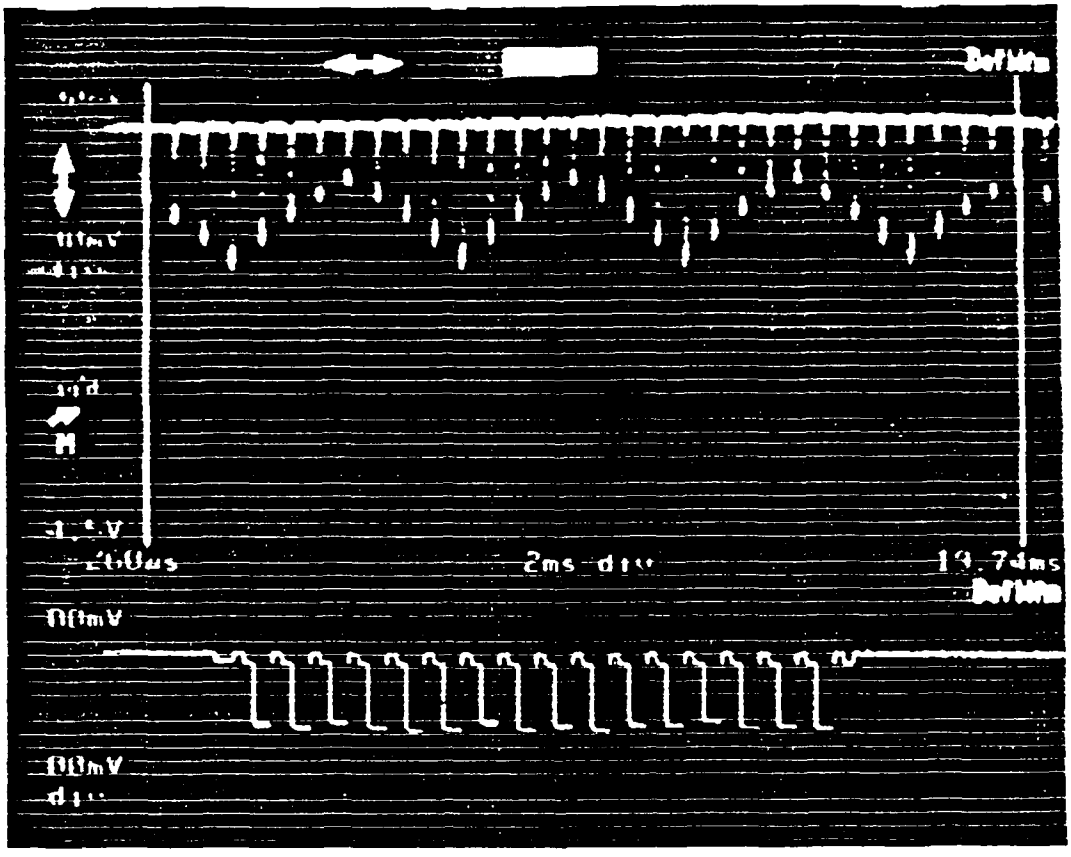

Figure 8. Triangle wave experiment output.

It should be emphasized that these tests were limited by the test equipment which did not allow analog matrix elements and limited the operating frequency to 1.5 MHz. A more advanced test station is now under construction.

B. The Phototransistor Neural Processor:

Architecture:

The phototransistor neural processor (PT NP) uses an array vertical bipolar transistors with a floating base region to sense the optically loaded weight matrix, $\underline{W}$. The PT NP can only be loaded optically since the weight ues (sensed as currents) must be continually generated. The system architecture is the same as shown in Figure 1.

The PT NP contains an array of vertical NPN bipolar transistors with the base region floating, shown in Figure 9. Light hitting the transistor creates a photocurrent in the base, producing an emitter current proportional to one element of the weight matrix, $\underline{W}$. This local current (W_{ij}) is switched by a MOSFET with its gate connected to V_j, thus performing a binary/analog

