Spatial Light Modulators As Parallel Memories For Optoelectronic Neural Networks

Aharon J. Agranat, Charles F. Neugebauer, Amnon Yariv

Aharon J. Agranat, Charles F. Neugebauer, Amnon Yariv, "Spatial Light Modulators As Parallel Memories For Optoelectronic Neural Networks," Proc. SPIE 1150, Spatial Light Modulators and Applications III, (22 May 1990); doi: 10.1117/12.962194

SPIE.

Event: 33rd Annual Technical Symposium, 1989, San Diego, United States

Spatial Light Modulators as Parallel Memories for Optoelectronic Neural Networks

Aharon J. Agranat, Charles F. Neugebauer and Amnon Yariv
Department of Applied Physics
California Institute of Technology
Pasadena, California 91125

ABSTRACT

A generic architecture for realizing neural networks is presented in which the synaptic interaction matrix is loaded in parallel into an electronic integrated circuit from a SLM. Three types of the electronic processors are described using CCD, CID and CMOS technologies respectively. The pros and cons of currently existing SLMs for this architecture are pointed out.

1. INTRODUCTION

It is now commonly accepted that further advancement of neural network (NN) models, as a competitive tool for artificial intelligence, will be largely determined by future success in developing efficient hardware realizations of these models. Conventional single processor and coarse grain multiprocessor simulators are not particularly suited to NN processing--a fine grain, low precision task.[1] A need therefore exists for fine grain hardware realizations of NN models based on existing technology that can be used immediately for applications development.

The optoelectronic realizations of NN models which are described below are aimed at satisfying this need. These systems consist of an electronic processor into which the synaptic interaction matrix is fed optically in parallel from a spatial light modulator (SLM). Thus these systems take advantage of the fact that signal processing in silicon is an advanced and mature technology, and incorporate optics where silicon fails--namely the interconnectivity problem.

In order to explain the underlying operational principles of the architecture, let us first describe the basic dynamics of a NN.[2,3]

Let $\underline{V} = (V_1 \ldots V_N)$ designate the state of a network with N neurons, where V_i is the state of the i'th neuron and let W designate the synaptic interaction matrix--i.e., W_{ij} is the strength of the interaction from the j'th neuron to the i'th neuron. The i'th neuron is updated periodically according to its total input I_i, where

$$I_i = \sum_j W_{ij} V_j \qquad [1]$$

and by using some decision process designated by

$$\tilde{V}_i = \phi(I_i) \qquad [2]$$

where $\tilde{V}_i$ is the next state of neuron i.

The basic idea of the architecture is presented schematically in Figure 1. The system consists of two main subassemblies: a 2D spatial light modulator (SLM), including its memory and control unit, and an integrated circuit which we shall henceforth refer to as the neural processor (NP).

Consider Figure 1. The synaptic interaction matrix W is stored in the SLM. Thus, by imaging the SLM contents onto an array detector, W can be loaded in parallel to the NP. The NP then updates the state of the network $\underline{V}$ by computing the inputs I_i in parallel and using the decision process [2].

In what follows we shall describe three types of neural processors, which are now at different stages of development.

1.1 CCD-NP (based on charge coupled devices--CCD):
These NPs implement synchronous semiparallel networks.

1.2 CID-NP (based on charge injection devices--CID):
These NPs implement fully parallel synchronous networks.

1.3 PT-NP (based on phototransistor arrays):
These NPs implement fully parallel continuous networks.

2. THE CCD-NEURAL PROCESSOR

2.1 Principle of Operation

The CCD-NP is a semiparallel synchronous device. For the sake of clarity we shall explain its principle of operation by describing an implementation of a network with analog synapses and binary neurons. It should be borne in mind that the CCD-NP is not limited to implementing such networks. Implementation of more complex networks will be presented further on. Consider Figure 2:

(SYNP) is a CCD array built of N self-connected rows, each N elements long. (The rightmost element feeds the leftmost element.)

(G) is a column of N analog switches with a common enable port in the case of binary neurons (Figure 2b), or a column of analog multipliers in the case of analog neurons.

(ACCUM) is a column of N integrators--each integrates the signal that flows into it from the respective element of G.

(F) is a column of N "decision functions" (defined schematically in [2]).

(X) is a one-dimensional binary CCD column. Each element of X can be fed directly from the respective element of (F), and the contents of (X) can be output sequentially through the top element.

The various control units, clocks, and input/output units are not included in Figure 2. A complete update cycle is as follows:

1. The synaptic interaction matrix W, and the initial state vector $\underline{V}^{(0)}$ are loaded into SYNP and X respectively.
2. After RESET, $V_i^{(0)}$ supplied by the upper element of X appears at the enable port of G while the first column of W, supplied by SYNP, appears to the respective inputs of G. The elements of ACCUM are set to zero.
3. After the first clock pulse the outputs of G, namely $W_{i1}V_1$:i=1...N, and $V_2^{(0)}$ appears at its enable port.
4. Similarly the consecutive terms $W_{ij}V_j$:j=3...N, are accumulated into ACCUM until after the N^{th} clock pulse . The i^{th} element of ACCUM contains the complete input to the i^{th} neuron (given by [1]).
5. The (N+1) clock pulse now activates the decision function, column F, which produces the updated values of $\underline{V}_i$'s.
6. Finally, the new updated state vector, $\underline{V}^{(i)}$, is transferred to X.