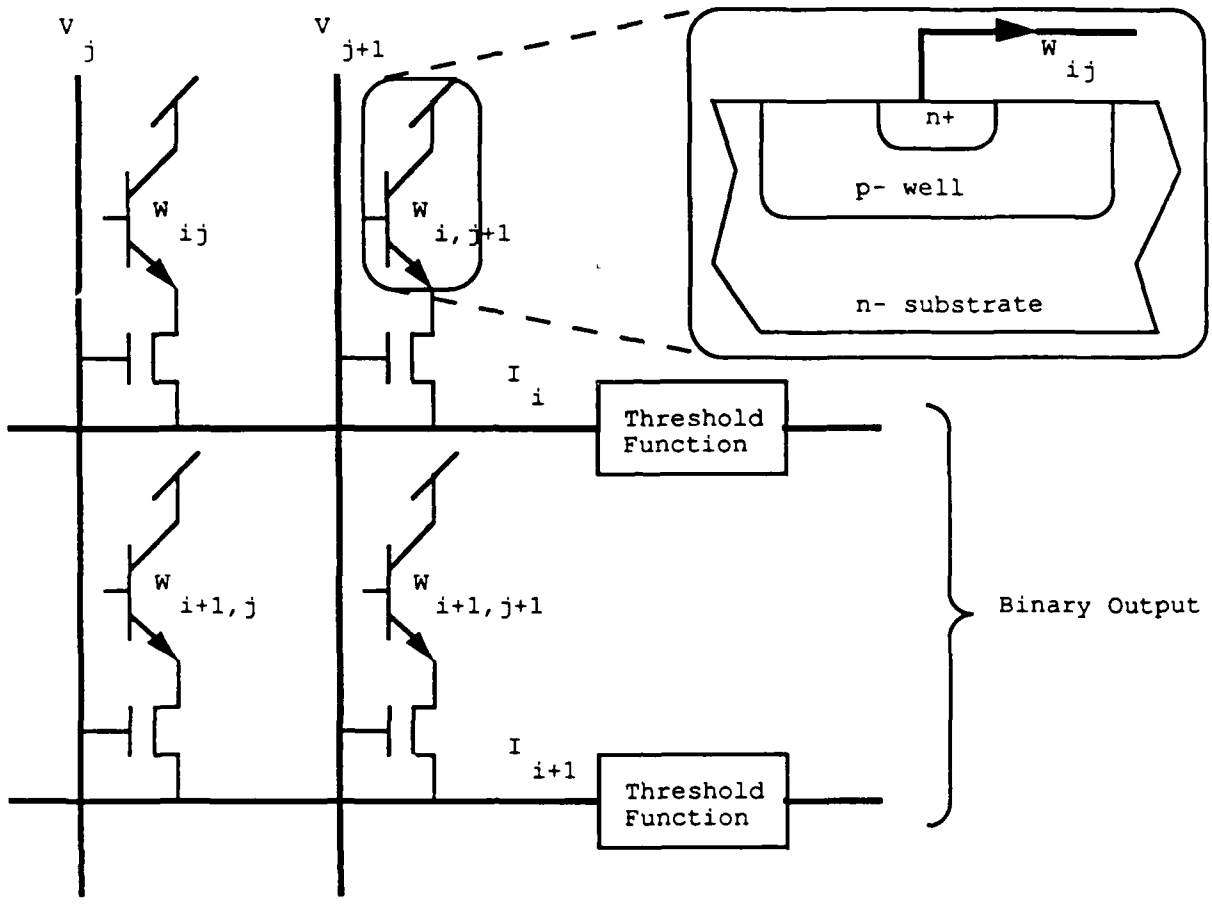

Figure 9. Phototransistor Network Architecture.

multiplication. The partial product current (proportional to $W_{ij}*V_j$) is summed horizontally with other partial products to form the matrix-vector product, I_i. A threshold function applied to the **I** vector finishes the calculation. Note that vectors (neurons) are binary, allowing the use of standard digital interface electronics.

Implementation:

An IC with 32 neurons (1024 connections) has been fabricated with a 3μm p-well CMOS technology. The connections, containing a phototransistor and a FET, occupy only 50μm by 50μm each, or 1.6 mm by 1.6 mm for the entire array, enabling the fabrication of much larger single chip systems with present technology.

Test Results:

The NP IC described above was tested using a binary magneto-optic SLM as the weight matrix input. An IBM PC was used to access the input and output vectors.

Simple linearity tests were performed on the matrix-vector multiplier. The linearity of the synaptic weighting computation with respect to the number of 'on' input vector elements was tested by uniformly illuminating the IC, digitally changing the input vector, and measuring the current in one of the horizontal collection lines (I_i), shown in Figure 10. A similar test was performed on the matrix -- the vector was fully 'on' while the number of 'on' columns in the matrix was varied, shown in Figure 11.

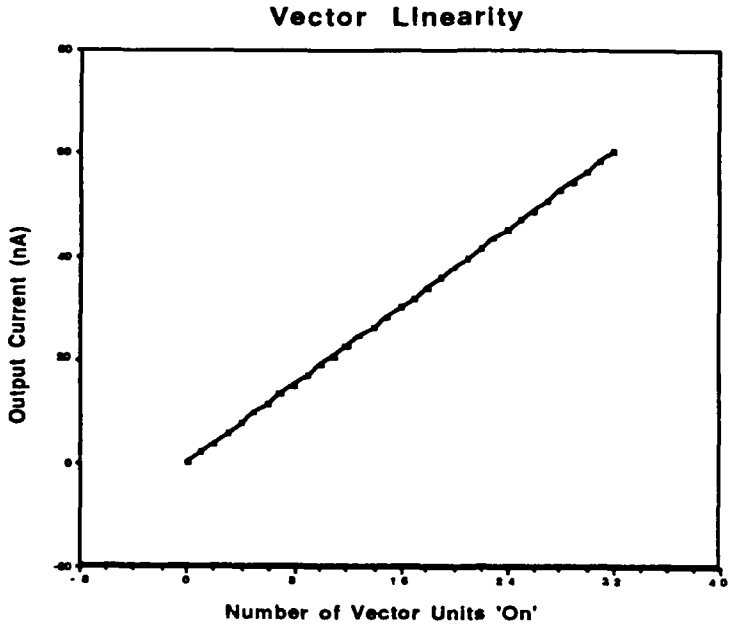

Figure 10. Vector Linearity

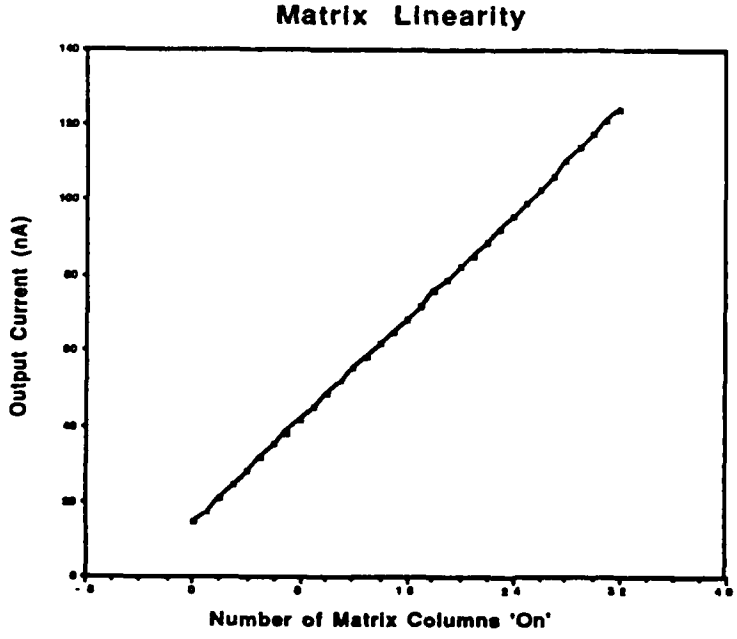

Figure 11. Matrix Linearity

A response time measurement is shown in Figure 12. The horizontal axis is the average photocurrent (proportional to average light intensity) while the vertical axis is the digital response delay to a change in the input intensity. The response delay of the circuit is seen to decrease with increasing intensity, as expected. Under normal laboratory illumination levels, the response delay was typically 1-10 microseconds.

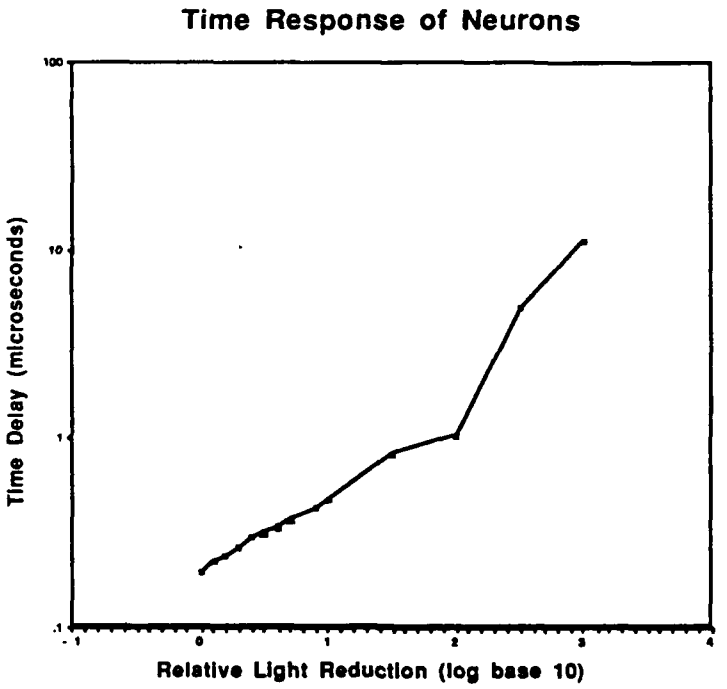

Figure 12. Time Response of Neurons

In addition, statistical analysis of measurements yielded a total nonuniformity (including phototransistor, threshold, and FET nonuniformities)

of approximately 5% at typical illumination levels. Also, circuitry for making the weights both positive and negative was included on the chip.