Thus the complete network is updated after N+2 clock cycles.

2.2 An Overview of the Design Considerations:

Careful consideration of the operation stages of the CCD-NP as described above reveals that the decision column remains inactive throughout most of the operation cycle. Moreover, at each clock pulse only one neuronic output is used to update the device (in the description above: at the ENABLE port of G). It therefore becomes natural to divide the CCD-NP into two separate modules: A CCD based module and a controller module.

The CCD module function is to compute the neuronic inputs I_i, using the neuron's state V_i as an input supplied by the controller module. The controller module function is to determine the next neuronic state V_i, using a decision module which realizes the decision process ϕ. The controller then outputs the V_i's into the CCD module. A schematic description of the two modules is presented in Figure 3.

The division of the CCD-NP into two separate modules has two main advantages:

2.2.1 The operations performed by the CCD module, namely the computations of the I_i's (as described in [1]), are common to most NN models. The details of the models differ in architecture (e.g., feedback networks, forward propagating networks, etc.) in the decision process ϕ, and in updating schemes. One CCD module can therefore serve for realizing many different NN models, the details of which are contained in the controller module. As the interconnectivity problem is taken care of by the CCD module, the controller module does not require massive parallelism and is therefore simple to design and fabricate.

2.2.2 The CCD module performs operations which are natural to perform in the charge domain. The controller performs operations which are natural to perform by digital/analog CMOS circuits. Combining CMOS devices with CCDs in the same integrated circuit decreases the yield of the fabrication process substantially. This problem is avoided by separating the CMOS devices from the CCD integrated circuit.

2.3 Expected Performance of the CCD-NP:

The state of the art of CCD technology enables the fabrication of arrays with $10^3 \times 10^3$ registers, each with 8 bit accuracy, that can be operated at 10 MHz. It can be easily seen that the interconnect update rate for a CCD-NP with $N=10^3$ operated at f=10 MHz will be

$$R_{CCD} = f \cdot N = 10^{10} \text{ interconn. updt./sec.} \quad [3]$$

The CCD-NP enables loading of the synaptic interaction matrix T both optically and electrically. Optical loading depends mainly on the SLM and will be discussed below. Electric loading, however, depends on the details of the CCD-NP and, in particular, on the number of the input lines, n_I. The time required to load the matrix is given by

$$\tau_L = N^2/(n_I \cdot f) \quad [4]$$

for $N=10^3$, f=10MHz and $n_I=32$ we get $\tau_L \simeq 3$msec.

Two effects are expected to limit the performance of the CCD-NP: the inefficiency of the charge transfer process and the thermally generated charge. It was verified in simulations that the required refresh rate of the main storage array (SYNP in Figure 2) is much lower than the loading rate ($1/\tau_L$), thus the CCD-NP will not be substantially affected by these effects.

3. THE CID NEURAL PROCESSOR

3.1 General Architecture

A fully parallel version of the CCD-NP can be built by using Charge Injection Devices[5] (CID), rather than CCD's. The general architecture of the CID-NP implementing an N dimensional NN with analog synapses and binary neurons,[6] is presented schematically in Figure 4.

(SYNP) is an NxN detector array where each detector is a CID pixel of the type described in References.
(AMP) is a column of N amplifiers each capable of sensing the charge flow into its respective row.
(DF) is a column of decision function circuits.
(X) is a column of binary registers containing the state vector of the network, $\underline{V}$.

A complete network update is performed as follows:
Initially the synaptic interaction matrix is imaged onto SYNP, so that charge proportional to W_{ij} is accumulated under the collecting electrode of the (i,j)'th pixel. After charge accumulation is completed, the charge in each pixel is then transferred to its respective row electrode provided its column neuron is on. (i.e., W_{ij} is transferred to the i'th row for all the column for which $V_j=1$). The charge transfer into each row is sensed by its respective amplifier in AMP. The output of these amplifiers (which is the neuronic input I_i), is used by the respective decision circuit in DF to determine the next state of the neurons. These new state values are then stored in X. Finally the device is reset. Two modes of resetting are

possible: destructive and nondestructive.[5] In destructive reset the charge under the sensing electrodes is flushed into the substrate and the device is ready for optical loading. In nondestructive reset the charge under each sensing electrode is returned to its respective collecting electrode, and the device is ready for the next update without the need for new optical loading.

3.2 Expected Performance of the CID-NP

The CID-NP should be comparable in size to the CCD-NP but the time required to read it is expected to be much shorter (typicaly τ_R=10 μsec per line[5]). The computation speed of the CID-NP is therefore, expected to be

$$R_{CID} = N^2/\tau_R = 10^9 - 10^{11} \text{ interconn. updt/sec} \qquad [5]$$

This speed can be maintained only by using nondestructive reset. The currently existing SLMs are not fast enough to enable reloading of W at each iteration at this speed.