An additional test was performed using the PT NP in a learning experiment. The test consisted of teaching a two layer neural network (shown in Figure 13) the XOR problem using a random optimization algorithm for adjusting the weights. Random optimization learning is an iterative learning rule that involves presenting the input patterns to the network, measuring the difference between the computed output and the desired output, changing the weight matrix with a random perturbation, then testing the network to see if the error has been reduced. If the error (the sum of the squares of the differences between computed and desired outputs) is reduced, the random perturbation is kept, otherwise the weight matrix remains the same. The system successfully learned the XOR problem within a few hundred pattern presentations as shown in Figure 14.

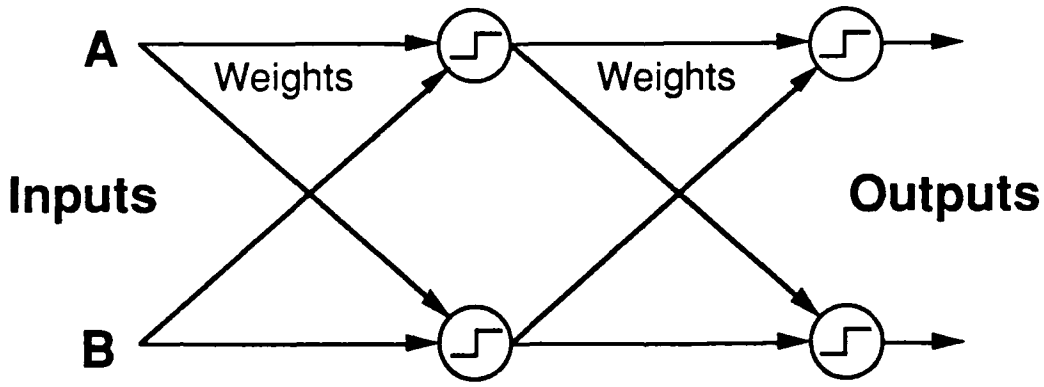

Figure 13. Two Layer XOR Network.

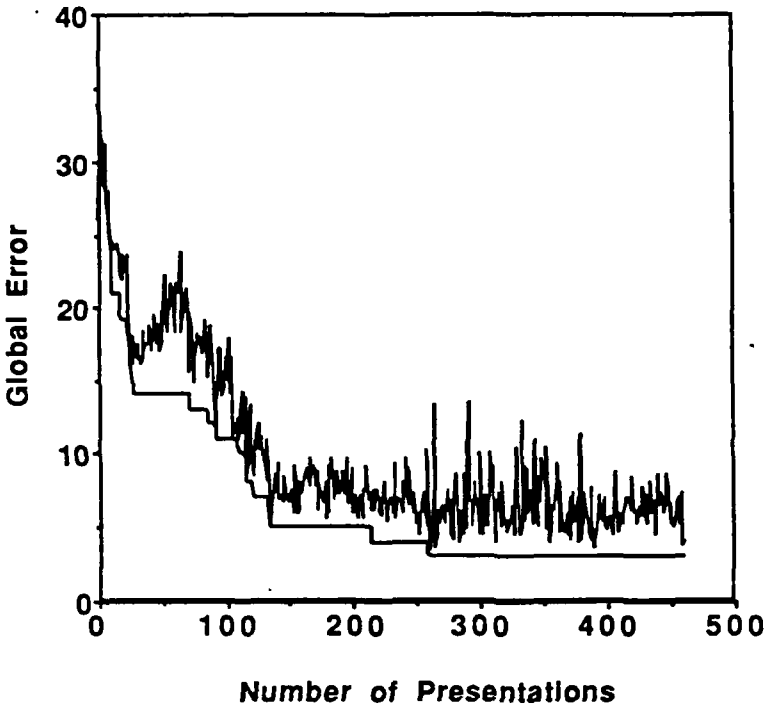

Figure 14. XOR Learning Error vs. Number of Training Periods.

III. References

1 SEITZ, C.L.: 'Concurrent VLSI Architectures',*IEEE Trans. on Computers*, 1984, **C-33**, p. 1247.

2 GRAF, H.P., and JACKEL, L.D., 'Analog Electronic Neural Network Circuits', *IEEE Circuits and Devices Mag.*, July 1989, pp. 44-55.

3 AGRANAT, A., NEUGEBAUER, C.F., and YARIV, A.: 'Parallel Optoelectronic Realization of Neural Network Models Using CID Technology', *Applied Optics*, 1988, **27**, pp. 4354-4355.