4. THE PHOTOTRANSISTOR NEURAL PROCESSOR

4.1 General Architecture

The underlying principal of the phototransistor (PT) neural processor is described schematically in Figures 5 and 6. The weight input is achieved with an array of phototransistors. The synaptic interaction matrix W, which is emitted from the SLM as a spatial distribution of light intensity, is incident continuously on this array and, thus, the W_{ij}'s are transformed into currents. Each node of the array, in addition to the photodetector, is a multiplier. One input of each multiplier is connected to the respective photodetector and the second input is connected to the column input line. The potential of each column line is the state of the respective neuron--that is, the potential of the j'th column is V_j. The multipliers at each synapse of the NN compute the partial product $W_{ij}V_i$ and produce a continuous current output proportional to $W_{ij}V_j$. Figure 5 describes the simplest synapse--the multiplier (a single FET) is an on-off switch (binary multiplication). The output of each multiplier is connected to the respective row line, where Kirchoff current summing occurs (i.e. partial products are summed), and thus each row line carries a current proportional to I_i. Each row line is connected to a circuit which realizes the decision function.[2] The output of each of these circuits can be connected to the respective column to achieve a feedback network. In addition, the PT-NP contains circuitry for subtracting weight and input offsets for full four-quadrant multiplication with minimum node size. In contrast to the charge transfer devices, this network requires that the optically configured W_{ij} matrix be continuously illuminated. The device is also unclocked and limited in computation time only by the response of the devices.

4.2 Circuit Description and Preliminary Results

A typical p-well CMOS process contains a parasitic vertical bipolar transistor which can be used as a photodetector. A description of this

device is presented in Figure 8. As can be seen this is an NPN bipolar device in which the substrate (which acts as the collector) is tied to V_{dd}, the emitter is formed by a heavy N type diffusion at the surface of the substrate, and the base region (the p-well) is left floating. Thus photocurrent generated in the base will be multiplied by the current gain factor. (It was found that the standard MOSIS 3μ CMOS process produces a phototransistor with a typical current gain of over 200.)

A simple version of the PT-NP has been built and tested. The IC containing a 32 x 32 array of synapses of the type described in Figure 5, and 32 binary decision functions was fabricated in MOSIS's 3μm p-well process. The synapse size was approximately 50 x 50 μm^2. The binary decision function circuits were found to have thresholds within 5-10% uniformity across the chip. The phototransistors themselves could resolve the minimum SLM changes in input intensity, putting their sensitivity >45 dB. The sensitivity of the entire neuron (synapses and decision functions) was measured at 35 dB, which translates into 5-6 bit accuracy. The settling time of the network depends on the illumination strength as shown in Figure 7. A simple two-layer inverting XOR was implemented using a low power CRT as the SLM. Six weights and three neurons were used to demonstrate the operation shown in Figure 9.

In summary, this device has 10^3 interconnects with 5-6 bit accuracy and converges between 1μsec to 10 μsec, thus the interconnect update rate here is 10^8-10^9 inerconnects updates per second (depending on the illumination level).

4.3 Expected Performance of the PT-NP

The device described in the last section is a first prototype, and as such its performance is far from the optimal performance of the PT-NP. The device is very small (less than 3mm on the side), and therefore scaling-up to a network with 100-500 neurons can be easily done. This will give an interconnect update rate of

$$R_{pt} = 10^{10-11} \text{ interconn. updates/sec.}$$

depending on the illumination level. It should be also noted that the PT-NP is not limited to realizing networks with analog synapses, and binary neurons. A more complex synapse than the one described in Figure 5 which incorporates an analog multiplier has been built and is now being tested. The new synapse enables the construction of networks with analog synapses and analog neurons.

Finally it should be noted that PT-NP has two limitations resulting from the fact that it is a current device (unlike the CCD-NP and the CID-NP which are charge devices).

4.3.1 The fact that in the PT-NP the synaptic strength W_{ij} is proportional to the light intensity emerging from the (i,j)'th pixel of the SLM requires that the PT-NP be illuminated continuously. This requires the SLM to output constant intensity (per pixel) after it is refreshed. This is true only for the magnetooptic SLMs which are binary, and for one type of analog

liquid crystal SLM[ ]. (The latter refresh time is too long rendering this device impractical for our purposes.) Alternatively, the update cycle of the PT-NP can be synchronized with the refresh cycle of the SLM. This slows down the PT-NP considerably and is therefore not an attractive solution.

This problem was found to be a major limitation of this approach by an AT&T group developing neural processing based on amorphous silicon photoconductive arrays.[7]

4.3.2 It was found experimentally that crosstalk between adjacent synapses (i.e., phototransistor) is large, if for one of them the respective neuron is off.

Consider Figure 6. If, for example, $V_2=0$, then the photocurrent W_{32} produced at the respective PT cannot flow normally into row 3, because the respective FET is off. Some of this photocurrent will flow into adjacent PT's and will affect the input into adjacent neurons. This problem is also characteristic to the PT-NP alone.

5. DISCUSSION

It is commonly accepted that the field of hardware realizations of NN is still in its infancy. Current efforts are aimed to be more exploratory in nature rather than to serve as an ultimate solution. Moreover, at this stage very few NN algorithms have been identified as solutions for "real-life" applications. In this context, namely for the purpose of serving as a research tool, the approach presented here seems to be very competitive.

The computational speed of other analog networks lies between 10^9-10^{11} interconnect updates/second, but it is usually argued that in general there is a tradeoff between the complexity of the interconnections and their size.[8] This is not the case here. The use of the SLM as a short term memory that can be loaded in parallel into the device simplifies the structure of the synapse considerably. We are currently developing a fully analog synapse (that can accept analog neuron states) with 4-6 transistors per per synapse. In particular the CCD-NP has a very simple structure and yet enables very complex networks to be realized. This results from the fact that the operation of the CCD-NP is semiparallel. Thus the complexity of the neuron affects one circuit in the controller module only.[9]

Although the CCD-NP is the slowest of the three NPs (but well within the range of 10^9-10^{11} updated/sec), it seems to be at this point to be the most promising of the three. This is unfortunately due to the fact that it can be operated without the SLM, using the CCD array as a short term memory.

The currently available SLMs are either nonstationary between loading (e.g., the CRT), or with a small dynamic range (e.g., the magnetooptic SLM). Thus the full potential of the architecture proposed here cannot be realized based on existing SLMs.

6. ACKNOWLEDGEMENTS

This research was supported by the Defense Advanced Research Projects Agency and the Air Force Office of Scientific Research.

7. REFERENCES

1. C. L. Seitz, "Concurrent VLSI Architectures," IEEE Trans. on Computers, C-33, p. 1247, December 1984.
2. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," Proc. Natl. Acad. Sci., vol. 79, pp. 2554-2558, 1982.
3. J. J. Hopfield, "Neurons with Graded Response Have Collective Computational Properties Like Those of Two State Neurons," Proc. Natl. Acad. Sci., Vol. 81, pp. 3038-3052, 1984.
4. A. Agranat and A. Yariv, "Semiparallel Microelectronic Implementation of Neural Network Models Using CCD Technology," Elect. Lett, Vol. 23, pp. 380-581 1987.
5. G. R. Sims and M. Bonner, J. Denton, "Spatial Pixel Crosstalk in a Charge Injection Device," Opt. Eng., Vol. 26, pp. 999-1007, 1987.
6. A. Agranat, C. F. Neugebauer, and A. Yariv, "Parallel Optoelectronic Realization of Neural Networks Models Using CID Technology," Appl. Opt., Vol. 27, pp. 4354-4355, 1988.
7. E. A. Rietman, R. C. Frye, C. C. Wong, and D. C. Kornfeld, "Amorphous Silicon Photoconductive Arrays for Artificial Neural Networks," Appl. Opt., Vol. 28, pp. 3474-3478, Aug. 1989.
8. H. P. Graf and L. D. Jackel, "Analog Electronic Neural Network Circuits," IEEE Circuits and Devices Mag., pp. 44-55, July 1989.
9. A. Agranat and A. Yariv, "A New Architecture for a Microelectronic Implementation of Neural Network Models," IEEE First International Conference on Neural Networks, Vol. III, pp. 405-409, June 1987.

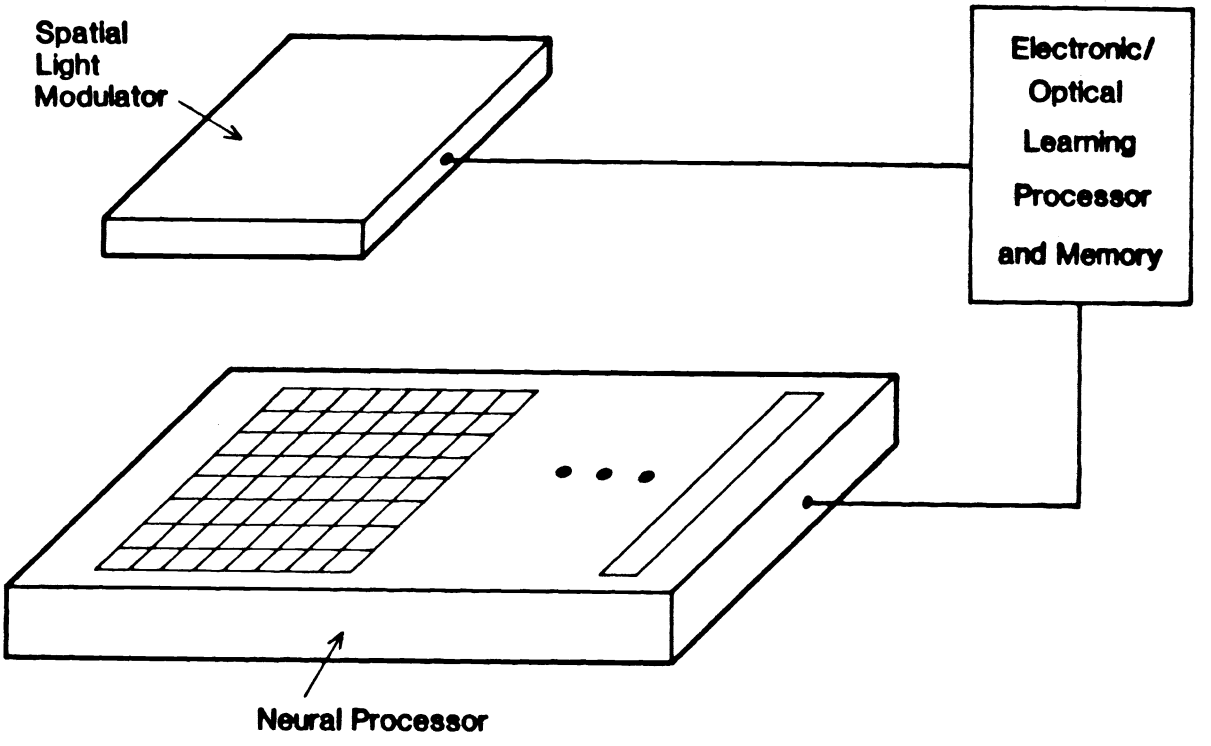

Fig. 1. Schematic description of the architecture.

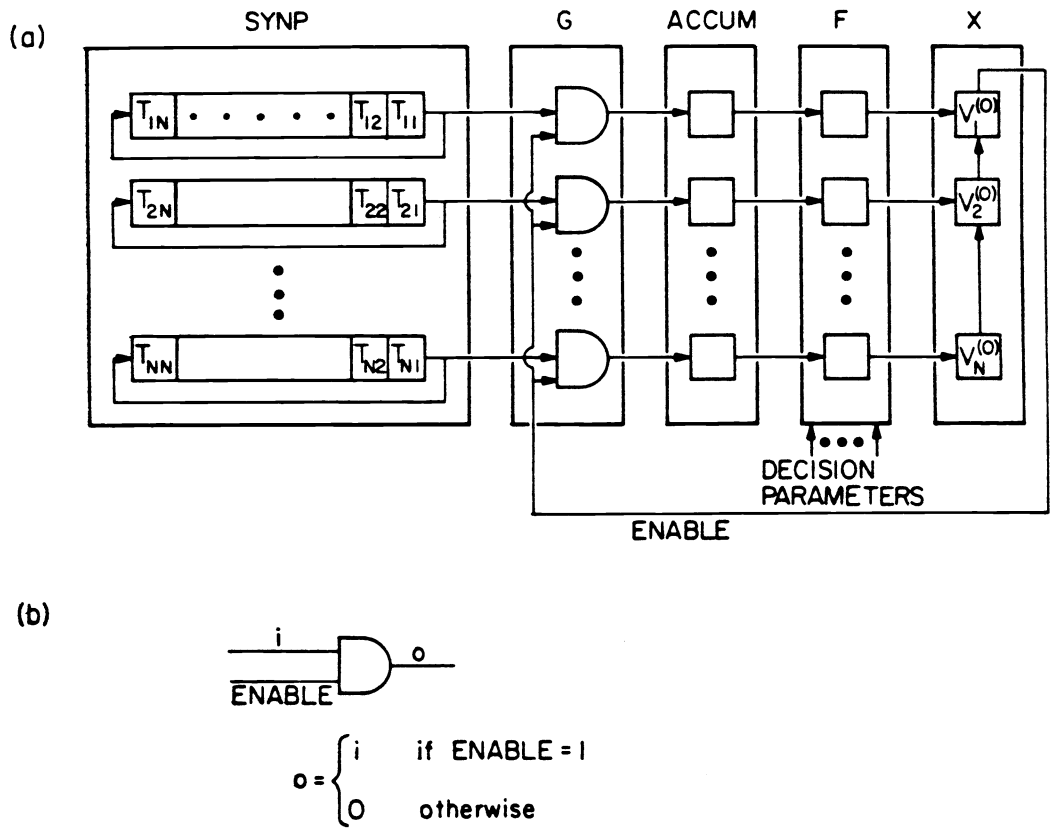

Fig. 2. (a) Basic Architecture of the CCD-NP

(b) One of the switches of (G) in Figure 2(a): an analog switch with binary enable port.

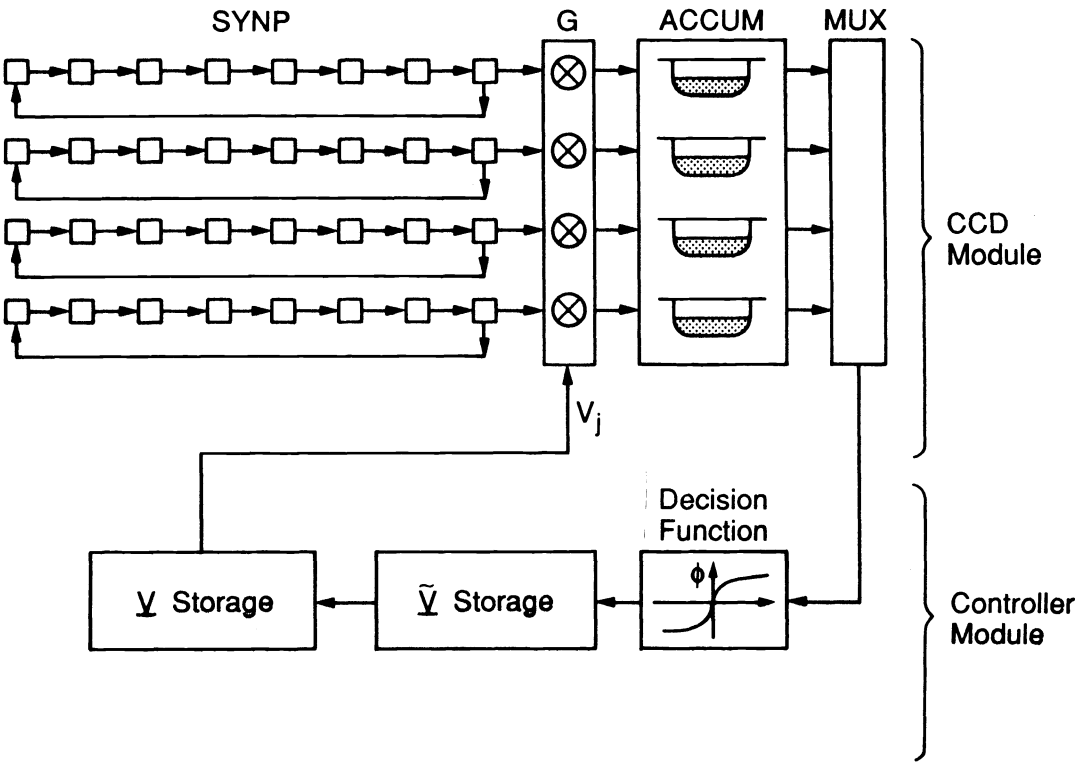

Fig. 3. The separation of the CCD-NP into a CCD module and a controller module.

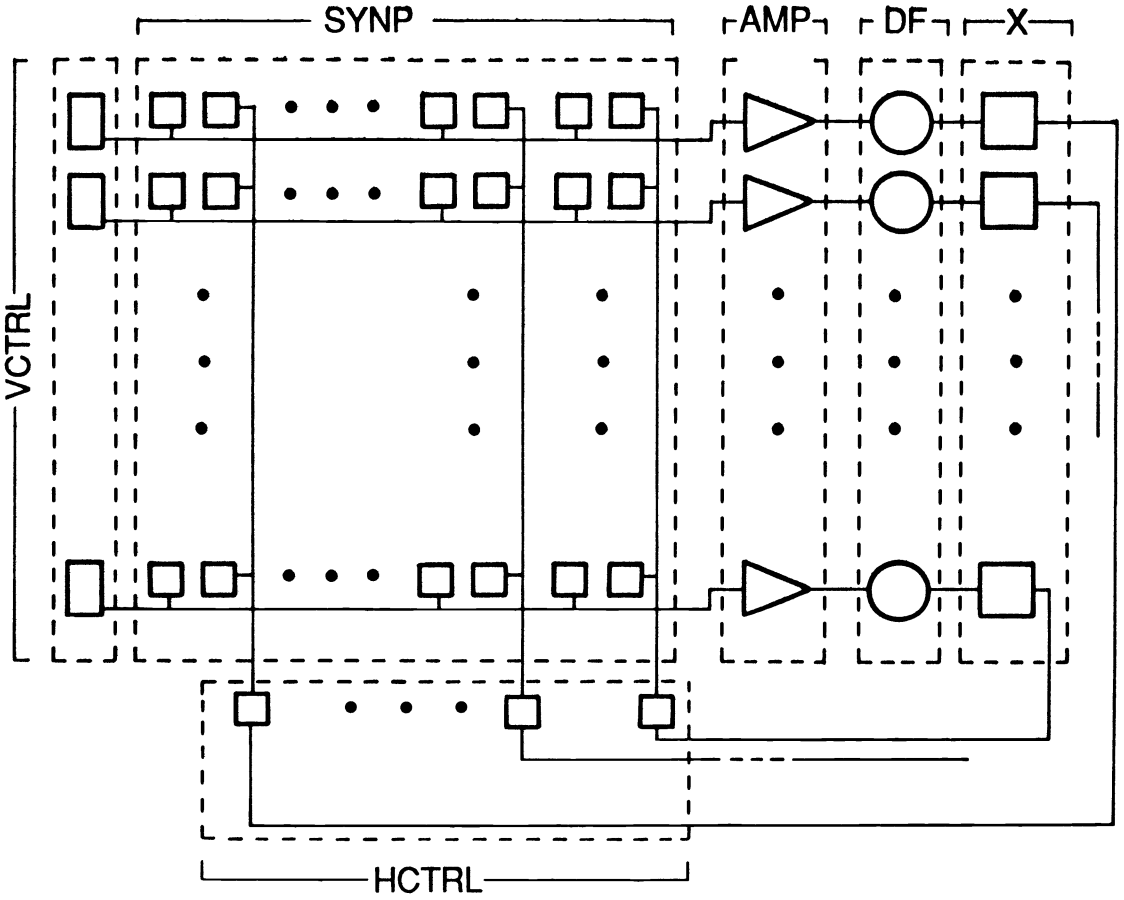

Fig. 4. Schematic layout of the CID neural processor.

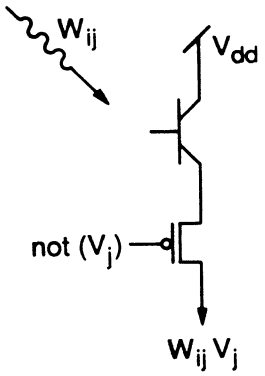

Fig. 5. Circuit description of one synapse in the PT-NP with analog synapses and binary neurons.

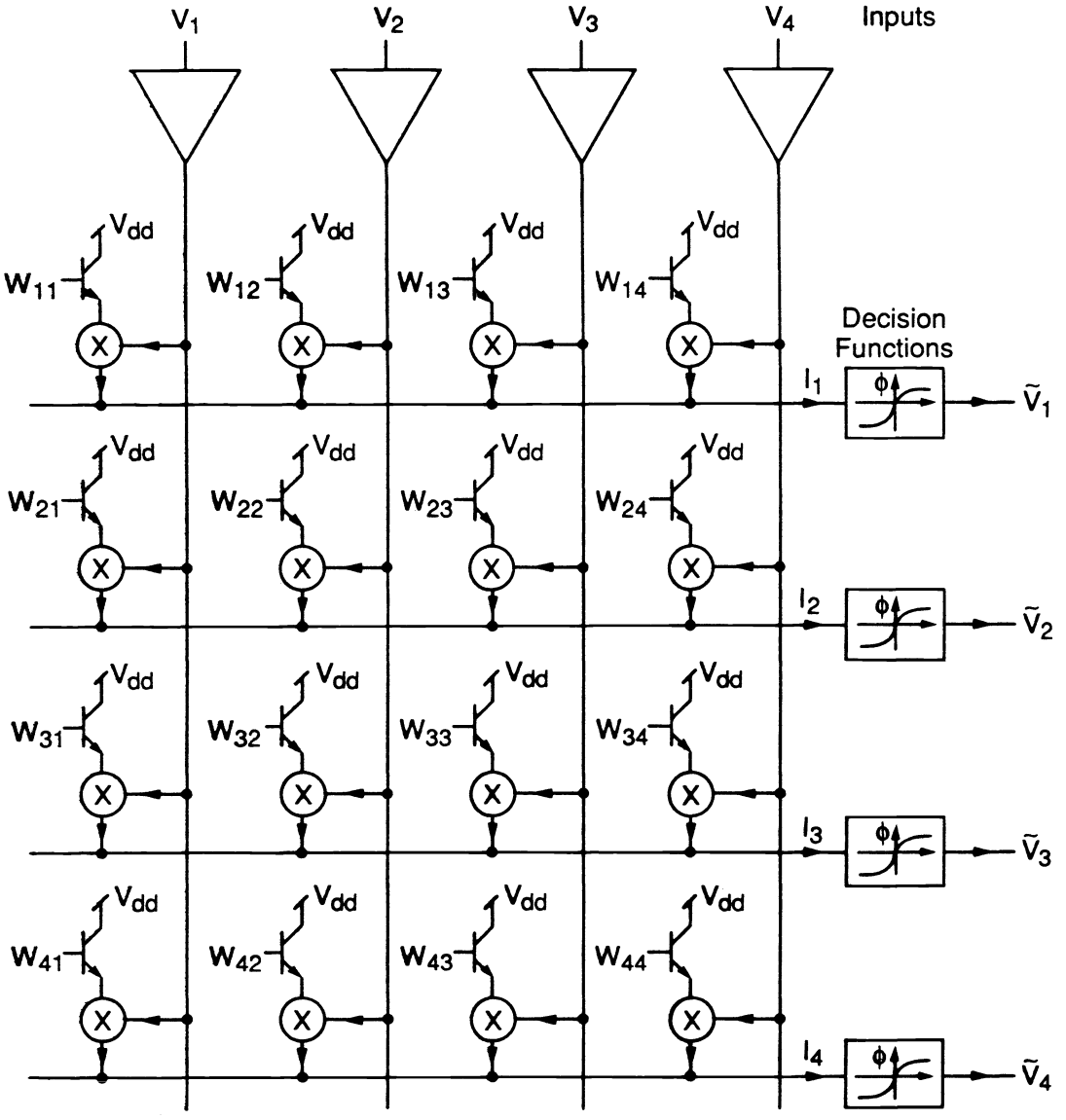

Fig. 6. Schematic description of a PT-NP with 4 neurons.

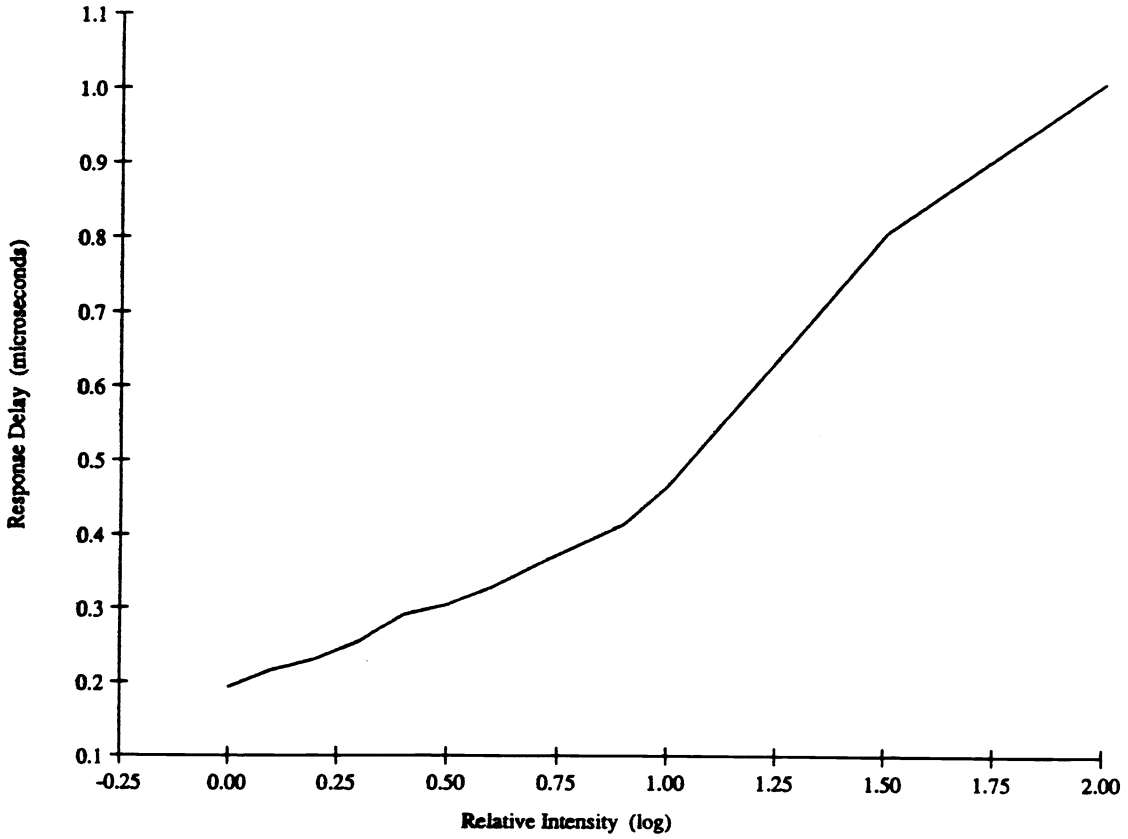

Fig. 7. The Neuron Response Time vs. Illumination intensity.

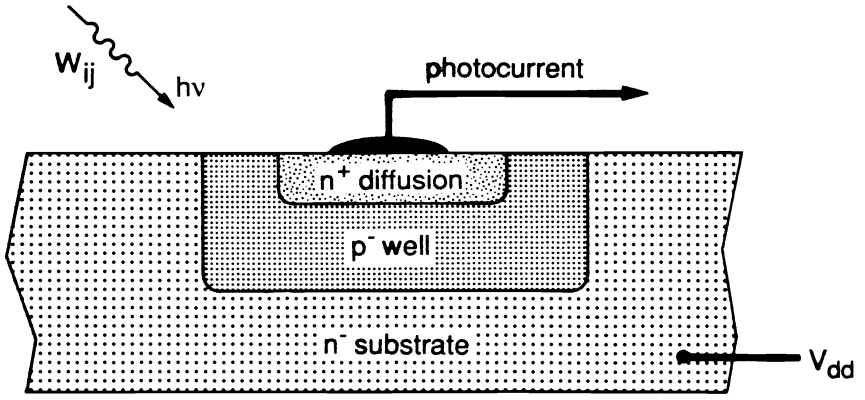

Fig. 8. Parasitic vertical bipolar phototransistor.

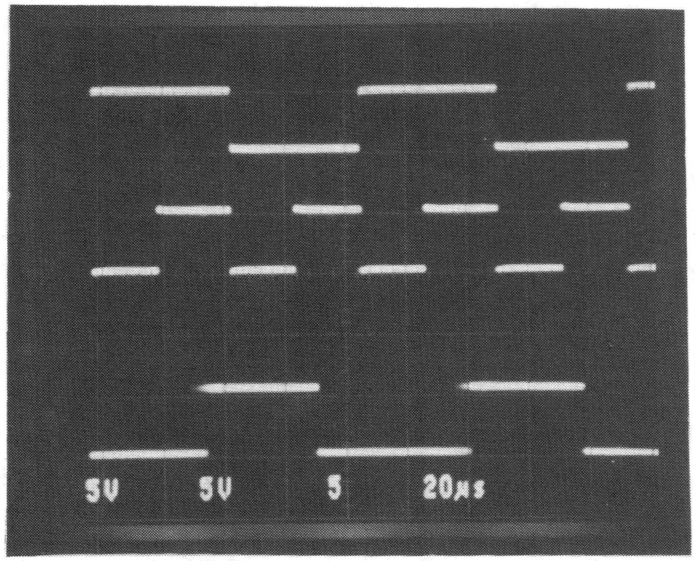

Fig. 9. Inverse XOR operation implemented by the PT-NP.